Image Classification of Marine-Terminating Outlet Glaciers using Deep Learning Methods

MELANIE MAROCHOV

Melanie Marochov

Image Classification of Marine-Terminating Outlet Glaciers using Deep Learning Methods

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. Despite the burgeoning availability of medium resolution satellite data, the manual methods traditionally used to monitor change of marine-terminating outlet glaciers from satellite imagery are time-consuming and can be subjective, especially where a mélange of icebergs and sea-ice exists at the terminus. To address this, recent advances in deep learning applied to image processing have created a new frontier in the field of automated delineation of glacier termini. However, at this stage, there remains a paucity of research on the use of deep learning for pixel-level semantic image classification of outlet glacier environments. This project develops and tests a two-phase deep learning approach based on a well-established convolutional neural network (CNN) called VGG16 for automated classification of Sentinel-2 satellite images. The novel workflow, termed CNN-Supervised Classification (CSC), was originally developed for fluvial settings but is adapted here to produce multi-class outputs for test imagery of glacial environments containing marine-terminating outlet glaciers in eastern Greenland. Results show mean F1 scores up to 95% for in-sample test imagery and 93% for out-of-sample test imagery, with significant improvements over traditional pixel-based methods such as band ratio techniques. This demonstrates the robustness of the deep learning workflow for automated classification despite the complex characteristics of the imagery. Future research could focus on the integration of deep learning classification workflows with platforms such as Google Earth Engine (GEE), to classify imagery more efficiently and produce datasets for a range of glacial applications without the need for substantial prior experience in coding or deep learning.

Image Classification of Marine-Terminating Outlet Glaciers using Deep Learning Methods

Melanie Marochov

Thesis submitted for the degree of Master of Science

Department of Geography
Durham University

December 2020

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1D	One-dimensional
2D	Two-dimensional
4D	Four-dimensional
AIS	Antarctic Ice Sheet
ANN	Artificial Neural Network
ASAR	Advanced Synthetic Aperture Radar
cCNN	Compact Convolutional Neural Network
CNN	Convolutional Neural Network
CSC	CNN-Supervised Classification
DEM	Digital Elevation Model
EO	Earth Observation
FCN	Fully Convolutional Neural Network
GEE	Google Earth Engine
GEEDiT	Google Earth Engine Digitisation Tool
GIS	Geographical Information Systems
GrIS	Greenland Ice Sheet
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IT	Inferotemporal (in reference to the inferotemporal cortex)
MaQiT	Margin change Quantification Tool
MLP	Multilayer Perceptron
MODIS	Moderate Resolution Imaging Spectroradiometer
NIR	Near Infrared
ReLU	Rectified Linear Unit
RGB	Red, Green, and Blue (in reference to image bands)
SAR	Synthetic Aperture Radar
SGLs	Supraglacial Lakes
TL	Transfer Learning
V1	Primary visual cortex

Additional Useful Terminology

Tile	Primarily refers to the subdivided Sentinel-2 images which represent samples of pure class used to train phase one models in the deep learning workflow outlined in this thesis. The term is also used to refer to larger (3000 x 3000 pixel) images on which the workflow is tested.
Tile Size	Refers to the size of tiles used to train phase one models (height x width, measured in pixels). Specifically, tile sizes of 50x50, 75x75, and 100x100 pixels are used for model training in this thesis.
Patch	Refers to the window (or kernel) used to detect the class of a central pixel in the phase two patch-based model (i.e., the cCNN).
Patch Size	Refers to the window size of patches used in phase two models (height x width, measured in pixels)
In-Sample	Refers to the study area that was used to train phase one models in the workflow, but also includes the surrounding landscape which composes the remainder of the Sentinel-2 scene not used in training. The in-sample test imagery used in this thesis is acquired on a different date, meaning it has never been 'seen' during training.
Out-of-Sample	Refers to the study area that was not used during training of the deep learning models. Thus, the out-of-sample test imagery used in this thesis helps indicate how spatially transferable the workflow is.

Statement of Copyright

Acknowledgements

Firstly, I would like to thank Dr Patrice Carbonneau and Prof Chris Stokes for their sage advice and continued support throughout the development of this project. Pat - thinking back to first year lectures when I sat listening to you talk about fish, I never imagined that four years down the line I would be working on a deep learning project with you, so thank you for sharing your wide-ranging wisdom with me. Chris - as one of the people that first instilled in me a passion for glaciology, I cannot thank you enough.

I would also like to thank my fellow Masters students, and everyone else in the department for an enjoyable year with plenty of new experiences. I must also express my gratitude towards my friends, especially Alex, Henry, and James for laughing at my frustration over code bugs, and Fi for her continued support despite the distance between us, I would not have made it through this pandemic without you. Finally, thanks also go to my family for their constant support and encouragement.

1 Introduction

1.1 Significance of Ice Sheets and Marine-Terminating Outlet Glaciers

The Greenland and Antarctic ice sheets act as large reservoirs which store 7.4 m and 58 m of potential sea-level rise, respectively (Fretwell *et al*., 2013; Morlighem *et al*., 2017). Alongside this, their interconnections with global atmospheric, oceanic, and biological systems makes them particularly important to monitor (Hawkings *et al*., 2014; Beaird *et al*., 2018; Cape *et al*., 2019; Catania *et al*., 2020). Observations show that the Earth's ice sheets have been losing mass at an increasing rate over the past several decades in response to climate forcing (Rignot *et al*., 2008, 2011, 2019; Csatho *et al*., 2014; Velicogna *et al*., 2014; Shepherd *et al*., 2018; Mouginot *et al*., 2019). This has resulted in sea-level contributions of 10.8 ± 0.9 mm from the Greenland Ice Sheet (GrIS), and 7.6 ± 3.9 mm from the Antarctic Ice Sheet (AIS) since 1992 (Shepherd *et al*., 2018, 2020). Moreover, mass loss has predominantly been concentrated at the ice sheet margins, where acceleration, thinning and retreat of marine-terminating outlet glaciers has been initiated and subsequently transmitted to the interior of the ice sheets (Nick *et al*., 2009; Felikson *et al*., 2017). According to mass balance reconstructions between 1972 and 2018, ice discharge (by iceberg calving) from marine-terminating glaciers alone caused ~66% of mass loss from the GrIS (Mouginot *et al*., 2019). Similarly, accelerated ice discharge has had a considerable impact on mass loss elsewhere in the Arctic (Carr *et al*., 2017) and in several regions of Antarctica (Joughin *et al*., 2003; Rignot, 2008; Rignot *et al*., 2008; Miles *et al*., 2013, 2017; Cook *et al*., 2014; Mouginot *et al*., 2014). As a result, a wealth of research has focused on elucidating the key drivers of marine-terminating outlet glacier retreat, acceleration and thinning (Vieli and Nick, 2011; Bevan *et al*., 2012; Rignot *et al*., 2014; Carr *et al*., 2017; Catania *et al*., 2018; Miles *et al*., 2021).

The terminus regions of marine-terminating outlet glaciers provide an important interface between ice and the ocean-climate system. Furthermore, since dynamic changes in ice discharge have been linked to terminus retreat (Vieli and Nick, 2011; Hill *et al*., 2018), terminus position monitoring is frequently used as a key method to analyse the driving mechanisms of dynamic outlet glacier change (Lea *et al*., 2014). Resulting observations have shown that marine-terminating outlet glaciers are sensitive to internal and external drivers over periods of weeks to months (Howat *et al*., 2005; Carr *et al*., 2013; King *et al*., 2018). These drivers include: 1) submarine melt (Sutherland *et al*., 2019), induced by both localised runoff-driven plumes (Carroll *et al*., 2016), and interaction with warm ocean currents (Chauché et al., 2014; Jenkins et al., 2010); 2) reduced buttressing due to loss of sea-ice and

ice mélange (a mixture of sea-ice and icebergs) (Amundson *et al*., 2010; Miles *et al*., 2017; Robel, 2017; Bevan *et al*., 2019); 3) changes in fjord and bed geometry (Bunce *et al*., 2018; Catania *et al*., 2018); and 4) temporary drainage changes at the ice-bed interface (Juan *et al*., 2010; Tuckett *et al*., 2019). In addition, these mechanisms are heterogeneous across local and regional scales (Carr *et al*., 2017; Shepherd *et al*., 2020), with significant spatial variability in thinning (Porter *et al*., 2018), velocity (Bevan *et al*., 2012), and terminus retreat (Motyka *et al*., 2017) which remains largely unexplained (Catania *et al*., 2018).

Due to the range of temporal scales on which these processes operate and influence outlet glacier behaviour, a growing body of literature has focused on measuring glacier termini at high temporal resolution (from daily to monthly satellite data) to measure seasonal changes as well as inter-annual and decadal trends (Fried *et al*., 2018; King *et al*., 2018). However, since mapping the ice fronts of marine-terminating outlet glaciers continues to rely on labour-intense and time-consuming manual digitisation (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), datasets tend to be spatially or temporally constrained (Seale *et al*., 2011). Thus, while recent efforts to examine seasonal changes in outlet glacier termini have helped elucidate our understanding of these drivers, the spatio-temporal limits of datasets resulting from methodological drawbacks are problematic, especially when extrapolating results for use in data-driven ice sheet models (Catania *et al*., 2020).

1.2 Challenges of Mapping Marine-Terminating Glaciers

Well-established, semi-automated techniques such as image band ratios which are used to map mountain glaciers or ice caps for glacier inventories (e.g. Bolch *et al*., 2010; Frey *et al*., 2012; Rastner *et al*., 2012; Guo *et al*., 2015; Stokes *et al*., 2018) are not suitable for mapping marine-terminating glaciers. This is largely due to the presence of seasonally variable areas of a spectrally similar mélange of sea-ice and icebergs near their termini (e.g. Amundson *et al*., 2020), where the use of locally varying and image-dependent threshold values produces inadequate results. Consequently, even manual digitisation can be challenging, and often requires prior expertise. Likewise, the time-consuming nature of manual digitisation, and the growing requirement of high-resolution datasets, highlights the rising need for more efficient methods to quantify glacier change in an era of increasingly available satellite data.

1.3 Automated Mapping of Marine-Terminating Glaciers

To confront the challenge of manual digitisation, some automated pixel-based techniques for extracting outlet glacier termini have been developed, exemplified in a small number of studies (Sohn and Jezek, 1999; Liu and Jezek, 2004a, b; Seale et al., 2011; Krieger and Floricioiu, 2017; Yu et al., 2019). These methods primarily use semantic segmentation, and edge detection image processing tools (described in Chapter 2.2). However, they require substantial pre- and post-processing, and have only been used for terminus delineation in a few studies (e.g. Joughin *et al*., 2008; Christoffersen *et al*., 2012). In general, techniques which rely solely on individual pixel values often miss contextual, class representative shapes and textures. Moreover, in land cover classification, traditional pixel-based approaches (e.g. Maximum Likelihood) commonly result in noisy classifications (Blaschke *et al*., 2000; Li *et al*., 2014). More recently, deep learning methods have been developed to overcome these drawbacks and utilise contextual data to extract the boundaries between 1) glaciers/ice shelves and ocean in Antarctica (Baumhoer *et al*., 2019), and 2) marine-terminating outlet glaciers and mélange in Greenland (Mohajerani *et al*., 2019; Zhang *et al*., 2019) (see Chapter 2.3). While these methods are incredibly useful for extracting glacier terminus outlines and quantifying fluctuations over time, they rely on binary classifications and perhaps overlook the ability of deep learning methods to create highly accurate multi-class outputs (i.e., not just ice and no-ice areas).

Detecting multiple semantic classes in a marine-terminating glacial landscape in combination with terminus position delineation may provide a greater holistic understanding of processes and interactions controlling outlet glacier behaviour. This would be particularly useful considering the wide range of processes occurring at the interfaces between ice, ocean, and atmosphere, which vary on both local and regional scales (Csatho *et al*., 2014; King *et al*., 2018; Catania *et al*., 2020). By capturing multiple classes in a landscape, the outputs could be used to quantify changes in a specific class over a range of timescales. For instance, to monitor changes in the area and extent of mélange (Foga *et al*., 2014; Moon *et al*., 2015; Cassotto *et al*., 2015), which has been found to impact the advance and retreat of marine-terminating outlet glaciers at seasonal timescales (Howat *et al*., 2010; Carr *et al*., 2013; Todd and Christoffersen, 2014). Similarly, classifying water with and without icebergs may help elucidate spatial and temporal patterns of iceberg flux and the resulting impacts on fjord water properties and circulation (Moon *et al*., 2018). This is important because changes in fjord water properties may influence the temporal and spatial distribution of submarine melt on outlet glaciers and have potential implications for glacier retreat (Moon *et al*., 2018;

Motyka *et al*., 2011). Alternatively, in the same way that terminus position delineation has relied on detection of class boundaries in previous work (e.g., Baumhoer *et al*., 2019; Zhang *et al*., 2019), multi-class outputs would permit monitoring of changes in other class boundaries. For example, to detect changes in snowline position and quantify ablation area changes (Noël *et al*., 2019). Additionally, classification at the scale of overall landcover types would allow the isolation of 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). Thus, multi-class outputs could provide the opportunity to monitor several glacial processes concurrently and understand how they interact in relation to outlet glacier behaviour at the scale of an entire landscape.

The use of deep learning in glaciology is still in its infancy (Figure 1.1), but given the abundance of available satellite imagery, it could be a significant aid in the automation of image processing of marine-terminating glacial settings. 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). However, image classification of entire marine-terminating outlet glacier environments has not yet been tested using deep learning. Apart from the clear potential to reduce labour-intensive manual methods, it could facilitate automated analysis in numerous research areas. In other words, aside from terminus delineation, a method which quickly produces accurate multi-class image classifications of complex and seasonally variable marine-terminating outlet glacier environments could provide an efficient and holistic way to further elucidate processes such as calving events, mélange evolution, subglacial plumes, and supra-glacial lakes at high temporal resolution. The compatibility of deep learning image classification methods with platforms such as GEE (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).

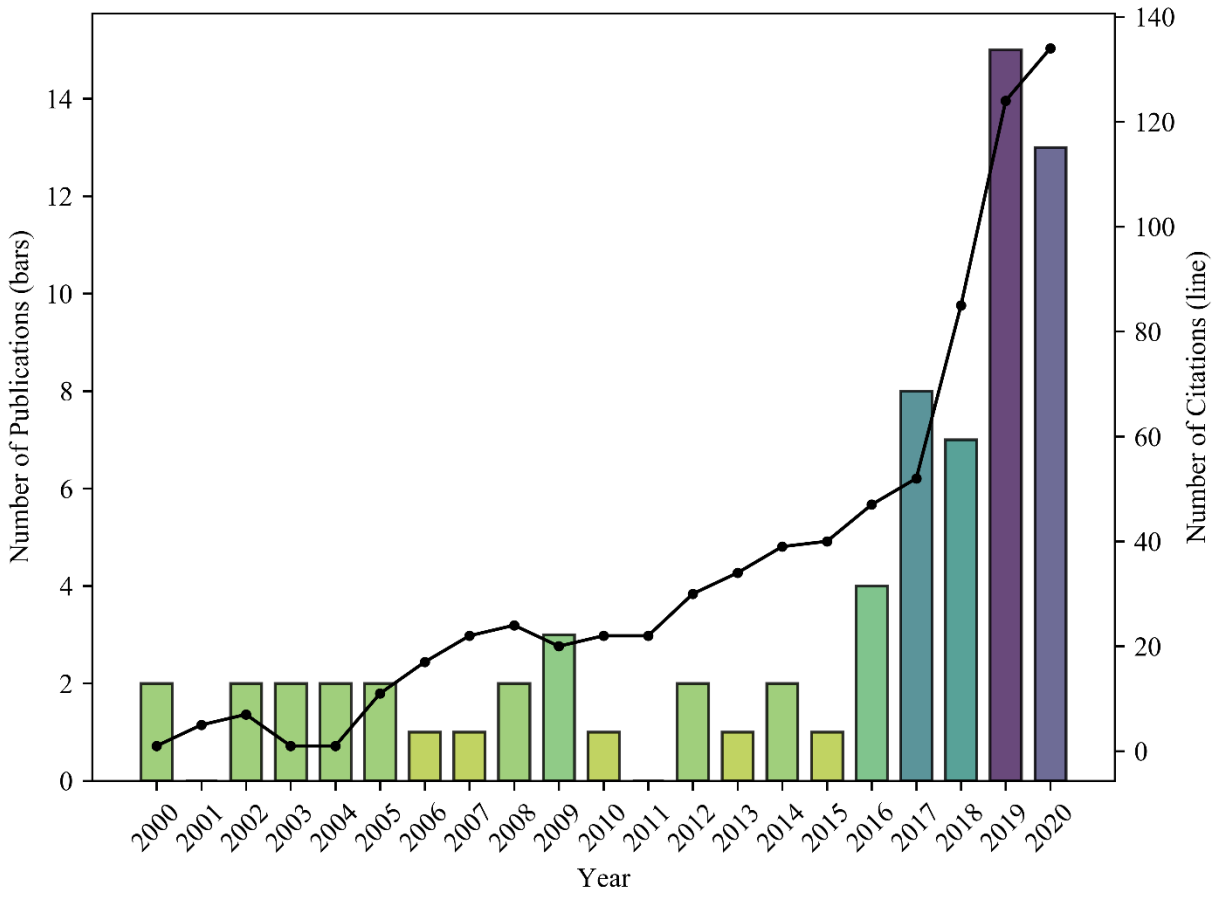


Figure 1.1: The number of published studies (bars) and citations (line) of research relating to deep learning in glaciology. Publications were identified using a systematic search of related terms included within publication titles. These terms contained the following key words/phrases: "Deep learning", "Neural Network(s)", or "CNN(s)", and "Glacier(s)", "Glacial", "Ice Sheet(s)", "Ice Shelf", "Sea-ice", "Ice Front(s)", or "Calving Front(s)". Data were obtained from the Web of Science Core Collection on 1st December 2020.

1.4 Thesis Aims and Objectives

This project aims to ***establish and evaluate a deep learning workflow for multi-class image classification of marine-terminating glacial environments in Greenland which can be accessed and used rapidly without having specialised knowledge of deep learning or the need for time-consuming generation of new training data.*** To achieve this, the following objectives were devised:

- To adapt a deep learning method developed in fluvial settings for airborne image classification and test it on satellite imagery of marine-terminating outlet glaciers in Greenland.
- To overcome problems associated with seasonal variability/spectral similarity in imagery by including seasonally variable model training data.
- To assess the sensitivity of the workflow to different band combinations, training techniques, and model parameters.
- To provide a preliminary evaluation of the spatial transferability of the workflow by applying it to unseen marine-terminating glacier environments in SE Greenland.
- To exceed the current state-of-the-art and advance accuracy levels (F1 scores >90%) for pixel-level image classification of glacial environments which contain complex marine-terminating outlet glaciers.

1.5 Thesis Outline

This chapter has outlined the importance of ice sheets and marine-terminating outlet glaciers, specifically in relation to quantification of glacier change for increased understanding of processes operating at multi- spatial and temporal scales. Chapter 2 appraises the relevant literature on previous methods for mapping marine-terminating outlet glaciers. Chapter 3 describes the methods, outlining a novel approach to classification of Sentinel-2 imagery containing landscapes with marine-terminating outlet glaciers in Greenland. Chapter 4 outlines the key results. Chapter 5 discusses the results in relation to the key aims and objectives, and Chapter 6 concludes the thesis. Chapter 7 provides information and links for fundamental code and data repositories. A revised version of this work has also been submitted to *The Cryosphere* which is under review at: https://doi.org/10.5194/tc-2020-310.

2 A Review of Previous Methods for Mapping Marine-Terminating Ice Fronts

2.1 Manual Digitisation

Manual digitisation is the most common method used to delineate the fronts of marine-terminating outlet glaciers and ice shelves from Synthetic Aperture Radar (SAR) and optical satellite imagery. However, it generally relies on prior expertise, and its labour-intense, time-consuming nature often limits the spatial and temporal resolution of datasets (Seale *et al*., 2011). In effect, studies which use manual digitisation are prone to either: 1) a high number of (daily to monthly) measurements for a limited number of glaciers or over short observational periods (< 10 years) (e.g. Schild and Hamilton, 2013; Moon *et al*., 2014, 2015; Kehrl *et al*., 2017); or 2) a limited number of (annual, interannual or decadal) measurements over larger spatial areas or observational periods (> 10 years) (e.g. Moon and Joughin, 2008; Miles *et al*., 2013). Additionally, manual digitisation frequently necessitates pre-processing steps for image enhancement (e.g. Schild and Hamilton, 2013), especially to overcome the challenges of digitising glacier or ice shelf margins near to spectrally similar areas of mélange, sea-ice, and icebergs. This is also particularly important where glacier termini are densely crevassed (Moon and Joughin, 2008), or where there is significant shadow due to topography and seasonal variations in solar illumination (Yu *et al*., 2019).

Manual digitisation also normally requires the download, storage, and processing of large numbers of images which further restricts user accessibility. However, the recent development of tools for more efficient manual digitisation has reduced such computational demands (Lea, 2018). Lea (2018) developed the Google Earth Engine Digitisation Tool (GEEDiT) and Margin change Quantification Tool (MaQiT) to allow more rapid digitisation of glacier and ice shelf fronts, without the need to download and process satellite imagery. These tools have since been applied successfully to digitise ice front positions and evaluate glacier change in several studies (Brough *et al*., 2019; Holmes *et al*., 2019; Tuckett *et al*., 2019; Amaral *et al*., 2020). For example, Brough *et al*. (2019) assessed the retreat of Kangerlussuaq Glacier in East Greenland using the GEEDiT and MaQiT tools, though problems arose where glacier ice could not be differentiated from areas of mélange. Similarly, in some cases manual digitisation has been combined with relatively simple automated approaches to increase the efficiency of terminus delineation. For example, Miles *et al*. (2017) used an automated classification method to map around 65% of terminus positions from Envisat Advanced Synthetic Aperture Radar (ASAR) imagery for outlet glaciers in Porpoise Bay, East Antarctica. Areas of glacier ice and sea-ice were classified using a threshold based on pixel statistics, and the boundary between classes was extracted

as the glacier terminus position. Nevertheless, variability in backscatter characteristics resulting from glacier surface melt during the austral summer impeded the automated method and prompted the remaining terminus delineations to be obtained manually.

While sufficient levels of accuracy can be achieved using manual digitisation, irrespective of data or sensor type (Baumhoer *et al*., 2018), several factors can impact accuracy, including georeferencing error, user bias, and the spatial resolution of imagery. In general, errors in manual digitisation range from approximately 0.5 to 2 pixels (i.e., on the order of tens of metres) (e.g. Miles *et al*., 2018). This small margin of error is usually deemed insignificant in relation to the large size of outlet glaciers (i.e., of the order of kilometres) monitored over decadal timescales in Antarctica and Greenland (Miles *et al*., 2013, 2018). However, this level of uncertainty becomes acutely important when monitoring outlet glacier change at higher temporal resolutions (i.e., over seasonal, and annual timescales). This necessitates efforts to develop consistent, automated tools to map the complex marine-terminating outlets of ice sheets. Indeed, transferable methods which produce results with comparable accuracy to manual digitisation, independently of seasonal variations or spectrally similar surface types, would be beneficial for high resolution analysis of outlet glaciers and ice shelves.

2.2 Image Segmentation and Edge Detection

Prior to the recent development of automated deep learning methods, most approaches for semi-automated digitisation of marine-terminating ice fronts have relied on image segmentation and edge detection techniques. 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. Edge detection identifies areas in an image with abrupt changes in pixel brightness, for example between glacier ice and darker areas of water and iceberg rich water, or lighter areas of mélange, presenting a foundation for boundary delineation in satellite data (Chen and Hong Yang, 1995). These image processing methods have been applied to both SAR and optical satellite data to delineate the marine-terminating margins of the AIS and GrIS (Sohn and Jezek, 1999; Liu and Jezek, 2004a; Seale *et al*., 2011; Krieger and Floricioiu, 2017; Yu *et al*., 2019) (Table 2.1). Extracting terrestrial ice sheet margins generally involves segmenting images into areas of ‘ice’ and ‘no-ice’ to create a binary classification. This is usually followed by applying edge detection algorithms, either to the binary image or directly to satellite data (Table 2.1) to highlight ice margin pixels.

Table 2.1: Selected previous studies which developed (semi-) automated techniques to extract the boundaries between land-based ice and water (including mélange, sea-ice, and icebergs) at the edges of the Greenland and Antarctic ice sheets.

Study and Area	*Imagery and Year(s) of Extraction*	*Summary of Methods*		
		Pre-processing	***Processing***	***Post-processing***
Sohn and Jezek, 1999 Western GrIS (Jakobshavn Glacier)	SPOT, ERS-1 (SAR) *1988, 1992*	• Images geocoded • Edge enhanced, and texture images created	• Segmentation for binary image classification using local dynamic thresholding • Noise removal • Region growing • Edge detection	• Removal of edge segments below a certain length
Liu and Jezek, 2004a, 2004b Antarctica	Radarsat-1 (SAR) *1997*	• Orthorectification • Noise and speckle reduction • Edge enhanced image created	• Segmentation for binary image classification using adaptive thresholding • Region growing • Class labelling • Removal of noisy objects • Edge detection	• Editing and manual correction of erroneous segments • Segments merged
Seale et al., 2011 Eastern GrIS (32 glaciers)	MODIS (optical) *Seasonal measurements between 2000-09*	• Image cropped to glacier front and rotated • Images with significant cloud cover, sensor noise, or missing data removed	• Sobel edge detection and brightness profiling algorithm applied • Peak frequencies identified from Gaussian distribution	• Automated removal of erroneous glacier front points using filter
Krieger and Floricioiu, 2017 North-eastern GrIS (Zachariæ Isstrøm Glacier)	Sentinel-1, TerraSAR-X (SAR) *2016*	• Images geocoded and sampled to 10 m spatial resolution	• Canny edge detection algorithm	
Yu et al., 2019 Antarctica	Sentinel-1, ENVISAT (SAR), Landsat 7 and 8 (optical) *2005, 2010, 2017*	• Geocoding, backscatter calibration, and terrain correction of SAR imagery • Landsat 7 SLC failure mitigation • Landsat 8 images mosaiced	• Image noise reduction using smoothing filter • Image gradient calculations • Canny edge detection with adaptive thresholding	• Noise removal with median filter • Geographic coordinates assigned to edge pixels • Segments merged and smoothed

In Greenland, Sohn and Jezek (1999) applied image segmentation and edge detection methods using SAR imagery for automated delineation of the glacier terminus and surrounding ice sheet limits at Jakobshavn Glacier. In pre-processing steps, the SAR imagery was geocoded using a digital elevation model (DEM), and the product was used to derive both edge-enhanced, and texture images. To segment the images and produce a binary classification, local dynamic thresholding algorithms were applied. Local dynamic thresholding allowed images with small physical variations to be classified successfully. Nonetheless, the lack of testing over larger spatial areas or study sites provided no indication of its transferability (Baumhoer *et al*., 2018). Thresholding was followed by noise removal and a region growing algorithm to produce more continuous class outlines. Finally, for the extraction of an ice sheet outline, the Roberts edge detection algorithm (Pratt, 1978) was applied to the binary image in combination with an algorithm to remove noisy edge segments below a specified length. Later, Seale *et al*. (2011) applied an edge detection algorithm directly to Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery, for seasonal observation of terminus change at 32 marine-terminating outlet glaciers along the eastern margin of the GrIS. Again, this involved a series of pre-processing steps, including 1) image cropping to within a specified width of the terminus, 2) cloud classification and removal of cloudy or noisy images, 3) conversion of coordinate projections, and 4) image rotation for consistent glacier flow direction. To extract the glacier terminus outlines, the Sobel edge detection algorithm (Sobel and Feldman, 2015) was applied and was followed by removal of erroneous results. In contrast to Sohn and Jezek (1999), Seale *et al*. (2011) applied this workflow to 32 glaciers, suggesting an increased level of spatial transferability. More recently, Krieger and Floricioiu (2017) applied the Canny edge detector (Canny, 1986) directly to SAR imagery for automated terminus delineation of Zachariæ Isstrøm Glacier in Northeast Greenland. However, this technique was only tested on one glacier and its transferability was not evaluated.

Similarly in Antarctica, Liu and Jezek (2004a) used image segmentation and edge detection techniques to extract the boundaries between ice/land and water classes for the whole AIS using an orthorectified SAR image mosaic. To automate the process, they applied a series of algorithms for pre-processing, segmentation, and post-processing (Figure 2.1). The pre-processing stage consisted of reducing noise in the data and applying an anisotropic diffusion operator to preserve prominent edges in the imagery (edge enhancement) (Figure 2.1b). In the segmentation stage, a series of steps were applied to segment the image and classify it into areas of land/ice and water, using local adaptive thresholding and the application of the Canny edge detector. Several further algorithms were applied in the post-processing stage.

This allowed removal of misclassified areas to reduce noise, for example, by changing misclassified water areas (i.e., rock, snow, frozen lakes, and radar shadows) to ice/land, and changing areas misclassified as ice/land (i.e., sea-ice, icebergs, and islands) to water (Figure 2.1d and e). Finally, an edge tracing algorithm was used to produce vector outlines which were corrected for errors and merged (Figure 2.1f). This resulted in an outline of the AIS, including land-based ice, rock, and ice shelves from SAR data collected during September and October 1997.

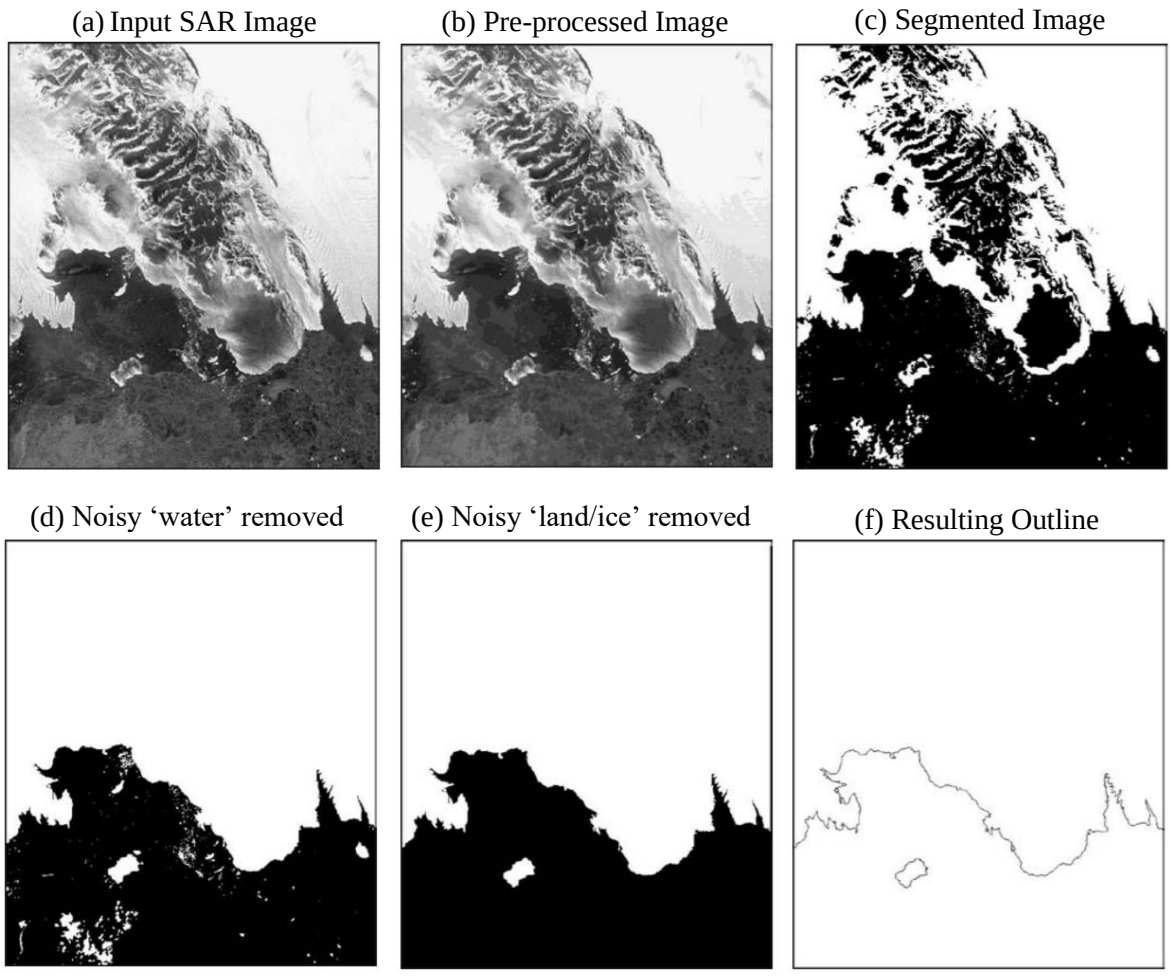


Figure 2.1: The outputs of each processing step in Liu and Jezek (2004a) which used image segmentation and edge detection methods to extract the boundaries between ice/land and water in Antarctica. a) Input SAR image. b) Pre-processed image (noise removal and edge enhancement). c) Image segmented using locally adaptive thresholding. d) and e) Removal of small noisy objects; and f) resulting vector outline extracted. Modified from Liu and Jezek, 2004a.

Yu *et al*. (2019) updated this with AIS outlines for 2005, 2010, and 2017 using Landsat 7 and 8 imagery as well as SAR data with the Canny edge detection algorithm (Figure 2.2). As in Liu and Jezek, 2004a, post-processing steps were applied to remove noise and merge segments extracted from different images.

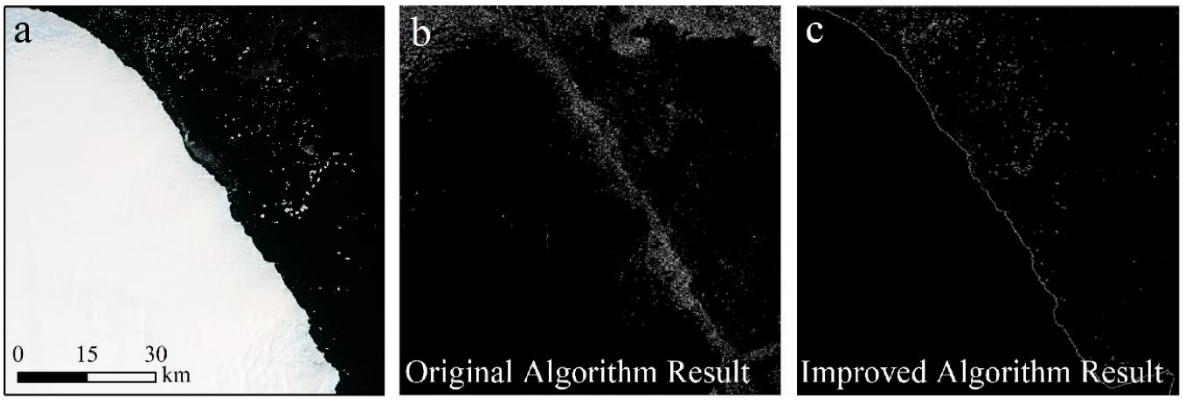


Figure 2.2: Outputs of edge detection techniques applied in Yu *et al.* (2019), showing a) the input Landsat 7 image (acquired 25/01/2017), and a comparison of edge detection outputs for b) fixed (maximum and minimum) thresholds; and c) adaptive thresholds (applied to extract the coastline of Antarctica for 2005, 2010, and 2017). Notably, adaptive thresholding produces less noisy outputs compared to the fixed thresholding technique. Source: Yu *et al.*, 2019.

Despite the acceptable levels of accuracy achieved using these automated techniques, they have a series of limitations, perhaps explaining the general preference for manual digitisation. For example, thresholding and region growing are common techniques used in image segmentation and are useful for creating continuous edges (e.g. Sohn and Jezek, 1999; Liu and Jezek, 2004a, 2004b; Yu *et al.*, 2019). However, they generally require numerous, time-consuming, processing steps. For instance, Sohn and Jezek (1999) produced seven different image products before arriving at a final outline between ice/land and water. Similarly, Liu and Jezek (2004a) produced a series of image derivatives before producing final outlines, including 1) an image with noise/speckle reduction and enhanced edges, 2) a binary classification image, 3) an image with noisy 'water' objects removed, and 4) an image with noisy 'land' objects removed (Figure 2.1). In contrast, edge detection techniques may require fewer steps but are more likely to produce discontinuous boundaries (Liu and Jezek, 2004b). They therefore require computationally expensive post-processing steps to remove insignificant edge segments and merge edges which represent ice fronts (e.g. Seale *et al.*, 2011).

