CLOUD MASKING OVER GLACIAL SNOW COVER USING SENTINEL-2 CLOUD PRODUCTS AND SEMI-SUPERVISED IMAGE CLASSIFICATION IN THE INDIAN WESTERN HIMALAYAS

SOMDUTTA MISHRA
August, 2022

SUPERVISORS:
Dr. Ir. Janneke Ettema
Dr. Harald van der Werff

CLOUD MASKING OVER GLACIAL SNOW COVER USING SENTINEL-2 CLOUD PRODUCTS AND SEMI-SUPERVISED IMAGE CLASSIFICATION IN THE INDIAN WESTERN HIMALAYAS

SOMDUTTA MISHRA
Enschede, The Netherlands, August, 2022

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Applied Remote Sensing

SUPERVISORS:
Dr. Ir. Janneke Ettema
Dr. Harald van der Werff

THESIS ASSESSMENT BOARD:
Dr. C.A. Hecker (Chair)
Prof. Dr. Ir. C. Persello (External Examiner, Faculty ITC)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Cloud contamination in satellite images are a major obstacle for using these for glacial studies. A number of cloud masking algorithms using rule based or machine learning techniques are developed and still an active field of research. However, one common challenge in cloud making algorithms is the detection of clouds over bright surfaces like snow. The recent study on intercomparison exercise (CMIX) on cloud masking algorithms found that the popular cloud products of Sentinel-2, scene classification layer (SCL) and S2cloudless suffers from accurate detection of clouds over snow cover. These cloud masks are frequently used by the remote sensing community for cloud screening before image analysis.

This study evaluates the three cloud masking methods available for Sentinel-2 images: level-1C cloud masks, SCL and S2cloudless for their cloud masking capabilities over the snow cover in high mountain glaciers of the Indian Western Himalayas. The results show that these cloud masking methods fail to distinguish snow from clouds in winter images. Level-1C performed the worst compared to SCL and S2cloudless. These cloud products also show wide overlap in their spectral signals in the Green and SWIR wavelengths which might explain the poor cloud masking by these products.

The study also attempted to develop a cloud mask using the spectral properties of landcover features in the study area. The convex and non-convex clustering methods were used to find spectral classes belonging to cloud and landcover features in the Green and SWIR wavelength feature space. These cluster labels were then used to classify the images using nearest-neighbouring classifier. The resulting classified images were assessed for their accuracy over manually labelled cloud pixels. The non-convex cluster labels of Spectral clustering showed >85% cloud detection in both a summer and winter image.

ACKNOWLEDGEMENTS

The completion of this thesis marks the pre-official end of my time at ITC. I hope the official end – my thesis defence goes well!

I am fortunate enough to come to ITC and do my Masters in Geo-information Science under the guidance of Dr. Ir. Janneke Ettema and Dr. Harald van der Werff. I thank them for their patience and support during my academic journey. The lessons I have learned will surely come to fruition one day.

I am grateful to the entire ITC Faculty for maintaining such high academic standards and especially to the Faculty in AES department for imparting knowledge in the MSc electives. I am thankful to the library staff for maintaining a silent space in the building where I can do my research work. I was also lucky to be in ARS. I am going to cherish the bonds I made with my classmates.

Life at ITC was rocky in the beginning due to Covid-19 but was made easier by the people I met here. I am forever grateful to have such people around me. Shruti, who has lovingly stood by me since day one, I am extremely grateful for her. Prashant has been my family here, always feeding me and checking in on me. Geethanjali, for always being a kind friend and offering her help every time. Malihe for being a supporter in times of need, Lynette for tirelessly listening to me and Siddhant for being there regardless of my shortcomings.

A special thanks to Aakash, Manish and Swati for their moral support and always being a call away to make me feel lighter.

I would like to give a special shout out to Miss Mahnoor Ahmed for agreeing to do a project together and ended up finishing my thesis.

Finally, I would like to thank my family for being the ultimate support system throughout this and many experiences of my life. Their patience, love and unconditional support has never come short.

I also thank the almighty for blessing me with good health and opportunities in life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

1.1.1. Glaciers

Glacial snow and ice act as freshwater reserves on the Earth surface. The High Mountain Asia (HMA), also known as the third pole or water tower of Asia, is the origin of major river systems like Indus, Ganga and Brahmaputra (Immerzeel et al., 2010). Apart from rainfall, these rivers are also fed from glacial snow and ice meltwater (Lutz et al., 2014). Snowmelt contributes nearly 50% to the annual freshwater requirements for nearly 700 million people in the plains of HMA across South Asia (Sarangi et al., 2020). Glacial snow and ice melt modulate the seasonal river flow patterns and provide water for irrigation in the absence of rainfall (Biemans et al., 2019). A stable glacier provides sustainable and long term discharge to river flow (Cuffey and Paterson, 2010).

Global glacier changes are widely accepted as a climate change indicator (Scherler et al., 2011). Satellite remote sensing is crucial for monitoring and mapping glacier changes (Winsvold et al., 2016). The earth observation satellites now provide large volume of data for many applications that allows for consistent monitoring of the terrestrial systems. The terrestrial monitoring studies are mainly based on the optical remote sensing wavelengths (0.443 - 2.190 µm). It, however, suffers from the presence of cloud cover (Coluzzi et al., 2018). Satellite images have shown that clouds occupy nearly two-thirds of the global surface area (Boucher et al., 2013). Cloud presence is more frequently observed on mountains than in flat terrains (Winsvold et al., 2016). The occurrence of clouds in satellite images restrict the utilization of measured sensor parameters (López-Puigdollers et al., 2021). Clouds are thus, considered noise in the input satellite images and require to be removed for glacier studies (Mahajan and Fataniya, 2020).

1.1.2. Clouds

Clouds are a visible collection of suspended particles in the atmosphere made up of minute water droplets, ice crystals or a combination of both (Lutgens et al., 2010). They are formed when the water vapour in an air parcel attains saturation and condenses (Lutgens et al., 2010). The air parcel can saturate by various combination of pressure, temperature and humidity conditions. However, the most common process of cloud formation is the adiabatic expansion and cooling in a rising air parcel (Boucher et al., 2013; Lutgens et al., 2010). Clouds appear in many shapes and forms in the sky and can be broadly classified on the basis of form and height (Lutgens et al., 2010) as shown in Figure 1.1.

The clouds show three distinct forms of cirrus, cumulus and stratus (Lutgens et al., 2010). Cirrus clouds are white in colour and form thin, wispy patches in the sky. Cumulus clouds appear as mixture of lumpy or spherical clouds. Stratus clouds, as the name suggest, appear as strata (layers) in the sky. Based on height of formation above the Earth surface, clouds can be described as low (< 2,000m), middle (2,000-6,000 m) and high (> 6,000 m) level clouds (Lutgens et al., 2010). The altitudinal values used for the classification are not rigid and vary based on latitudes and seasons (Lutgens et al., 2010). Besides these clouds, there are also special clouds like cumulonimbus, mammatus etc. that cannot be put in these categories.

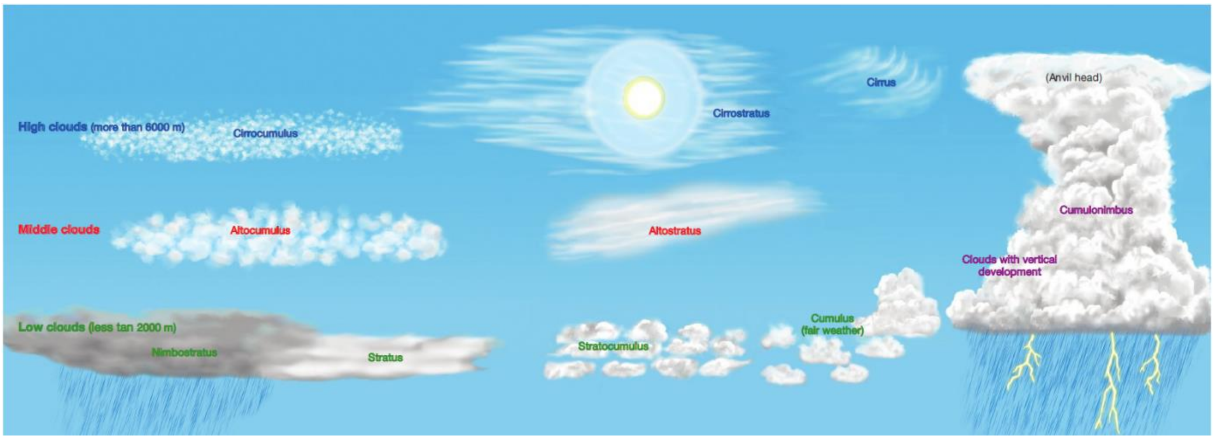


Figure 1.1 Cloud types classification based on form and height. Source: (Lutgens et al., 2010)

1.1.3. Spectral properties of clouds and snow

Clouds can occur in various shades of white and grey in the sky for human eyes (Figure 1.1). Cirrus clouds are always white while rain bearing cumulonimbus clouds are darker. Clouds usually appear white due to the high reflectance in the visible wavelength but also possess high reflectance values in other wavelength ranges of the electromagnetic spectrum (Figure 1.2). In the optical remote sensing, clouds are classified either as dense (also can be called opaque) clouds or cirrus (also can be called translucent) clouds (Coluzzi et al., 2018). These clouds differ from each other on the basis of constituent particles. Dense clouds are made of water droplets and cirrus clouds are made up of ice crystals (Lutgens et al., 2010). Dense clouds are also found at low-medium altitudes and have high reflectance in the visible wavelength range. Cirrus clouds are high level clouds and are translucent in the visible wavelength. However, both of them have relatively high reflectance in the shortwave infrared wavelength (hereafter, SWIR) wavelength (Warren, 1982). Snow, on the other hand, only shows high reflectance in the visible wavelength (Figure 1.2). Coarser grained snow has much lower reflectance than fine-grained snow (Figure 1.2) in the near-infrared and reaches very low values in the SWIR wavelength (Warren, 1982).

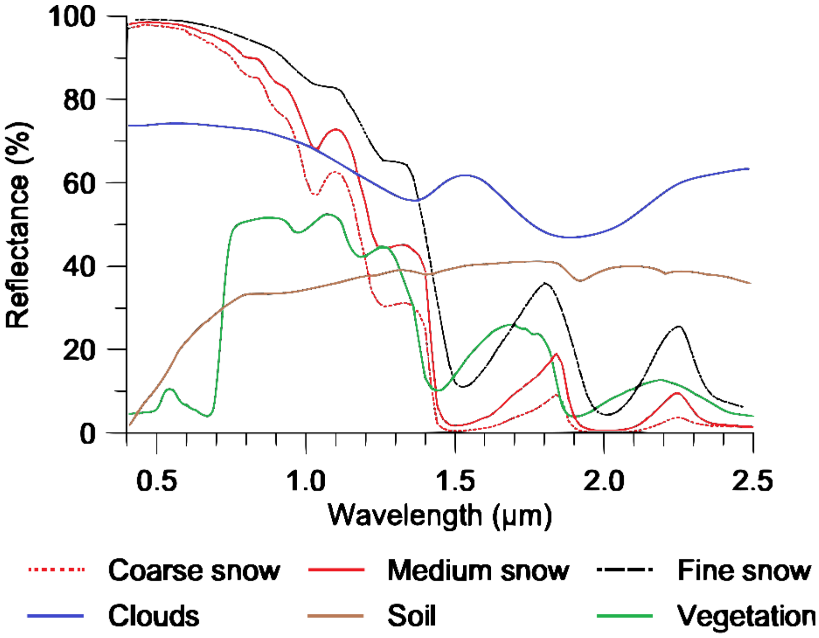


Figure 1.2 Spectral reflectance curve of snow, soil, clouds and vegetation in the visible, near-infrared and short-wave infrared wavelength. Source: (Dong, 2018).

The spectral behaviour of clouds in the SWIR wavelengths (1.4-3.0 µm) is governed by their composition – water droplets, ice crystals or a mixture of both. Clouds have higher reflectance in SWIR wavelengths compared to snow (Figure 1.2 & 1.3). Reflectance values are highest for cumulus clouds (subset of dense clouds) followed by cirrus clouds and snow (Figure 1.3). Cirrus clouds have smaller ice crystals than snow and hence higher albedo. Cumulus clouds have the highest reflectance among the three due to thicker cloud structure, relatively low absorption by water molecules and even smaller size of droplets than ice crystals of snow and cirrus clouds (Warren, 1982).

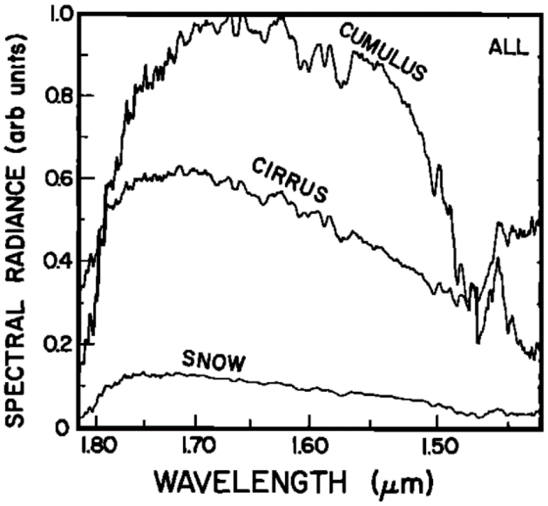


Figure 1.3 Spectral reflectance curve of snow and cloud (cirrus & cumulus) in the SWIR wavelength. Source: (Warren, 1982).

1.1.4. Satellite detection of clouds

The literature suggest multiple algorithms for detecting clouds in satellite images. The cloud detection algorithms can be divided into single-date algorithms or multi-temporal algorithms based on the number of images used (Zhu and Helmer, 2018). Single-date algorithms uses one image and can be further categorized as rule-based or machine learning based methods (López-Puigdollers et al., 2021; Qiu et al., 2019; Skakun et al., 2022). Rule-based methods are also known as threshold methods. These methods utilize the reflectance measurement in different spectral bands to identify the physical properties of clouds, such as 'white', 'bright', 'cold' and 'high' (López-Puigdollers et al., 2021; Qiu et al., 2019). The different spectral bands are integrated together to create spectral indices. Then, clouds are detected based on applying appropriate thresholds, either constant or dynamic, on these spectral indices (Qiu et al., 2019; Zhu and Helmer, 2018). For example, Luo et al. (2008) detected clouds in MODIS image using fixed reflectance threshold values. Some examples of rule-based adaptive threshold algorithms includes Function of mask (Fmask) and Sen2Cor (Zhu et al., 2015). On the other hand, the machine learning algorithms are simpler compared to the rule-based methods (Qiu et al., 2019). The algorithm requires manually annotated training data to build a statistical model for cloud detection (Hollstein et al., 2016; López-Puigdollers et al., 2021). The model classifies image pixels with possible cloud presence as 'cloud class' using a specific classifier. Examples of classifiers are 'decision trees', 'neural networks', 'support vector machines' etc. (Qiu et al., 2019).

Multi-temporal cloud detection algorithms focuses on finding clouds as anomaly to the otherwise gradual landcover reflectance change (Zhu and Woodcock, 2014). By comparing the satellite image to a cloud-free reference image, the cloud pixels can be identified. Each multi-temporal algorithm has its own way of detecting cloud anomalies between target and reference image. For example, Wang et al. (1999) used brightness changes, Lyapustin (2008) used low covariance values, Hagolle et al. (2010) used differences in blue band and Goodwin et al. (2013) used reflectance change from the smoothed time series values for detecting clouds.

1.1.5. Sentinel-2 cloud products

There are three ready to use cloud products with Sentinel-2 images (hereafter, standard cloud products). These are Level-1C cloud mask (hereafter, QA60 band), Scene Classification Layer (hereafter, SCL) cloud classes and S2cloudless map (Coluzzi et al., 2018; Main-Knorn et al., 2017; Zupanc, 2017). These standard cloud products are widely used for cloud screening in terrestrial monitoring studies. A detailed description of the data is given in Section 2.2.2.