In general, image rotation, cropping, edge enhancement, and noise removal are commonly required pre-processing steps when applying these techniques (Table 2.1). Indeed, the noisy nature of SAR data has resulted in a heavy reliance on noise removal steps during pre- and post-processing stages (e.g. Yu *et al.*, 2019). Despite this, erroneous detection of edge segments has been noted to occur, particularly in areas where sea-ice or mélange is close or connected to glacier and ice shelf fronts (Liu and Jezek, 2004a; Yu *et al.*, 2019). This has also necessitated the use of filtering algorithms and manual correction in most studies (Sohn

and Jezek, 1999; Liu and Jezek, 2004a, b; Seale et al., 2011; Yu et al., 2019). As a result, the often numerous and time-consuming processing steps required to use image segmentation and edge detection, which also frequently rely on specialized knowledge and expertise, reduce the transferability of existing automated methods for glacier and ice shelf margin delineation.

Moreover, like manual digitisation, spectral similarity and seasonal variability in the physical environment can cause difficulties when applying these methods. For example, it can be challenging to differentiate between ice shelves or glacier ice and spectrally similar areas of icebergs, mélange, and landfast or drifting sea-ice using these methods. Additionally, aside from wind roughening in SAR imagery and cloud cover in optical imagery, variations in snow melt, sea-ice formation, and iceberg cover can impact the backscatter and spectral reflectance characteristics of satellite imagery. Variability between images and within individual classes directly impacts techniques such as thresholding. For example, Sohn and Jezek (1999), Liu and Jezek (2004a), and Yu *et al*. (2019) applied adaptive thresholding across images instead of using fixed thresholds (Figure 2.2). This was due to different levels of image contrast, for instance resulting from changes in water roughness, ice surface deformation and snow cover properties (Yu *et al*., 2019). Thus, the data dependency of automated thresholding techniques and local adaptations potentially reduces transferability to new images, time periods, or study areas. Thus, while providing increased levels of automation, workflows based on image segmentation and edge detection for ice front delineation have not successfully overcome all the problems associated with manual digitisation.

The adaptive thresholding method was also primarily used to produce binary classifications, removing the opportunity to extract information beyond ice/water boundaries. Indeed, Sohn and Jezek (1999) note that using multiple classes may elucidate other important processes occurring in complex glacial environments (i.e. for ice contact lakes and outwash plains), while potentially improving boundary delineation. Thus, using methods which produce multi-class outcomes with meaningful class descriptions may allow a more holistic approach to quantification of glacier change.

In summary, the development of (semi-) automated techniques which apply image segmentation and edge detection methods have advanced efforts towards more efficient mapping of marine terminating glaciers and ice shelves. However, they still require numerous processing steps and expertise, without necessarily overcoming the delineation problems resulting from seasonal variations and spectral similarity within and across images.

In contrast, recent advances in the field of deep learning (described in section 2.3.1) and the ability of deep learning methods to create temporally and spatially transferable multi-class outputs provides a new avenue to combat these challenges and build on existing automated methods.

2.3 Deep Learning

2.3.1 Overview of Deep Learning and Convolutional Neural Networks (CNNs)

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). While the field of deep learning has been evolving since the 1940s (Goodfellow *et al*., 2016), the discipline has experienced significant advances over the past few decades alongside computer vision. This has resulted from the increasing availability and size of training datasets, and the improvement of computer hardware and software (LeCun *et al*., 2015). Numerous fields have helped shape the development of contemporary deep learning, including contributions from neuroscience, engineering, and fundamental mathematical principles such as probability theory (see Goodfellow *et al*., 2016 for a detailed review).

Several of the earliest designs of deep learning architectures were inspired by, and attempted to replicate, learning procedures in the mammalian brain, whereby layers of computational 'neurons' interact to acquire knowledge from an input (Goodfellow *et al*., 2016). For example, Fukushima (1980) developed a neural network for pattern recognition in images called the neocognitron. The model was based on the organisation of neurons used for visual perception, elucidated by early studies of the visual system in cats (Hubel and Wiesel, 1962). It was designed to correspond to the ventral stream of the visual cortex which processes a retinal image using a hierarchy of cells from the eye to the primary visual cortex (V1), visual areas V2 and V4, and the inferotemporal (IT) cortex (Hubel and Wiesel, 1962; Serre, 2013). Neurons in each progressive level of the hierarchy can identify increasingly complex features ranging from simple edges in the V1 visual area to complex combinations composing entire patterns and objects in the IT visual area (Felleman and Van Essen, 1991). Alongside this, neurons in higher stages of the hierarchy are shown to be increasingly tolerant to small changes in the scale and position of input images (Serre, 2013). This increase in image processing and neuron invariance represented by progressive layers in the visual hierarchy was also a key inspiration for the convolutional and pooling layers in the more recent CNN (LeCun *et al*., 1989, 1998).

CNNs are deep learning models specifically designed to process multiple two-dimensional (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 shape and texture, in the same way a human operator would. CNNs are usually arranged in a series of layers containing convolutional, non-linearity, and pooling functions (LeCun *et al*., 2015). The input data is converted into an array of features (called a feature map) in each convolutional stage using a locally weighted sum which represents an array of parameters (weights) adjusted by the model learning algorithm (Goodfellow *et al*., 2016). Initial convolutional layers learn low-level features such as lines and edges which compose the high-level features extracted by deeper convolutional layers, allowing the model to extract textures and shapes representative of image classes (Cheng *et al*., 2017). The outputs pass through a non-linear activation function such as the rectified linear unit (ReLU) (which allows the network to learn complex data by non-linear transformation) and then go through a pooling layer to introduce some invariance to the features, meaning the model can detect features with small variations such as differences in orientation (Goodfellow *et al*., 2016). There are typically several of these stages in a CNN, creating a hierarchy similar to that of the mammalian visual system, allowing the model to learn features from an image and output a prediction of class for each pixel. As a result of this, one of the main benefits of CNNs is that they remove the need for prior feature extraction or thresholding for image classification (Längkvist *et al*., 2016). The CNN used for image classification in this study falls into the category of supervised learning (Goodfellow *et al*., 2016). This means the CNN is trained using labelled pixels and tested based on its ability to predict the class of pixels in unseen imagery. The ability of a model to accurately predict the class of pixels in an unseen image is called generalisation (Goodfellow *et al*., 2016) and determines the transferability of the model.

CNNs were popularised in 2012 when Krizhevsky *et al*. (2012) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a CNN called AlexNet. They have since been applied to a broad range of disciplines, improving tasks in object detection (Zhao *et al*., 2019), speech recognition (Abdel-Hamid *et al*., 2014), and numerous medical imaging applications (Lundervold and Lundervold, 2019). They are also increasingly being used for a variety of remote sensing applications (Buscombe and Ritchie, 2018), including classification of fluvial scenes (Carbonneau *et al*., 2020a), land-use classification (e.g. Luus *et al*., 2015), and automated detection of geological features on Mars (Palafox *et al*., 2017).

2.3.2 Deep Learning for Automated Delineation of Marine-Terminating Ice Fronts

In glaciology, 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). The application of deep learning models in workflows for automated delineation of marine-terminating glacier termini and ice shelf fronts has also been effective, resulting in accuracy comparable to conventional manual methods (Baumhoer *et al*., 2019; Mohajerani *et al*., 2019; Zhang *et al*., 2019).

For example, Mohajerani *et al*. (2019) used a type of CNN architecture called a Fully Convolutional Neural Network (FCN) to classify ice front pixels and non-ice front pixels in Landsat imagery containing marine-terminating outlet glaciers in Greenland. The previous success of FCN architectures trained on small datasets with the help of augmentation methods justified its use for application in marine-terminating outlet glacier environments, where training and validation data production relies on manual digitisation. The FCN architecture was trained using Landsat 5 (green band), 7, and 8 (panchromatic bands) imagery. A series of pre-processing steps were applied to the imagery to improve the FCN performance, including cropping (to within 300 m of the terminus), rotation, normalisation, grey-scale intensity equalisation, smoothing, and edge enhancement (Figure 2.3). Dataset augmentation was applied to increase the number of training samples by flipping each image. Images were also plotted on 200 x 300-pixel grids with ice flow in the y direction. Therefore, instead of using the original Landsat spatial resolution (15/30 m), this resulted in images with different resolutions for each glacier, and consequently errors were also dependent on different spatial resolutions for each study site.

Collection of training and validation data involved manual digitisation of terminus positions, which were rasterised into pixel-wide lines to train the model. Due to the small proportion of the images inhabited by the rasterised terminus outline (Figure 2.3), the FCN was particularly prone to class imbalance, whereby high accuracy could be obtained by simply excluding the terminus outline class. Therefore, custom sample weights were applied to avoid this issue. In post-processing steps, fjord boundaries were manually digitised in order to apply a least-cost path method for extraction of the pixels which most likely represent the terminus position. This workflow achieved similar levels of accuracy to manual digitisation, with most error noted to occur at the edges of the glacier termini. Mean distance from manually digitised fronts was 96.3 m for Helheim glacier (the test glacier).

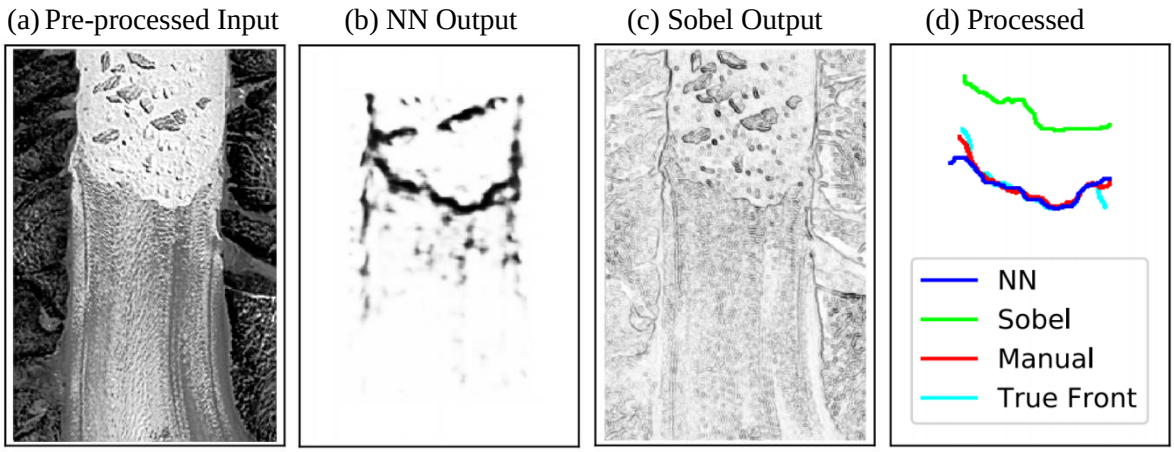


Figure 2.3: The results of the deep learning method developed by Mohajerani *et al.* (2019) showing b) the classified output of the modified FCN model compared to c) results from the Sobel edge detection method when applied to a) a pre-processed satellite image. d) Shows the processed delineations of both methods in addition to manually derived terminus positions. Adapted from Mohajerani *et al.* (2019).

Similarly, Zhang *et al*. (2019) also used an FCN to extract terminus positions from TerraSAR-X imagery of Jakobshavn Glacier, central west Greenland. A total of 159 images from 2009-2015 were used to train the model, which classified images into 'mélange' and 'non mélange' areas. Image pre-processing involved speckle reduction, multilooking, and georeferencing. Images were also subdivided into 960 x 720 pixel tiles, and edge enhanced, normalised, and augmented (flipping and rotation) before model training. In post-processing, the binary classification was converted to vector format and small, erroneous polygons were removed before terminus extraction. The transferability of this approach was not tested as it was only applied to one study site. Indeed, the use of 'mélange' and 'non-mélange' classes also suggests it can only be applied to glaciers with mélange adjacent to the terminus. However, its ability to classify multitemporal data suggested it overcame problems with seasonal variations across imagery. Overall, the technique resulted in a mean difference of 38 m from manually delineated terminus positions.

Finally, Baumhoer *et al*. (2019) used an FCN to classify the boundaries of land-based ice and ocean in Antarctica (Figure 2.4). Baumhoer *et al*. (2019) trained the FCN using different SAR polarisations derived from Sentinel-1 data in combination with a TanDEM-X DEM. Image pre-processing involved applying the Orbital File to SAR data, thermal noise removal, radiometric calibration, geometric terrain correction using the DEM, and stacking of HH, HV, HV/HH polarisations with the DEM. A total of 38 pre-processed images from four training sites were tiled into 780 x 780 pixel samples, normalised and augmented (rotation

and flipping) to train the model. Like Zhang *et al*. (2019), resulting class predictions were binary and consisted of land ice and ocean areas. The resulting classifications were filtered and vectorised prior to extraction of class boundaries which represented glacier terminus and ice shelf outlines (Figure 2.4). The use of several training sites and testing of the model on four areas suggests it is also transferable to other areas in Antarctica, with mean deviations from manually digitised fronts of 108 m in test areas. In terms of classification accuracy, Baumhoer *et al*. (2019) achieved mean F1 scores of 89 to 90% for training sites and 90 to 91% for test sites (Greenland-based studies did not provide classification F1 scores).

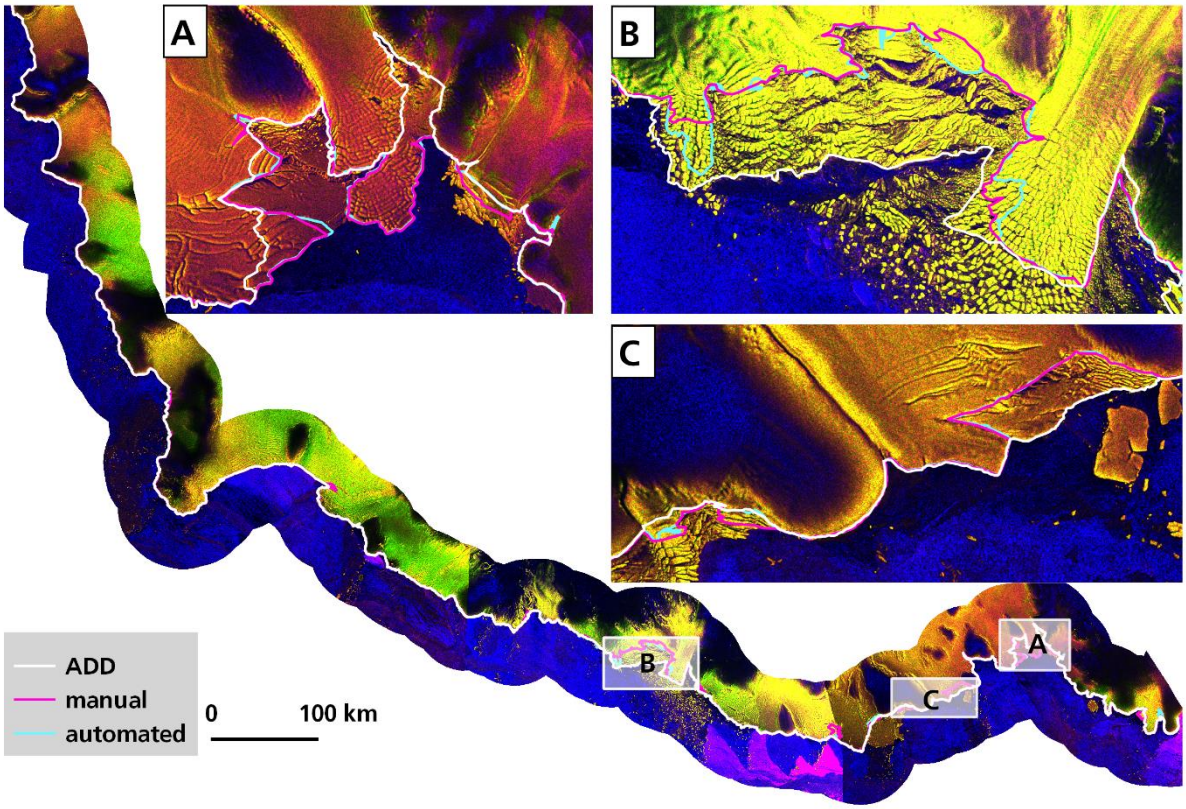


Figure 2.4: Results of the deep learning method developed by Baumhoer *et al*. (2019) for extraction of the boundaries between ice/land and water in Antarctica, showing comparisons between the deep learning method (automated), manual digitisation (manual), and Antarctic Digital Database (ADD) delineations. Insets (a-c) show magnified sections of Marie Byrd Land where methods deviated significantly. Background imagery: Sentinel-1 scenes acquired from 18/06/2018 – 23/06/2018. Source: supplementary materials of Baumhoer *et al*. 2019.

In summary, all three deep learning methods applied a fully convolutional architecture which was adapted to create binary classifications. Mohajerani *et al*. (2019) used the glacier terminus front itself as the primary class, producing results with a similar appearance to edge detection methods (Figure 2.3). Meanwhile, Zhang *et al*. (2019) and Baumhoer *et al*. (2019) applied an FCN to produce similar results to automated segmentation methods, whereby they

classified images into two classes and extracted the boundary between classes as the terminus position (Figure 2.4). The deep learning methods are promising due to their transferability across seasons (Baumhoer *et al*., 2019; Zhang *et al*., 2019) and spatial areas (Baumhoer *et al*. 2019), which is especially important for mapping complex marine-terminating outlet glaciers at high temporal resolution. However, there is substantial potential to widen the scope of deep learning methods for classification of marine-terminating glacial environments. Such advancements include producing deep learning workflows with multi-class outputs that could be used in a variety of applications, without numerous pre-processing steps or the need for specialised prior experience. Indeed, the methods presented below aim to deliver a deep learning workflow for multi-class outputs with simple pre-processing steps and the capacity to accurately detect spectrally similar surface types, using only limited training data composed of three to four optical satellite image bands. Moreover, the deep learning workflow adapted here is trained and tested on outlet glaciers in south east Greenland with a pre-defined set of image classes. In future work the workflow may be applicable to mapping outlet glaciers in other regions of the GrIS and elsewhere in the world, dependant on further testing, suitable adaptations to training data inputs and additional fine-tuning.

3 Methods

3.1 Introduction

This chapter explains the steps and data involved in developing a deep learning workflow for classification of imagery containing marine-terminating outlet glaciers in Greenland. In summary, the workflow is composed of two deep learning phases (Carbonneau *et al*., 2020a). First, a well-established CNN called VGG16 (Simonyan and Zisserman, 2015) was modified and trained using labelled image tiles from 13 seasonally variable Sentinel-2 images of Helheim Glacier, south east Greenland (Figure 3.1). In the first phase of the workflow, an unseen image from an outlet glacier environment is tiled, and the pre-trained CNN is applied to detect the class of each tile in the image. The resulting class predictions are then used as training data for a second pixel-level model which is specific to the unseen input image. In phase two, the second deep learning model uses the class predictions of the phase one CNN and input image features to determine a final pixel-level classification. The methods developed here are primarily tested on marine-terminating outlet glaciers in SE Greenland, providing a preliminary test of transferability. To determine whether the method is applicable for classifying marine-terminating glaciers elsewhere in Greenland, a larger number of test sites from different regions of the GrIS would be required. Similarly, since the pre-trained CNN was only trained on Helheim Glacier, additional training data would be required to classify landscapes with significantly different characteristics, for example to classify glacial landscapes in Antarctica.

3.2 Study Areas

3.2.1 Training Area: Helheim Glacier, SE Greenland

The area chosen to train the phase one CNN in the deep learning workflow spans 68.8 x 37.2 km (Figure 3.1c) and includes Helheim Glacier (66.4° N, 38.8° W), a major outlet of the south-eastern GrIS. 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) (Figure 3.1c). Inspection of available satellite imagery reveals that the area of mélange varies seasonally

with monthly variations in extension and composition (Andresen *et al*., 2012, 2013). For example, observations from February through to April 2019 show that the area of mélange was relatively small and consisted primarily of sea-ice, with fewer large icebergs in comparison to later months. Fjord waters were also dominated by sea-ice in various stages of development with few icebergs. From May through to August 2019, the mélange area expanded to cover a larger proportion of the fjord surface and its composition became dominated by icebergs, reflecting a change to iceberg-dominant fjord waters and a reduction in sea-ice. 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 as previously observed (Straneo *et al*., 2011).

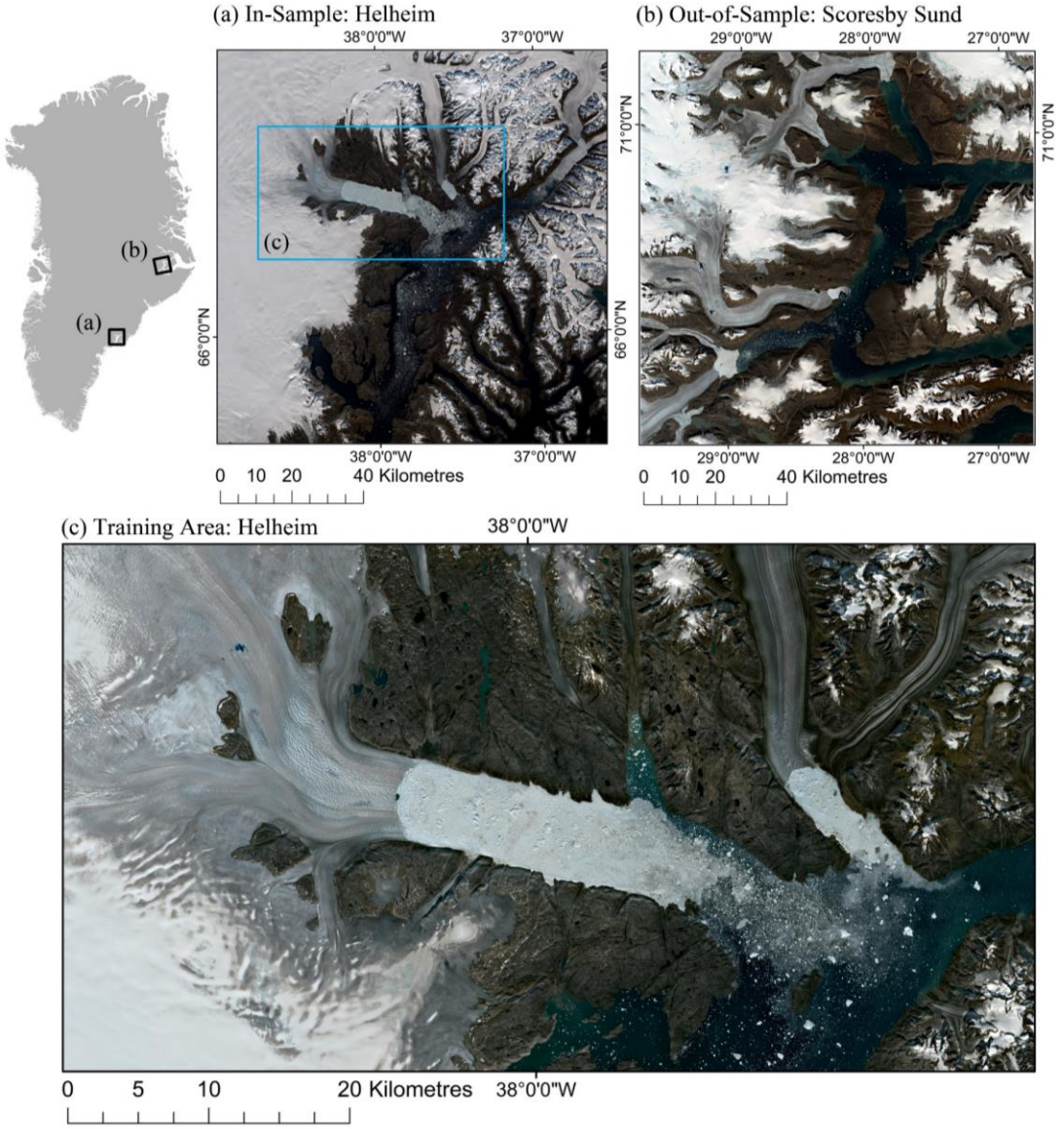


Figure 3.1: Location of outlet glacier environments used for training and testing the deep learning workflow. (a) Sentinel-2 tile of Helheim Glacier (acquired 13/09/2019) used for testing the workflow (in-sample), with inset which shows the specific area used to create training data. (b) Sentinel-2 tile of Scoresby Sund area (acquired 01/08/2019) used for testing the workflow (out-of-sample). (c) Model training area (acquired 07/08/2019). Note the substantial area of ice mélange and an active plume at the terminus of Helheim Glacier.

The glacier, fjord, and surrounding landscape provide an ideal test 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. These characteristics include 1) seasonal variations in the degree of surface meltwater ponding 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. It is worth noting that since some elements of marine-terminating outlet glacier landscapes are not abundantly represented within the Helheim training area (e.g., off-glacier vegetation, or medial moraines), further testing and fine-tuning of the workflow with inclusion of representative training data would be required to classify imagery containing these elements.

3.2.2 Test Areas: Helheim Glacier and Scoresby Sund, SE Greenland

The deep learning workflow was trained at Helheim Glacier and then tested on two areas (Figure 3.1a and b) using: 1) a previously unseen Sentinel-2 tile of Helheim Glacier and the surrounding landscape, acquired on 13/09/2019 (in-sample), and 2) a Sentinel-2 image of the glacial landscape in the area of Scoresby Sund, ~600 km north of Helheim, which features several smaller outlet glaciers and was acquired on 01/08/2019 (out-of-sample). This area was chosen as an ideal test site because it encompasses all the classes used in model training (including mélange which is not always present at glacier termini). Both unseen Sentinel-2 tiles used for testing were divided into nine smaller image tiles spanning 3000x3000 pixels, resulting in 18 test images for processing by the deep learning workflow.

3.3 Imagery

Remote sensing studies which apply deep learning to image classification usually use high resolution (sub-metre) imagery (Sharma *et al.*, 2017) and typically require large datasets (Krizhevsky *et al.*, 2012). Acquiring high resolution imagery of outlet glacier landscapes can

be expensive and challenging, especially over large spatial areas. Therefore, the abundance of widely available medium resolution satellite imagery (10 - 60 m), often used for remote sensing applications in glaciology, provides an ideal data source for training and testing the deep learning workflow. Here Sentinel-2 bands 2 (blue), 3 (green), 4 (red), and 8 (Near Infrared (NIR)) at 10 m spatial resolution were used to train and test the approach (Table 3.1). The red, green, and blue bands were chosen because they are commonly used in image classification with deep learning architectures such as VGG16, making existing, pre-trained, models easily transferable for the purpose of this study. The NIR band was chosen due to its common use in remote sensing of glacial environments, for example in band ratios to automatically identify glacier outlines (e.g. Alifu *et al*., 2015).

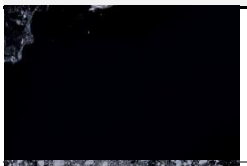
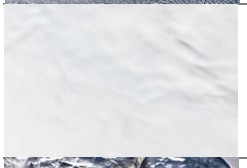
Table 3.1: List of Sentinel-2 images used for training and testing the deep learning workflow.

	Study Area	*Acquisition Date*	*Satellite*
Training	**Helheim**	08/02/2019	S2A
		10/02/2019	S2B
		07/03/2019	S2A
		10/03/2019	S2A
		15/03/2019	S2B
		04/04/2019	S2B
		29/05/2019	S2A
		15/06/2019	S2A
		05/07/2019	S2A
		07/08/2019	S2A
		01/09/2019	S2B
		28/09/2019	S2B
		26/10/2019	S2A
Testing		13/09/2019	S2A
	Scoresby	01/08/2019	S2A

Examination of available Sentinel-2 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 detailed criteria for each in Table 3.2). 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 taken between February and October 2019 were acquired (Table 3.1). Level-2A images were downloaded at no cost from Copernicus Open Access Hub (available at: https://scihub.copernicus.eu/dhus/#/home, last accessed: 20/07/20). The atmospherically

corrected red, green, blue and NIR bands were combined into composite four band images and cropped to the training area (Figure 3.1c). Two Sentinel-2 tiles of the unseen Helheim (Figure 3.1a) and Scoresby Sund (Figure 3.1b) study areas were also acquired (Table 3.1), and the corresponding composite band images were created.

Table 3.2: Example image samples and descriptions of each of the seven semantic classes used to train and validate the phase one convolutional neural networks in the deep learning workflow. Total number of tiles refers to the total number of tiles used for training and validation in each of the three datasets used to test model sensitivity to tile size after the tiling process described in Figure 3.5. Note that the open water, mélange, and bedrock classes have the smallest representation of all classes, despite the aim of producing equally represented class samples.

Example Image of Class	*Class Number and Label*	*Class Description*	*Total number of Tiles*		
			50x50 (total: 354,668)	**75x75** (total: 319,292)	**100x100** (total: 293,720)
	**1 Open Water**	Open water with no icebergs	14,312	12,024	10,520
	**2 Iceberg Water**	Water with varying amounts of icebergs or disintegrated mélange	48,668	44,084	41,212
	**3 Mélange**	Mixture of sea-ice, and icebergs of varying sizes	25,540	23,396	21,192
	**4 Glacier Ice**	Glacier ice, with seasonally variable surface meltwater	84,356	77,584	71,040
	**5 Snow on Ice**	Snow/ice with a smooth appearance	88,412	79,540	77,004
	**6 Snow on Rock**	Bedrock with varying amounts of snow cover	63,180	55,052	47,588
	**7 Bedrock**	Bedrock with no snow cover	30,300	27,612	25,164

3.4 Classification Workflow, Model Architectures and Training

3.4.1 CNN-Supervised Classification (CSC)

The classification workflow used here is termed CNN-Supervised Classification (CSC), and was originally developed and tested on airborne imagery (<10 cm resolution) of fluvial environments (Carbonneau *et al.*, 2020a). CSC is a novel two-phase workflow (Figure 3.2) which uses a pre-trained CNN to replace the human operator's role in labelling training areas for final pixel-level classification. In the first phase of the workflow, a pre-trained CNN is used to predict the classes of a tiled input image. The image tiles are then reassembled to create a class raster which is used as training data for the second model in phase two of the workflow. In the second phase, the reassembled class raster and image features are vectorised and used to train a second model specific to the input image. The predictions of the second model result in a final pixel-level classified image output (Figure 3.2).

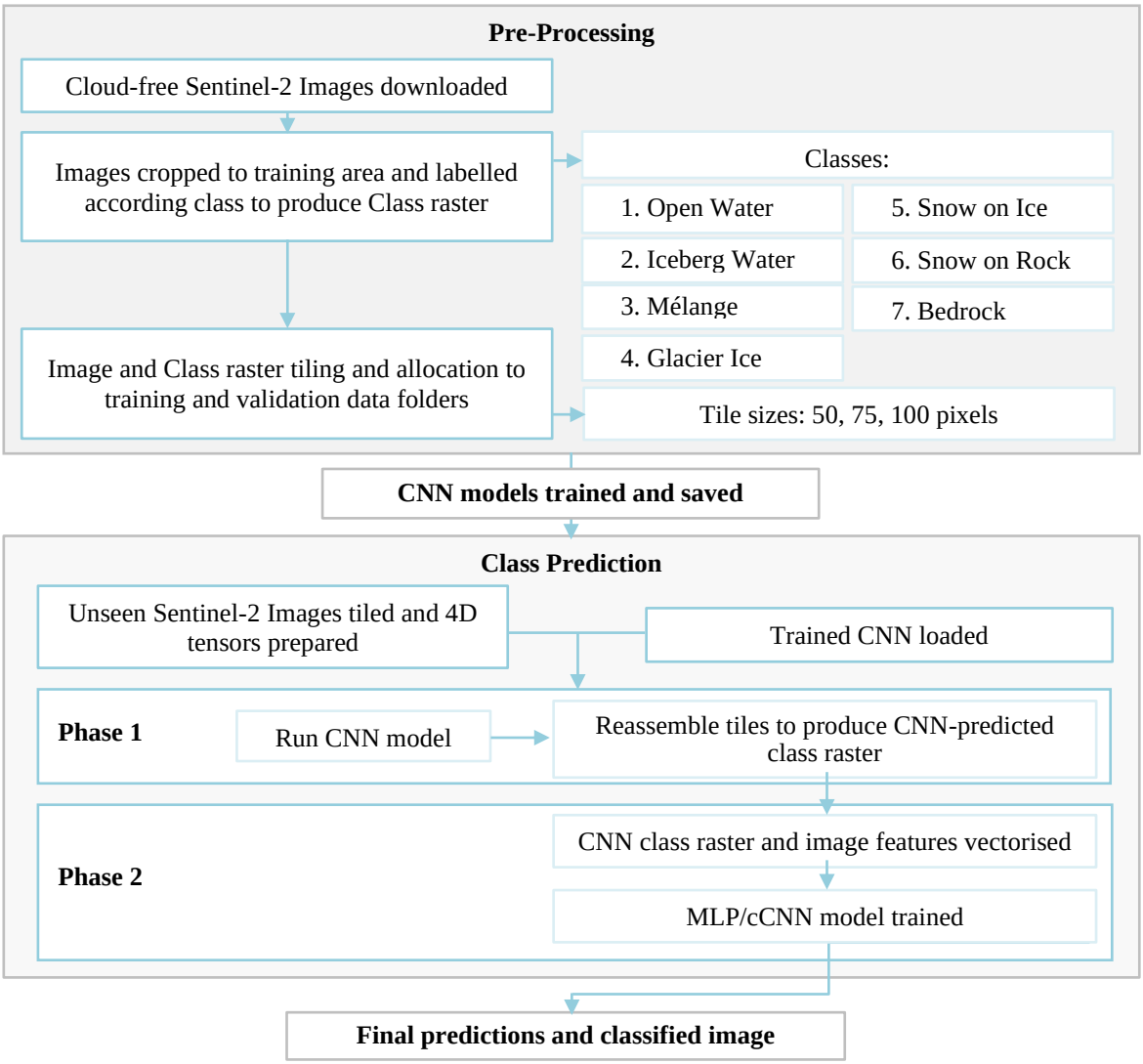


Figure 3.2: Image classification workflow showing pre-processing steps, convolutional neural network training and 2-phase final classification steps.

3.4.2 Phase 1: Model Architecture and Training

For the base architecture of the pre-trained CNN used in phase one, a well-established CNN called VGG16 (Simonyan and Zisserman, 2015) which outperformed the state-of-the-art performance of AlexNet in the ILSVRC 2014 was adapted. The VGG model used consists of five stacks of 13 2D convolutional layers which have filters with a size of 3x3 pixels (Figure 3.3). A filter is an array of numbers (which are also known as weights). The filter spatially convolves over the input image to create a feature map using the filter weights. For example, if we have a single band input image of 7x7 pixels, a 3x3 filter would convolve across each available pixel within the 7x7 image and produce a 5x5 pixel feature map. The CNN learns input features to detect the classes in an image by adjusting these filter weights in each convolutional layer (Goodfellow et al., 2016). In the VGG model used here, the dimensions of the output filters increase from 64 in the first stack of convolutional layers to 512 in the last (Figure 3.3). So, in the first convolutional layer, since there are 64 filters, this produces 64 individual feature maps which become the input to the next convolutional layer, and so on. This allows a hierarchy of features to be detected in deepening convolutional layers.

All the convolutional layers in the VGG base use ReLU activation and are interspersed with five max-pooling layers. ReLU is a conventional and computationally efficient non-linear activation function which allows non-linear transformation of the input data to make it separable for classification. The pooling function reduces the size of each feature map to make outputs more computationally manageable while retaining important information (Goodfellow et al., 2016). 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. L^2 regularization was used in this top neural network to reduce overfitting, which occurs when a model is unable to generalize between training and test data (Goodfellow *et al*., 2016). Adam gradient-based optimisation, a common optimisation algorithm used in deep learning, was also used to update the weights in the network (Kingma and Ba, 2017). This fully connected neural network 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 Softmax layer allocates the outputs of the CNN to a set of normalised probability scores. In effect, each input image is assigned a probability score for each class, so the final class label for the image is that which has the highest probability of membership (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 the impact of tile size was tested by using three datasets with different tile sizes of 50x50, 75x75, and 100x100 pixels. Thus, the input image size was adjusted so it matched the three tile sizes (Figure 3.3 shows an example of an input tile size of 100), and the number of input channels was also adjusted depending on the number of image bands used for training (i.e., three or four). Each of these image tiles was fed into the phase one CNN in the form of a four-dimensional (4D) tensor which contains multiple tiles (Dimensions: [tiles, x, y, bands]).