(Coluzzi et al., 2018) performed the first assessment on the Level-1C cloud masks and discussed the systematic underestimation of cloud detection in conditions of high atmospheric water vapour content (specifically over rain-forest areas) and over-detects in low atmospheric water vapour content over mountainous terrain. Specifically, snow and other bright surfaces (sand, buildings) were identified as clouds.

The SCL map was evaluated over a test site in Antarctica with flat and mountainous topographic conditions, using images from both the satellite missions of Sentinel-2 (Sentinel 2-A and 2-B). The results obtained from this showed that the classes of water, snow and high cloud probability pixels were perfectly identified by Sen2Cor processor with both user's and producer's accuracy higher than 96% (Main-Knorn et al., 2017).

The validation performance of S2cloudless with Sen2Cor is shown in Table 1.1 retrieved from the work by (Zupanc (2017). It can be observed that S2cloudless classifier (Sentinel Hub) performs better than Sen2Cor in classifying clouds and snow. Moreover, the misclassification rates are lower for S2cloudless across all labels except for shadow areas.

Table 1.1 Cloud detection performance (in percent) of Sen2Cor and S2cloudless for 108 manually labelled reference datasets by (Hollstein et al., 2016). Source: (Zupanc, 2017)

		Fraction of classifications as clouds	
		Sen2Cor	**S2cloudless**
True Label	**Cloud**	97.5%	99.4%
	Cirrus	87.7%	83.8%
	Land	5.7%	2.2%
	Water	0.0%	0.1%
	Snow	30.7%	13.5%
	Shadow	3.9%	5.8%

1.1.6. Image classification for cloud detection

The image classification process mainly consist of two steps. The first step is the identification of spectral classes in the feature space followed by assigning spectral class labels to pixels using a classifier. Based on the priori information available for identifying spectral classes, the classification process can be categorized as supervised learning (Koutroumbas and Theodoridis, 2008). The prior knowledge refers to the availability of 'training data' which are also known as labelled data. The training data is the ground-truth generated through field-work or manual image processing (Richards and Jia, 2005).

In the absence of prior knowledge about the spectral classes, statistical methods is used to cluster similar but unlabelled pixel points in the feature space. This is known as un-supervised learning (Koutroumbas and Theodoridis, 2008). Semi-supervised classification is a mixed approach that progresses in a similar way to the supervised learning except that there are unknown classes along with training data (Koutroumbas and Theodoridis, 2008). In the absence or limited access to labelled data, semi-supervised pattern recognitions in images can be of great importance (Koutroumbas and Theodoridis, 2008). Labels are generated for the unlabelled data by clustering algorithms that respect certain constraints. The constraints are set to cluster similar pixels and keep dis-similar pixels separated. In this regard, the classification process imparts a priori information in the semi-supervised learning (Koutroumbas and Theodoridis, 2008).

Moving towards clustering methods, they can be broadly divided into convex and non-convex approaches. The convex methods like K-Means and Mean-Shift generate compact and spherical clusters due to the nature of distance algorithm used (Arthur and Vassilvitskii, 2007; Comaniciu and Meer, 2002). All data points within a distance 'd' in the feature space are grouped together, leading to spherical shape of clusters. On the other hand, the Spectral clustering method is a non-convex method (Yu and Shi, 2003). It reduces the dimensionality of the data by finding new eigen vectors. These eigen vectors are then clustered using K-means or Mean-Shift. This reduction of data dimension leads to non-spherical nature of clusters.

1.1.7. Problem statement

The standard cloud products are widely used for cloud screening in terrestrial monitoring studies (Coluzzi et al., 2018; Main-Knorn et al., 2017; Zupanc, 2017). In glacier studies, the removal of clouds is a necessary pre-processing step. However, the existing standard cloud products are not efficient in distinguishing clouds and bright surfaces like snow (Skakun et al., 2022). Snow, on the glacier surface, shows a range of reflectance values due to changes in grain size, presence of impurities, thickness of snow etc. It makes cloud and snow separation difficult using multi-spectral wavelength bands due to the varying reflectance of snow (Zhu and Helmer, 2018). Moreover, ice clouds also have very similar spectral signatures as snow (Zhu and Woodcock, 2014), which adds to the problem of distinguishability.

The existing cloud products, such as the output of the QA60 band, has been explained to suffer from a large number of undetected clouds (Coluzzi et al., 2018). Sen2Cor is considered to be a powerful algorithm that reduces affects by meteorological conditions and the sun angle but showed inability to detect cloud boundaries (Skakun et al., 2022). Moreover, S2cloudless provides a computationally fast classification where the processing is done per pixel of input image (Skakun et al., 2022). However, it produced errors when discriminating bright objects like snow.

Automatic accurate cloud detection is also difficult due to complexities in cloud types and inadequate spectral bands to decipher cloud physical properties (Zhu and Woodcock, 2014). The earlier Landsat missions struggled to detect clouds due to their limited bands (Zhu and Helmer, 2018).

Properties of clouds make it difficult to be detected for instance thicker clouds block solar radiation while thinner clouds retrieve a mixed spectral behaviour with the underlying landcover (Zhu and Woodcock, 2014). The single image cloud detection requires benchmarks, for both rule-based and machine learning based methods, in order to evaluate and refine the algorithms and boost its performance (Hollstein et al., 2016; López-Puigdollers et al., 2021). However, synchronous ground truth measurements of clouds, aligning with the satellite revisit time, can be difficult due to large areas covered instantly by satellite sensors and other complexities regarding sensor conditions and time constraints (Hollstein et al., 2016). Therefore, the manual visual-interpretation and labelling of cloud pixels is considered an accurate cloud detection method (Hollstein et al., 2016; López-Puigdollers et al., 2021; Qiu et al., 2019; Skakun et al., 2022). The manual cloud classification is however, time consuming and not feasible for large image collections. In such cases, multi-temporal cloud algorithms would be preferred. However, a serious drawback of such an algorithm is the assumption that the landcover remain stable in its reflectance values and does not change appreciably (Zhu and Woodcock, 2014), which may not be the case in reality.

The description of cloud detection obstacles mentioned above is the motivation behind defining the objectives of this study. The main aim of this study is twofold. First, the focus is on inspecting the cloud masking capability of standard products in different seasons over a glacierized catchment of Indian Western Himalayas, where clouds and snow (or ice) are expected to coexist. The capabilities will be explored in a spatial, temporal and spectral context. Secondly, the study aims to develop a cloud mask based on semi-supervised image classification techniques as a possible improvement to the currently available cloud masks.

1.2. Research objective

The main objective of this thesis is "To evaluate the cloud masking capability of Level-1C cloud mask, Scene Classification Layer, S2cloudless map and develop a cloud mask using semi-supervised image classification techniques for the glacial snow cover in the Indian Western Himalayas".

1.2.1. Sub-objective 1

To assess the cloud masking capability of Level-1C cloud mask, Scene Classification Layer and S2cloudless map in discriminating clouds over a glacierized catchment of Indian Western Himalayas.

Research questions

1.1. How does the visual interpretation differ for cloud masking by standard cloud products in a cloudy summer and winter image?

1.2. How does the visual interpretation differ for cloud masking by standard cloud products in a cloud-free summer and winter image?

1.3. What is the time-series behaviour of standard cloud products for glacial and non-glacial regions in the study area?

1.4. How does the standard cloud products discriminate spectral signals of permanent snow cover and bare rock areas in the reflectance values of Green and SWIR wavelengths?

The first research question evaluates the mono-temporal (fixed-date) images for cloud masking capabilities. The second and third research question evaluates the temporal and spectral characteristics of the standard cloud products.

1.2.2. Sub-objective 2

To develop a cloud mask over glacial snow cover by applying semi-supervised image classification techniques using Green and SWIR wavelength bands of Sentinel-2.

Research questions

2.1. What is the spectral behaviour of glacial and non-glacial areas in the Green and SWIR wavelengths?

2.2. How does clouds exhibit their spectral signals in the Green and SWIR wavelength reflectance of glacial and non-glacial areas?