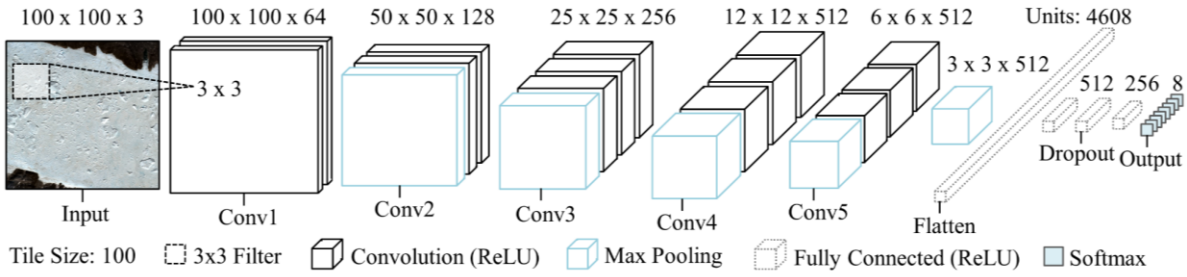


Figure 3.3: Architecture of phase one convolutional neural network, adapted for three tile size datasets from the original VGG16 model architecture (Simonyan and Zisserman, 2015). Diagram shows an example using a tile size of 100 pixels. There are five stacks of 2D convolutional layers (labelled 'Conv#') which extract features from input tiles using a 3 x 3 filter. The convolutional stacks are followed by a fully connected neural network and Softmax activation for final class predictions used as localised training data for phase two models.

Three approaches for training the phase one CNN were tested using the three image tile datasets to test the sensitivity of each approach to tile size, resulting in a total of nine trained CNNs. The three approaches of model training were as follows: 1) only three image bands (RGB) were used; 2) the NIR band was used in addition to the three RGB bands (RGB+NIR), and 3) three image bands (RGB) were used in combination with transfer learning (RGB+TL). The transfer learning approach trained the model using pre-existing weights from the ImageNet database which contains over 14 million labelled images (Deng *et al.*, 2009). Only the weights in the final layers of the CNN were re-trained specifically to classify glacial scenes, making it quicker to train than standard full CNN architectures (Buscombe and Ritchie, 2018). Transfer learning has been shown to decrease training time and reduce the volume of data needed to produce similar levels of accuracy to non-transfer learning techniques (Kunze *et al.*, 2017). As a result, with a tile size of 100 the transfer learning model had 9,572,616 trainable parameters of a total 17,207,880 trainable parameters if the VGG16 model was trained without transfer learning, and weights in all layers were adjusted. For

each of the nine models, training hyperparameters were kept constant, with training occurring over 15 epochs, with a batch size of 50 images and a learning rate of 0.0001. Following training of the phase one CNNs, they were saved for application on unseen images in phase two without further training.

3.4.3 Phase 2: Model Architectures and Training

To classify airborne imagery of fluvial scenes using the CSC workflow, Carbonneau *et al*. (2020a) applied a pixel-based approach using a multilayer perceptron (MLP) in the second phase of the workflow, achieving high levels of accuracy (90-99%). This project proposes 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. Therefore, a patch-based approach was adopted, 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) used 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 a single convolutional layer 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 both pixel- and patch-based approaches were tested using an MLP and cCNN in the second phase of the workflow (the architectures and application of which are detailed in the following sections 3.4.3.1 and 3.4.3.2). Specifically, four patch sizes of 1x1 (pixel-based), 3x3, 7x7, and 15x15 pixels were tested. In combination with the phase one CNNs using different tile sizes and bands, this resulted in the testing of 36 model workflows overall which were subsequently tested on in-sample and out-of-sample test images.

3.4.3.1 Multilayer Perceptron (MLP)

For the pixel-based classification in phase two, an MLP was used (Figure 3.4a). 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 first and final layers of an MLP are called the input and output layers, respectively. The layers in between are 'hidden layers' used to apply weights to the input data, which is then fed forward to units in other hidden layers (Atkinson and Tatnall, 1997). The MLP used here has five layers consisting of four fully connected (dense) layers and one batch normalisation layer (Figure 3.4a). 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 by adjusting the activations in the network to add noise. This is followed by two more dense layers with 32 and 16 filters, respectively. 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. All the layers use ReLU activation except the output layer which uses Softmax activation to produce a vector of class probability scores. For both the MLP and cCNN, model training hyperparameters were kept constant (150 epochs, learning rate of 0.001, and subsamples size of 100,000). Since the MLP is pixel-based, the number of parameters was smaller compared to the patch-based model, with 3,128 trainable parameters.

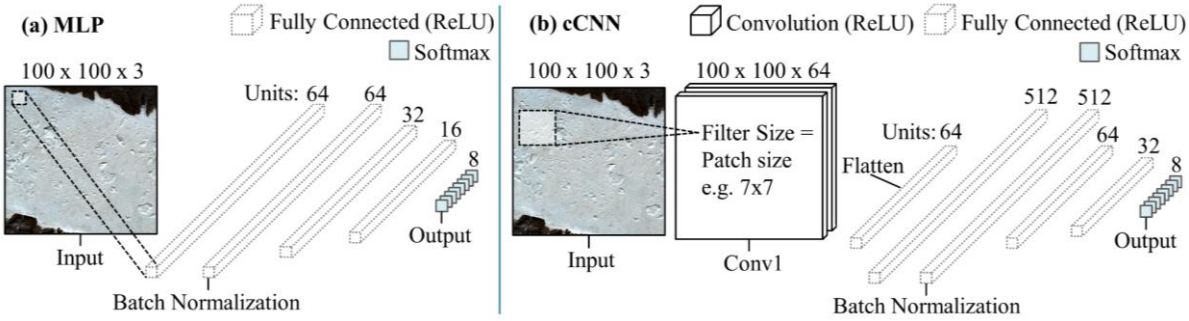


Figure 3.4: Architecture of phase two models. (a) Shows the Multilayer Perceptron used for the pixel-based classification of new input images. (b) Shows the compact convolutional neural network used for patch-based classification of new input images. The size of the filter in the cCNN changes according to the patch size being tested. For example, as shown in (b) the filter size is 7 x 7 for testing a patch size of 7 pixels.

3.4.3.2 Compact Convolutional Neural Network (cCNN)

For the patch-based classification in phase two, a cCNN was used (Figure 3.4b). This model architecture is referred to as a compact CNN (cf. Samarth *et al*., 2019) because it only contains one convolutional layer and is much smaller than conventional CNNs (Figure 3.4b). This model is comprised of a 2D convolutional input layer which extracts features from the input image using a small window of pixels called a filter. The input layer has 64 filters with a kernel (window) size which is modified dependant on patch size (i.e., for testing a patch

size of 7x7 pixels, the kernel size is 7) (exemplified in Figure 3.4b). As with the phase one CNN, the input shape is a 4D tensor with the dimension of [patches, x, y, bands]. This is followed by a flatten layer which converts the inputs into a one-dimensional feature vector to be fed into the following four fully connected (dense) layers. The first dense layer has 512 filters and is followed by a batch normalisation layer. The following three dense layers have 64, 32, and 8 filters, respectively. As with the MLP, all the layers use ReLU activation except the output layer. As with all the models used in the workflow, the final layer comprises the same number of units as output classes and results in a vector of probability scores used to predict class. The cCNN had 71,272 trainable parameters with a patch size of 3, 78,952 trainable parameters with a patch size of 7, and 112,744 trainable parameters with a patch size of 15.

3.4.4 Training and Validation Data Preparation

The CNNs used in phase one of the workflow were trained using image tiles which represent image subsamples of each individual class. These tiles were processed by the model in the form of 4D tensors consisting of multiple image bands (consistent with conventional data formatting designed for training CNNs for multiband image classification). To create training and validation data for the model, the composite images were manually labelled according to the seven training classes using QGIS 3.2 digitising tools. Vector polygons labelled by class number were rasterised to produce a class raster with the same geometry as the input image. Both the input image and class raster were then tiled using a specified size (height and width in pixels) and stride (number of pixels the window moves before extracting another tile) (Figure 3.5). Three different tile sizes were used to test model sensitivity and its ability to identify landscape features at the scale of the 10 m resolution imagery. This resulted in three datasets containing tile sizes of 50x50, 75x75, and 100x100 pixels (Table 3.2). A stride of 35 pixels was used to allow overlap between tiles, and any tiles occupied by less than 95% pure class were rejected, removing tiles containing mixed classes. The image tiles were then rotated in increments of 90° to augment the dataset and saved to separate class folders. 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). In addition to data augmentation, tile rotation 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 workflow transferability.

Each tile was normalised by 16,384 (a maximum integer value drawn from satellite imagery) to reduce bit depth to a scale from 0 to 255. This adjusts the range of pixel values to make them compatible with RGB imagery for processing by the CNN. The tiles were divided into training and validation datasets whereby 95% of tiles were randomly allocated to a training data folder and the remaining 5% were allocated to a validation data folder (Figure 3.5). It is common when training deep learning models for image classification applications to have an 80/20% split of training and validation data (Carbonneau *et al*., 2020a). However, here a 95/5% split is appropriate as the 'in-sample' data we used to test the workflow is a new satellite image of the training area and surrounding landscape, previously unseen by the model during training, making it a more stringent test. Overall, this resulted in three datasets containing 354,768 tiles of 50x50 pixels, 319,292 tiles of 75x75 pixels, and 293,720 tiles of 100x100 pixels for training and validating the phase one CNNs (Table 3.2). These datasets were extracted from only 13 cropped images of Helheim Glacier and are much larger compared to those used in previous work to train and validate CNNs for glacier boundary delineation. For example, Mohajerani *et al*. (2019) used only 123 tiles of 152x240 pixels obtained from three different glacier study sites. Baumhoer *et al*. (2019) opted for larger tile sizes and used a dataset of 19,576 tiles of 780x780 pixels derived from 38 scenes from four study sites. Finally, Zhang *et al*. (2019) used 36,414 tiles with a larger size of 960x720 pixels using 75 images from one glacier.

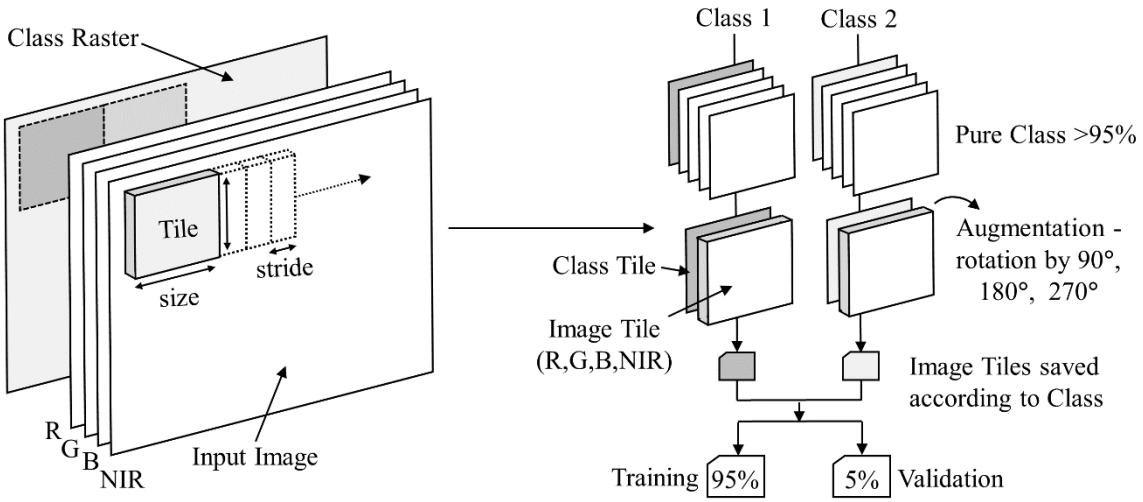


Figure 3.5: Conceptual diagram of tiling process used to create training and validation data. A specified tile size (of 50, 75 or 100 pixels) and stride (of 35 pixels) are used to extract tiles from the class raster and image bands. These tiles are filtered and augmented and saved to individual class folders using a 95/5 % split for training and validation data.

3.5 Sensitivity Analysis: Training Epochs

Since CNNs are sensitive to the number of epochs used in training, we applied the epoch tuning method used by Chollet (2017) with 95% of our data used to train the model and 5% used for validation. The term epochs refers to the iterations over training data in the CNN (Chollet, 2017). Each epoch is used to adjust weights and improve accuracy in the CNN based on training loss. Training loss is the error in CNN predictions compared to validation data and is quantified using a loss function. Categorical cross entropy was used as the loss function in all models and is a common loss function used in deep learning for multi-class classification (Goodfellow et al., 2016). The VGG16 models were run for 25 epochs, and training accuracy, training loss, validation accuracy, and validation loss for each individual epoch was saved. Similarly, the MLP and cCNN models were run for 500 epochs and the same values were saved. These were plotted against number of epochs (Figures 3.6 and 3.7). The number of epochs used to train the final set of models was then determined by the point of divergence between training and validation data. Where a gap between training and validation data appears, the model begins to overfit and its ability to generalise is reduced. The epoch tuning graph of the VGG16 model (Figure 3.6) begins to diverge slightly after 15 epochs, so the model was trained for 15 epochs for optimal accuracy and training time. The epoch tuning graphs for the phase two models revealed that the optimum number of training epochs was 150 (Figure 3.7).

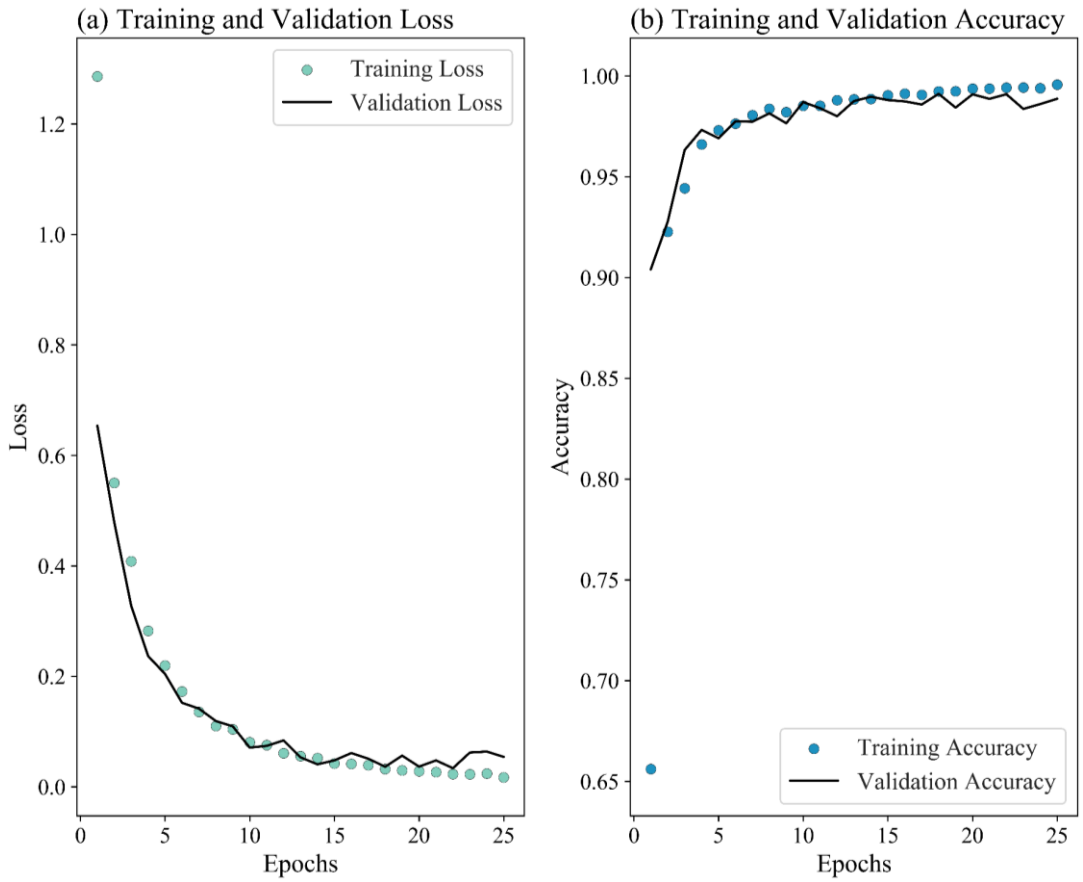


Figure 3.6: Epoch tuning graph for phase one model (trained using 50x50 pixel RGB tiles).

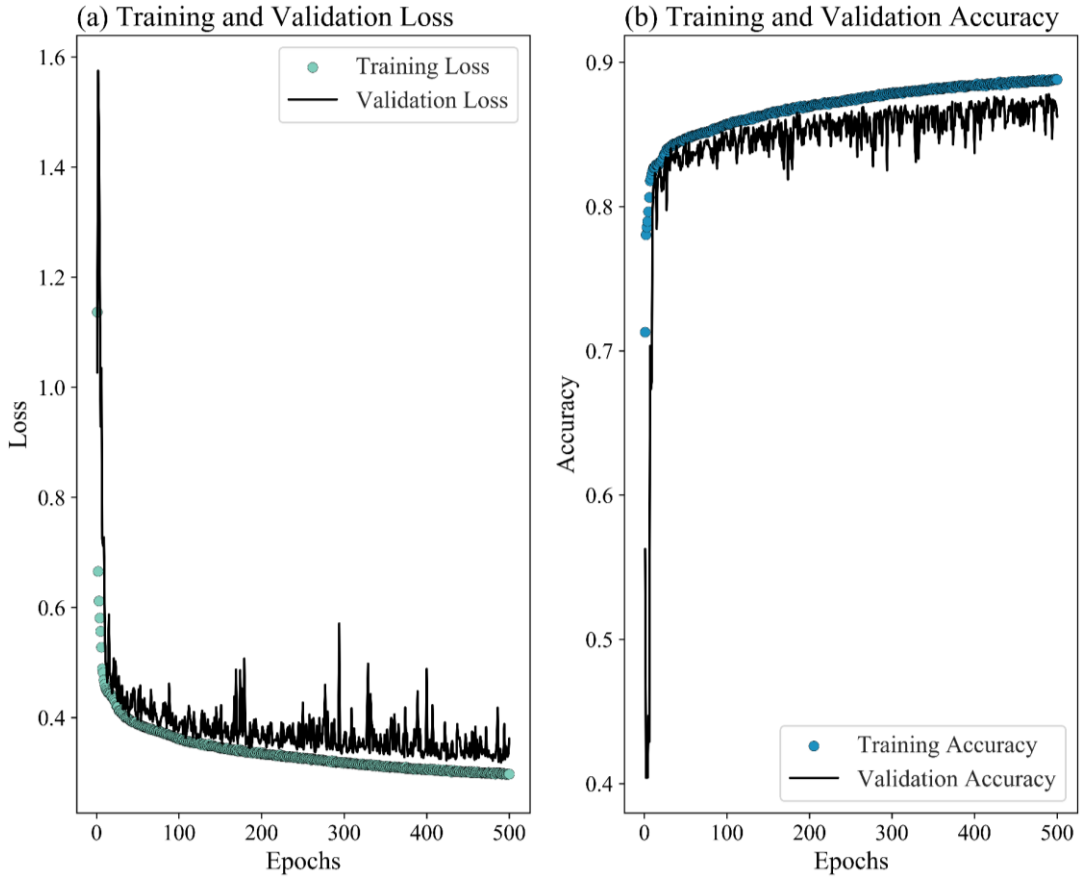


Figure 3.7: Epoch tuning graph for phase two model (using RGB model with tile size of 50 and patch size of 7).

3.6 Model Performance

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 that also take into account confusion between predicted classes (Goodfellow *et al*., 2016; Carbonneau *et al*., 2020a). This project used 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 1,724 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%. F1 scores were plotted against patch and tile sizes to show workflow sensitivity for each of the three training approaches. Confusion matrices were also plotted to show agreement between predicted classes and manually delineated validation data in the final classification outputs.

Cohen's Kappa was also used as a secondary performance metric which is a coefficient of agreement (Cohen, 1960). This compares the agreement between the model class predictions and manually determined classes (validation data). Cohen's Kappa accounts for the chance occurrence of true positives in class predictions (i.e., correctly guessing the class). It is a useful complement to metrics such as accuracy and F1 because it better reflects the performance of models with class imbalance. It removes the problem of overshadowing in prediction performance for a smaller class by that of a larger class. Cohen's Kappa is a normalised statistic, so it ranges from -1 to 1. A set of arbitrary thresholds were determined by Landis and Koch, (1977) to interpret the agreement statistic (Table 3.3).

Table 3.3: Arbitrary thresholds used to interpret Cohen's Kappa measure of agreement (Landis and Koch, 1977).

Cohen's Kappa Statistic	*Strength of Agreement*
<0.0	Poor
0.0 – 0.2	Slight
0.21 – 0.4	Fair
0.41 – 0.6	Moderate
0.61 – 0.8	Substantial
0.81 – 1.0	Almost Perfect

3.7 Comparison to Traditional Mapping Techniques

For a comparison of effectiveness between the CSC workflow, and pixel-based techniques such as band ratio methods, a test image tile of Helheim Glacier was classified using a band ratio technique. To create the band ratio image, the Sentinel-2 band 4 (red) was divided by band 11 (Shortwave Infrared) (Paul *et al*., 2016). A series of thresholds were used to classify the resulting band ratio image into three classes including glacier ice, snow on ice and bedrock. Classifying the band ratio image using all seven classes utilised in the CSC workflow was not possible. This is because the band ratio method did not detect changes between all the different classes such as mélange, iceberg water and open water. For comparison to the CSC classifications, an overall F1 score was produced for the resulting band ratio classification using the same validation labels used to produce F1 scores for the CSC classification.

4 Results

4.1 CNN-Supervised Classification

4.1.1 Performance of Phase 1 CNNs and Tile Size Sensitivity

The performance of the phase one VGG16 models in classifying unseen Sentinel-2 image tiles of the Helheim and Scoresby Sund study areas are shown in Figure 4.1. With the exception of the transfer learning model (RGB+TL) in the Scoresby Sund study area, all models produced accurate classifications (F1 Scores ≥ 88%). The best performing model on the Helheim study area was the RGB transfer learning model (RGB+TL) with a tile size of 50 pixels. The model predictions produced classifications with an overall F1 score of 93% (Figure 4.1a) and Kappa value of 0.9 (Figure 4.2). This indicates that the model class predictions are highly accurate and have almost perfect agreement with manually delineated validation data (see Table 3.3). The highest performing models for the Scoresby Sund study area were the RGB models which scored slightly lower F1 scores of 90% irrespective of tile size (Figure 4.1b). This shows that the model produces slightly improved classification performance on in-sample data compared to out-of-sample data. However, the RGB model performance on the Scoresby Sund image remains high and indicates that the phase one model is transferable to outlet glacier landscapes in SE Greenland which were not used in training.

Overall, the performance of non-transfer learning models does not appear to be greatly sensitive to tile size, with RGB and RGB+NIR models resulting in F1 scores ranging from 90 to 92% for in-sample (Helheim) data and 88 to 90% for out-of-sample (Scoresby) data. However, the transfer learning models were greatly impacted by tile size for both test areas, with tile sizes of 75 and 100 pixels producing lower F1 and Kappa scores compared to models trained with a tile size of 50 pixels (Figure 4.1a and Figure 4.2). The transfer learning models also performed substantially worse on out-of-sample data (Figure 4.1b). The addition of the NIR band in both study areas did not appear to improve classification results.

In summary, while the best performing phase one CNN for in-sample data used transfer learning, the transfer learning approach was highly sensitive to tile size and did not perform well on out-of-sample data, suggesting it is less transferable compared to non-transfer learning approaches of model training. Additionally, both models trained using RGB and RGB+NIR tiles were only slightly sensitive to tile size, but the addition of the NIR band did not improve model performance, suggesting that the RGB models are the most transferable while providing high levels of classification accuracy.

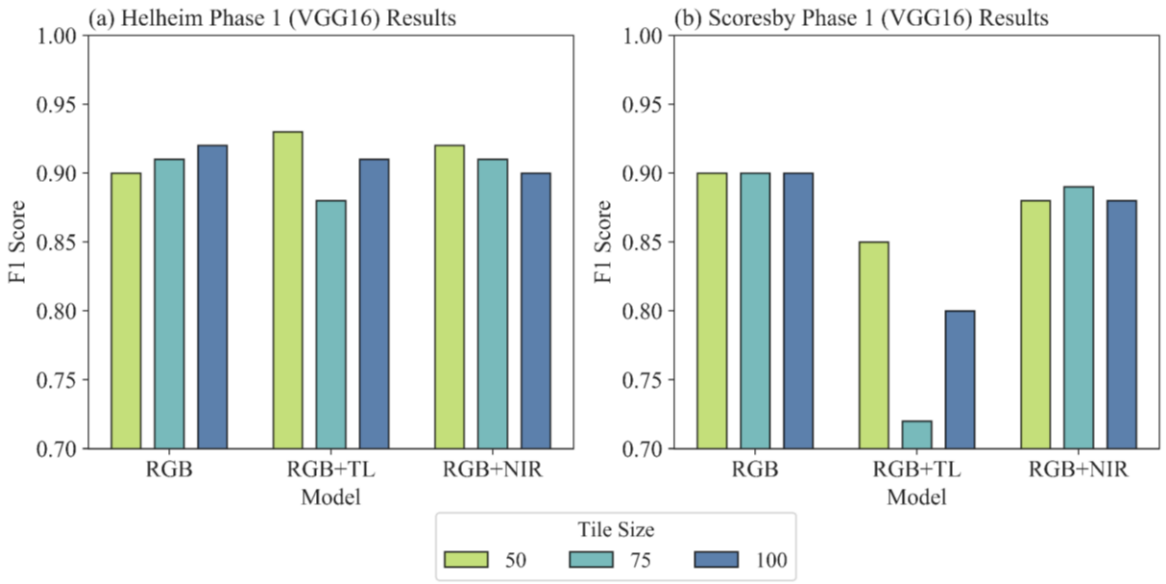


Figure 4.1: The F1 scores of the phase one (VGG16) model classifications used to produce training data for phase 2 of the CSC workflow. Showing results for (a) the Helheim test area (in-sample) and (b) the Scoresby Sund test area (out-of-sample). Note the low sensitivity of RGB and RGB+NIR models to tile size (with a range in F1 scores of 2 % for both (a) and (b)). Also note the high sensitivity of transfer learning approaches to tile size and lower transferability to out-of-sample data compared to non-transfer learning approaches.

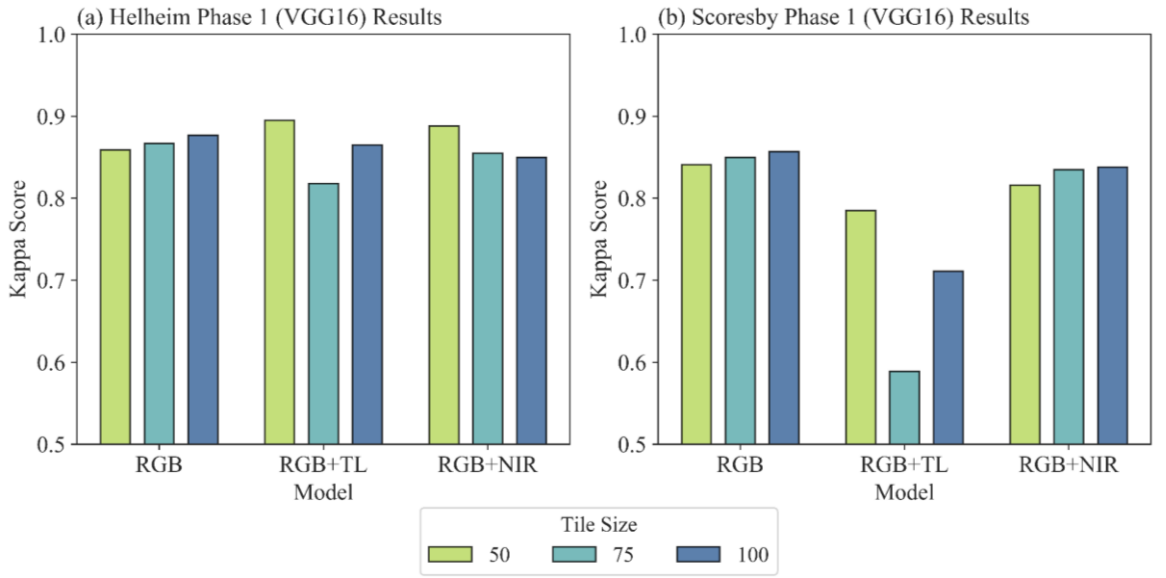


Figure 4.2: Kappa scores resulting from phase one model classification outputs. Note the relatively high kappa scores showing good agreement between model results and manually digitised truth data used for validation, with the exception of transfer learning model results on out-of-sample Scoresby data.

4.1.2 Performance of Phase 2 Models and Patch Size Sensitivity

4.1.2.1 Helheim (In-sample)

Figure 4.3 shows the overall F1 scores of the CSC (CNN + MLP/cCNN) results, demonstrating the impact of patch size. In general, the results of applying CSC to the Helheim study area showed a clear sensitivity to patch size with a patch size of 1 pixel yielding lower F1 scores and Kappa values than larger patch sizes in all models. Larger patch sizes of 3, 7, and 15 pixels either produced F1 scores consistent with phase one CNN outputs or improved upon classification performance by 1 to 2%. A patch size of 7 pixels yielded the best results in all models with the highest F1 scores of 92% in the RGB+NIR model (Figure 4.3e), 93% in the RGB model (Figure 4.3a), and 95% in the RGB transfer learning model (Figure 4.3c).

Specifically, the CSC results of the RGB models yielded F1 scores from 82 to 93% (Figure 4.3a) and Kappa values of 0.75 to 0.89 (Figure 4.4). RGB models with tile sizes of 75 and 100 pixels scored highest and had correspondingly high Kappa scores (≥0.8: see Figure 4.4). In terms of patch size, the RGB models using a cCNN patch size of 7 improved on the results of the phase one CNNs by 1%. RGB models using a cCNN patch size of 3 and 15 also performed well, either producing the same F1 score as phase one CNNs or improving classification results (Figure 4.3a).

The CSC results of the RGB transfer learning models yielded F1 scores from 84 to 95% (Figure 4.3c) and Kappa values of 0.85 to 0.92 (Figure 4.4). RGB+TL models with a tile size of 50 were highest performing with F1 scores of 94 to 95% for patch sizes of 3 to 15 pixels (Figure 4.3c). As with the RGB models, the use of a cCNN with a patch size of 7 was the best, consistently improving on phase one results by 2%.

The RGB+NIR models had F1 scores ranging from 85 to 92% (Figure 4.3e) and Kappa values of 0.77 to 0.88 (Figure 4.4). The results of phase one RGB+NIR models with a tile size of 50 were not improved by the addition of a patch-based cCNN. However, RGB+NIR models with tiles sizes of 75 and 100 and a cCNN patch size of 3 and 7 were consistent with or improved upon phase one classification results. As with the pixel-based approach, the phase two model which used a patch size of 15 did not improve phase one RGB+NIR results.

Overall, this suggests that the pixel-based CSC workflow is outperformed by the patch-based CSC workflow for in-sample classification, with a patch size of 7 pixels producing the optimal results. It also suggests that with optimal patch size, phase one model classifications are improved upon by phase two model results.

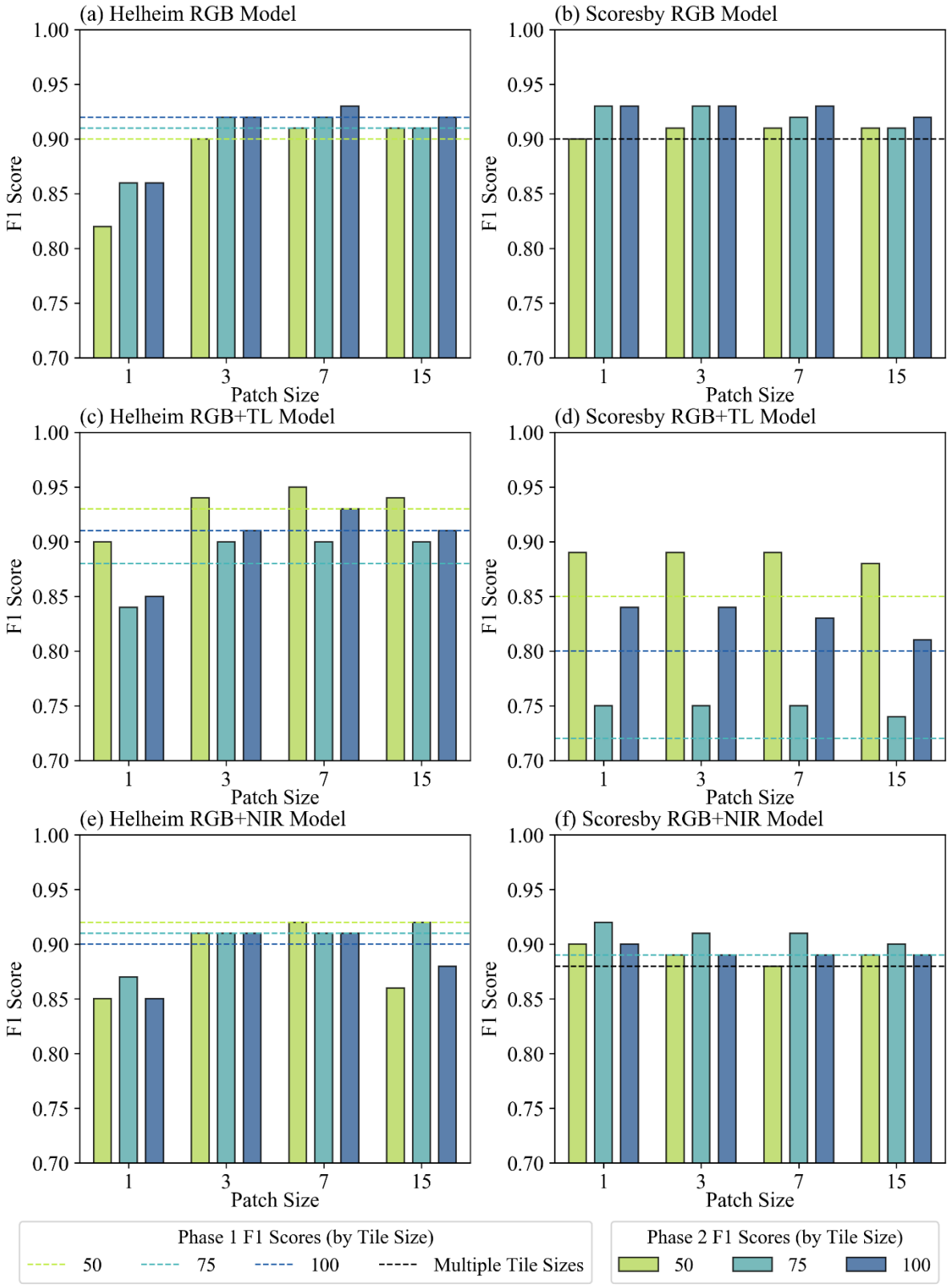


Figure 4.3: The F1 scores of the phase two classifications following the CSC workflow for the Helheim test image (a, c, e) and Scoresby test image (b, d, f). Note in some cases phase two results outperform phase one results. One prominent exception is the pixel-based approach for in-sample data. The patch-based approach performs well for in-sample data, with a patch size of 7 creating optimal results. The pixel-based approach performs better on out-of-sample data compared to in-sample data.

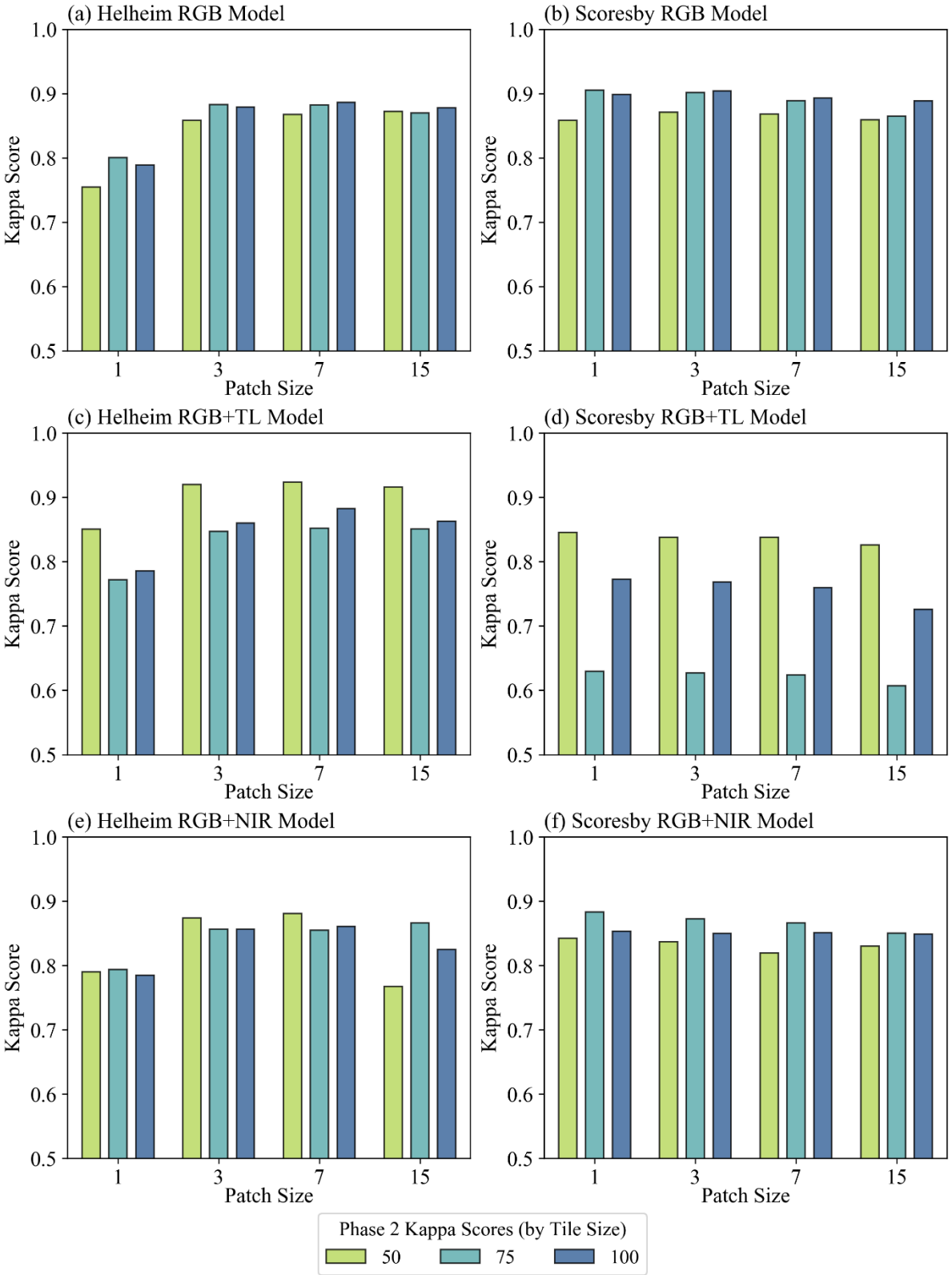


Figure 4.4: Kappa scores resulting from phase two model classification outputs. Note that kappa scores are generally similar to phase one results, with high levels of agreement for most models. Notable exceptions are the pixel-based in-sample results and out-of-sample results from the transfer learning model.

Figure 4.5 shows the in-sample CSC outputs for the best performing phase one models using RGB (Figure 4.5c), RGB+TL (Figure 4.5d), and RGB+NIR (Figure 4.5e) training approaches. All models in the figure used a cCNN patch size of 7 pixels and are applied to a 3000x3000 pixel image tile of Helheim glacier (Tile S2A5: 5 of 9 extracted from the test image). The RGB model produced an F1 score of 94% (Figure 4.5c), while the RGB model with transfer learning (Figure 4.5d) and the RGB+NIR model (Figure 4.5e) both produced F1 scores of 97%. Visual comparison between the results suggests only small variations in classification outputs, corresponding to small variations in F1 scores (on the scale of 1 to 3%). Figure 4.6 shows the confusion matrices illustrating agreement between model-predicted classes and manually delineated classes for the three best workflows shown in Figure 4.5. In all three model workflows there is excellent agreement between class predictions and manually obtained truth data, perhaps with the exception of the RGB model which shows some confusion between open water and bedrock classes.

Taken together, these results indicate that for in-sample data the patch-based (CNN + cCNN) CSC workflow produces the best results. Specifically, the best performing model used a phase two cCNN with a patch size of 7 pixels.

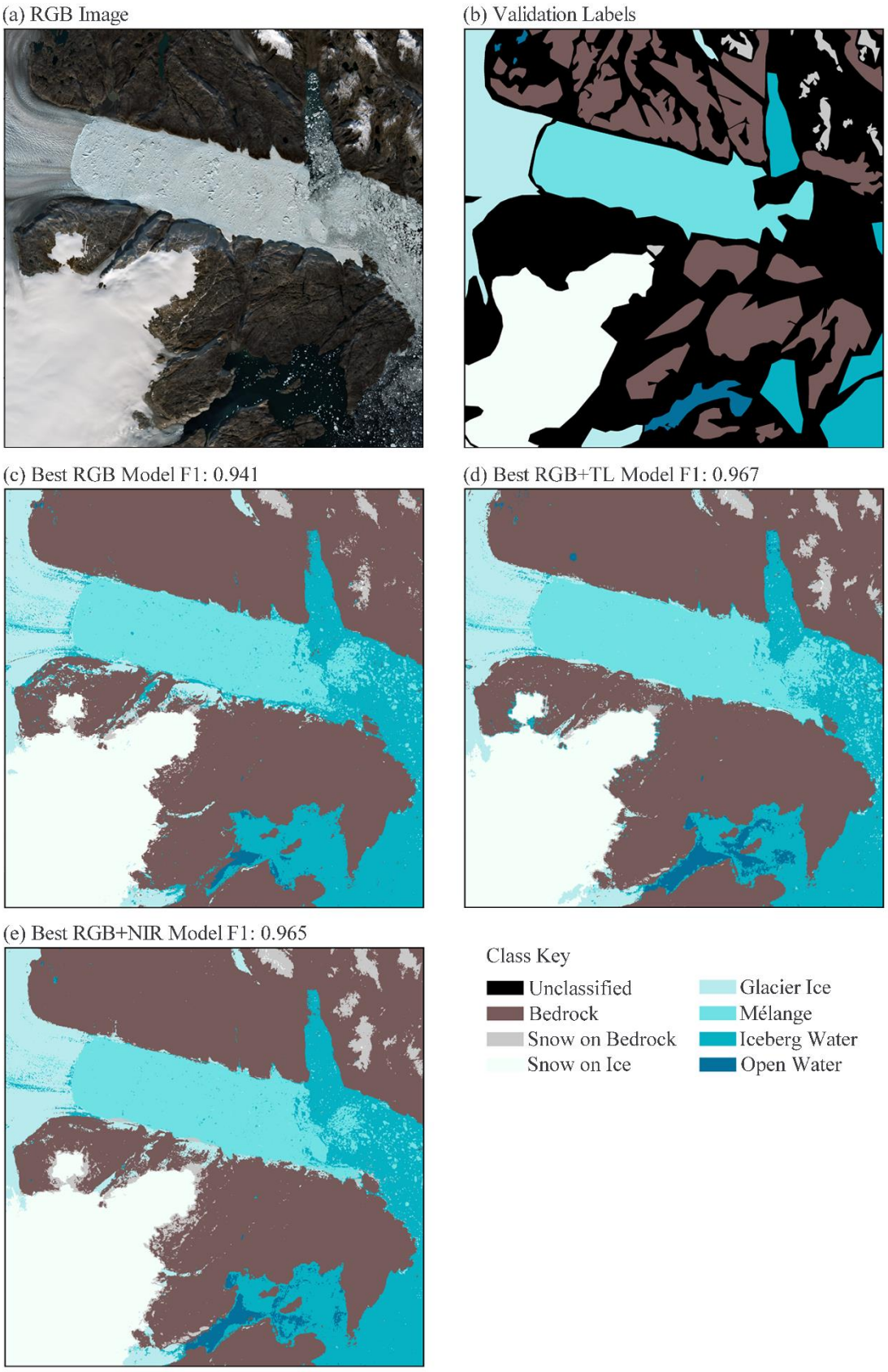


Figure 4.5: Best performing CSC results for tile 5 of 9 from the Helheim study area (07/08/2019). (a) RGB input image (composite Sentinel-2 bands 4, 3, and 2). (b) Validation raster composed of manually digitised 'ground truth' polygons. Showing workflow outputs using (c) the RGB model (tile size: 100 pixels, patch size: 7 pixels), (d) the RGB model with transfer learning (tile size: 50 pixels, patch size: 7 pixels), and; (e) the RGB+NIR model (tile size: 50 pixels, patch size: 7 pixels). Note all models produce highly accurate classification outputs with small variations between outputs and minimal noise.

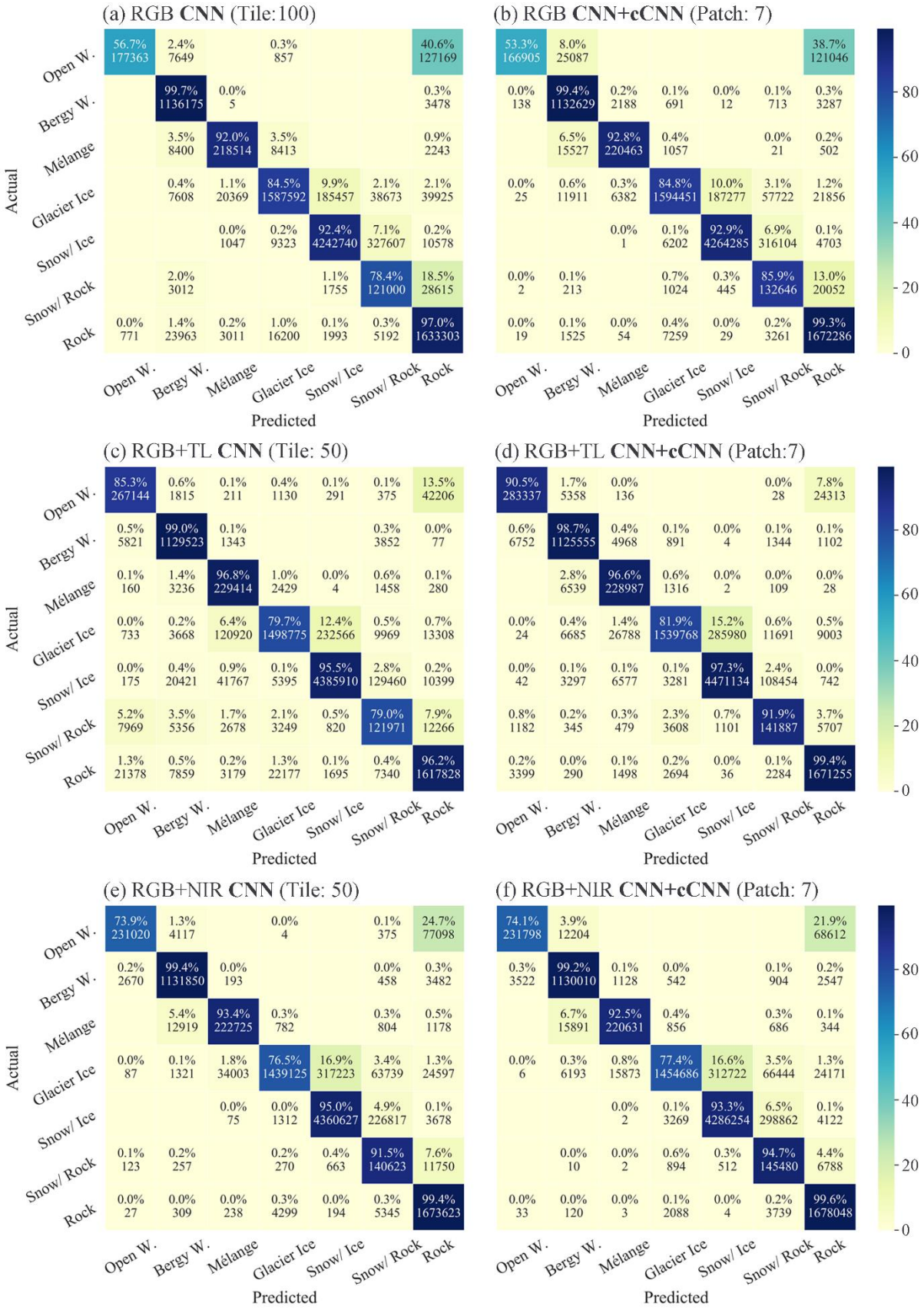


Figure 4.6: Confusion matrices for the results of the best performing in-sample models depicted in Figure 4.5. a) and b) Show the degree of class agreement for the RGB model which performs best with a tiles size of 100 and patch size of 7. C) and d) show the agreement for results of the best RGB+TL model, while e) and f) show the agreement for best RGB+NIR results.

4.1.2.2 Scoresby Sund (Out-of-sample)

In contrast to the Helheim study area, the performance of CSC in out-of-sample data was high for the pixel-based approach, in most cases with identical or improved F1 scores compared to the larger patch-sizes (Figure 4.3). For the most part, models using smaller patch sizes (1, 3, and 7) performed slightly (1 to 3%) better than those with a patch size of 15. However, it is important to note that for each of the nine individual models tested on out-of-sample data, F1 scores only varied by up to 3% for the four different patch sizes. The MLP and cCNN outputs also notably outperformed phase one classification results in most models, by up to 4% (Figure 4.3).

For the Scoresby Sund area the RGB models were the highest performing overall (Figure 4.3b), with F1 scores ranging from 90 to 93%. Kappa values also showed highest agreement for the RGB models, with a value of 0.9 for most RGB models with tile sizes of 75 and 100 (Figure 4.4). For patch size, the RGB model performed best with patch sizes of 1, 3, and 7.

While the transfer learning approach improved model performance for the Helheim study area, as with the phase one CNN, its performance in the Scoresby Sund area (out-of-sample) was substantially worse, with F1 scores ranging from 74 to 89% (Figure 4.3). The transfer learning approach was also more sensitive to tile size than other models, with a tile size of 50 pixels yielding the highest F1 scores of 88 to 89%, 75 pixels yielding 74 to 75%, and 100 pixels yielding 81 to 84%, mirroring phase one model results. However, it showed very little sensitivity to patch size. There was no variation in F1 score between patch sizes of 1, 3, and 7 for models trained on tile sizes of 50 and 75, and a patch size of 15 produced F1 scores 1% lower than smaller patch sizes. Despite this, the addition of the phase two model improved phase one transfer learning classification results consistently for the out-of-sample data.

RGB+NIR models had F1 scores ranging from 88 to 92% (Figure 4.3f) and Kappa values from 0.81 to 0.88 (Figure 4.4), with a tile size of 75 yielding the best results. As with all models tested on out-of-sample data, the RGB+NIR models showed limited sensitivity to phase two patch size, but the pixel-based approach (patch size: 1) was consistently 1 to 2% better than the patch-based approach.

Overall, these results show that out-of-sample data is less sensitive to patch size compared to in-sample data, with the pixel-based approach performing substantially better in out-of-sample data. The results also suggest that phase two model predictions are highly dependent on the quality of classification outputs resulting from phase one predictions which are subsequently used as localised training data.

Figure 4.7 shows a visual comparison of the out-of-sample CSC outputs for the best performing RGB, RGB with transfer learning, and RGB+NIR models when used on a 3000x3000 pixel image tile extracted from the Scoresby Sund area (Tile S2A8: 8 of 9 extracted from the Sentinel-2 test tile). Figure 4.7c shows the output of the RGB model, with an F1 score of 97%. The transfer learning model produced an F1 score of 89% (Figure 4.7d) while the RGB+NIR model produced an F1 score of 94% (Figure 4.7e). In the case of the example tile, most confusion appears to occur between open water and iceberg water classes (Figure 4.7).

The confusion matrices for overall results produced by the three best performing models for out-of-sample data (shown in Figure 4.7) are illustrated in Figure 4.8. The confusion between iceberg water and open water seen particularly in Figures 4.7d and e are also clear from the confusion matrices. However, a higher degree of confusion is noted to occur between snow-covered rock and bare bedrock classes. In general, the models with the lowest performance experience confusion between one or more classes (see Figures A14-24). For example, the application of the transfer learning model with a tile size of 75 (the lowest performing model) to out-of-sample data resulted in confusion between open water and iceberg water classes, as well as confusion between snow on rock and glacier ice classes (see Figures A17-20). Since phase one results are used to train phase two models, high amounts of class confusion in phase one models can be transmitted to phase two results. However, some class confusion in phase one is overcome in phase two, as indicated by the consistent improvement of phase two results over phase one classifications.

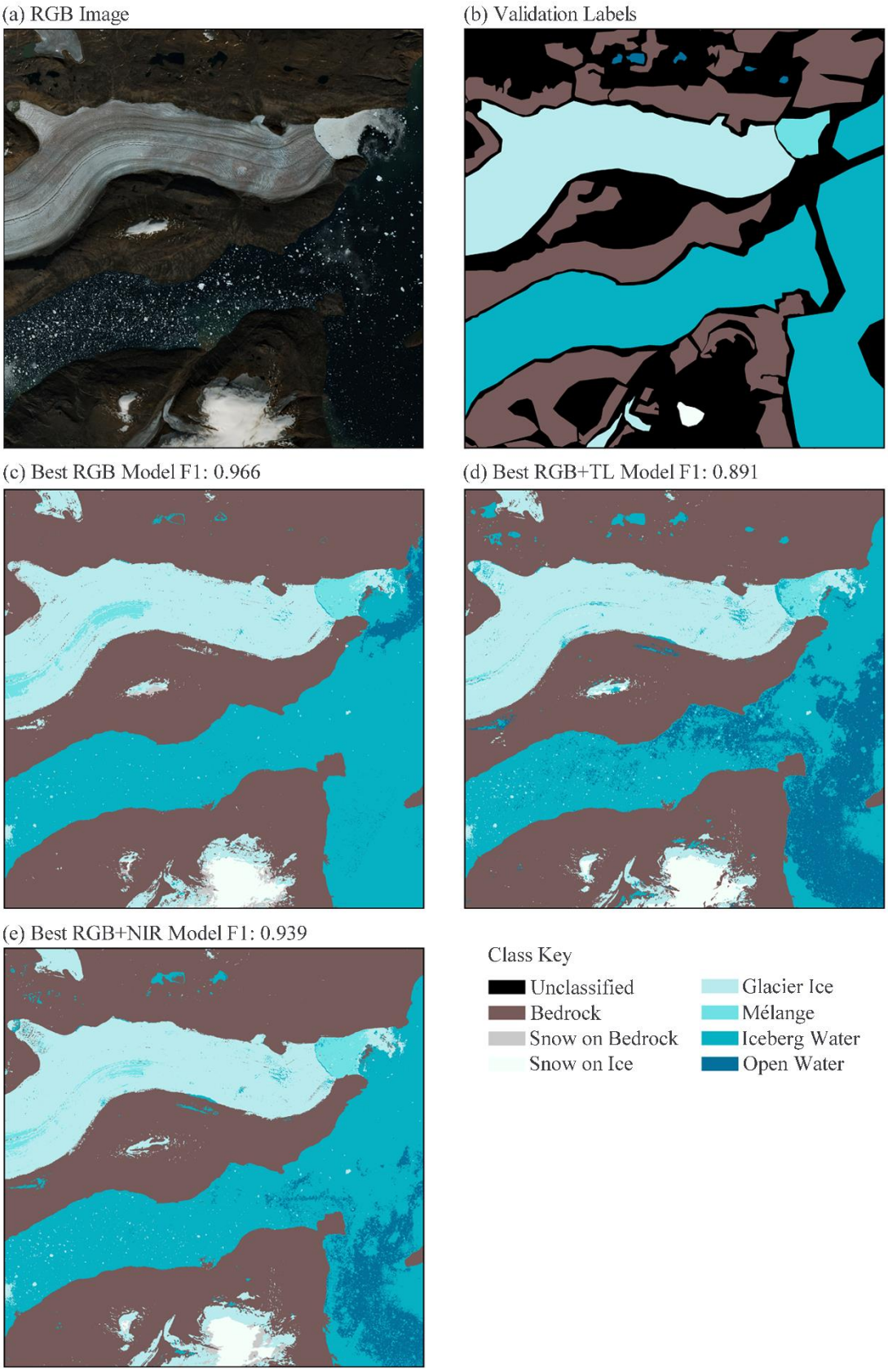


Figure 4.7: Best performing CSC results for tile 8 of 9 extracted from the Scoresby Sund study area (01/08/2019). (a) shows the RGB input image (composite Sentinel-2 bands 4, 3, and 2); and (b) shows the validation raster composed of manually digitised 'ground truth' polygons. Showing workflow outputs using: (c) the RGB model (tile size: 100 pixels, patch size: 3 pixels), (d) the RGB model with transfer learning (tile size: 50 pixels, patch size: 1 pixel), and (e) the RGB+NIR model (tile size: 75 pixels, patch size: 1 pixel). Note that most class confusion occurs between open water and iceberg water classes.

Figure 4.8: Confusion matrices for the results of the best performing out-of-sample models depicted in Figure 4.7. a) and b) Show the degree of class agreement for the RGB model; c) and d) show the agreement for results of the best RGB+TL model, while e) and f) show the agreement for best RGB+NIR results.

4.2 Spatial and Temporal Transferability

Figures 4.9 to 4.11 show CSC predictions for several 3000x3000 pixel example tiles extracted from the Sentinel-2 test images. These results indicate that the deep learning workflow is capable of classifying marine-terminating landscapes not 'seen' during training and suggest that the method is spatially transferable to glacial landscapes elsewhere in SE Greenland. In some cases, there are small areas of erroneous class predictions, particularly relating to areas of bedrock that are shadowed, supraglacial debris, supraglacial lakes (SGLs), small lakes in bedrock areas, and small pockets of fjord water which appear to contain high volumes of suspended sediment. For example, shadowed areas of bedrock were often misclassified as open water. This is most likely because they have similar spectral characteristics and only small areas of shadowed bedrock would have been included in model training data. Likewise, other areas of misclassification included surface types that were not included within the seven semantic classes outlined in Chapter 3 (e.g., SGLs, sediment rich water, supraglacial debris). Model predictions of these areas may have improved if the number of classes was expanded to be inclusive of these surface types. However, these areas generally tend to occupy small portions of imagery and do not significantly impact overall performance. Moreover, due to the small portion of imagery containing these classes, extracting sufficient volumes of training data would be challenging and potentially lead to class imbalance problems. Similarly, given that the phase one CNN operates at the scale of 50x50 to 100x100 pixel tiles, features which span only a few pixels (e.g., small lakes) would be too small to be labelled and subsequently be lost in training data for the phase two models. This suggests that CSC can not be used to detect smaller features than the input tile size, and that CSC could be improved to classify larger scale classes such as sediment rich water by including more diverse training data.

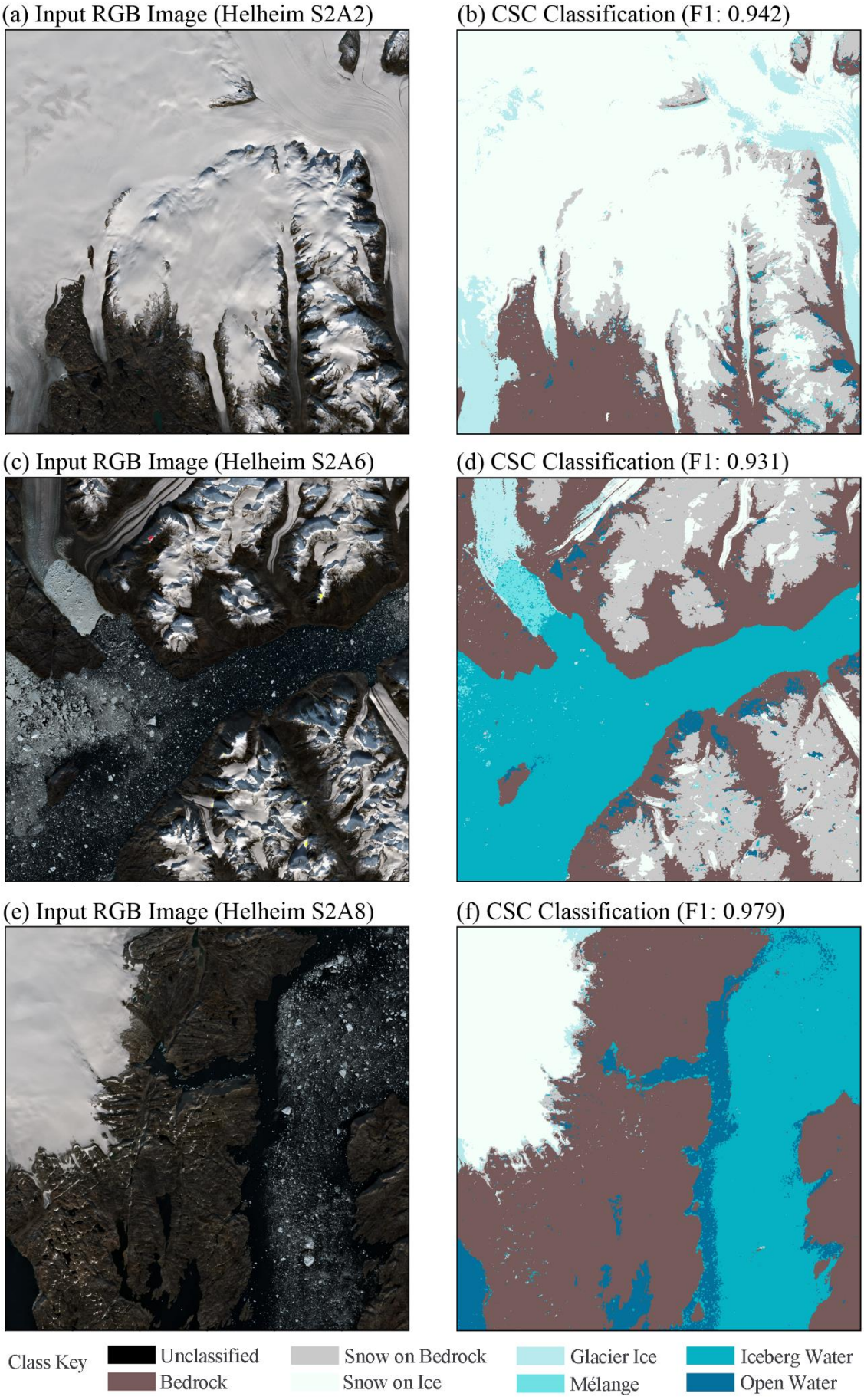


Figure 4.9: CSC (RGB+TL, Tile:50, Patch:7) results for three example tiles from the test Sentinel-2 image of Helheim. Note that CSC performs well in classifying both land- and marine-terminating glaciers, even in the presence of mélange and large volumes of icebergs. However, note that some small areas of shadowed bedrock are misclassified as open water.

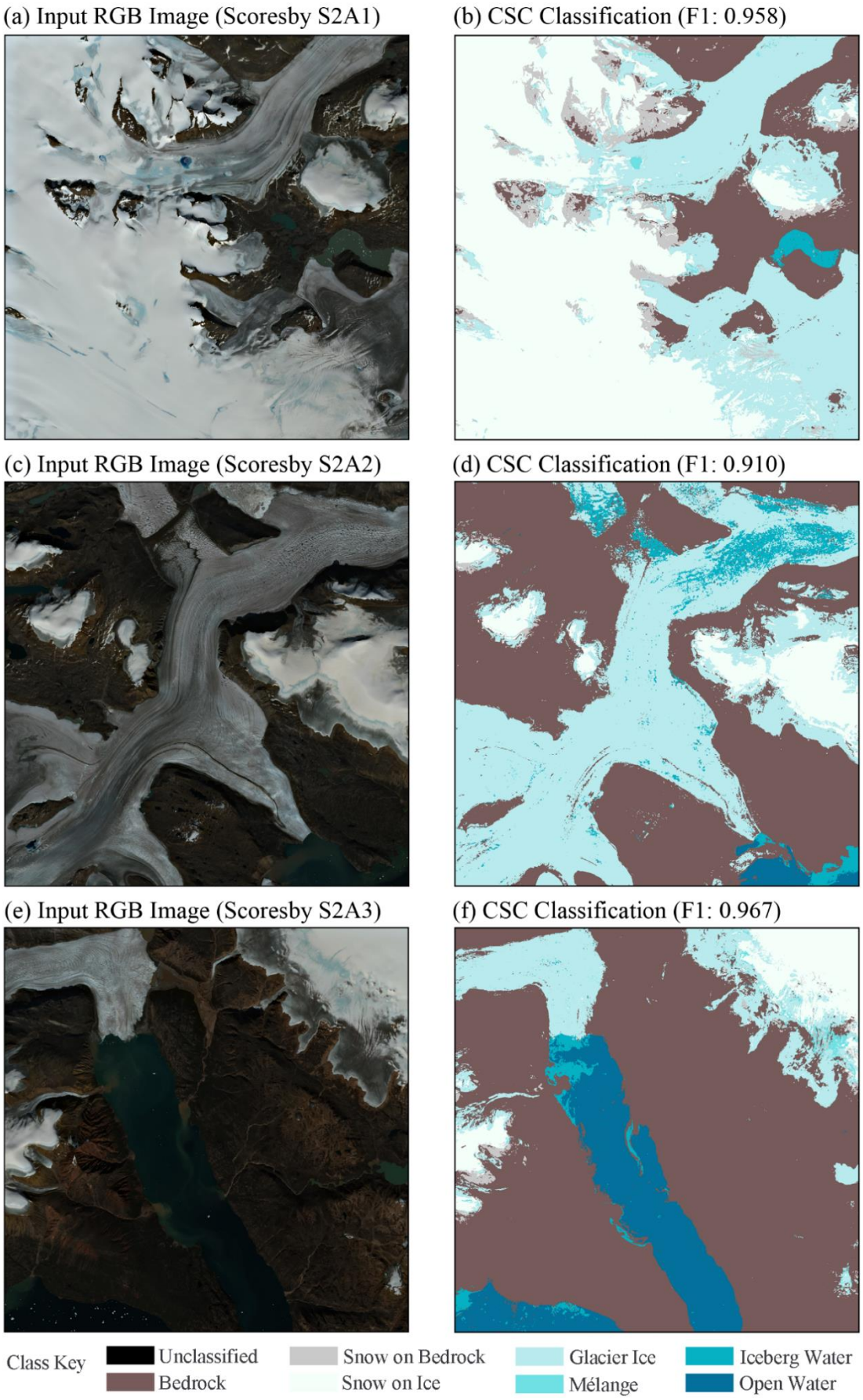


Figure 4.10: CSC (RGB, Tile:100, Patch:7) results for first three tiles extracted from the test Sentinel-2 image of the Scoresby Sund study area. Note that the supraglacial lake in (a) is classified as mélange in (b) and some small lakes in bedrock areas are missed by the model. In (c) large areas of SGLs resulted in misclassification of glacier ice as iceberg water in (d). Also note the areas of sediment-rich fjord water in (c) and (e) which are misclassified as bedrock.

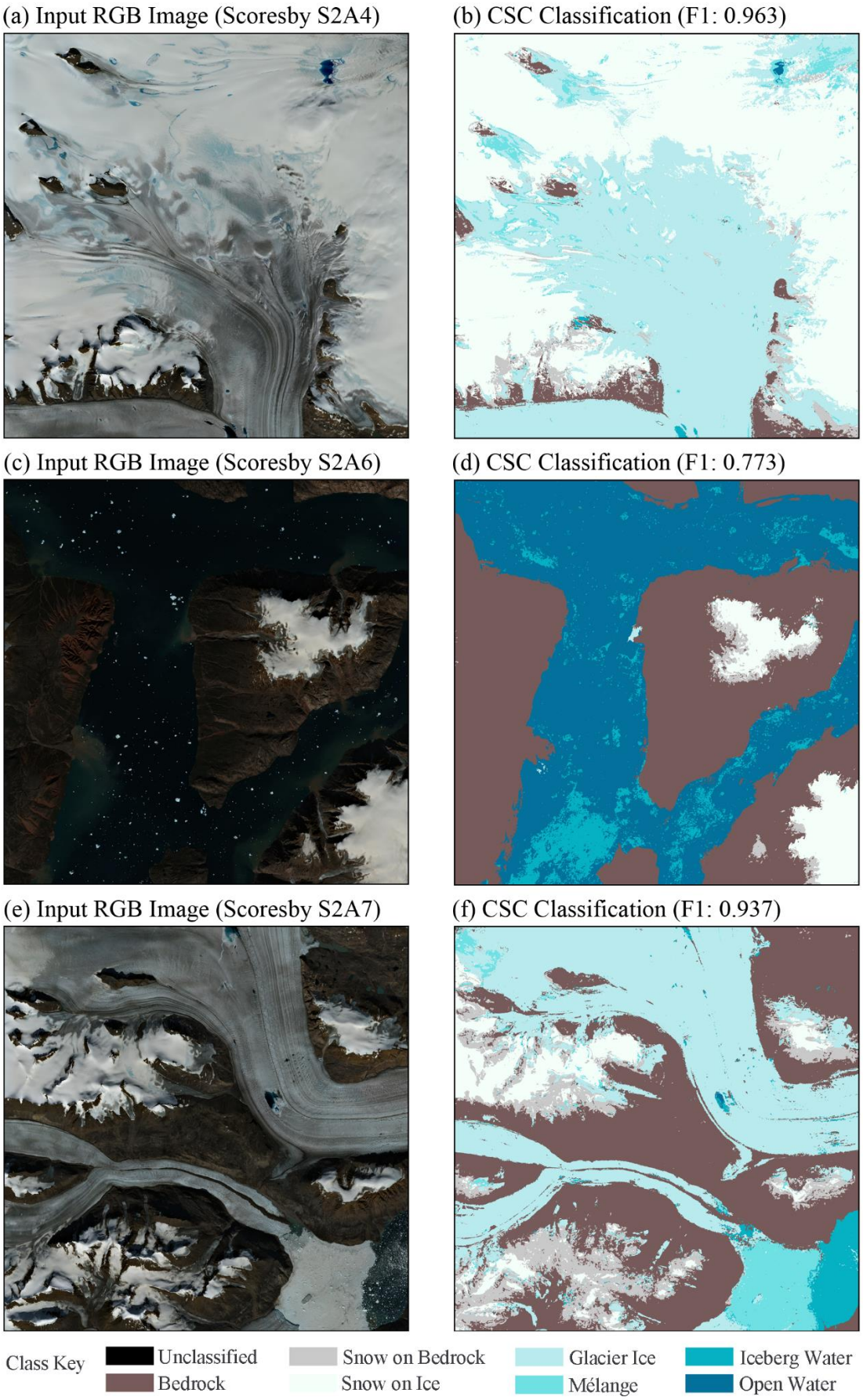


Figure 4.11: CSC (RGB, Tile:100, Patch:7) results for three more example tiles extracted from the test Sentinel-2 image of the Scoresby Sund study area. Note in (a), (b), (e), and (f) several more SGLs are classified as mélange, iceberg water or open water. The lower F1 score in (d) is due to confusion between predicted water types in comparison to manual truth labels. In (f) some areas of smooth 'snow on ice' areas have been classified as 'snow on rock'.