2.3. How does the standardization of reflectance data affect the clustering process?

2.4. How does the optimum number of clusters vary for different clustering methods?

2.5. What are the cloud classes identified using different clustering methods in nearest-neighbour image classification?

2.6. Which clustering method performs better at cloud masking over glacial snow cover?

The first five research questions relate to the development of cloud masks using semi-supervised image classification while the last research question assesses the accuracy of cloud masks generated through different clustering methods.

1.3. Thesis structure

This MSc thesis consist of five chapters including this Introduction chapter. The fulfilment of thesis objectives required separate methods, results and discussion sections for cloud masking by existing standard cloud products and development of a new cloud mask using semi-supervised image processing. This lead to defining chapters for sub-objectives and hence, this format is used.

Chapter 1 is this introduction itself. It provides the context of the study, the state-of-the-art on cloud masking techniques before introducing the problem statement and research objectives.

Chapter 2 introduces the study area, data and software used in this study. The climatology, geographical location and surface characteristics are provided for the study area. The data section details the sensor characteristics of Sentinel-2 and its spatial and temporal resolution. The standard cloud products are also introduced. The software used for data processing is also mentioned.

Chapter 3 evaluates the cloud masking by the standard cloud products of Sentinel-2 on various parameters. The chapter looks at the problems and potential associated with the use of standard cloud products. It builds on the need for cloud free images for terrestrial studies. The work looks at the capability of these cloud products to detect clouds and especially discriminate clouds from bright snow surfaces.

Chapter 4 looks at manual cloud identification by semi-supervised image classification using Green and SWIR wavelength bands of Sentinel-2. It involves unsupervised statistical separation of clouds and landcover signals using different clustering methods. The clustering labels are then used for supervised nearest-neighbour image classification. The results are evaluated for best clustering labels for cloud class detection. This class is used to mask clouds from images.

Chapter 5 synthesizes the results and observations made in the Chapter 3 and 4 to provide the overall summary of the work done in this thesis along with the limitation s posed and recommendations made for future work.

.

2. STUDY AREA, DATASET AND SOFTWARE

2.1. Study area

The Indian Western Himalayas (IWHs) are home to one of the largest river systems of the world - the Indus and the Ganges rivers (Frey et al., 2012). This region contains Ladakh, Zanskar and Pir Panjal ranges of the Himalayan mountain system. The southern slopes of Pir Panjal ranges are heavily forested due to the orographic precipitation by moisture laden monsoonal winds (Frey et al., 2012). In contrast, the northern regions of these mountain ranges are arid and non-vegetated. The Chandra basin lies north of the Pir Panjal ranges. The basin is heavily glacierized and the glacial discharge is the source of Chandra river, which is a tributary of the Indus river (Azam et al., 2014). Chhota Shigri Glacier (hereafter, CSG) lies within the Chandra basin on the northern slopes of Pir Panjal ranges (Figure 2.1(a)).

The glacier lies on the transition between monsoon-arid climatic zone (Azam et al., 2014). The precipitation regime is characterized by two dominant and independent atmospheric circulations. The glacier receives precipitation by the Indian Summer Monsoon (hereafter, ISM) during summer months (July-September) and by Mid Latitude Westerlies (hereafter, MLW) during winter months (January-April) (Azam et al., 2014; Bookhagen and Burbank, 2010). Precipitation measurements in the year 2012-13 showed that the contribution by the MLW and ISM was 80% and 20%, respectively on CSG, making it a winter accumulation type glacier (Azam et al., 2014).

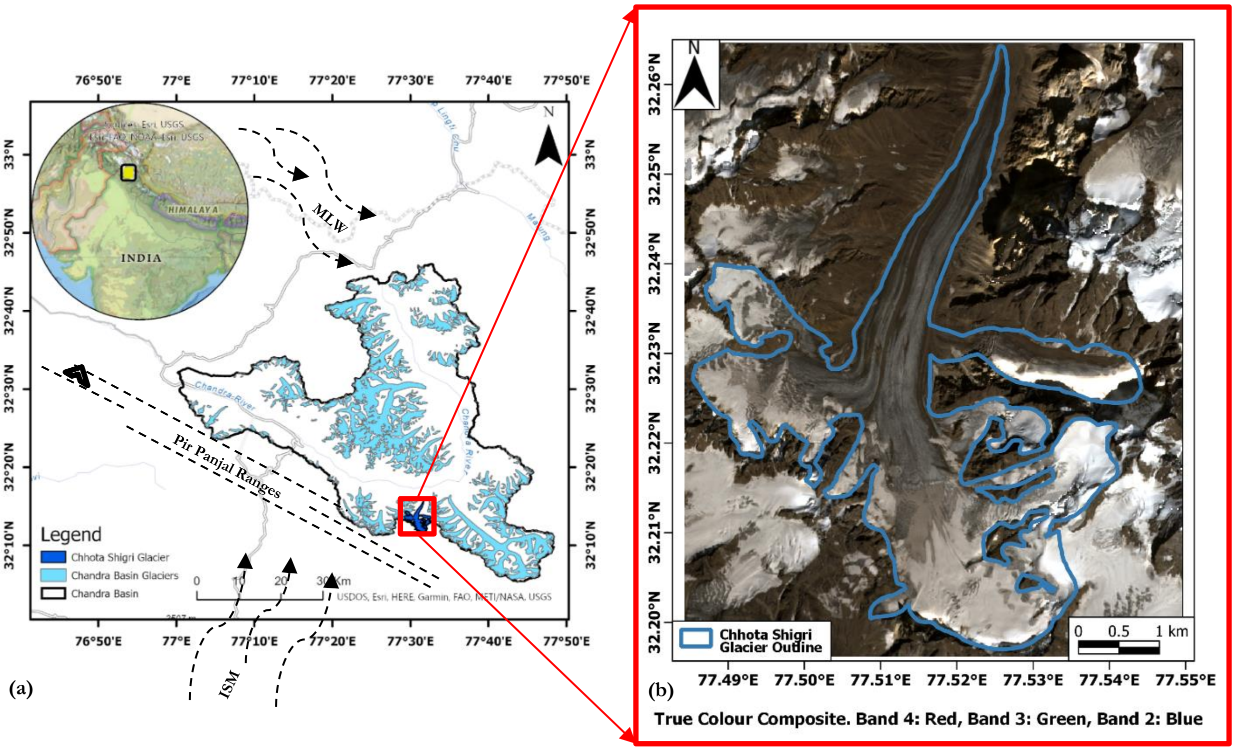


Figure 2.1 CSG in the Chandra basin of Lahaul and Spiti valley in the Indian state of Himachal Pradesh, Western Himalayas. Figure (a) shows the glaciers (light blue) of Chandra basin (black outline) in the Indian subcontinent (in inset), including the CSG (dark blue). The Pir Panjal ranges and the general direction of dominant precipitation systems ISM and MLW are also shown. Figure (b) shows the enhanced glacierized catchment of CSG (blue outline). The true colour image is dated 21-09-2022. The glacier outline is taken from the Randolph Glacier Inventory. Source: RGI Consortium (2017) https://www.glims.org/RGI/).

The CSG (32.280 N, 77.580 E) is a valley glacier mainly oriented along the north-south direction (Figure 2.1(b)). The mass accumulation zone at the higher elevation containing permanent snow cover lies to the south with mass ablation/melting zone in the north. The glacier occupies a surface area of 15.7 km2 and the main glacier body in the north-south direction is approximately 9 km long (Wagnon et al., 2007). The glacier is fed by many tributaries of varying orientation (Figure 2.1(b)). The northward flow of tributaries creates medial moraines (catchment debris) after joining the main glacier body. These medial moraines can be seen as long stretches of catchment rocks within the glacier body (Figure 2.1(b)). The lower reaches of the glacier (in the north) are also partially covered by catchment debris accounting for 3.4% of the total surface area (Vincent et al., 2013).