4.3 Comparison of CNN-Supervised Classification to Traditional Band Ratio Methods

Figure 4.12 shows a visual comparison between a traditional band ratio technique (as described in Paul *et al*. (2016)) and the result of the CSC workflow using the best performing model on tile 5 of the 9 tiles extracted from the test image of Helheim. CSC successfully identifies areas of mélange, glacier ice and iceberg rich fjord waters as different classes (Figure 4.12b). The band ratio method allows clear identification of rock, and land terminating ice margins. However, the technique struggles to distinguish boundaries between glacier ice, mélange, and iceberg water (Figure 4.12c). In the example shown, the abundant spectral variation and noise makes using a series of thresholds to extract margins in a mélange-filled fjord almost impossible. This is reflected by an F1 score of 53% for the band ratio technique which is substantially outperformed by CSC with a corresponding F1 of 97%. This comparison and preliminary tests of CSC transferability suggests it is more robust than traditional techniques and does not rely on the requirement of identifying threshold values to extract classes.

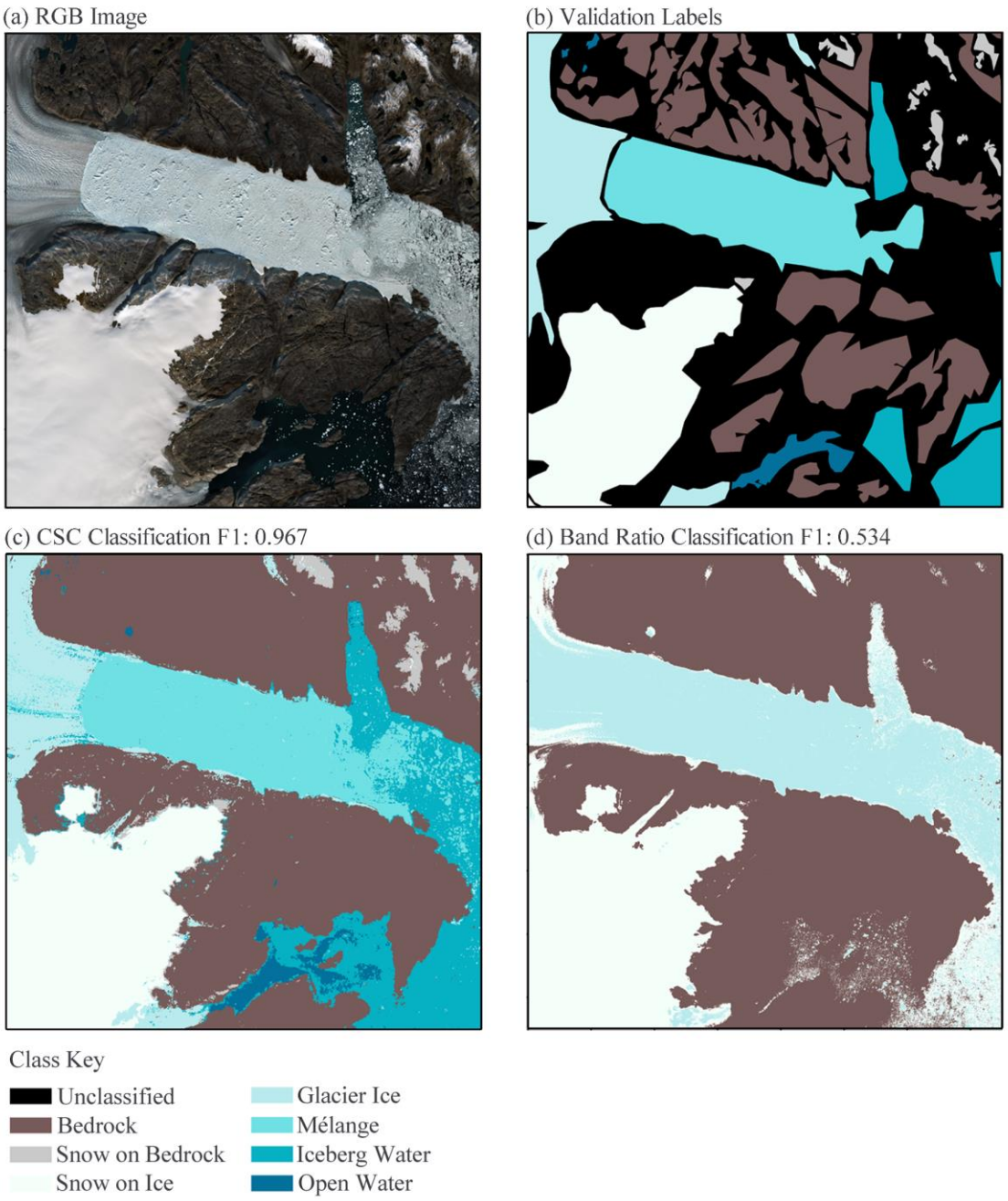


Figure 4.12: Comparison of methods used on tile 5 of 9 (a) extracted from the Helheim study area (07/08/2019), including (b) Validation labels used to create F1 scores, (c) the CSC classification, and (d) a band ratio classification using Sentinel-2 bands 4 (red) and 11 (SWIR). Note that only three classes could be extracted from the band ratio image due to significant noise and no contrast between glacier ice, mélange, and iceberg water classes.

5 Discussion

5.1 CSC Performance in Marine-Terminating Outlet Glacier Environments

The results reported here demonstrate novel multi-class satellite image classification of complex outlet glacier image scenes using deep learning. The CSC workflow adapted for glacial settings in Greenland produced mean F1 scores up to 95% for in-sample test imagery and 93% for out-of-sample test imagery, with corresponding Kappa values of 0.92 and 0.9, respectively. The method created multi-class outputs in contrast to the binary classification outputs used by Mohajerani *et al*. (2019), Zhang *et al*. (2019), and Baumhoer *et al*. (2019) for automated delineation of marine-terminating ice fronts. Despite this difference in output classification type, mean F1 scores of classifications by Baumhoer *et al*. (2019) were 89 to 90% for in-sample training sites and 90 to 91% for out-of-sample test sites, suggesting the CSC workflow advances the state-of-the-art in image classification of complex marine-terminating glacial environments using deep learning.

In addition to advancing the state-of-the-art for marine-terminating glacial settings, the multi-class outputs of the CSC workflow widen the scope of image classification for a variety of research applications, beyond just automated delineation of calving fronts from binary classification outputs. Moreover, the ability of the CSC deep learning workflow to classify images previously unseen by the model for both training and testing areas to a similarly high level of accuracy suggests good generalisation and highlights the transferability of CSC to other marine-terminating outlet glacier environments in SE Greenland.

5.2 Comparison to Previous Work

The results of this project build on the work of deep learning-based classification methods for glacier delineation (Baumhoer *et al*., 2019; Mohajerani *et al*., 2019; Zhang *et al*., 2019; Xie *et al*., 2020), with several key innovations and variations of note. In particular, the volume, type, and number of input channels of training data used in this workflow differs from those of previous work. Furthermore, there are substantial differences in the deep learning architectures and classification approach tested in this project.

5.2.1 Volume of Training Data

In terms of training data volume, fewer training images (i.e., 13) were used here compared to the number of training images used by Baumhoer *et al*. (2019), Mohajerani *et al*. (2019), and Zhang *et al*. (2019) (i.e., 38, 123, and 75, respectively). In terms of the number of training samples, 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 (< 360,000) 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, the adapted VGG16 architecture in this project 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 the CSC workflow, reducing the time required for initial creation of manually labelled training data. Furthermore, the CSC workflow does not require the same pre-processing steps such as manually rotating images so that glacier flow directions are consistent or cropping input images to a specified buffer width encompassing glacier calving fronts as in Mohajerani *et al*. (2019). As such, CSC has the advantage of needing fewer satellite acquisitions for training and simpler pre-processing steps.

5.2.2 Type of Training Data

In relation to the type of data used to train the deep learning models, Baumhoer *et al*. (2019) and Zhang *et al*. (2019) used Sentinel-1 and TerraSAR-X SAR data, respectively. Specifically, Baumhoer *et al*. (2019) used different SAR polarisations with the addition of a DEM to train the FCN. In contrast, Mohajerani *et al*. (2019) used Landsat 5, 7 and 8 imagery for FCN training, in particular using the 'green' band from Landsat 5 data and 'panchromatic' band from Landsat 7 and 8 data. In this project, Sentinel-2 optical data was used which is generally easier to process in comparison to SAR data and requires less specialised knowledge for pre-processing. For example, common pre-processing steps to implement SAR data include noise removal, radiometric calibration, and geocoding correction (in addition to training area cropping and tiling for model training) (Baumhoer *et al*., 2019; Zhang *et al*., 2019). While SAR data is not limited by clouds or polar night, using L2A Sentinel-2 imagery eliminates the need to incorporate DEM data and removes SAR

pre-processing steps from the deep learning workflow, allowing cloud-free imagery to be downloaded, cropped, and tiled more quickly for model training.

5.2.3 Dimensions of Training Data

In terms of input dimensions, Zhang *et al*. (2019) and Mohajerani *et al*. (2019) used one-dimensional (1D) training inputs while Baumhoer *et al*. (2019) used data with four input channels. The input channels of the CSC workflow (i.e., 3 or 4 input bands) were analogous with those of Baumhoer *et al*. (2019) but with different input data types (i.e., multispectral data vs. SAR and DEM data). At the opposite end of the spectrum, Xie *et al*. (2020) used 17 input channels, incorporating all 11 Landsat 8 bands, a DEM, and five layers derived from the DEM to produce binary classifications for debris-covered land-terminating glaciers. They produced results with F1 scores up to 94% using a CNN trained and tested on images of glaciers in the Karakoram region, and 90% for a transfer learning approach using the model initially trained on the Karakoram region, with weights adjusted using new training data from Nepal. Xie *et al*. (2020) note that using fewer input channels in experimental CNN training resulted in lower levels of accuracy. However, despite the large difference in input dimensionality between the CNN used here and that of Xie *et al*. (2020), resultant F1 scores show that the use of only three Sentinel-2 bands produces classifications with similar levels of accuracy. However, it is important to note that the CSC workflow was applied to a markedly different glacial landscape compared to Xie *et al*. (2020). Nevertheless, the results presented here show that using only three input Sentinel-2 bands is sufficient for producing accurate classifications in scenes containing complex marine-terminating glaciers.

5.2.4 Deep Learning Model Architectures

Likewise, further variations in comparison to previous work are apparent in the model architectures, number of models used, and training approaches tested for classification. All previous deep learning classification methods for marine-terminating glacial environments (Baumhoer *et al*., 2019; Mohajerani *et al*., 2019; Zhang *et al*., 2019) use the U-Net architecture (Ronneberger *et al*., 2015). Whilst U-Net architectures have reached state-of-the-art performance in computer vision tasks, their application in complex natural landscapes is not necessarily optimal given the intrinsic assumptions of U-Net models. For example, U-Net architectures perform exceedingly well at delineating people in imagery (Xie *et al*., 2018; Wang and Bai, 2019). In such cases, skin colour and clothing colour must not be considered as identifying features. However, in Earth Observation (EO) images of natural

landscapes, there is a much stronger correlation between colour and landform. Furthermore, the U-Net architecture will learn shapes that have a limited variability of both form and scale. For example, people have similar dimensions in imagery used in self-driving vehicles and their location in the image is limited to a horizontal zone across the field of view. In contrast, natural landforms can vary in scales over several orders of magnitude and be located anywhere in an image. Therefore, it can be argued that more evidence is needed before considering the use of U-Net architectures as the *de facto* algorithm for glacial landscape classification. Moreover, the results presented here show that a deep learning approach based on a combination of local spectral and spatial properties determined by a compact CNN architecture has exceeded the results derived from U-Net architectures.

5.3 Comparison to Traditional Glacier Mapping Methods using Band Ratios

In contrast to previous work, this project also assessed the workflow performance in comparison to a traditional band ratio method for classification of an image containing a marine-terminating glacier (Figure 4.12). The results show that CSC is better at identifying classes which are spectrally similar such as mélange, glacier ice and iceberg water. This suggests that the method outperforms traditional pixel-based classification techniques, similar to findings from classification of fluvial image scenes (Carbonneau *et al*., 2020a). CSC is also more robust because it is able to classify new unseen images in SE Greenland without further training and does not require additional steps to determine an optimal threshold value for outlining class boundaries. Moreover, the method has the ability to pick out textures and patterns in the same way a human operator would, irrespective of variations in illumination, weather conditions or seasonal changes to the landscape and individual classes. This also highlights the benefit and transferability of CNNs over purely pixel-based techniques for classifying complex image scenes with substantial seasonal variations.

5.4 Evaluation of Training Methods

To evaluate the workflow, three different approaches of training the phase one CNNs were tested by using two different band combinations and a transfer learning technique. The results show that the addition of the NIR band did not significantly improve classification accuracy. Further testing of alternative band combinations or the addition of different satellite data types (e.g., SAR data) may be advantageous. However, using RGB bands produces satisfactory results and adding additional image bands is likely to increase

processing time without necessarily improving on the overall accuracy, as also suggested by Xie *et al*. (2020).

In terms of transfer learning, the technique has been applied successfully in previous image classification studies which use remotely sensed satellite imagery, allowing reduced training times for smaller datasets (Hu *et al*., 2015; Pires de Lima and Marfurt, 2020). These studies highlighted that CNNs pre-trained on ImageNet data may be transferable to remote sensing imagery by fine-tuning the last layers in the CNN for dataset specific feature extraction, regardless of disparities in input image properties (e.g., angle of acquisition). However, Pires de Lima and Marfurt, (2020) recognise that, in contrast to training all the layers of a CNN on remote sensing data, the difference between ImageNet data and remote sensing data in some cases has resulted in transfer learning techniques overfitting and reducing the ability of models to learn. In this study, while transfer learning performed exceptionally well for in-sample data, its performance degraded substantially when applied to out-of-sample test imagery, suggesting reduced transferability. The strong performance of the transfer learning approach on in-sample data supports findings that it is a powerful deep learning tool (Xie *et al*., 2020). However, it is suggested that the high-level features representative of diverse, seasonally variable image elements and classes are not as successfully detected using transfer learning in comparison to full CNN training. This indicates that transfer learning techniques would require further efforts to fine-tune for improved transferability to marine-terminating outlet glacier environments.

5.5 The Impact of Tile Size on Model Performance

The impact of tile size (height and width of image samples used for training and validation) on model performance was also evaluated. For models to learn the features which represent diverse image elements, class representative features need to fit within an individual tile, thus making careful choice of tile size especially important. It is also important to consider that the number of tiles produced to compose a training dataset changes based on tile size. With the same source imagery, a large number of small tiles can be produced compared to fewer larger tiles (e.g., Table 3.2). Thus, the selection of tile size is dependent on the desired information content of a tile and the number of tiles needed to sufficiently train a CNN. It was for this reason that tiles sizes of 50, 75 and 100 pixels were tested.

Results show that non-transfer learning phase one CNNs were not substantially sensitive to tile size, with models trained on all three tile sizes producing F1 scores within a 2% range

for both in-sample and out-of-sample test data. Following the full CSC workflow, the RGB models (without transfer learning) trained on larger tile sizes produced slightly better classifications with tile sizes of 100 and 75 outperforming tile sizes of 50 pixels by up to 3%. This suggests that using fewer larger tiles (e.g., size of 100 pixels) slightly improves RGB model performance, specifically for the scale of features in outlet glacier landscapes in Greenland.

In contrast, the transfer learning phase one CNNs had increased sensitivity to tile size, producing F1 scores with a range of 5% for in-sample test data and 13% for out-of-sample test data. This was mirrored in the CSC classifications, with large differences in F1 scores depending on tile size, whereby the smallest tile size of 50 pixels produced the best results, but model performance deteriorated with tiles sizes of 75 and 100 pixels. It is interesting to note that the transfer learning technique benefited from using a larger number of smaller tiles compared to the preferred smaller number of large tiles for the fully trained CNN. These results suggest that, for classification of outlet glacier landscapes, fully trained CNNs are more invariant to tile size for both in- and out-of-sample data, whereas transfer learning models produce a larger variability of F1 scores for different tile sizes, especially when applied to out-of-sample data. This supports the assertion that the most transferable CSC workflow for outlet glacier image classification uses a phase one CNN with all weights trained using RGB bands and larger tile sizes (of 75 or 100 pixels).

5.6 The Impact of Patch Size on Model Performance

In addition to testing the influence of tile size during training for phase one CNNs, the sensitivity of phase two model performance was tested by using pixel- and patch-based techniques. Specifically, four patch sizes of 1x1 (pixel-based), 3x3, 7x7, and 15x15 pixels (patch-based) were tested. The reason for testing pixel- and patch-based techniques is due to the use of medium resolution 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). The best performing patch size may also vary depending on the type of medium resolution satellite imagery (Sharma *et al*., 2017) making it an important testable parameter. It was 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). The results showed that for the most part this was true, especially for in-sample test imagery.

When CSC was applied to in-sample test data, the workflow performance was clearly sensitive to patch size, with the pixel-based approach producing classifications with lower F1 scores compared to the patch-based technique. The optimal patch-size for in-sample test data was 7x7 pixels. The in-sample results support the hypothesis that pixel-based approaches do not perform as well on medium-resolution imagery compared to the patch-based approach. This also validates similar findings that patch-based CNNs outperform standard pixel-based neural networks and CNNs (Sharma *et al*., 2017). In contrast, for out-of-sample data, the pixel-based approach performed substantially better than for the in-sample test data, and smaller patch sizes of 1, 3, and 7 generally outperformed a larger patch size of 15. However, in general CSC results for out-of-sample data were less sensitive to patch size, producing a range of F1 scores that varied by only 3 % between all four patch sizes (per the three individual models). Therefore, it is suggested that testing a range of patch sizes would be beneficial before applying the workflow to a new dataset.

5.7 Limitations, Transferability, and Implications for Future Work

The performance of the CSC workflow is dependent on the success of the pre-trained CNNs to identify image-specific training areas in phase one. The performance of the phase one models can be impacted by the size and class representation of training data. It is noted that the data used to train the phase one CNNs was extracted from only one outlet glacier in Greenland, and that producing a larger training dataset from a wider array of imagery in other similar settings may be beneficial to increase model performance and transferability in future work. The lack of a benchmark dataset specifically for marine-terminating outlet glacier settings means that the application of deep learning in this field initially relies on labour-intensive manual labelling of training data. Despite this limitation, the deep learning workflow here produced highly accurate classifications for both images of the glacier used in training and of different locations not ‘seen’ by the model. Furthermore, once the phase one models are trained and weights are saved, no further training is required to apply the workflow to other marine-terminating outlet glaciers in SE Greenland.

In addition to the size of the training dataset, class imbalance can impact model performance. Marine-terminating outlet glacier environments have classes which are naturally less abundant in imagery. For example, there were smaller areas of mélange compared to glacier ice or snow-covered ice in most full satellite scenes of the Helheim and Scoresby study areas. So, despite efforts to balance the training dataset, certain classes such as open water (without icebergs), mélange, and bedrock were represented by a smaller number of training samples

compared to more prominent classes (Table 3.2). This can lead to confusion between classes as was found in some of the experiments. In models with lower performance, confusion occurred between open water and iceberg water classes, as well as between bedrock and snow on bedrock classes. Furthermore, class imbalance may be problematic for classifying large images (e.g., entire Sentinel-2 images that have not been tiled) which contain only small areas of a single class. For example, if only a small area of mélange was classified following the application of the phase one model to a Sentinel-2 image, only a small proportion of image data would be available for image-specific training in phase two. Consequently, if removing the mélange class altogether would reduce model loss, the output classification would have increased class confusion due to the misclassification of the absent mélange class. However, in the case of the smaller 3000x3000 pixel test tiles used in this study, confusion between classes in output classifications was minimal for most of the 36 models tested.

Additionally, there are numerous hyperparameters (e.g., learning rate, batch size, etc.) that could be tested and tuned for improved workflow implementation. Variations in such parameters are likely to impact model performance but require significant time to test and substantially increase the dimensions of model outputs (Carbonneau *et al.*, 2020a). In addition, only one model architecture was tested for the phase one pre-trained model. VGG16 has a relatively simple architecture but the state-of-the-art for image classification is constantly evolving, with consistent and rapid development of new CNN architectures. As a result, there are a myriad of variations in CNN characteristics that could be tested in future work, such as CNN depth and filter sizes. Moreover, other well-established pre-trained model architectures such as GoogLeNet (Szegedy *et al.*, 2014), and NASNet (Zoph *et al.*, 2018) have also been successfully applied to remote sensing applications (Ostankovich and Afanasyev, 2018; Carbonneau *et al.*, 2020a) and could be explored for use in the CSC workflow for marine-terminating outlet glacier image classification. Thus, there are several avenues for expanding the use of deep learning for image classification of marine-terminating outlet glacier landscapes in future work. Nevertheless, the implications of this study suggest that the adapted CSC workflow is transferable to unseen landscapes in SE Greenland and capable of maintaining a high level of performance.

Further integration of the workflow with GIS platforms could provide an efficient tool for processing large amounts of imagery at high temporal resolution. In addition, since the workflow was implemented in Python 3.7, it is compatible with GEE (Gorelick *et al.*, 2017), a cloud-based geospatial platform for processing and analysing large-scale datasets

(Tamiminia *et al*., 2020). GEE allows processing of Landsat and Sentinel-2 imagery without the need to download large volumes of data and has been used effectively in glacial applications such as automated mapping of glacial lakes (Shugar *et al*., 2020). Therefore, there is scope to implement CSC within the GEE platform and build on existing tools for automated glacier margin extraction (e.g. Lea, 2018) and classification without the need for expertise in coding or glaciology. With such integration, classification outputs could be rapidly produced and used to efficiently generate vector datasets from boundaries between classes, for wide-ranging applications and analysis in outlet glacier landscapes.

6 Conclusions

In this study a workflow for image classification of seasonally variable marine-terminating outlet glacier environments using deep learning was developed and evaluated. The development of deep learning methods for automated classification of outlet glaciers is an important step towards monitoring important processes at high temporal and spatial resolution (e.g., changes in frontal position, calving events, plume development, supraglacial lake development and drainage). 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. The two-phase CSC workflow was adapted for classification of medium resolution Sentinel-2 imagery of outlet glaciers in south east 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 accurate training data without the requirement of time-consuming manual digitisation. In phase two, the workflow produces a pixel-level classification according to seven semantic classes characteristic of complex outlet glacier settings. Alongside an evaluation of various parameters and training methods on model performance, the workflow was applied and tested on two new Sentinel-2 tiles containing marine-terminating outlet glaciers, previously unseen by phase one CNNs during training.

Exemplified by resulting overall F1 scores of up to 95% for in-sample data and 93% for out-of-sample data, the workflow establishes a state-of-the-art in multi-class image classification for outlet glacier environments in Greenland. Additionally, when compared to traditional pixel-based techniques, the results of CSC clearly outperform those of image band ratio methods. These results demonstrate the transferability and robustness of the approach, and although the CSC workflow was applied and tested on outlet glaciers in Greenland, it may also be transferred to outlet glacier landscapes in other glaciated regions with additional testing and fine-tuning.

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). The deep learning workflow presented here offers an efficient tool for glaciologists to analyse the dynamics of marine-terminating outlet glaciers, without significant prior experience in coding or deep learning.

7 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 nine pre-trained VGG16 models are available for download from this institutional repository: http://doi.org/10.15128/r2gh93gz51k 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.

Appendix

Helheim Confusion Matrices

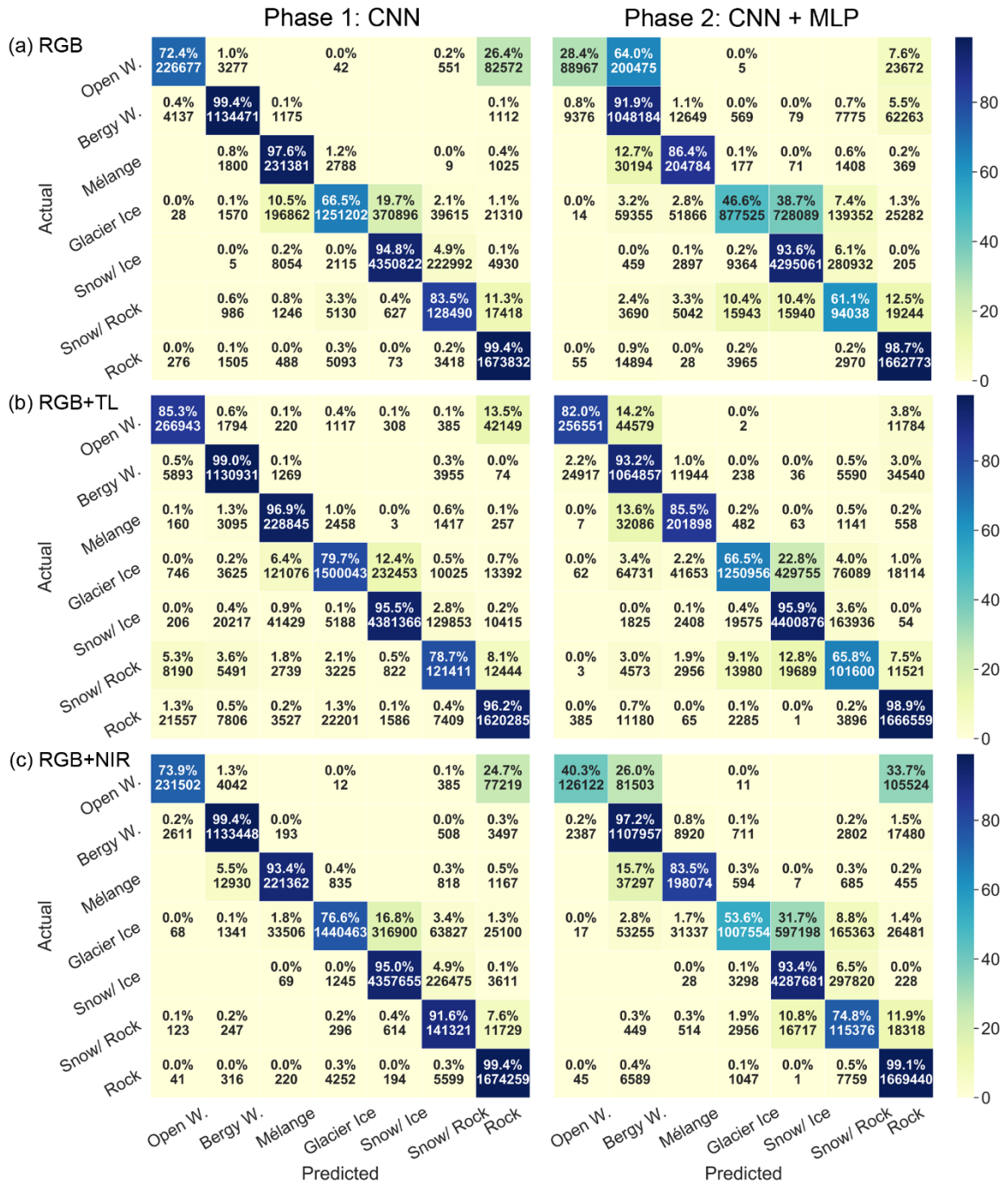


Figure A1: Confusion matrices for CSC results on the Helheim test image using 50x50 pixel tiles and a pixel-based approach (patch size: 1).

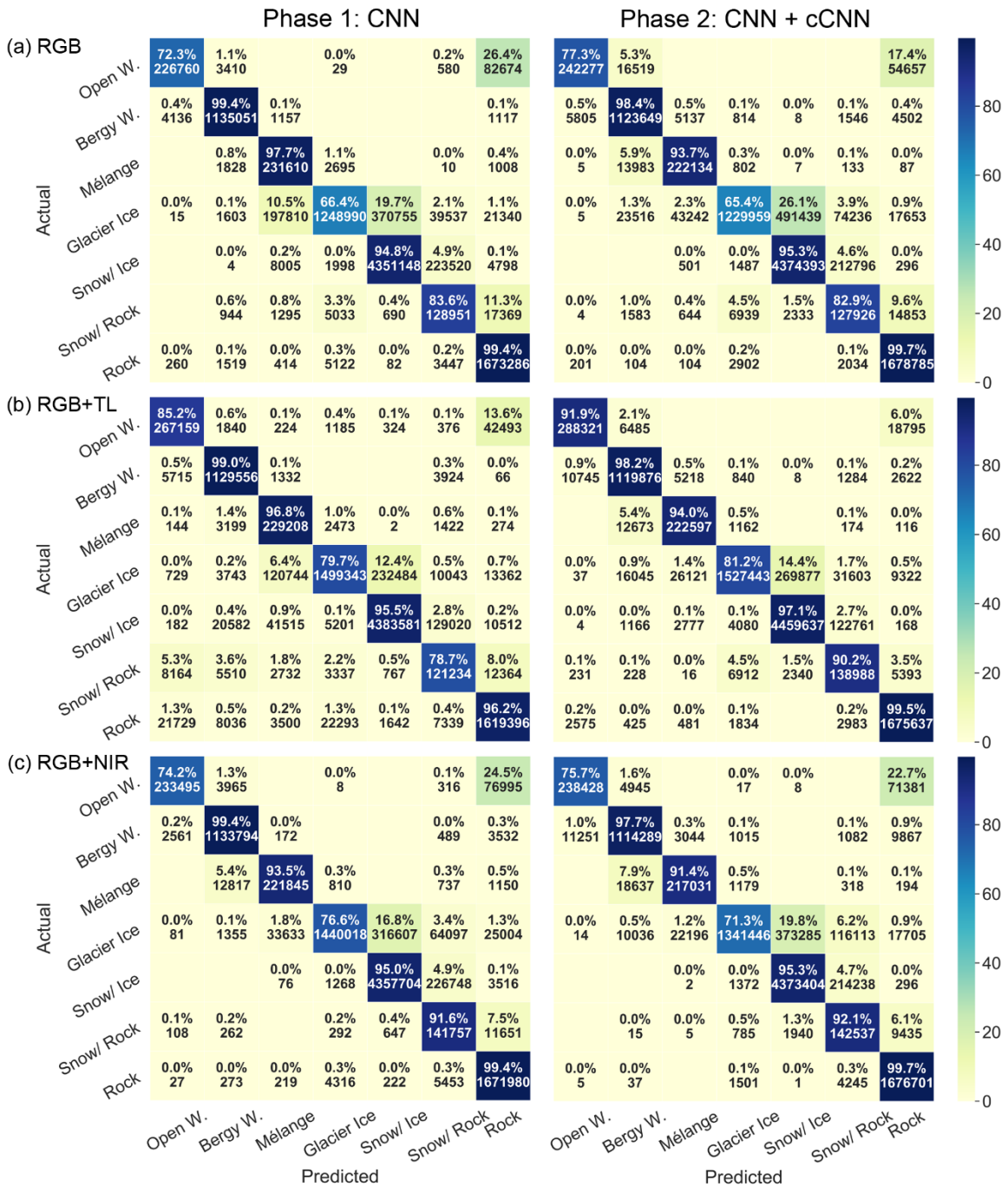


Figure A2: Confusion matrices for CSC results on the Helheim test image using 50x50 pixel tiles and a patch-based approach (patch size: 3).

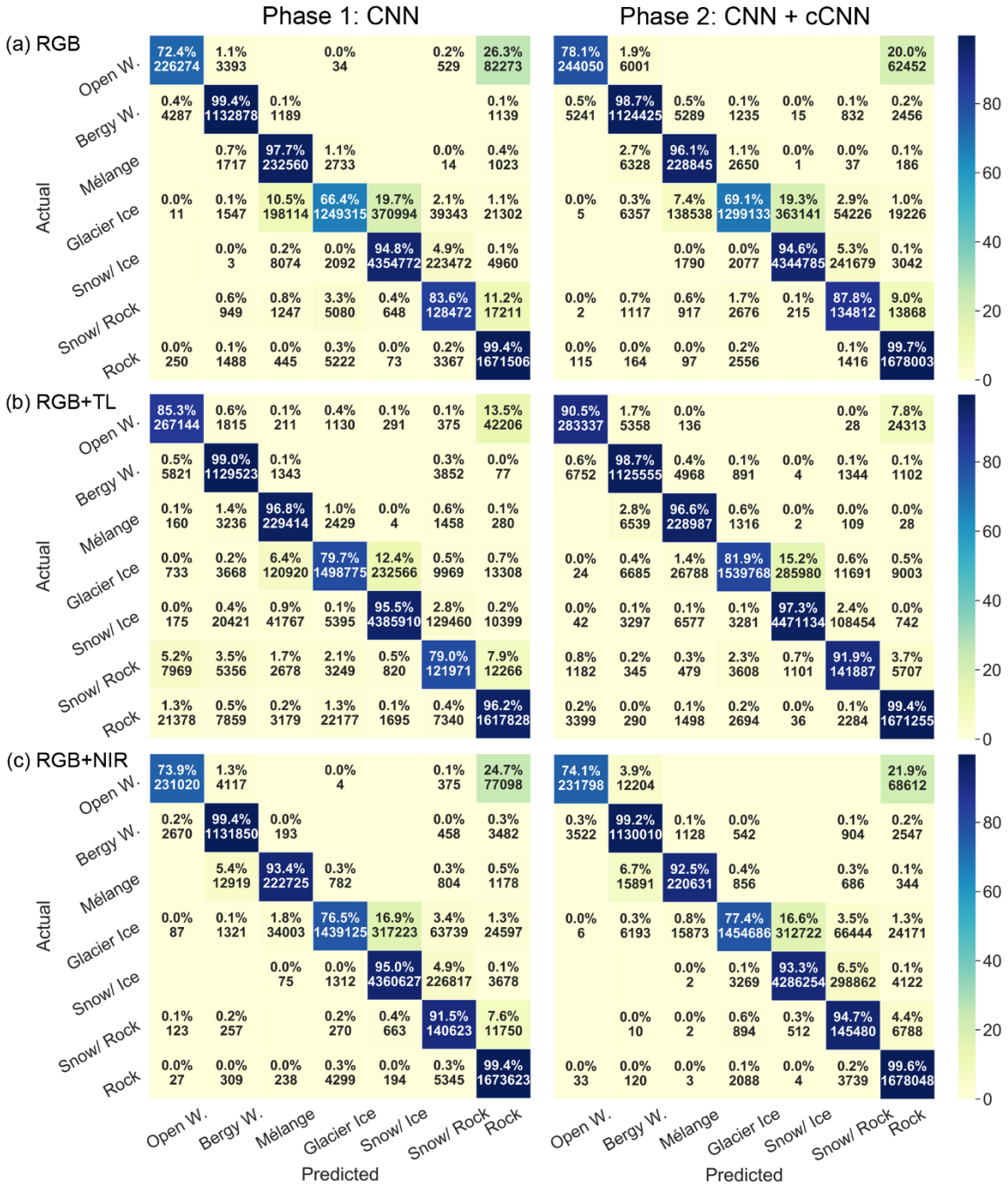


Figure A3: Confusion matrices for CSC results on the Helheim test image using 50x50 pixel tiles and a patch-based approach (patch size: 7).