2.2. Dataset

2.2.1. Sentinel-2 surface reflectance product

Sentinel-2 is a constellation of two polar-orbiting sun-synchronous satellite mission launched by the European Space Agency (ESA) (Drusch et al., 2012). The first satellite Sentinel-2A was launched on 23rd June, 2015 followed by the second satellite Sentinel-2B on 7th March, 2017 (Main-Knorn et al., 2017). The twin-satellites host a Multi-Spectral Instrument (MSI). The MSI measures the reflected radiance from the Earth surface in 13 spectral bands. It has 4 bands in visible, 6 in near-infrared and 3 in short-wave infrared wavelengths (SWIR) (Table 2.1). The reflectance values, theoretically between 0-1, are scaled by 10,000 in all the 13 spectral bands of Sentinel-2 images.

Table 2.1 Spatial and spectral resolution of Sentinel-2 satellite missions. Source: (https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/msi-instrument, last access 01-08-2022).

Band Number	Band Name	S2A		S2B		Spatial resolution (m)
		Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	
1	Coastal aerosols	442.7	20	442.3	20	60
2	Blue	492.7	65	492.3	65	10
3	Green	559.8	35	558.9	35	10
4	Red	664.6	30	664.9	31	10
5	Vegetation Red Edge	704.1	14	703.8	15	20
6	Vegetation Red Edge	740.5	14	739.1	13	20
7	Vegetation Red Edge	782.8	19	779.7	19	20
8	Near-infrared (NIR)	832.8	105	832.9	104	10
8a	Vegetation Red Edge	864.7	21	864	21	20
9	Water vapour	945.1	19	943.2	20	60
10	SWIR -Cirrus	1373.5	29	1376.9	29	60
11	SWIR	1613.7	90	1610.4	94	20
12	SWIR	2202.4	174	2185.7	184	20

Sentinel-2 provides open access data in moderate spatial (10,20 and 60 m) and spectral resolution (Table 2.1). The satellite revisit time is 10 days at the equator and with two satellites, every 5 days an image is available. Sentinel-2 also provides two levels of data. Level-1C is top-of-atmosphere data whereas Level-

2A is bottom-of-atmosphere (or surface reflectance) data. Level-2A is atmospheric corrected product of Level-1C using Sen2Cor algorithm (Main-Knorn et al., 2017). The Level-2A images comes with additional maps of Aerosol Optical Thickness (AOT), Water Vapour (WP) and Scene Classification Layer (SCL). It also provide probability maps for snow and cloud classes at 60 m spatial resolution. Sentinel-2A bottom-of-atmosphere data is used in this study. The broad properties of the image collection is shown in Table 2.2. The at sensor radiance values are converted to reflectance values. One of the image in the image collection partially cover the study area. This image is unusable for any analysis and hence dropped from the image collection. The image is dated 25-11-2021 (Image ID: 20211125T053141_20211125T053958_T43SGR).

Table 2.2 Overview of Sentinel-2 image collection used in this study

Total number of images used in the study	257
First image date	16-12-2018
Last image date	23-06-2022
Data type	Reflectance
Band value range	0-10,000 (scaled by 10,000)

2.2.2. Sentinel-2 cloud products (standard cloud products)

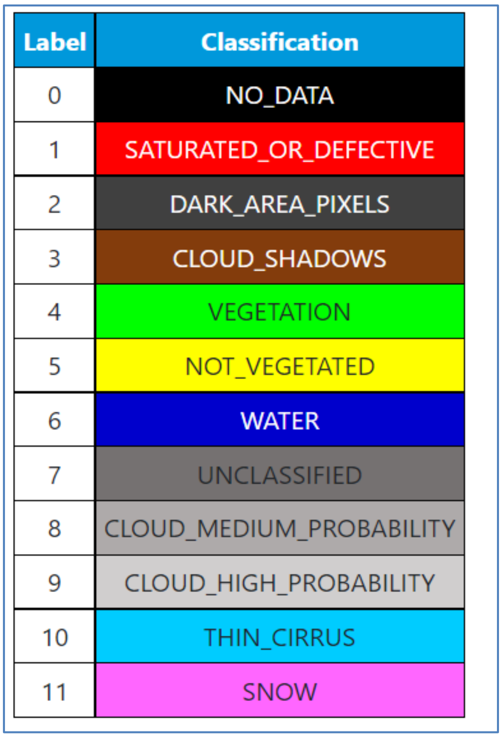
The three open-source cloud products for Sentinel-2 are discussed in detail in the following sections:

2.2.2.1. The Level-1C cloud mask (QA60 band)

The Level-1C cloud mask is a vector layer in Geography Markup Language (GML) (Coluzzi et al., 2018). The vector band is associated with each raster image as a band named 'QA60'. The cloud masks gives dense and cirrus cloud classes calculated using Sentinel-2 Level-1C (top-of-the-atmosphere) reflectance products (https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-1c/cloud-masks, last access 01/08/2022). The bands used for cloud masking are resampled at 60m spatial resolution. For dense clouds, a threshold on the blue wavelength band B2 (0.490 µm) of Sentinel-2 is applied. The thresholds are variable in nature due to scene-dependence of reflectance values. Snow and clouds are separated using SWIR wavelength bands B11 (1.610 µm) and B12 (2.190 µm). For cirrus clouds, additional Sentinel-2 Band 10 (1.375 µm) is used. It is the water vapour band of Sentinel-2 and only high altitude icy clouds reflect solar radiation while the rest of radiation is absorbed by the atmosphere (Drusch et al., 2012). In the QA60 band, the value 0, 1024 and 2048 represent cloud-free, dense cloud and cirrus cloud respectively.

2.2.2.2. The SCL cloud map

The SCL algorithm generates an 11 class classification layer that include three cloud classes (Main-Knorn et al., 2017). SCL classes are not landcover classification in the conventional sense, since the classes are derived after applying thresholds to values of different wavelength bands, band ratios and normalized indices (Main-Knorn et al., 2017). This SCL map is a product of Sen2Cor processor which corrects atmospheric effects on the Level-1C (top-of-the-atmosphere) images to produce Level-2A (bottom-of-the-atmosphere) images (Main-Knorn et al., 2017). The processor first detects cloud and generates SCL map followed by Aerosol Optical Thickness (AOT) and water vapour retrieval. The resulting SCL classes and their labels are shown in Figure 2.2. In the final SCL map, there are three cloud classes '8', '9' and '10' of medium, high and cirrus clouds, respectively. Cloud shadows also have their own class '3'. The Sen2Cor processor is robust and takes less than 5 mins for processing a full image tile (Skakun et al., 2022). It also provides cloud mask at a moderate spatial resolution of 20m, sufficient for cloud screening in terrestrial monitoring applications (Main-Knorn et al., 2017).



Label	Classification
0	NO_DATA
1	SATURATED_OR_DEFECTIVE
2	DARK_AREA_PIXELS
3	CLOUD_SHADOWS
4	VEGETATION
5	NOT_VEGETATED
6	WATER
7	UNCLASSIFIED
8	CLOUD_MEDIUM_PROBABILITY
9	CLOUD_HIGH_PROBABILITY
10	THIN_CIRRUS
11	SNOW

Figure 2.2 SCL map values for each label. Source (https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-1c/cloud-masks, last access 01/08/2022).

2.2.2.3. The S2cloudless map

The S2cloudless is a single scene algorithm for pixel based cloud detection in Sentinel-2 images using machine learning techniques (Zupanc, 2017). The research team at Sinergise (https://www.sinergise.com/, last access 01/08/2022) is credited with the development of S2cloudless algorithm. S2cloudless algorithm gives a per pixel (10, 20 or 60m) cloud cover probability (in %) based on the spectral response of the pixel (Zupanc, 2017). The ten wavelength bands of Level-1C products are used in the algorithm. These are B1, B2, B4, B5, B8, B8A, B9, B10, B11 and B12. The application of the algorithm is independent of the image resolution (Skakun et al., 2022). A cloud mask can be made by converting the cloud probability map to a binary map by thresholding. The default value is set at 0.4 (40%) to minimize omission errors (Skakun et al., 2022). The algorithm also allows additional processes for e.g. morphological operations to improve cloud detection capabilities.