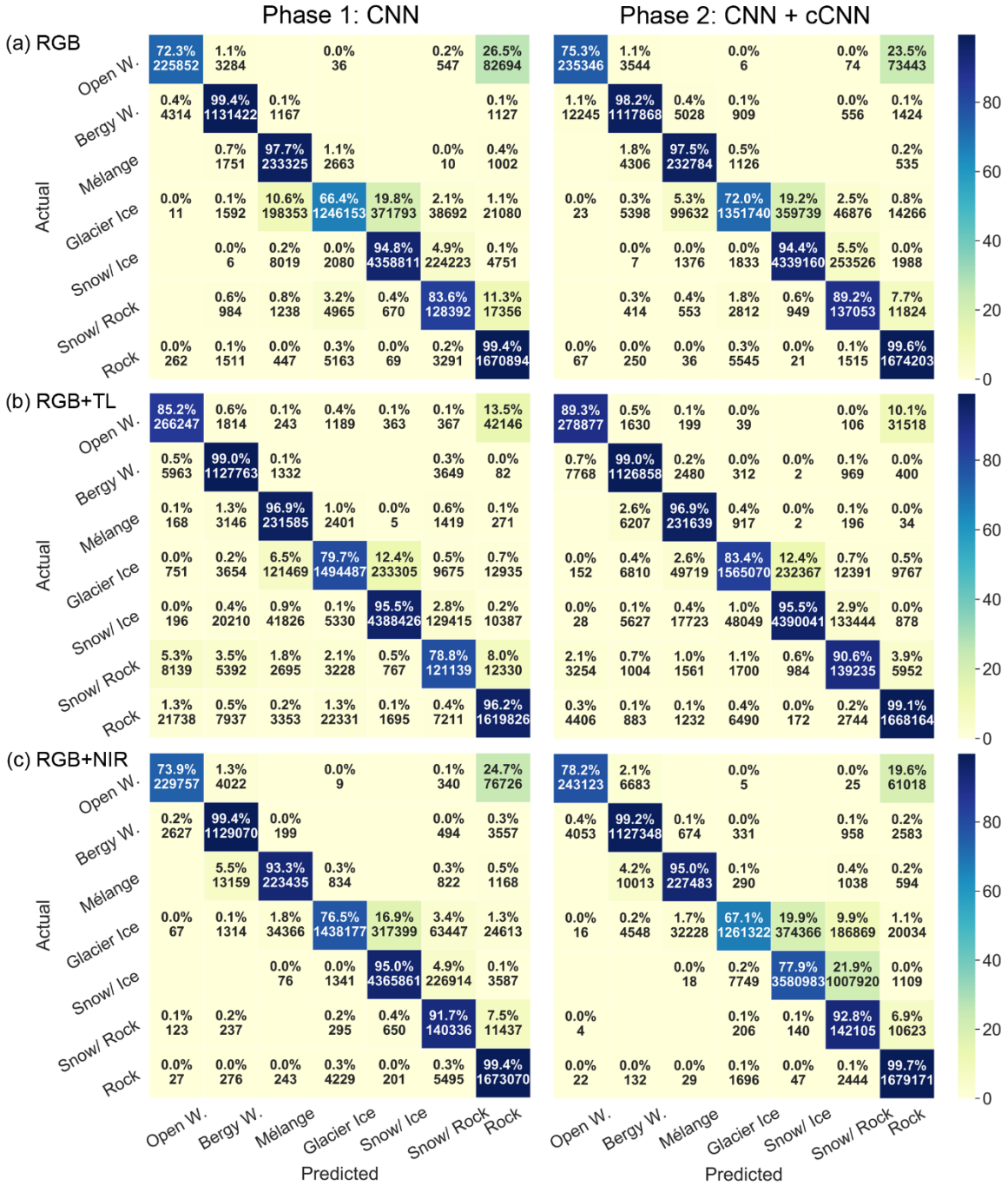


Figure A4: Confusion matrices for CSC results on the Helheim test image using 50x50 pixel tiles and a patch-based approach (patch size: 15).

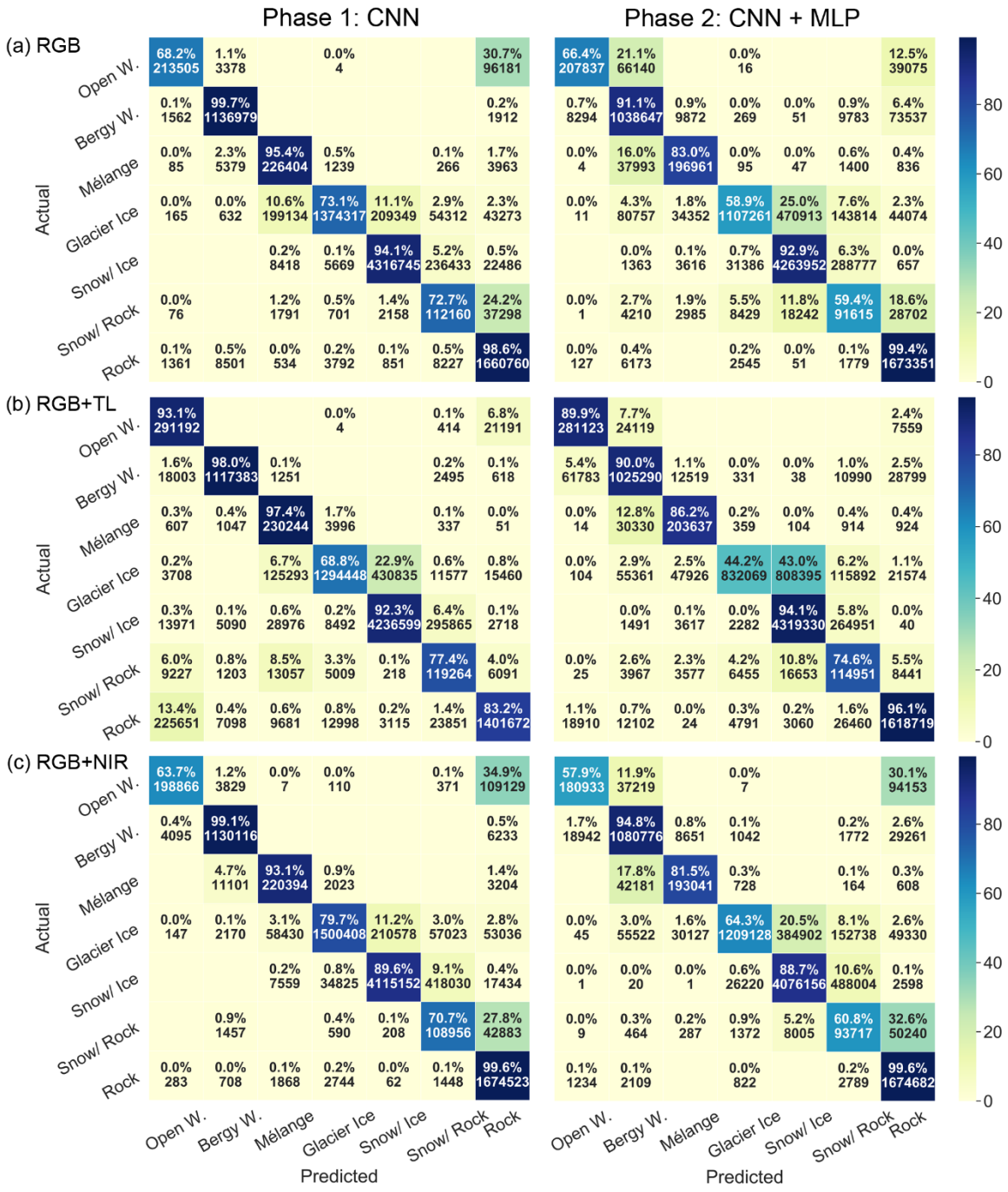


Figure A5: Confusion matrices for CSC results on the Helheim test image using 75x75 pixel tiles and a pixel-based approach (patch size: 1).

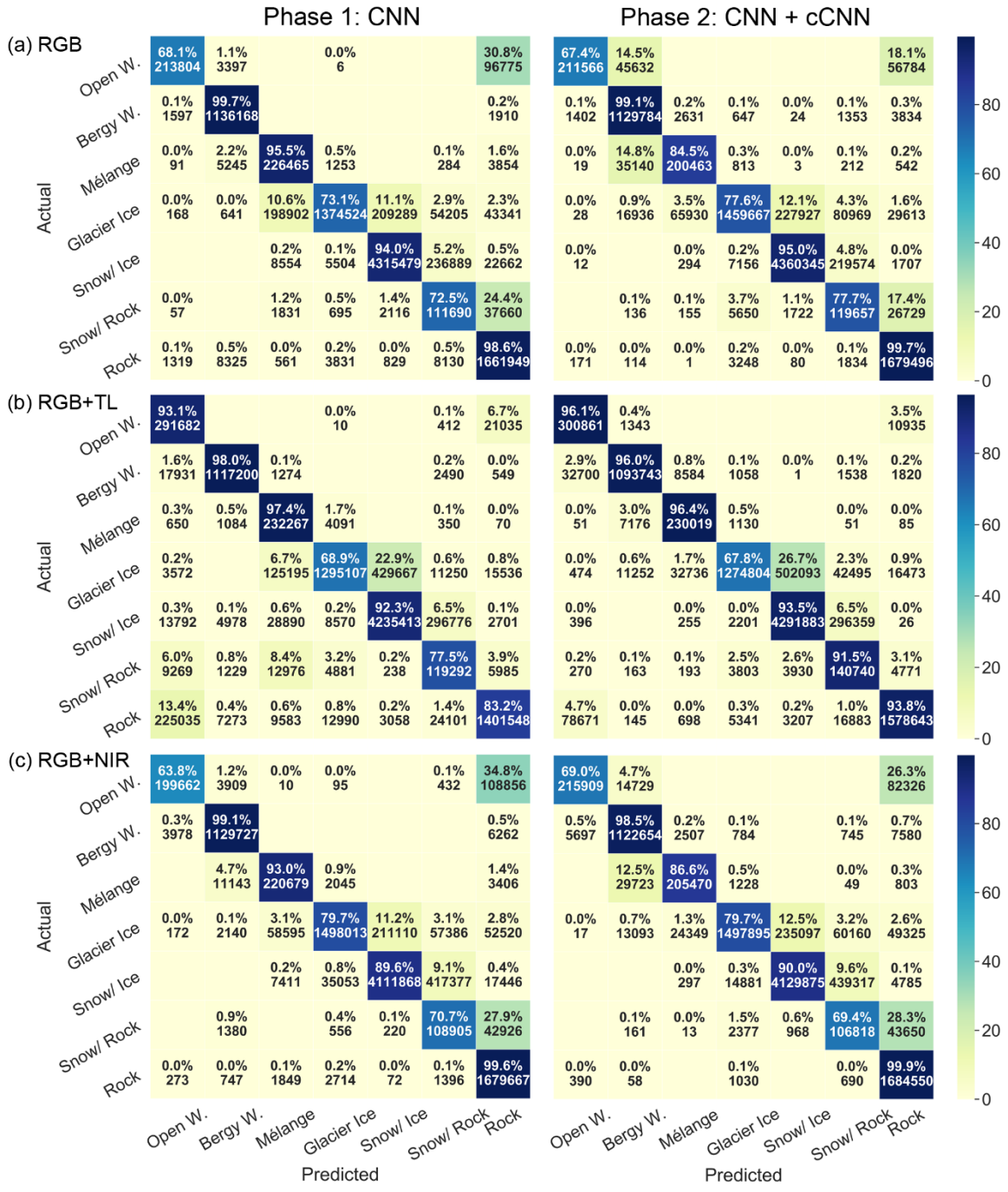


Figure A6: Confusion matrices for CSC results on the Helheim test image using 75x75 pixel tiles and a patch-based approach (patch size: 3).

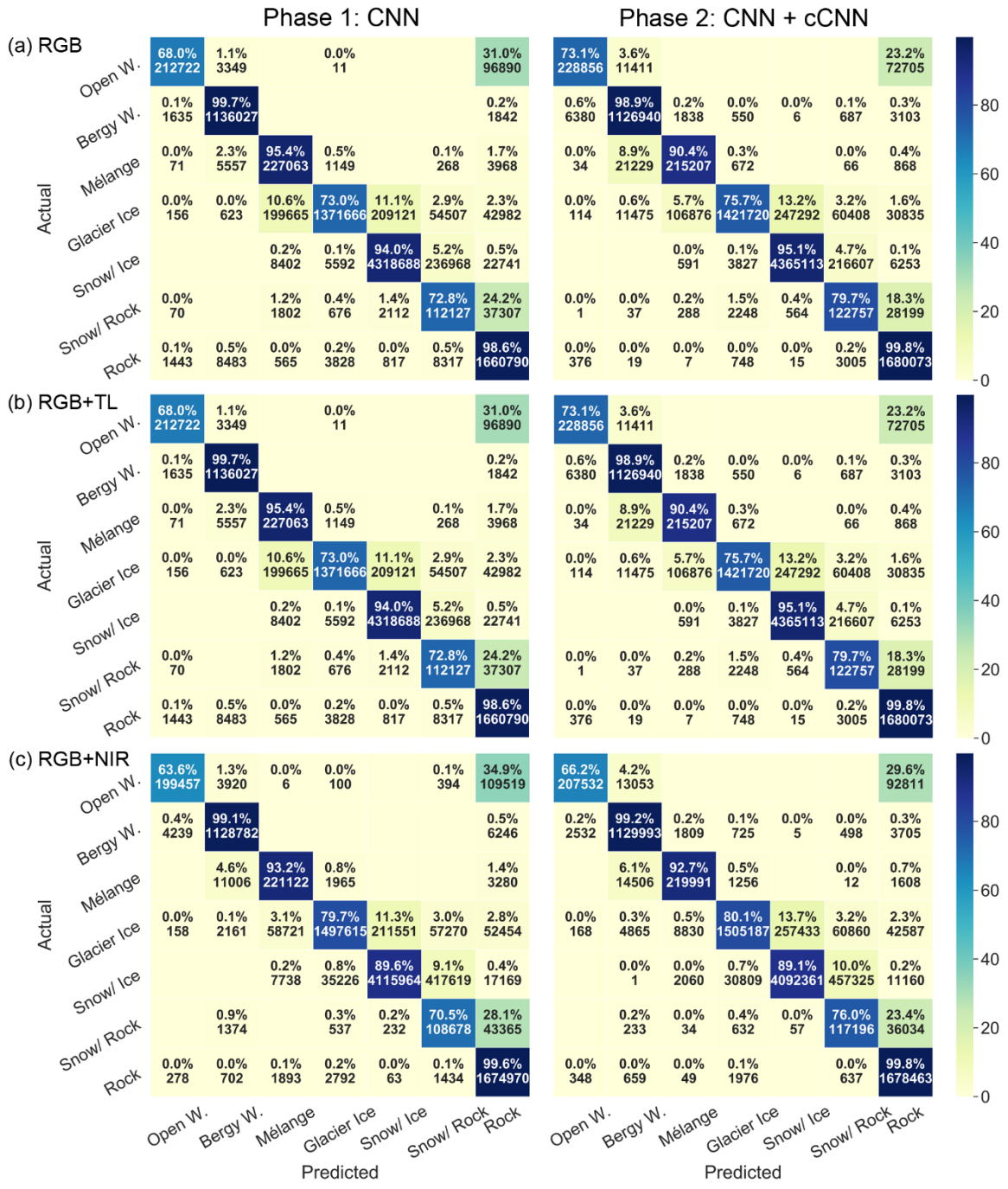


Figure A7: Confusion matrices for CSC results on the Helheim test image using 75x75 pixel tiles and a patch-based approach (patch size: 7).

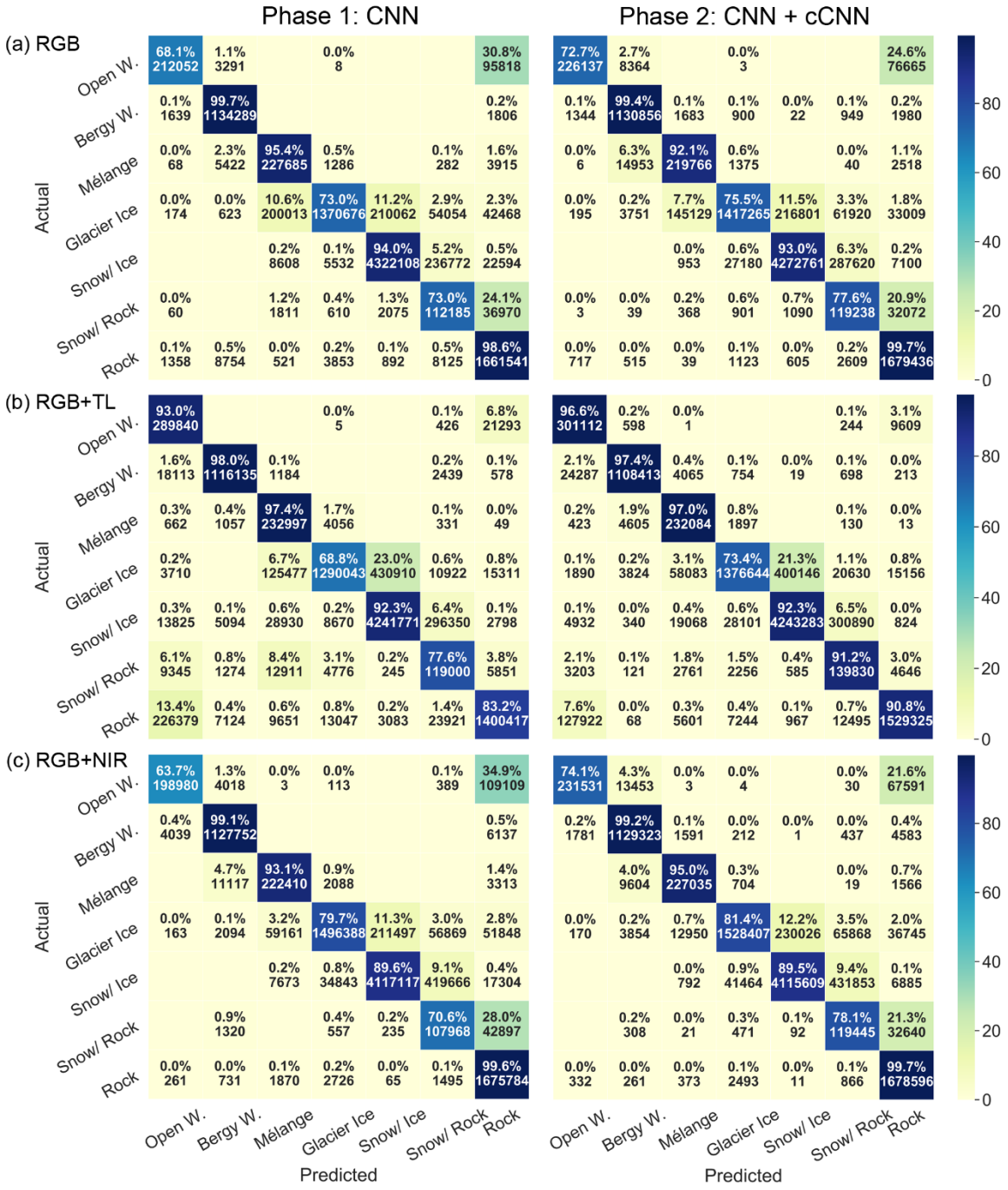


Figure A8: Confusion matrices for CSC results on the Helheim test image using 75x75 pixel tiles and a patch-based approach (patch size: 15).

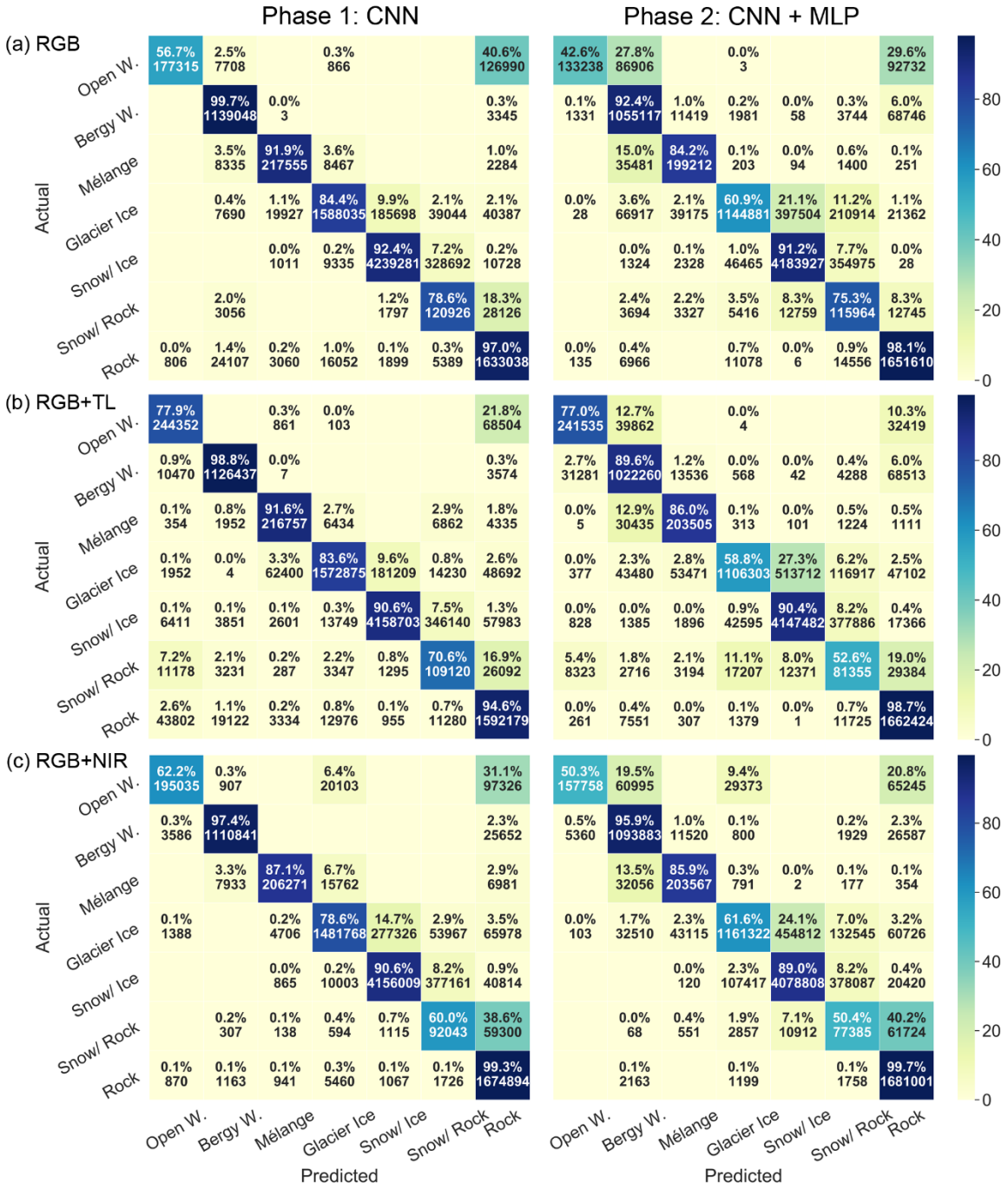


Figure A9: Confusion matrices for CSC results on the Helheim test image using 100x100 pixel tiles and a pixel-based approach (patch size: 1).

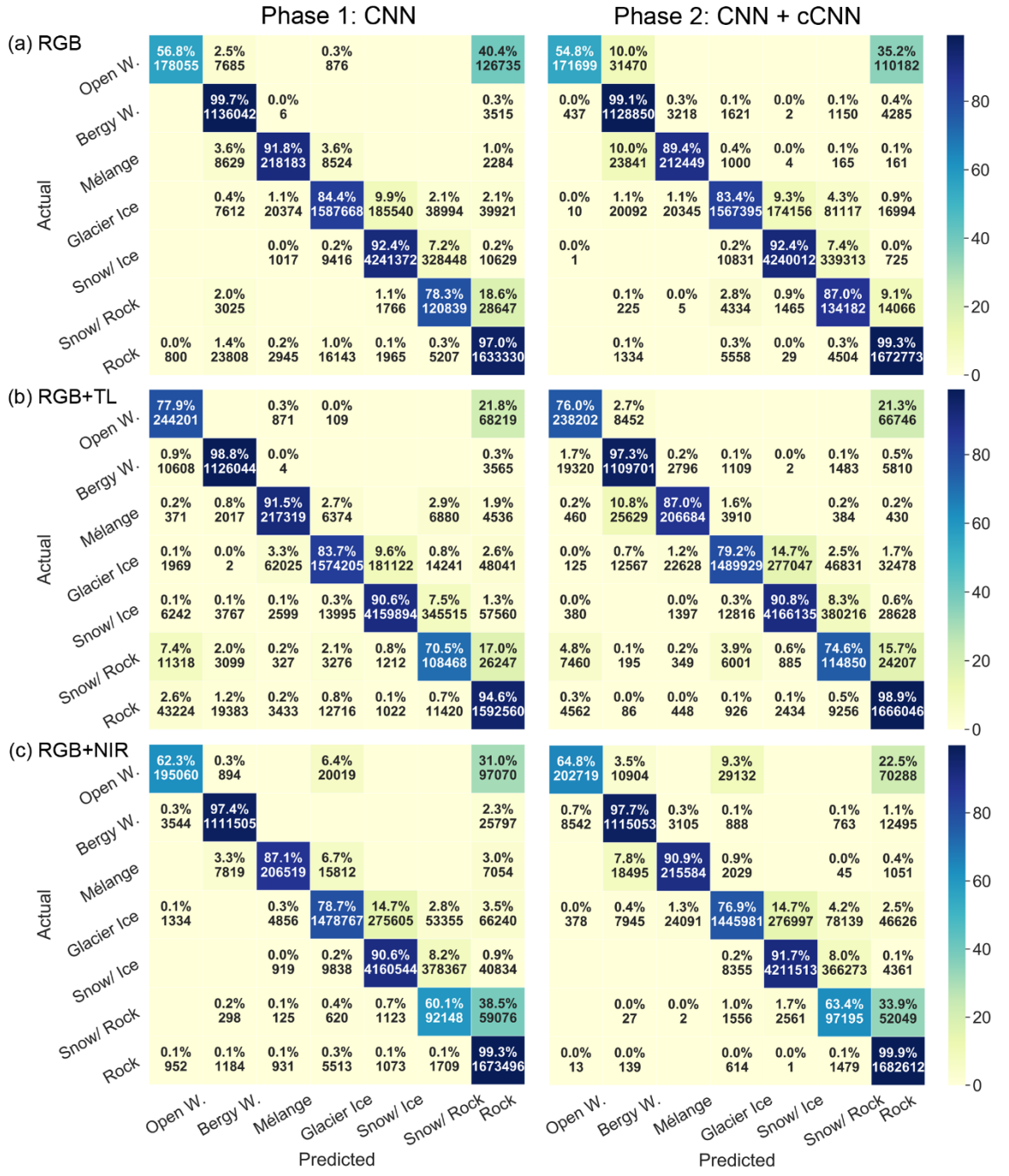


Figure A10: Confusion matrices for CSC results on the Helheim test image using 100x100 pixel tiles and a patch-based approach (patch size: 3).

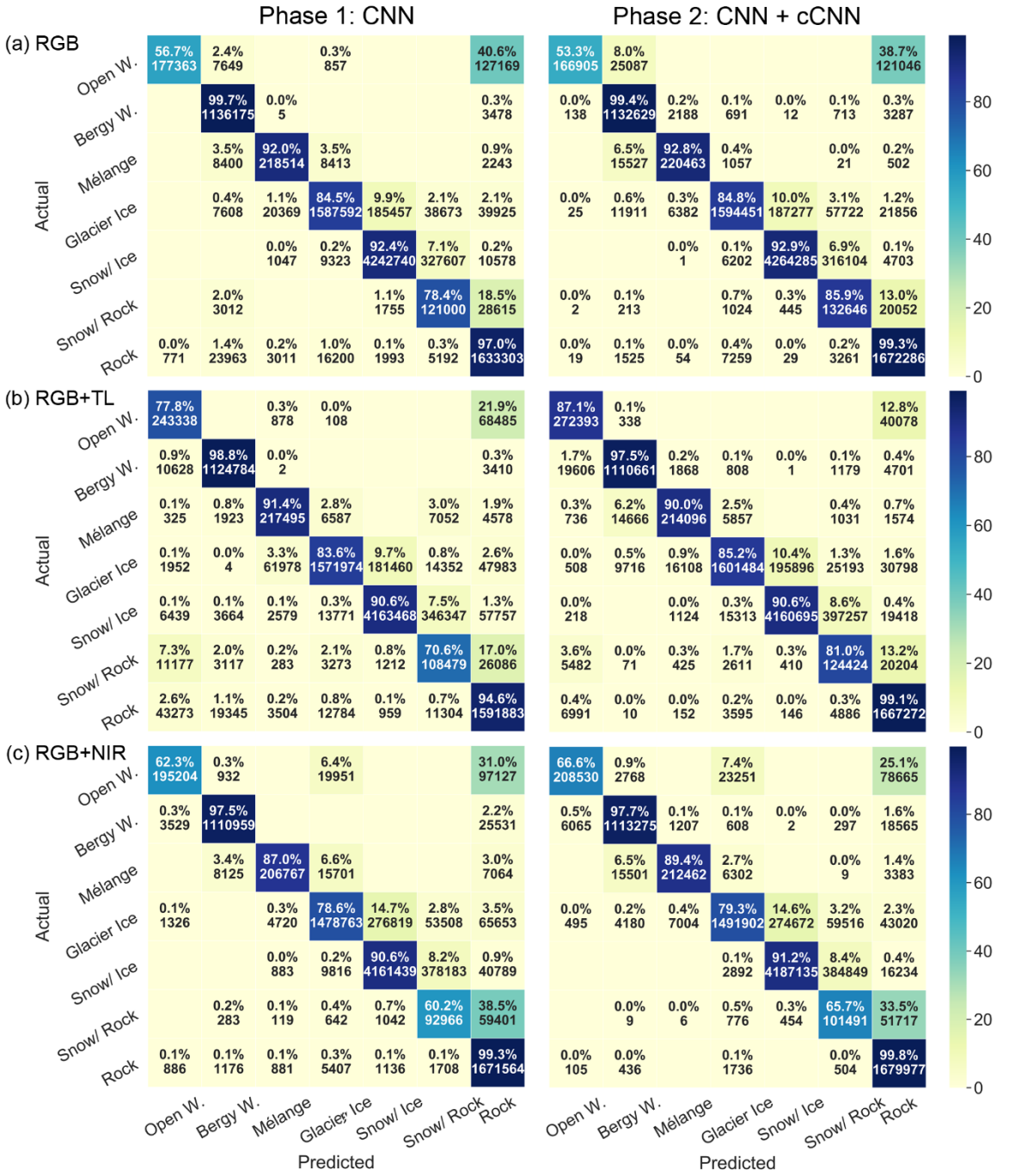


Figure A11: Confusion matrices for CSC results on the Helheim test image using 100x100 pixel tiles and a patch-based approach (patch size: 7).

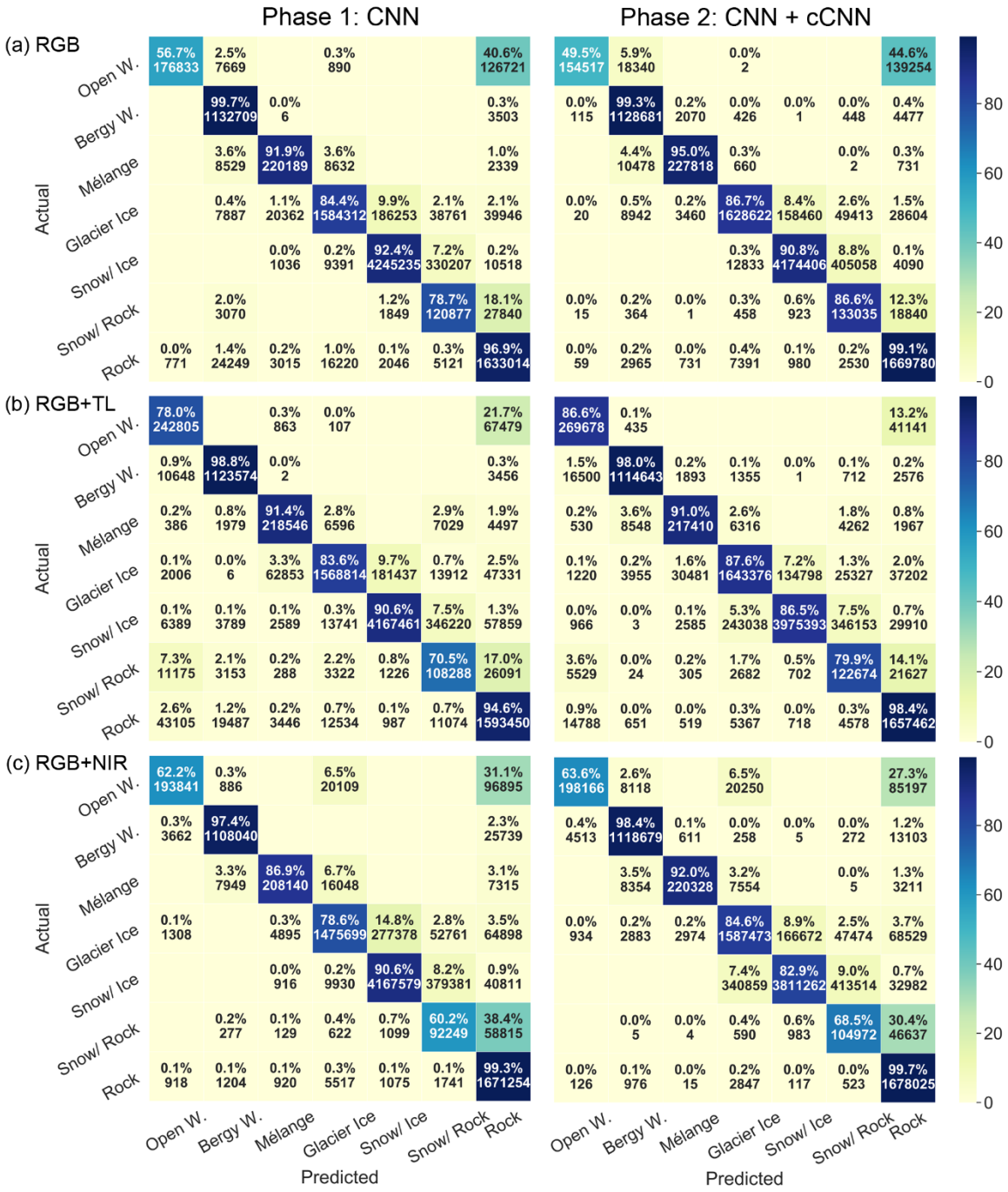


Figure A12: Confusion matrices for CSC results on the Helheim test image using 100x100 pixel tiles and a patch-based approach (patch size: 15).

Scoresby Confusion Matrices

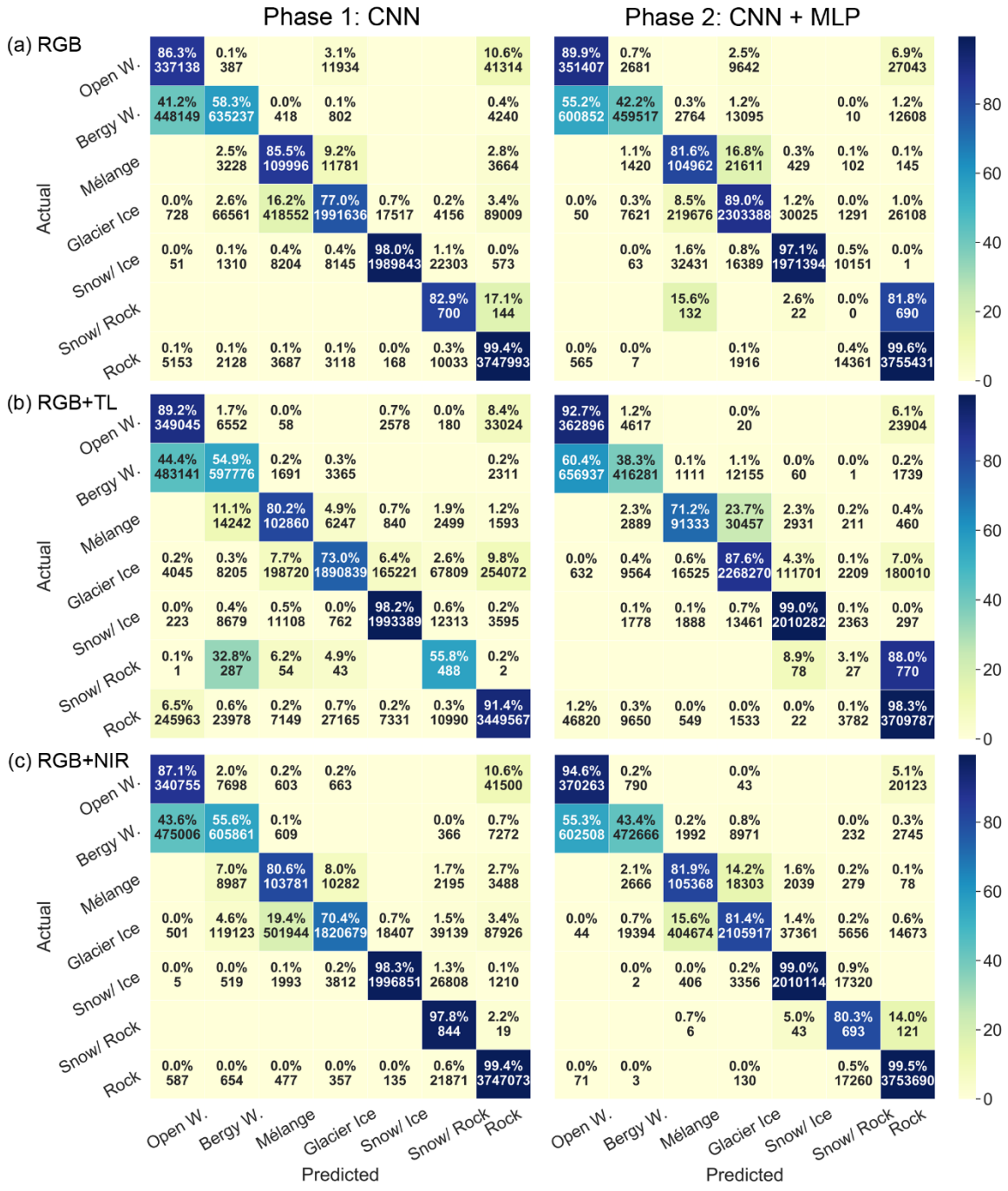


Figure A13: Confusion matrices for CSC results on the Scoresby test image using 50x50 pixel tiles and a pixel-based approach (patch size: 1).