2.3. Software

The software are used for data downloading, processing and creating visual outputs. Owing to the large size of image collection, a system with high computational power is required. The cloud computing facility of Google Earth Engine (hereafter, GEE) is used to solve this issue. On the other hand, the clustering methods were applied on a small dataset and hence were performed on the local system. The geographic maps were created using QGIS, an open source geographic information system (QGIS.org, 2022).

The web version of GEE is limited in the availability of image processing algorithms. However, it provides customization through different Application Programming Interface (hereafter, API). Introduction to GEE, along-with the Python API ('geemap') used to access the platform is given in Section 2.3.1. Its application is shown in Chapter 3 and 4 (Section 3.2 and 4.2, respectively). The clustering algorithms were accessed from an open-source Python module 'Scikit-learn'. The clustering algorithms are used in Chapter 4 (Section 4.2).

2.3.1. Google Earth Engine

GEE is a cloud computing platform by Google which is capable of "planetary-scale geospatial analysis" (Gorelick et al., 2017). GEE provides a wide collection of geo-spatial and satellite data in its multi-petabyte repository. It can be accessed on a web-based interactive environment or through Application Programming Interface (API) services. The 'geemap' is an interactive Python package that utilizes the API services of GEE (Wu, 2020). The 'geemap' python package allows access to GEE using pre-defined codes and algorithms which can be modified based on requirements.

The interactive mapping environment of 'geemap' relies on the 'Jupyter' notebook (Wu, 2020). The 'Jupyter' notebook is an interactive web based computational notebook for programming languages (https://jupyter.org/, last access 01-08-2022), in the Python distribution 'Anaconda'. The 'Anaconda' distribution manages the packages of Python and R programming languages for scientific processes (https://www.anaconda.com/products/distribution, last access 01-08-2022). It contains 250 installed packages and over 7,500 additional open-source packages.

2.3.2. Scikit-learn

Scikit-learn provides a broad range of popular supervised and un-supervised machine learning algorithms (Pedregosa et al., 2011). It is an open-source, straight-forward interface built on Python programming language. It is designed to provide easy-to-use statistical analysis of data by non-computer science background people (Pedregosa et al., 2011). It is used to perform unsupervised classification methods of K-Means, Mean-Shift and Spectral clustering for image analysis in this study.

3. CLOUD MASKING OVER SNOW COVER USING SENTINEL-2 CLOUD PRODUCTS

3.1. Background

This chapter aims to evaluate the cloud masking capabilities of standard cloud products over glacial snow cover in a glacierized catchment of the Indian Western Himalayas. Sentinel-2 is equipped to address the clouds contamination issue in the optical satellite imagery for land surface studies. The Sentinel-2 missions are devoid of spectral bands in the thermal infrared wavelength unlike Landsat and ASTER missions, its 13 multi-spectral bands are useful enough for cloud screening in optical images (Coluzzi et al., 2018).

The Level-1C cloud masks suffer from a large number of undetected clouds especially for cirrus clouds. The cloud masks under-detects in conditions of high atmospheric water vapour content over rain-forests and over-detects in low atmospheric water vapour content over mountainous terrain (Coluzzi et al., 2018). Sen2Cor under-detects clouds on cloud edges and water bodies (Skakun et al., 2022). It also faces difficulty in distinguishing clouds and bright surfaces like buildings or snow areas. S2cloudless is also prone to misclassification of bright surfaces as clouds (Skakun et al., 2022). It cannot provide cloud shadow masks and does not take spatial information around pixels into account for cloud detection.

The evaluation of these cloud detection algorithms is also limited due to scarce availability of 'ground truths' dataset (Main-Knorn et al., 2017; Qiu et al., 2019). The cloud masks are validated against reference cloud and cloud-free scenes. These reference images are made by either expert interpretation of satellite images or generated through models (Qiu et al., 2019). In case of a large image collection like Sentinel-2, the validation procedure becomes even more difficult due to sparse geographical distribution of global reference dataset (Skakun et al., 2022).

The standard cloud products for Sentinel-2 are widely used for cloud screening in terrestrial monitoring studies. However, the limitations posed by standard cloud products makes them unsuitable for surface change studies over bright surfaces in mountainous terrain. Therefore, this study aims to evaluate the capability of standard cloud products in successful discrimination of clouds over glacial snow and ice in the Indian Western Himalaya.

3.2. Methods

This chapter focuses on the first sub-objective of this study "to assess the cloud masking capability of Level-1C cloud mask, Scene Classification Layer and S2cloudless map in discriminating clouds over a glacierized catchment of Indian Western Himalayas." There are three main steps followed in this methodology:

1. Cloud masking in mono-temporal images – to address Research Questions 1.1 and 1.2
2. Time-series behaviour of standard cloud products – to address Research Question 1.3
3. Spectral behaviour of standard cloud products – to address Research Questions 1.4

3.2.1. Cloud masking in mono-temporal images

1. Selection of cloud-covered and cloud-free images from the image collection.

- Required number of images – images are selected for months of September and December. September is the end-of-summer month when the snow cover is expected to be at minimum and all landcover features are usually exposed (Azam et al., 2014). The glacier starts receiving precipitation by December and becomes completely snow covered (Azam et al., 2019). Another reason for the selected months is to account for the variability in cloud types, associated with two independent and dominant summer and winter precipitation mechanisms (ISM and MLW, respectively) over the study area (see, Section 2.1). Spring and autumn months will likely have landcover transitioning between complete snow cover and exposed glacier hence, they are not preferred to be used for this study.
- Band combinations for cloud detection – the Sentinel-2 band combination used to identify clouds in the image collection are 11,8,3 and 4,3,2 corresponding to the SWIR, Near-infrared, Green and Red, Green, Blue wavelengths, respectively. The band combinations are chosen as they theoretically distinguish clouds and snow (Figure 1.2). Similar band combinations were used for manual cloud detection in Sentinel-2 images by (Hollstein et al., 2016). (Hollstein et al., 2016) used the Sentinel-2 band combination of 2,8,10 (among other band combinations) to detect and digitize the cloud cover in an example scene. These bands correspond to the Blue, Near-infrared and Cirrus wavelength bands of Sentinel-2 Level1-C data. However, the Sentinel-2 band 10 (Cirrus band) is a water vapour band and is not available for Level-2A products used in this study.
- Selection of cloud-covered and cloud-free images for summer and winter months – the cloud cover and cloud-free images are selected by looking at the image composites for the month of September and December in the image collections. Partial cloud covered images are selected to ensures that the performance of cloud masks in demarcating cloudy and non-cloudy areas can be better compared.

The image identification of partly cloud covered Sentinel-2 images selected are '20211001T052649_20211001T052746_T43SGR' (dated '01-10-2021') for the end-of-summer month and '20201205T053209_20201205T053212_T43SGR' (date '05-12-2020') for the winter month. The Sentinel-2 image identification of cloud-free summer and winter month images are '20200921T052651_20200921T053331_T43SGR' (dated '21-09-2020') and '20210109T053211_20210109T053213_T43SGR' (dated '09-01-2021') respectively.

2. Visual interpretation of images with the standard cloud products.
   - Digitization of cloud cover – The spectral band combinations mentioned above highlights the visual differences between clouds and landcover classes. The manual labelling of pixels was done using false colour composites of 11,8,3 and 4,3,2 bands and taking into account the 'spatial context' of clouds in the image scene. The spatial context is deemed crucial for accurate digitization of image features (Hollstein et al., 2016). It includes the location of image features and their relationship with neighbouring pixel features. For example, the cloud shadows lie adjacent to the cloud in the image and depends on the sun angle. It may also completely cover the landcover or in the case of cirrus clouds make the landcover signals attenuated (Zhu and Helmer, 2018). Cloud shadows may also come from neighbouring image scenes (Hollstein et al., 2016). The manual interpretation of clouds unfortunately imparts some subjectivity to the study. The results of digitization is shown in Appendix A.
   - Visual interpretation of cloud-covered and cloud-free images of summer and winter – the manually digitized cloud covered areas in false colour composite (11,8,3 band combination)

are compared with the cloud detected areas in standard cloud products. The cloud-free images are also used to inspect for any cloud detection in standard cloud products.

3.2.2. Time-series behaviour of standard cloud products

1. Creating glacial and non-glacial polygons – the time series of standard cloud product are studied for some representative areas within the study area. These representative areas are selected on the glacier (for snow and ice features) and around the glacier on seasonal bare rocks. The details are given in Appendix B. 4 glacial and 4 non-glacial polygons are made.

2. Extracting QA60 band, SCL band and S2cloudless map values for the glacial and non-glacial polygons from the image collection – this is done in Google Earth Engine.

3. Visual interpretation of time-series behaviour of standard cloud products taking the glacial and non-glacial polygons as focused areas. The trends observed by standard cloud products are then plotted and key observations are made.

3.2.3. Spectral behaviour of standard cloud products

1. Selecting glacial and non-glacial areas for examining the spectral behaviour in standard cloud products. In this case, glacial_1 is representing an area that is permanent snow cover, thus the corresponding attribute can either be snow or cloud for this polygon. Similarly, non_glacial_1 can be bare rock in summer, snow in winter and clouds whenever cloud are present over the polygon.

2. Extracting reflectance values from Green and SWIR wavelength bands for the reference polygons – Google Earth Engine is used to extract the reflectance values of Green and SWIR wavelengths for glacial_1 and non_glacial_1 polygons.

3. Time-series and scatterplot of reflectance values of Green and SWIR wavelength, colour coded with standard cloud product band values of cloud classes. The reflectance values are graphically presented and visually interpreted for the relationship between cloud occurrence with the reflectance values.

3.3. Results

Sub-objective 1: To assess the cloud masking capability of Level-1C cloud mask, Scene Classification Layer and S2cloudless map in discriminating clouds over a glacierized catchment of Indian Western Himalayas.

3.3.1. Standard cloud products in cloudy and non-cloudy images during summer and winter

Figure 3.1. shows the visual comparison between the three standard cloud products and a cloudy summer image in false colour composite. The false colour composite highlights the differences between cloud and landcover. The manually digitised cloud polygons (hereafter, true cloud cover) are also overlayed in all the images (Figure 3.1 a,b,c,d). As per visual inspection, it can be seen that S2cloudless and SCL band cover a similar cloud extent as the true cloud cover. However, QA60 band identifies cloud to some extent but the smaller cloud areas were not detected.

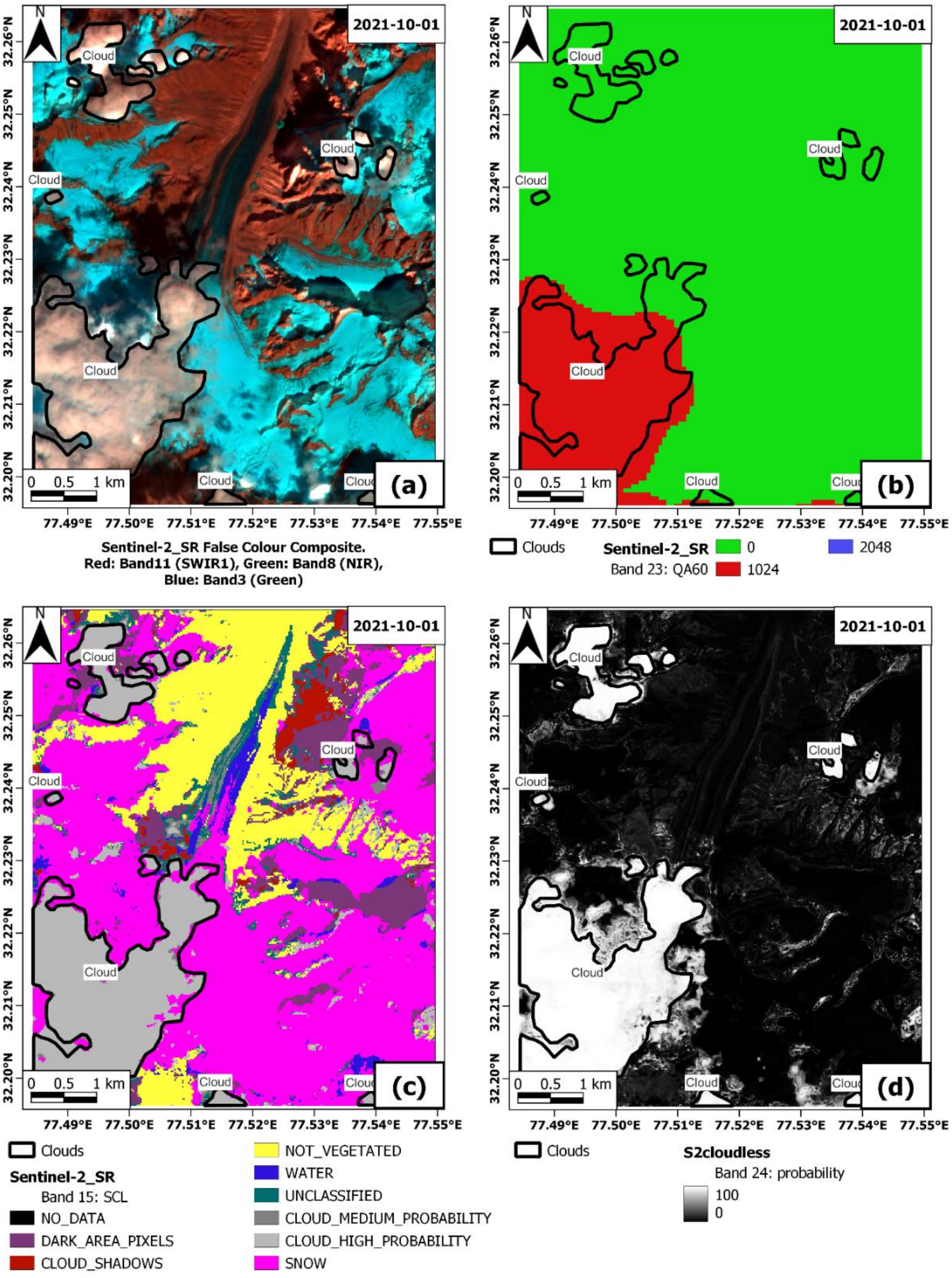


Figure 3.1 Cloud detection by standard cloud products in a summer cloudy scene. True cloud cover is shown by a polygon. Figure (a) shows the false-colour-composite. Figure (b), (c) and (d) shows the image in QA60 band, SCL and S2cloudless, respectively.

Similarly, Figure 3.2 exhibits the visuals for comparing standard cloud products and the true cloud cover for a winter cloudy image. Here, QA60 band (Figure 3.2 (b)) seems to fail in identifying cloud areas by classifying the entire image scene as cloud covered. SCL (Figure 3.2 (c)) does classify the bigger cloud polygon as high cloud probability but simultaneously assigns majority of the image with medium cloud probability. S2cloudless (Figure 3.2(d)) displays high probability value for area covered by clouds as well as other areas in the image scene. Though, SCL and S2cloudless are identifying the polygon in Figure 3.2 (a) as cloud, they also misclassify a large area of the scene as possible cloud cover.