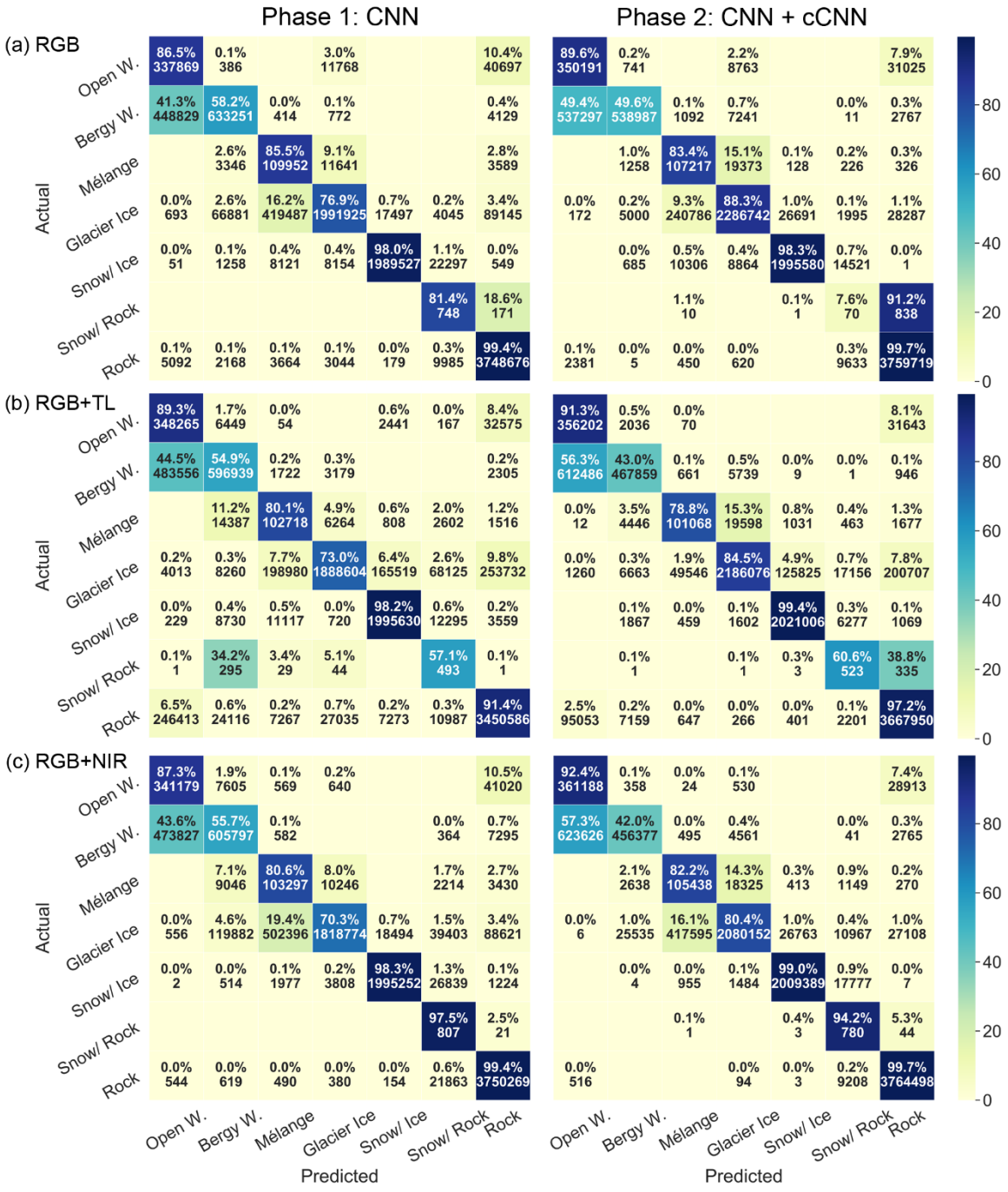


Figure A14: Confusion matrices for CSC results on the Scoresby test image using 50x50 pixel tiles and a patch-based approach (patch size: 3).

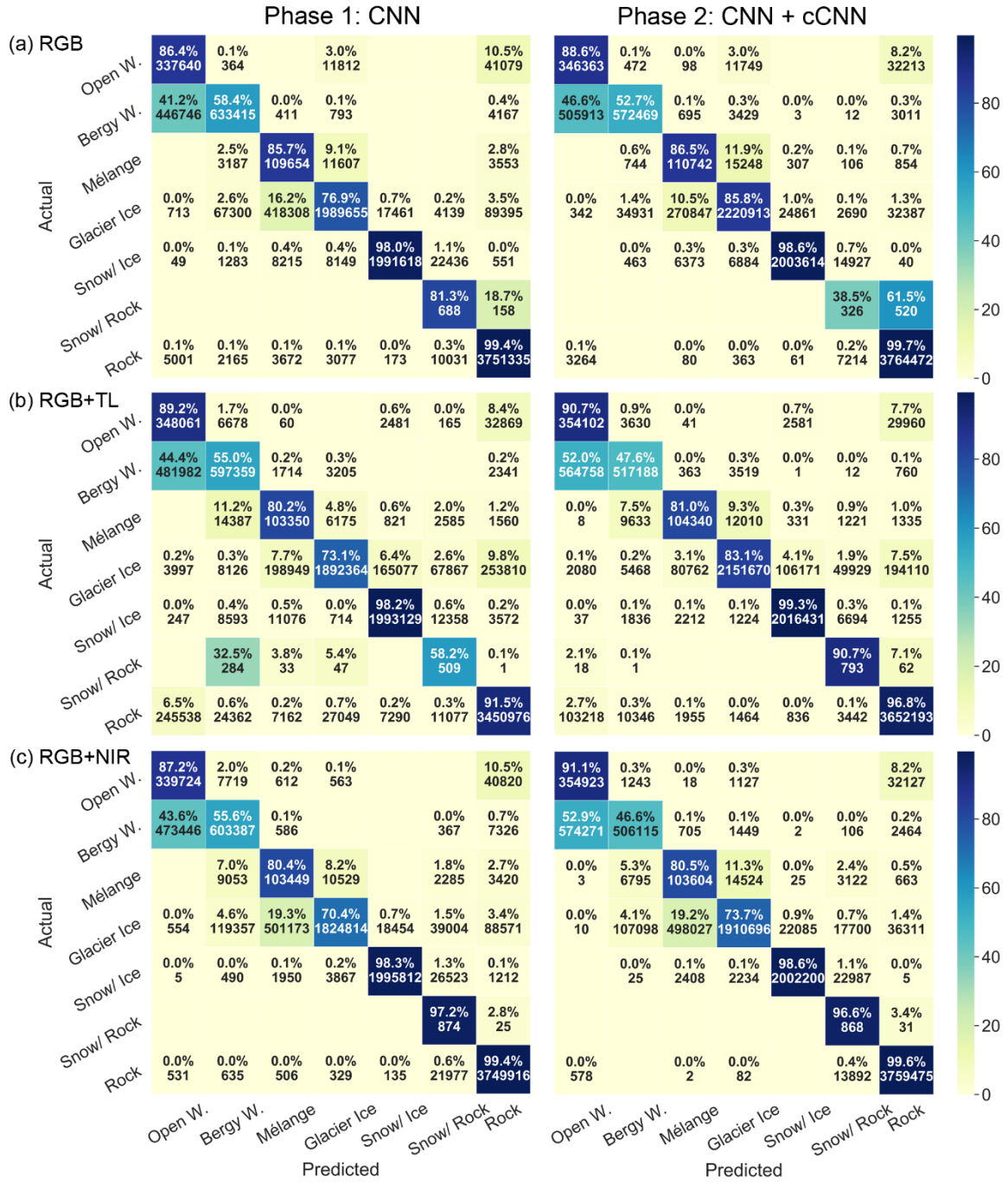


Figure A15: Confusion matrices for CSC results on the Scoresby test image using 50x50 pixel tiles and a patch-based approach (patch size: 7).

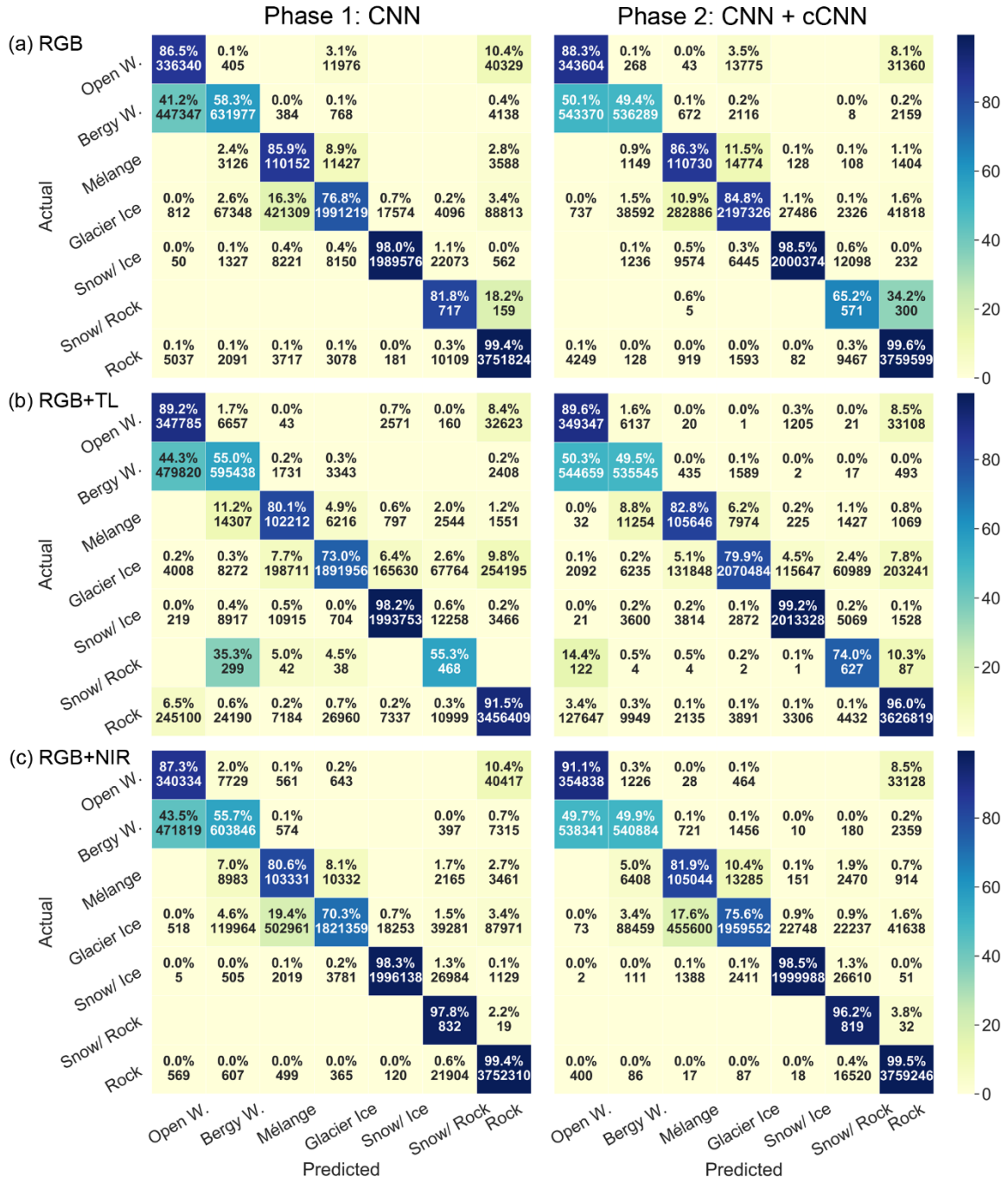


Figure A16: Confusion matrices for CSC results on the Scoresby test image using 50x50 pixel tiles and a patch-based approach (patch size: 15).

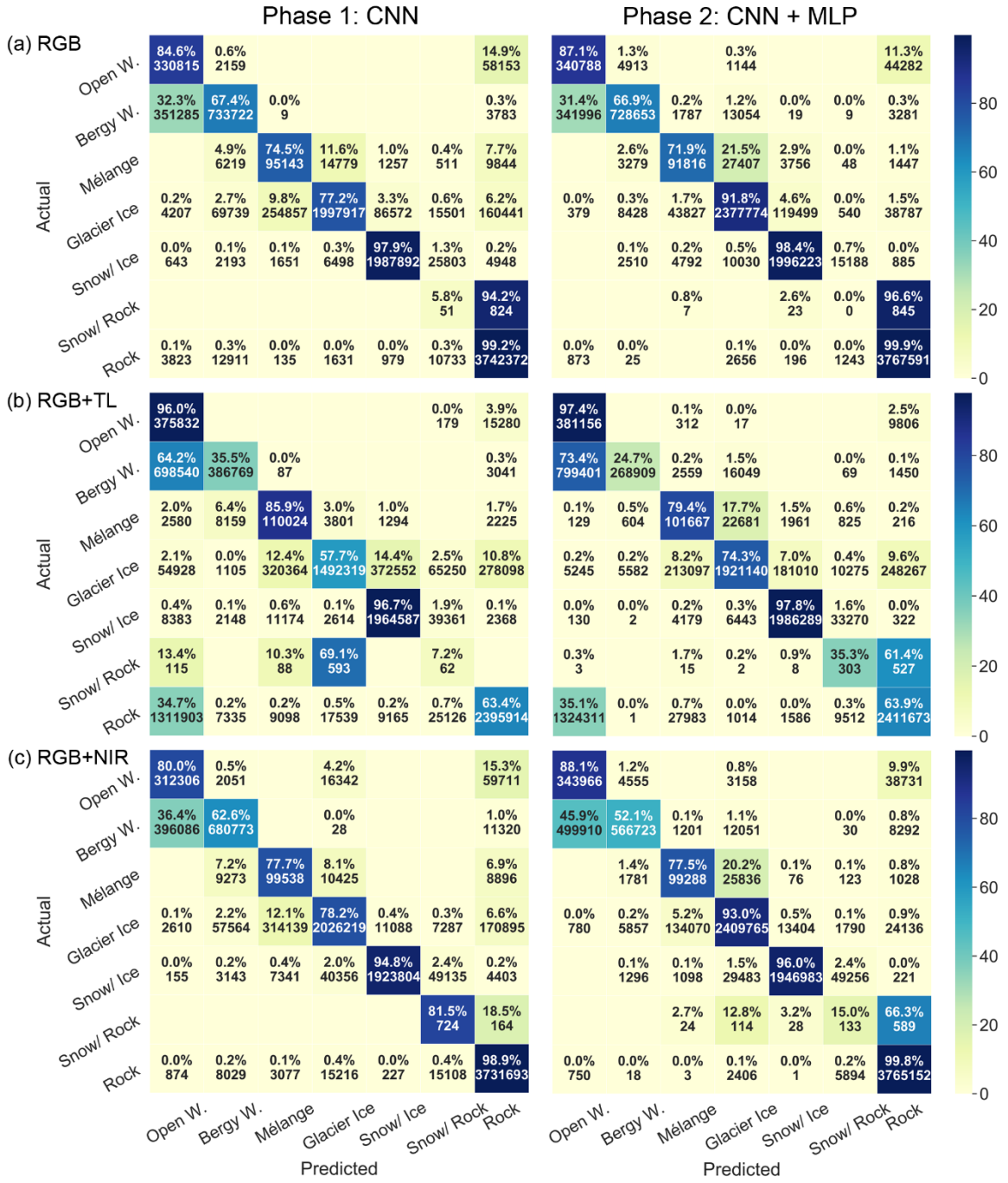


Figure A17: Confusion matrices for CSC results on the Scoresby test image using 75x75 pixel tiles and a pixel-based approach (patch size: 1).

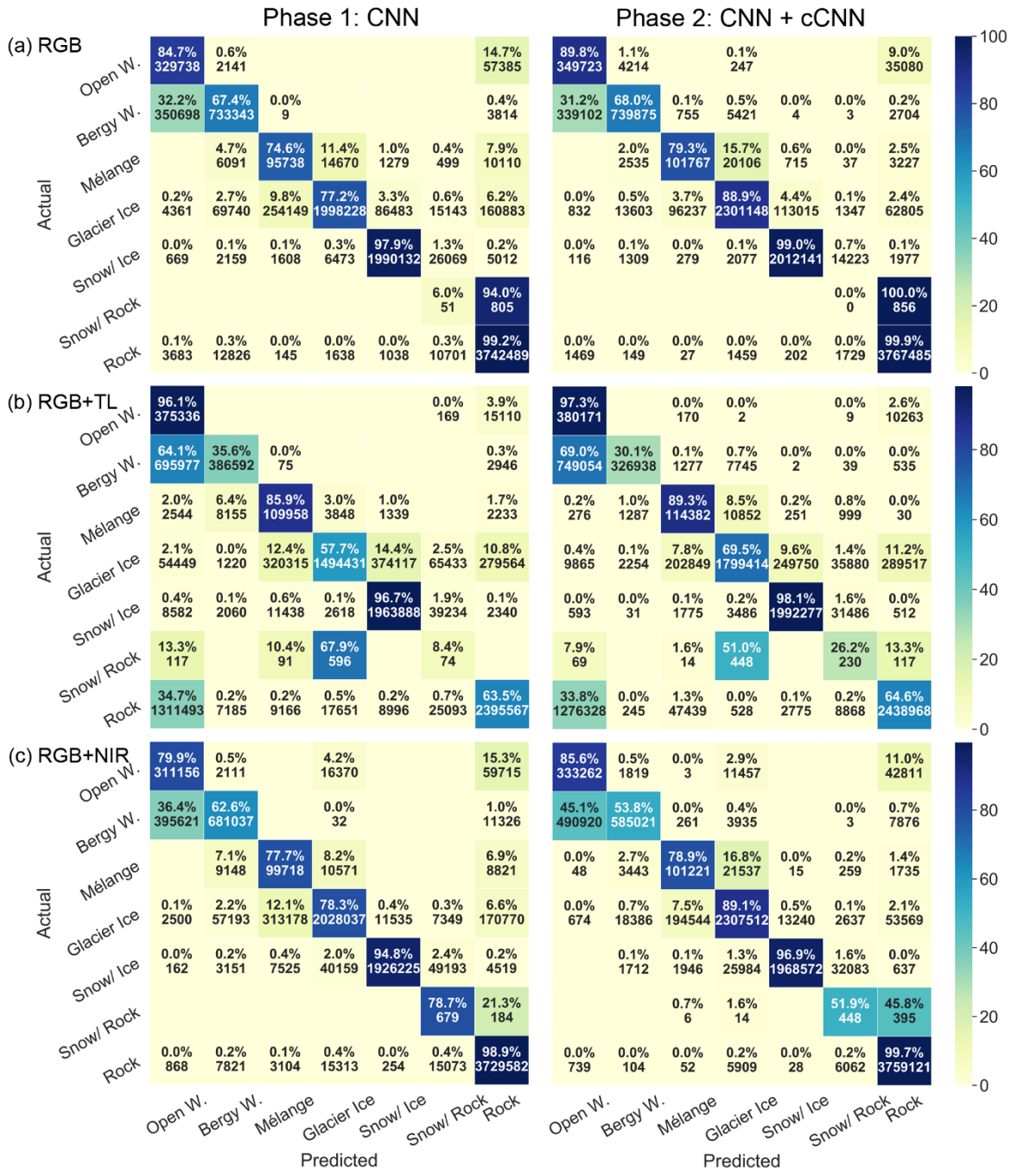


Figure A18: Confusion matrices for CSC results on the Scoresby test image using 75x75 pixel tiles and a patch-based approach (patch size: 3).

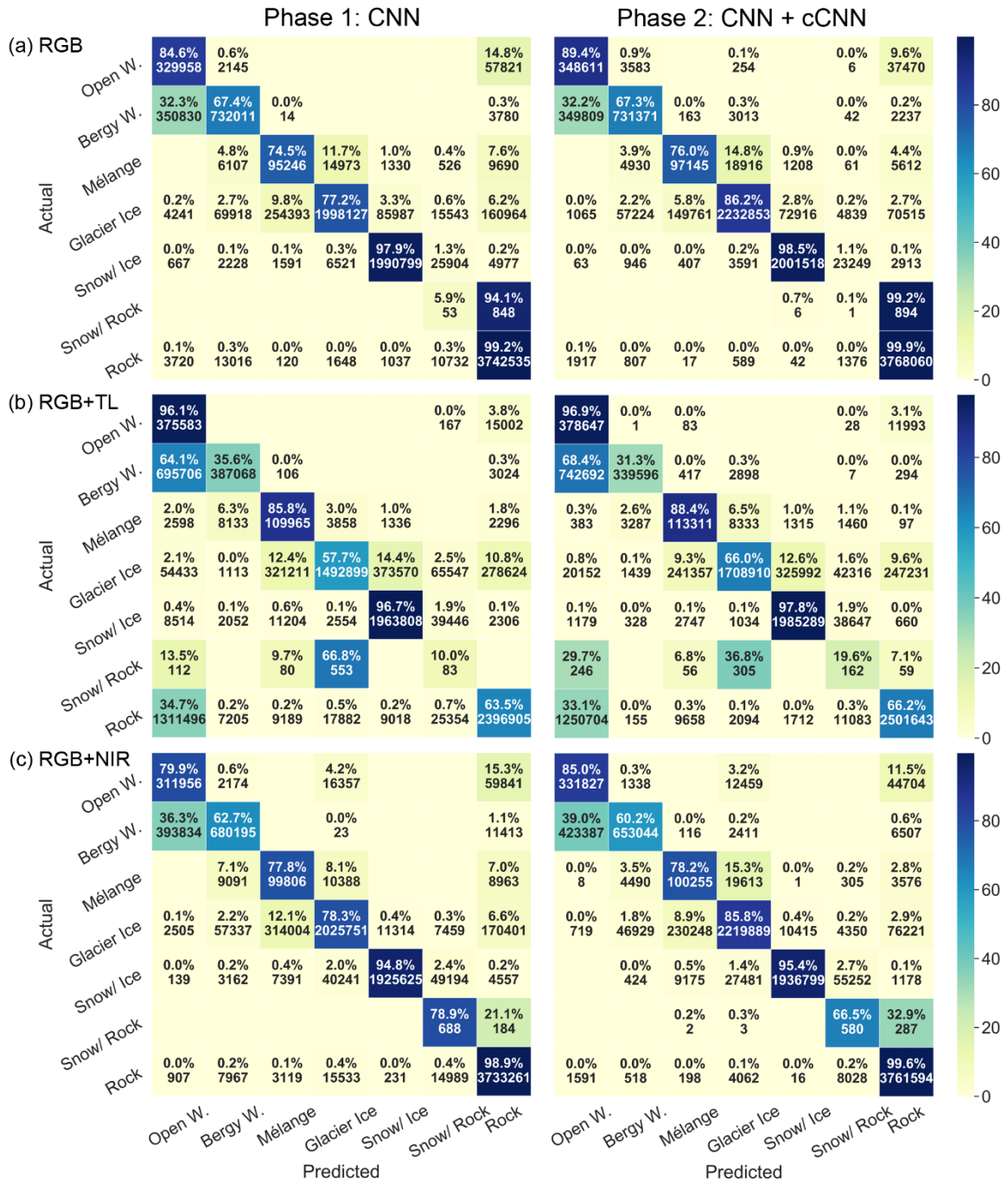


Figure A19: Confusion matrices for CSC results on the Scoresby test image using 75x75 pixel tiles and a patch-based approach (patch size: 7).

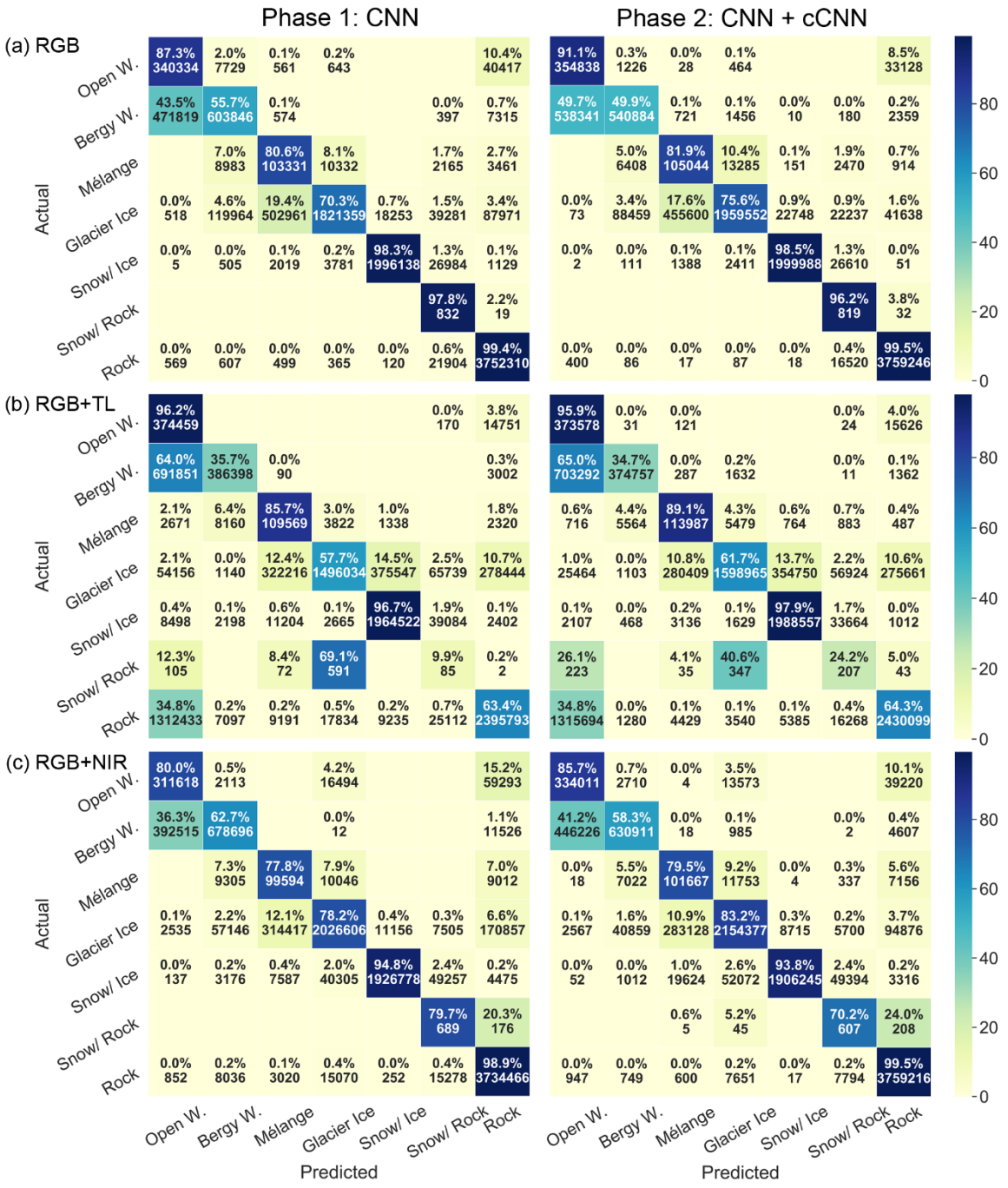


Figure A20: Confusion matrices for CSC results on the Scoresby test image using 75x75 pixel tiles and a patch-based approach (patch size: 15).

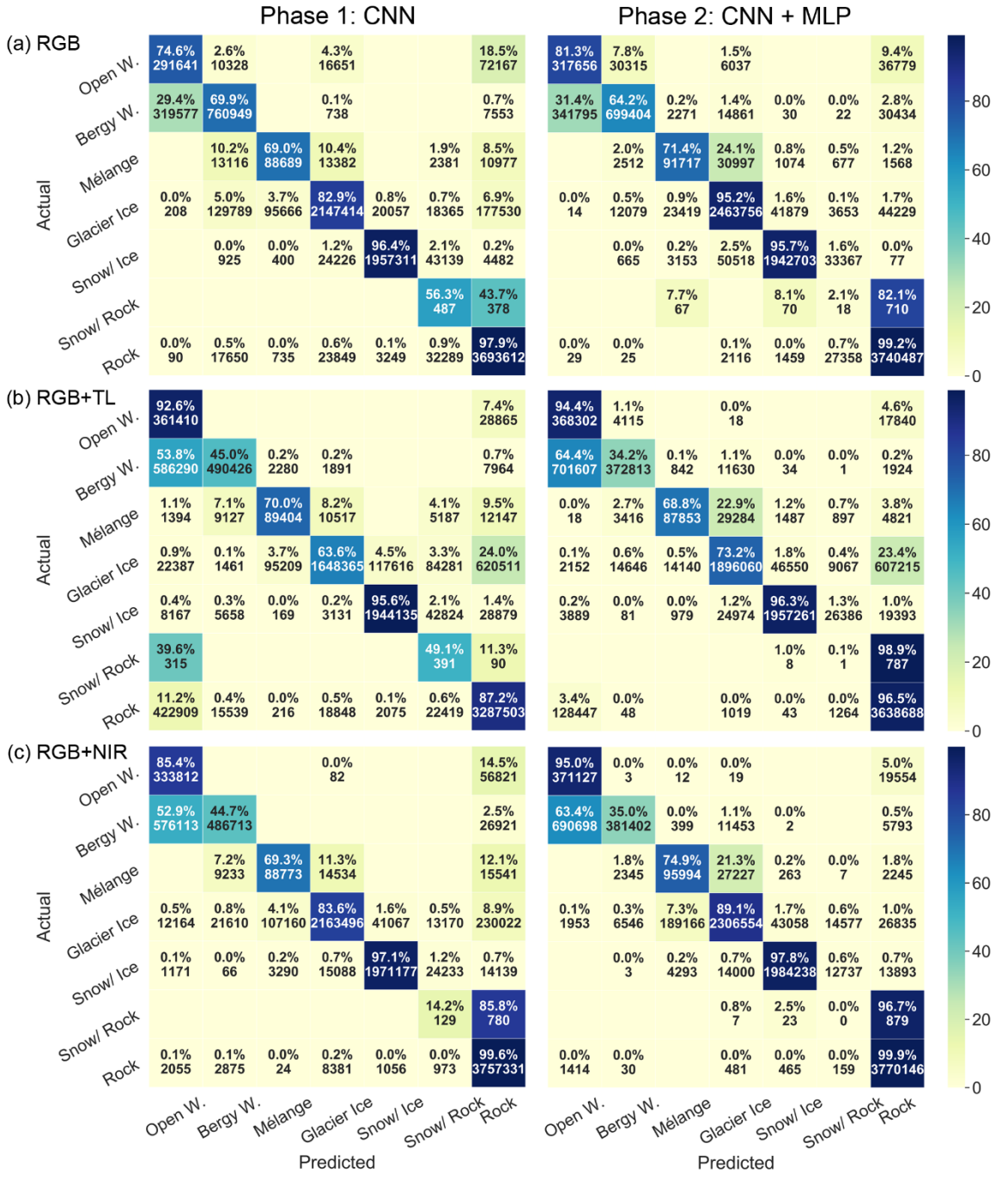


Figure A21: Confusion matrices for CSC results on the Scoresby test image using 100x100 pixel tiles and a pixel-based approach (patch size: 1).

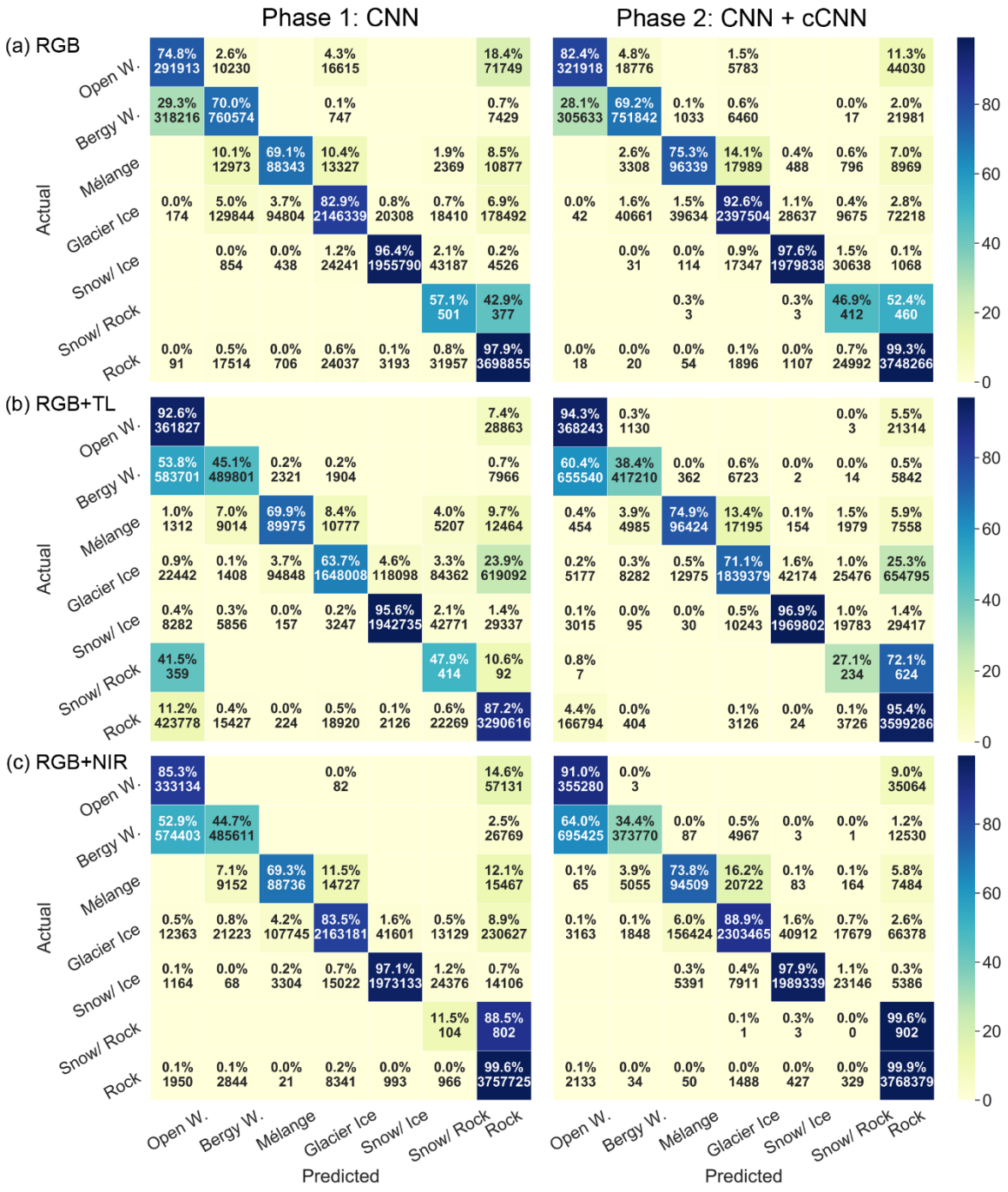


Figure A22: Confusion matrices for CSC results on the Scoresby test image using 100x100 pixel tiles and a patch-based approach (patch size: 3).

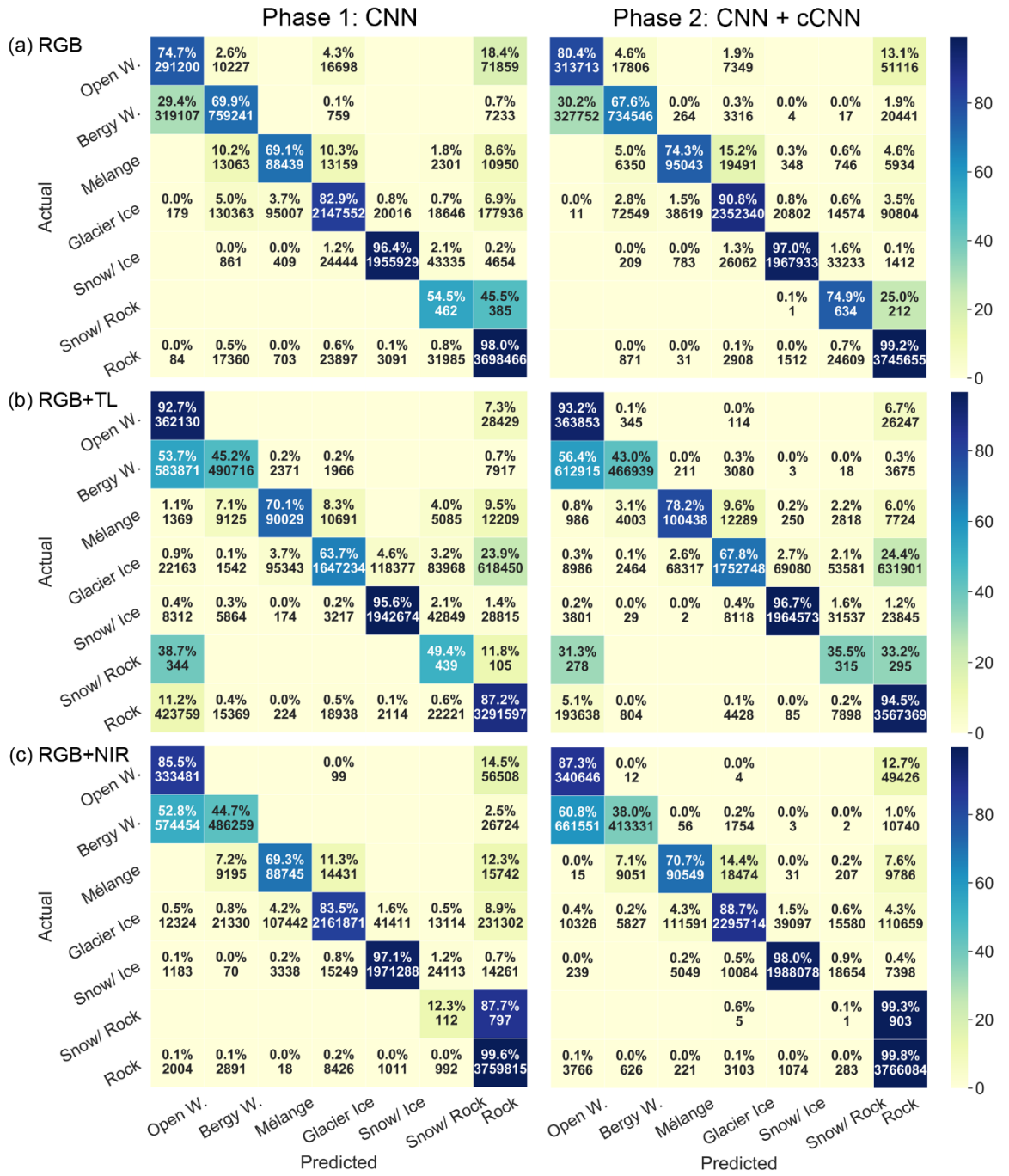


Figure A23: Confusion matrices for CSC results on the Scoresby test image using 100x100 pixel tiles and a patch-based approach (patch size: 7).

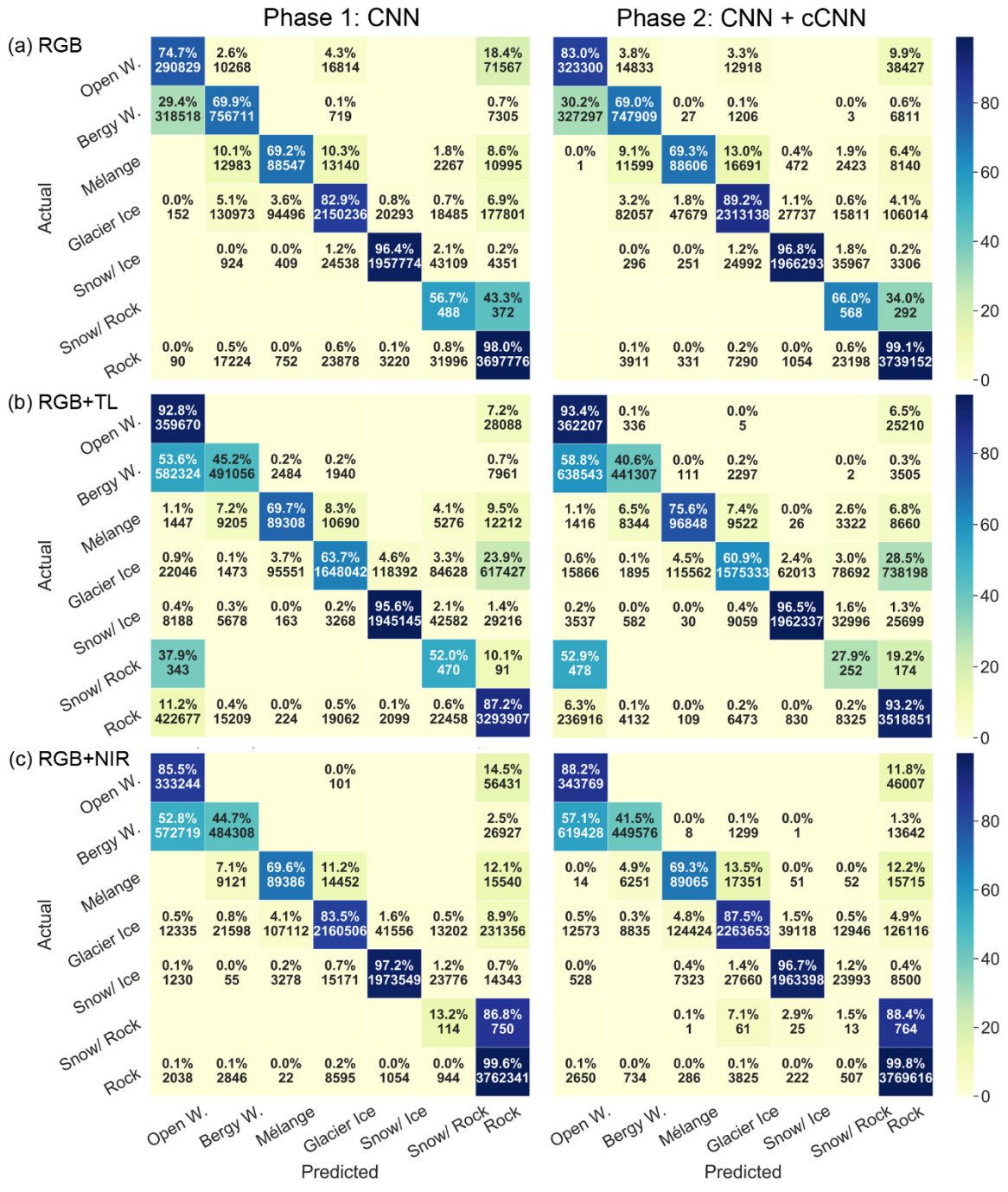


Figure A24: Confusion matrices for CSC results on the Scoresby test image using 100x100 pixel tiles and a patch-based approach (patch size: 15).

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., and Yu, D.: Convolutional Neural Networks for speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process., 22, 10, 1533–1545, https://doi.org/10.1109/TASLP.2014.2339736, 2014.

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, https://doi.org/10.1080/2150704X.2015.1034886, 2015.

Amaral, T., Bartholomaus, T. C., and Enderlin, E. M.: Evaluation of iceberg calving models against observations from Greenland outlet glaciers, Journal of Geophysical Research: Earth Surface, 125, e2019JF005444, https://doi.org/10.1029/2019JF005444, 2020.

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, https://doi.org/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, 66, 258, 577-590, https://doi.org/10.1017/jog.2020.29, 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 Geoscience, 5, 37–41, https://doi.org/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, https://doi.org/10.1016/j.csr.2013.10.003, 2013.

Atkinson, P. M. and Tatnall, A. R. L.: Introduction Neural networks in remote sensing, International Journal of Remote Sensing, 18, 699–709, https://doi.org/10.1080/014311697218700, 1997.

Baumhoer, C. A., Dietz, A. J., Dech, S., and Kuenzer, C.: Remote sensing of Antarctic glacier and ice-shelf front dynamics—a review, Remote Sensing, 10, 1445, https://doi.org/10.3390/rs10091445, 2018.

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, 2529, https://doi.org/10.3390/rs11212529, 2019.

Beaird, N. L., Straneo, F., and Jenkins, W.: Export of strongly diluted Greenland meltwater from a major glacial fjord, Geophysical Research Letters, 45, 4163–4170, https://doi.org/10.1029/2018GL077000, 2018.

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, 385–396, https://doi.org/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, 923–937, https://doi.org/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, 2303–2315, https://doi.org/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, ISPRS Journal of Photogrammetry and Remote Sensing,16, 2000.

Bolch, T., Menounos, B., and Wheate, R.: Landsat-based inventory of glaciers in western Canada, 1985–2005, Remote Sensing of Environment, 114, 127–137, https://doi.org/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, https://doi.org/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, Journal of Glaciology, 64, 523–535, https://doi.org/10.1017/jog.2018.44, 2018.

Buscombe, D. and Ritchie, A. C.: Landscape classification with deep neural networks, Geosciences, 8, 244, https://doi.org/10.3390/geosciences8070244, 2018.

Canny, J.: A computational approach to edge detection, PAMI-8, 679–698, https://doi.org/10.1109/TPAMI.1986.4767851, 1986.

Cape, M. R., Straneo, F., Beaird, N., Bundy, R. M., and Charette, M. A.: Nutrient release to oceans from buoyancy-driven upwelling at Greenland tidewater glaciers, Nature Geoscience, 12, 34–39, https://doi.org/10.1038/s41561-018-0268-4, 2019.

Carbonneau, P. E. and Marochov, M.: SEE_ICE: Glacial landscape classification with deep learning, Zenodo, https://doi.org/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, https://doi.org/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, https://doi.org/10.1002/esp.4955, 2020b.

Carr, J. R., Stokes, C. R., and Vieli, A.: Recent progress in understanding marine-terminating Arctic outlet glacier response to climatic and oceanic forcing: Twenty years of rapid change, Progress in Physical Geography: Earth and Environment, 37, 436–467, https://doi.org/10.1177/0309133313483163, 2013.

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, 72–91, https://doi.org/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, 9739–9748, https://doi.org/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, 76–88, https://doi.org/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, 2024–2038, https://doi.org/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, e2018JF004873, https://doi.org/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, 1457–1468, https://doi.org/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, 636–643, https://doi.org/10.1109/21.370194, 1995.

Cheng, G., Han, J., and Lu, X.: Remote sensing image scene classification: benchmark and state of the art, Proceedings of the IEEE, 105, 1865–1883, https://doi.org/10.1109/JPROC.2017.2675998, 2017.

Chollet, F.: Deep learning with Python, Manning Publications Co, Shelter Island, New York, 361 pp., 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, Annals of Glaciology, 53, 249–256, https://doi.org/10.3189/2012AoG60A087, 2012.

Cohen, J.: A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, 37–46, https://doi.org/10.1177/001316446002000104, 1960.

Cook, A. J., Vaughan, D. G., Luckman, A. J., and Murray, T.: A new Antarctic Peninsula glacier basin inventory and observed area changes since the 1940s, Antarctic Science, 26, 614–624, https://doi.org/10.1017/S0954102014000200, 2014.

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, https://doi.org/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, PNAS, 111, 18478–18483, https://doi.org/10.1073/pnas.1411680112, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei: ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255, https://doi.org/10.1109/CVPR.2009.5206848, 2009.

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, 866–872, https://doi.org/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, 13467, https://doi.org/10.1038/s41598-018-31875-8, 2018.

Felikson, D., Bartholomaus, T. C., Catania, G. A., Korsgaard, N. J., Kjær, K. H., Morlighem, M., Noël, B., van den Broeke, M., Stearns, L. A., Shroyer, E. L., Sutherland, D. A., and Nash, J. D.: Inland thinning on the Greenland ice sheet controlled by outlet glacier geometry, Nature Geoscience, 10, 366–369, https://doi.org/10.1038/ngeo2934, 2017.

Felleman, D. J. and Van Essen, D. C.: Distributed Hierarchical Processing in the Primate Cerebral Cortex, Cerebral Cortex, 1, 1–47, https://doi.org/10.1093/cercor/1.1.1, 1991.

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, 81–91, https://doi.org/10.4031/MTSJ.48.5.3, 2014.

Fretwell, P., Pritchard, H. D., Vaughan, D. G., Bamber, J. L., Barrand, N. E., Bell, R., Bianchi, C., Bingham, R. G., Blankenship, D. D., Casassa, G., Catania, G., Callens, D., Conway, H., Cook, A. J., Corr, H. F. J., Damaske, D., Damm, V., Ferraccioli, F., Forsberg, R., Fujita, S., Gim, Y., Gogineni, P., Griggs, J. A., Hindmarsh, R. C. A., Holmlund, P., Holt, J. W., Jacobel, R. W., Jenkins, A., Jokat, W., Jordan, T., King, E. C., Kohler, J., Krabill, W., Riger-Kusk, M., Langley, K. A., Leitchenkov, G., Leuschen, C., Luyendyk, B. P., Matsuoka, K., Mouginot, J., Nitsche, F. O., Nogi, Y., Nost, O. A., Popov, S. V., Rignot, E., Rippin, D. M., Rivera, A., Roberts, J., Ross, N., Siegert, M. J., Smith, A. M., Steinhage, D., Studinger, M., Sun, B., Tinto, B. K., Welch, B. C., Wilson, D., Young, D. A., Xiangbin, C., and Zirizzotti, A.: Bedmap2: improved ice bed, surface and thickness datasets for Antarctica, The Cryosphere, 7, 375–393, https://doi.org/10.5194/tc-7-375-2013, 2013.

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, https://doi.org/10.1016/j.rse.2012.06.020, 2012.

Fried, M. J., Catania, G. A., Stearns, L. A., Sutherland, D. A., Bartholomaus, T. C., Shroyer, E., and Nash, J.: Reconciling drivers of seasonal terminus advance and retreat at 13 Central West Greenland tidewater glaciers, Journal of Geophysical Research: Earth Surface, 123, 1590–1607, https://doi.org/10.1029/2018JF004628, 2018.

Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernetics, 36, 193–202, https://doi.org/10.1007/BF00344251, 1980.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, 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, https://doi.org/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, 357–372, https://doi.org/10.3189/2015JoG14J209, 2015.

Hawkings, J. R., Wadham, J. L., Tranter, M., Raiswell, R., Benning, L. G., Statham, P. J., Tedstone, A., Nienow, P., Lee, K., and Telling, J.: Ice sheets as a significant source of highly reactive nanoparticulate iron to the oceans, Nature Communications, 5, 3929, https://doi.org/10.1038/ncomms4929, 2014.

Hill, E. A., Carr, J. R., Stokes, C. R., and Gudmundsson, G. H.: Dynamic changes in outlet glaciers in northern Greenland from 1948 to 2015, The Cryosphere, 12, 3243–3263, https://doi.org/10.5194/tc-12-3243-2018, 2018.

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, 205, https://doi.org/10.3390/rs13020205, 2021.

Holmes, F. A., Kirchner, N., Kuttenkeuler, J., Krützfeldt, J., and Noormets, R.: Relating ocean temperatures to frontal ablation rates at Svalbard tidewater glaciers: Insights from glacier proximal datasets, Scientific Reports, 9, 9442, https://doi.org/10.1038/s41598-019-45077-3, 2019.

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, 2691–2710, https://doi.org/10.5194/tc-11-2691-2017, 2017.

Howat, I. M., Joughin, I., Tulaczyk, S., and Gogineni, S.: Rapid retreat and acceleration of Helheim Glacier, east Greenland, Geophysical Research Letters, 32, https://doi.org/10.1029/2005GL024737, 2005.

Howat, I. M., Box, J. E., Ahn, Y., Herrington, A., and McFadden, E. M.: Seasonal variability in the dynamics of marine-terminating outlet glaciers in Greenland, Journal of Glaciology, 56, 601–613, https://doi.org/10.3189/002214310793146232, 2010.

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, https://doi.org/10.1029/2011GL047565, 2011.

Hu, F., Xia, G.-S., Hu, J., and Zhang, L.: Transferring Deep Convolutional Neural Networks for the scene classification of high-resolution remote sensing imagery, Remote Sensing, 7, 14680–14707, https://doi.org/10.3390/rs71114680, 2015.

Hubel, D. H. and Wiesel, T. N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J Physiol, 160, 106-154.2, 1962.

Jenkins, A., Dutrieux, P., Jacobs, S. S., McPhail, S. D., Perrett, J. R., Webb, A. T., and White, D.: Observations beneath Pine Island Glacier in West Antarctica and implications for its retreat, Nature Geoscience, 3, 468–472, https://doi.org/10.1038/ngeo890, 2010.

Joughin, I., Rignot, E., Rosanova, C. E., Lucchitta, B. K., and Bohlander, J.: Timing of Recent Accelerations of Pine Island Glacier, Antarctica, Geophysical Research Letters, 30, https://doi.org/10.1029/2003GL017609, 2003.

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, https://doi.org/10.1029/2007JF000837, 2008.

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, https://doi.org/10.1029/2010GL043289, 2010.

Kehrl, L. M., Joughin, I., Shean, D. E., Floricioiu, D., and Krieger, L.: Seasonal and interannual variabilities in terminus position, glacier velocity, and surface elevation at Helheim and Kangerlussuaq Glaciers from 2008 to 2016, Journal of Geophysical Research: Earth Surface, 122, 1635–1652, https://doi.org/10.1002/2016JF004133, 2017.

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, 3813–3825, https://doi.org/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, 1–7, https://doi.org/10.1038/s43247-020-0001-2, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for Stochastic Optimization, arXiv:1412.6980, 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), 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2817–2820, https://doi.org/10.1109/IGARSS.2017.8127584, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems 25, edited by: Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., Curran Associates, Inc., 1097–1105, 2012.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., and Stober, S.: Transfer Learning for Speech Recognition on a Budget, arXiv:1706.00290, 2017.

Landis, J. R. and Koch, G. G.: The measurement of observer agreement for categorical data, Biometrics, 33, 159–174, https://doi.org/10.2307/2529310, 1977.

Längkvist, M., Kiselev, A., Alirezaie, M., and Loutfi, A.: Classification and segmentation of satellite orthoimagery using Convolutional Neural Networks, Remote Sensing, 8, 329, https://doi.org/10.3390/rs8040329, 2016.

Lea, J. M.: Google Earth Engine Digitisation Tool (GEEDiT), and Margin change Quantification Tool (MaQiT) - simple tools for the rapid mapping and quantification of changing Earth surface margins, Earth Surface Dynamics, 6, 551–561, 2018.

Lea, J. M., Mair, D. W. F., and Rea, B. R.: Evaluation of existing and new methods of tracking glacier terminus change, Journal of Glaciology, 60, 323–332, https://doi.org/10.3189/2014JoG13J061, 2014.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.: Backpropagation applied to handwritten zip code recognition, Neural Comput., 1, 541–551, https://doi.org/10.1162/neco.1989.1.4.541, 1989.

LeCun, Y., Bottou, L., Bengio, Y., and Ha, P.: Gradient-based learning applied to document recognition, 46, 1998.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, 521, 436–444, Nature, 521, 7553, https://doi.org/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, https://doi.org/10.1016/j.jag.2014.04.018, 2014.

Liu, H. and Jezek, K. C.: A complete high-resolution coastline of Antarctica extracted from orthorectified Radarsat SAR imagery, Photogramm eng remote sensing, 70, 605–616, https://doi.org/10.14358/PERS.70.5.605, 2004a.

Liu, H. and Jezek, K. C.: Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods, International Journal of Remote Sensing, 25, 937–958, https://doi.org/10.1080/0143116031000139890, 2004b.

Liu, X., Deng, Z., and Yang, Y.: Recent progress in semantic image segmentation, Artificial Intelligence Review, 52, 1089–1106, https://doi.org/10.1007/s10462-018-9641-3, 2019.

Lundervold, A. S. and Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI, Zeitschrift für Medizinische Physik, 29, 102–127, https://doi.org/10.1016/j.zemedi.2018.11.002, 2019.

Luus, F. P. S., Salmon, B. P., van den Bergh, F., and Maharaj, B. T. J.: Multiview deep learning for land-use classification, IEEE Geoscience and Remote Sensing Letters, 12, 2448–2452, https://doi.org/10.1109/LGRS.2015.2483680, 2015.

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), 2016 IEEE International Geoscience and

Remote Sensing Symposium (IGARSS), 5071–5074, https://doi.org/10.1109/IGARSS.2016.7730322, 2016.

Marochov, M., and Carbonneau, P.: Image classification of marine-terminating outlet glaciers using deep learning methods: pre-trained models [dataset], http://doi.org/10.15128/r2gh93gz51k, 2020.

Miles, B. W. J., Stokes, C. R., Vieli, A., and Cox, N. J.: Rapid, climate-driven changes in outlet glaciers on the Pacific coast of East Antarctica, Nature, 500, 563–566, https://doi.org/10.1038/nature12382, 2013.

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, https://doi.org/10.1126/sciadv.1501350, 2016.

Miles, B. W. J., Stokes, C. R., and Jamieson, S. S. R.: Simultaneous disintegration of outlet glaciers in Porpoise Bay (Wilkes Land), East Antarctica, driven by sea ice break-up, The Cryosphere, 11, 427–442, https://doi.org/10.5194/tc-11-427-2017, 2017.

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, 3123–3136, https://doi.org/10.5194/tc-12-3123-2018, 2018.

Miles, B. W. J., Jordan, J. R., Stokes, C. R., Jamieson, S. S. R., Gudmundsson, G. H., and Jenkins, A.: Recent acceleration of Denman Glacier (1972-2017), East Antarctica, driven by grounding line retreat and changes in ice tongue configuration, The Cryosphere, 15, 2, https://doi.org/10.5194/tc-15-663-2021, 2021.

Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Detection of glacier calving margins with Convolutional Neural Networks: A case study, Remote Sensing, 11, 74, https://doi.org/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, https://doi.org/10.1029/2007JF000927, 2008.

Moon, T., Joughin, I., Smith, B., Broeke, M. R. van den, Berg, W. J. van de, Noël, B., and Usher, M.: Distinct patterns of seasonal Greenland glacier velocity, Geophysical Research Letters, 41, 7209–7216, https://doi.org/10.1002/2014GL061836, 2014.

Moon, T., Joughin, I., and Smith, B.: Seasonal to multiyear variability of glacier surface velocity, terminus position, and sea ice/ice mélange in northwest Greenland, Journal of Geophysical Research: Earth Surface, 120, 5, 818–833, https://doi.org/10.1002/2015JF003494, 2015.

Moon, T., Sutherland, D. A., Carroll, D., Felikson, D., Kehrl, L., and Straneo, F.: Subsurface iceberg melt key to Greenland fjord freshwater budget, Nature Geoscience, 11, 49–54, https://doi.org/10.1038/s41561-017-0018-z, 2018.

Morlighem, M., Williams, C. N., Rignot, E., An, L., Arndt, J. E., Bamber, J. L., Catania, G., Chauché, N., Dowdeswell, J. A., Dorschel, B., Fenty, I., Hogan, K., Howat, I., Hubbard, A., Jakobsson, M., Jordan, T. M., Kjeldsen, K. K., Millan, R., Mayer, L., Mouginot, J., Noël, B. P. Y., O'Cofaigh, C., Palmer, S., Rysgaard, S., Seroussi, H., Siegert, M. J., Slabon, P., Straneo, F., van den Broeke, M. R., Weinrebe, W., Wood, M., and Zinglersen, K. B.: BedMachine v3: Complete bed topography and ocean bathymetry

mapping of Greenland from multibeam echo sounding combined with mass conservation, Geophysical Research Letters, 44, 11051-11061, https://doi.org/10.1002/2017GL074954, 2017.

Motyka, R. J., Truffer, M., Fahnestock, M., Mortensen, J., Rysgaard, S., and Howat, I.: Submarine melting of the 1985 Jakobshavn Isbræ floating tongue and the triggering of the current retreat, Journal of Geophysical Research: Earth Surface, 116, https://doi.org/10.1029/2009JF001632, 2011.

Motyka, R. J., Cassotto, R., Truffer, M., Kjeldsen, K. K., As, D. V., Korsgaard, N. J., Fahnestock, M., Howat, I., Langen, P. L., Mortensen, J., Lennert, K., and Rysgaard, S.: Asynchronous behavior of outlet glaciers feeding Godthåbsfjord (Nuup Kangerlua) and the triggering of Narsap Sermia's retreat in SW Greenland, Journal of Glaciology, 63, 288–308, https://doi.org/10.1017/jog.2016.138, 2017.

Mouginot, J., Rignot, E., and Scheuchl, B.: Sustained increase in ice discharge from the Amundsen Sea Embayment, West Antarctica, from 1973 to 2013, Geophysical Research Letters, 41, 1576–1584, https://doi.org/10.1002/2013GL059069, 2014.

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, 9239–9244, https://doi.org/10.1073/pnas.1904242116, 2019.

Nick, F. M., Vieli, A., Howat, I. M., and Joughin, I.: Large-scale changes in Greenland outlet glacier dynamics triggered at the terminus, Nature Geoscience, 2, 110–114, https://doi.org/10.1038/ngeo394, 2009.

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, 759–773, https://doi.org/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, https://doi.org/10.1126/sciadv.aaw0123, 2019.

Ostankovich, V. and Afanasyev, I.: Illegal Buildings Detection from Satellite Images using GoogLeNet and Cadastral Map, in: 2018 International Conference on Intelligent Systems (IS), 2018 International Conference on Intelligent Systems (IS), 616–623, https://doi.org/10.1109/IS.2018.8710565, 2018.

Palafox, L. F., Hamilton, C. W., Scheidt, S. P., and Alvarez, A. M.: Automated detection of geological landforms on Mars using Convolutional Neural Networks, Computers & Geosciences, 101, 48–56, https://doi.org/10.1016/j.cageo.2016.12.015, 2017.

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, https://doi.org/10.3390/rs8070575, 2016.

Pires de Lima, R. and Marfurt, K.: Convolutional Neural Network for remote-sensing scene classification: transfer learning analysis, Remote Sensing 12, 1, https://doi.org/10.3390/rs12010086, 2020.

Porter, D. F., Tinto, K. J., Boghosian, A. L., Csatho, B. M., Bell, R. E., and Cochran, J. R.: Identifying spatial variability in Greenland's outlet glacier response to ocean heat, Front. Earth Sci., 6, https://doi.org/10.3389/feart.2018.00090, 2018.

Pratt, W., K.: Digital Image Processing, Wiley, New York, NY, 1978.

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, https://doi.org/10.5194/tc-6-1483-2012, 2012.

Rignot, E.: Changes in West Antarctic ice stream dynamics observed with ALOS PALSAR data, Geophysical Research Letters, 35, 12, https://doi.org/10.1029/2008GL033365, 2008.

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

Rignot, E., Bamber, J. L., van den Broeke, M. R., Davis, C., Li, Y., van de Berg, W. J., and van Meijgaard, E.: Recent Antarctic ice mass loss from radar interferometry and regional climate modelling, Nature Geoscience, 1, 2, 106–110, https://doi.org/10.1038/ngeo102, 2008.

Rignot, E., Velicogna, I., Broeke, M. R. van den, Monaghan, A., and Lenaerts, J. T. M.: Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise, Geophysical Research Letters, 38, 5, https://doi.org/10.1029/2011GL046583, 2011.

Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H., and Scheuchl, B.: Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011, Geophysical Research Letters, 41, 10, 3502–3509, https://doi.org/10.1002/2014GL060140, 2014.

Rignot, E., Mouginot, J., Scheuchl, B., van den Broeke, M., van Wessem, M. J., and Morlighem, M.: Four decades of Antarctic Ice Sheet mass balance from 1979–2017, PNAS, 116, 1095–1103, https://doi.org/10.1073/pnas.1812883116, 2019.

Robel, A. A.: Thinning sea ice weakens buttressing force of iceberg mélange and promotes calving, Nature Communications, 8, 1, https://doi.org/10.1038/ncomms14596, 2017.

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, https://doi.org/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, Cham, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 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, 2019.

Schild, K. M. and Hamilton, G. S.: Seasonal variations of outlet glacier terminus position in Greenland, Journal of Glaciology, 59, 759–770, https://doi.org/10.3189/2013JoG12J238, 2013.

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, https://doi.org/10.1029/2010JF001847, 2011.

Serre, T.: Hierarchical Models of the Visual System, in: Encyclopedia of Computational Neuroscience, edited by: Jaeger, D. and Jung, R., Springer, New York, NY, 1–12, https://doi.org/10.1007/978-1-4614-7320-6_345-1, 2013.

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, https://doi.org/10.1016/j.neunet.2017.07.017, 2017.

Shepherd, A., Ivins, E., Rignot, E., Smith, B., van den Broeke, M., Velicogna, I., Whitehouse, P., Briggs, K., Joughin, I., Krinner, G., Nowicki, S., Payne, T., Scambos, T., Schlegel, N., A, G., Agosta, C., Ahlstrøm, A., Babonis, G., Barletta, V., Blazquez, A., Bonin, J., Csatho, B., Cullather, R., Felikson, D., Fettweis, X., Forsberg, R., Gallee, H., Gardner, A., Gilbert, L., Groh, A., Gunter, B., Hanna, E., Harig, C., Helm, V., Horvath, A., Horwath, M., Khan, S., Kjeldsen, K. K., Konrad, H., Langen, P., Lecavalier, B., Loomis, B., Luthcke, S., McMillan, M., Melini, D., Mernild, S., Mohajerani, Y., Moore, P., Mouginot, J., Moyano, G., Muir, A., Nagler, T., Nield, G., Nilsson, J., Noel, B., Otosaka, I., Pattle, M. E., Peltier, W. R., Pie, N., Rietbroek, R., Rott, H., Sandberg-Sørensen, L., Sasgen, I., Save, H., Scheuchl, B., Schrama, E., Schröder, L., Seo, K.-W., Simonsen, S., Slater, T., Spada, G., Sutterley, T., Talpe, M., Tarasov, L., van de Berg, W. J., van der Wal, W., van Wessem, M., Vishwakarma, B. D., Wiese, D., Wouters, B., and The IMBIE team: Mass balance of the Antarctic Ice Sheet from 1992 to 2017, Nature, 558, 7709, 219–222, https://doi.org/10.1038/s41586-018-0179-y, 2018.

Shepherd, A., Ivins, E., Rignot, E., Smith, B., van den Broeke, M., Velicogna, I., Whitehouse, P., Briggs, K., Joughin, I., Krinner, G., Nowicki, S., Payne, T., Scambos, T., Schlegel, N., A, G., Agosta, C., Ahlstrøm, A., Babonis, G., Barletta, V. R., Bjørk, A. A., Blazquez, A., Bonin, J., Colgan, W., Csatho, B., Cullather, R., Engdahl, M. E., Felikson, D., Fettweis, X., Forsberg, R., Hogg, A. E., Gallee, H., Gardner, A., Gilbert, L., Gourmelen, N., Groh, A., Gunter, B., Hanna, E., Harig, C., Helm, V., Horvath, A., Horwath, M., Khan, S., Kjeldsen, K. K., Konrad, H., Langen, P. L., Lecavalier, B., Loomis, B., Luthcke, S., McMillan, M., Melini, D., Mernild, S., Mohajerani, Y., Moore, P., Mottram, R., Mouginot, J., Moyano, G., Muir, A., Nagler, T., Nield, G., Nilsson, J., Noël, B., Otosaka, I., Pattle, M. E., Peltier, W. R., Pie, N., Rietbroek, R., Rott, H., Sandberg Sørensen, L., Sasgen, I., Save, H., Scheuchl, B., Schrama, E., Schröder, L., Seo, K.-W., Simonsen, S. B., Slater, T., Spada, G., Sutterley, T., Talpe, M., Tarasov, L., van de Berg, W. J., van der Wal, W., van Wessem, M., Vishwakarma, B. D., Wiese, D., Wilton, D., Wagner, T., Wouters, B., Wuite, J., and The IMBIE Team: Mass balance of the Greenland Ice Sheet from 1992 to 2018, Nature, 579, 7798, 233–239, https://doi.org/10.1038/s41586-019-1855-2, 2020.

Shugar, D. H., Burr, A., Haritashya, U. K., Kargel, J. S., Watson, C. S., Kennedy, M. C., Bevington, A. R., Betts, R. A., Harrison, S., and Strattman, K.: Rapid worldwide growth of glacial lakes since 1990, Nature Climate Change, 1–7, https://doi.org/10.1038/s41558-020-0855-4, 2020.

Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556, 2015.

Sobel, I. and Feldman, G.: An Isotropic 3x3 Image Gradient Operator, https://doi.org/10.13140/RG.2.1.1912.4965, 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, 3201–3216, https://doi.org/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, https://doi.org/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 Geoscience, 4, 322–327, https://doi.org/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, https://doi.org/10.1126/science.aax3528, 2019.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going Deeper with Convolutions, arXiv:1409.4842, 2014.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., and Brisco, B.: Google Earth Engine for geo-big data applications: A meta-analysis and systematic review, ISPRS Journal of Photogrammetry and Remote Sensing, 164, 152–170, https://doi.org/10.1016/j.isprsjprs.2020.04.001, 2020.

Todd, J. and Christoffersen, P.: Are seasonal calving dynamics forced by buttressing from ice mélange or undercutting by melting? Outcomes from full-Stokes simulations of Store Glacier, West Greenland, The Cryosphere, 8, 2353–2365, https://doi.org/10.5194/tc-8-2353-2014, 2014.

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, https://doi.org/10.1038/s41467-019-12039-2, 2019.

Velicogna, I., Sutterley, T. C., and Broeke, M. R. van den: Regional acceleration in ice mass loss from Greenland and Antarctica using GRACE time-variable gravity data, Geophysical Research Letters, 41, 22, 8130–8137, https://doi.org/10.1002/2014GL061052, 2014.

Vieli, A. and Nick, F. M.: Understanding and modelling rapid dynamic changes of tidewater outlet glaciers: issues and implications, Surv. Geophys., 32, 437–458, https://doi.org/10.1007/s10712-011-9132-4, 2011.

Wang, P. and Bai, X.: Thermal infrared pedestrian segmentation based on conditional GAN, IEEE Transactions on Image Processing, 28, 12, 6007–6021, https://doi.org/10.1109/TIP.2019.2924171, 2019.

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, https://doi.org/10.1029/2018GL078024, 2018.

Xie, H.-X., Lin, C.-Y., Zheng, H., and Lin, P.-Y.: An UNet-based head shoulder segmentation network, in: 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 1–2, https://doi.org/10.1109/ICCE-China.2018.8448587, 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, https://doi.org/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, https://doi.org/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, https://doi.org/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, 1729–1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X.: Object detection with deep learning: a review, IEEE Transactions on Neural Networks and Learning Systems, 30, 11, 3212–3232, https://doi.org/10.1109/TNNLS.2018.2876865, 2019.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V.: Learning Transferable Architectures for Scalable Image Recognition, arXiv:1707.07012, 2018.