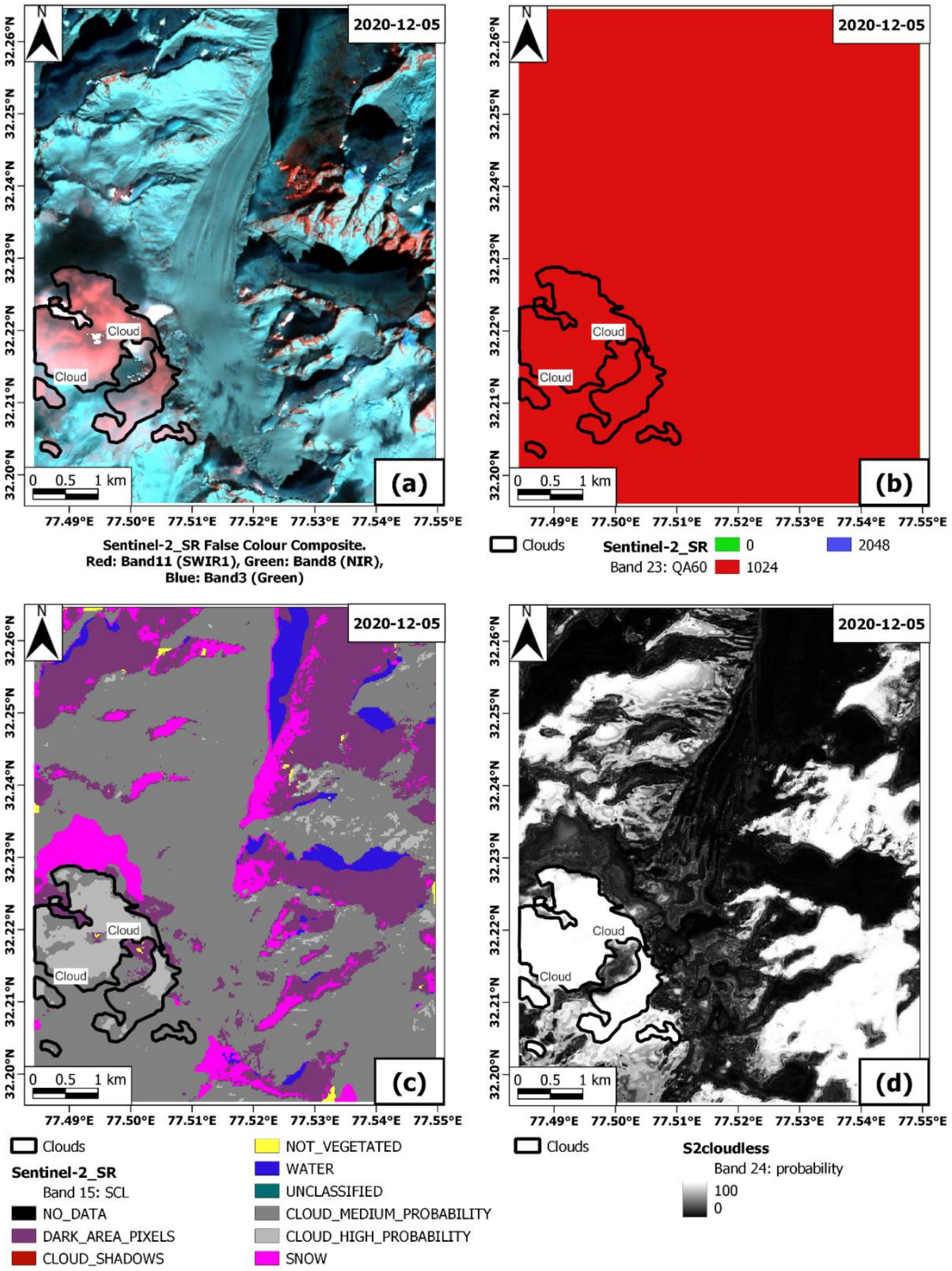


Figure 3.2 Cloud detection by standard cloud products in a winter cloudy scene. True cloud cover is shown by a polygon. Figure (a) shows the false-colour-composite. Figure (b), (c) and (d) shows the image in QA60 band, SCL and S2cloudless, respectively.

Figure 3.3 shows standard cloud products compared with the image in false colour composite for a summer cloud-free image. In Figure 3.3 (b), QA60 band labels significant area as cloud in the scene exhibiting a randomised pattern. SCL (Figure 3.3 (c)) classifies areas with high reflectance in SWIR as clouds. Figure 3.2 (d) referring to S2cloudless also assigns high cloud probability along thinly-defined patterns in some parts of the scene.

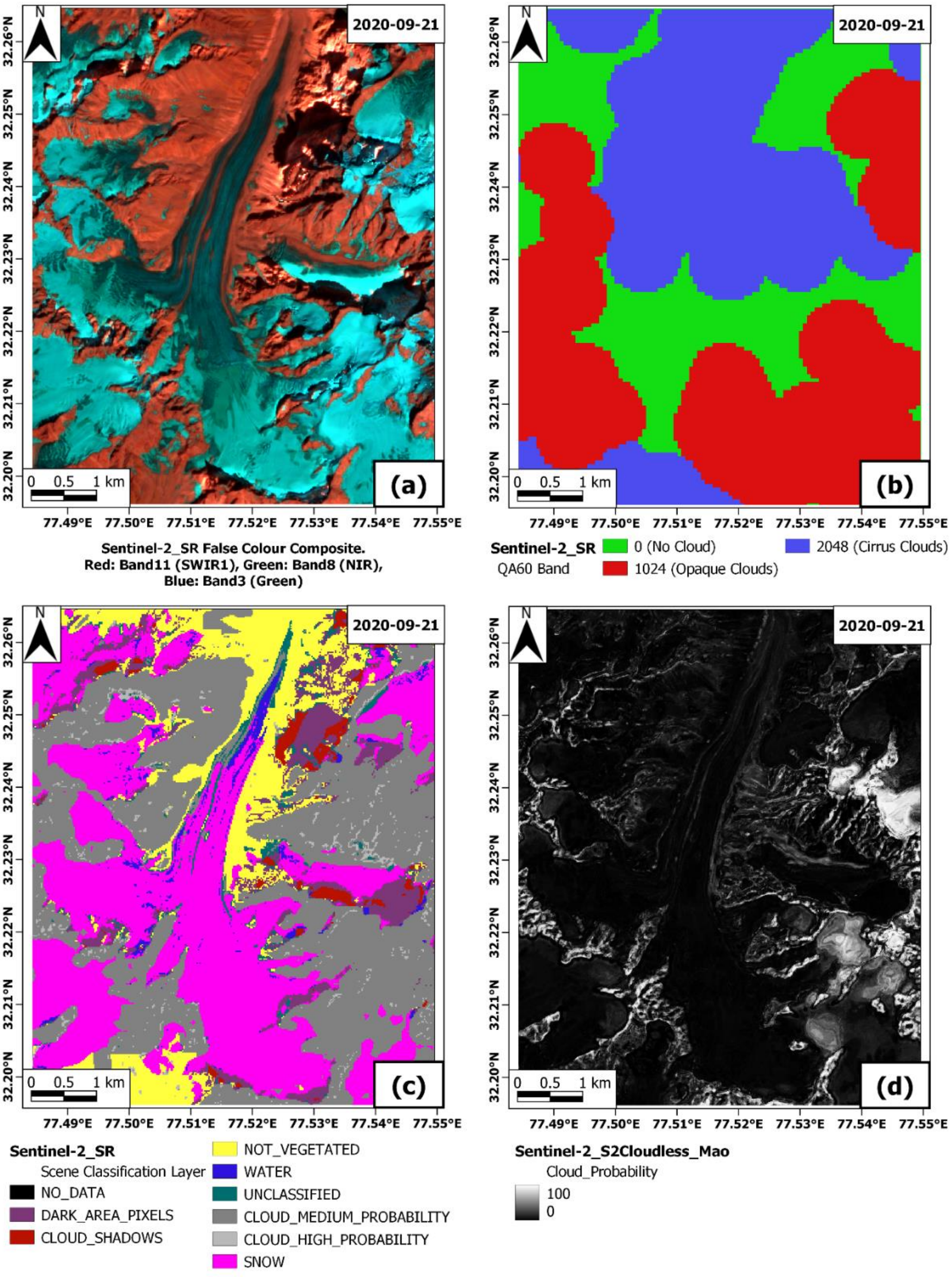


Figure 3.3 Cloud detection by standard cloud products in a summer cloud-free scene. True cloud cover is shown by a polygon. Figure (a) shows the false-colour-composite. Figure (b), (c) and (d) shows the image in QA60 band, SCL and S2cloudless, respectively.

Figure 3.4 displays standard cloud products comparing to the winter cloud-free image in false colour composite. The visual comparison suggests an overall poor performance by standard cloud products. QA60 band classifies the entire scene as cloud in a cloud-free image. SCL along with S2cloudless also classify large portions of the scene as high probability for cloud presence.

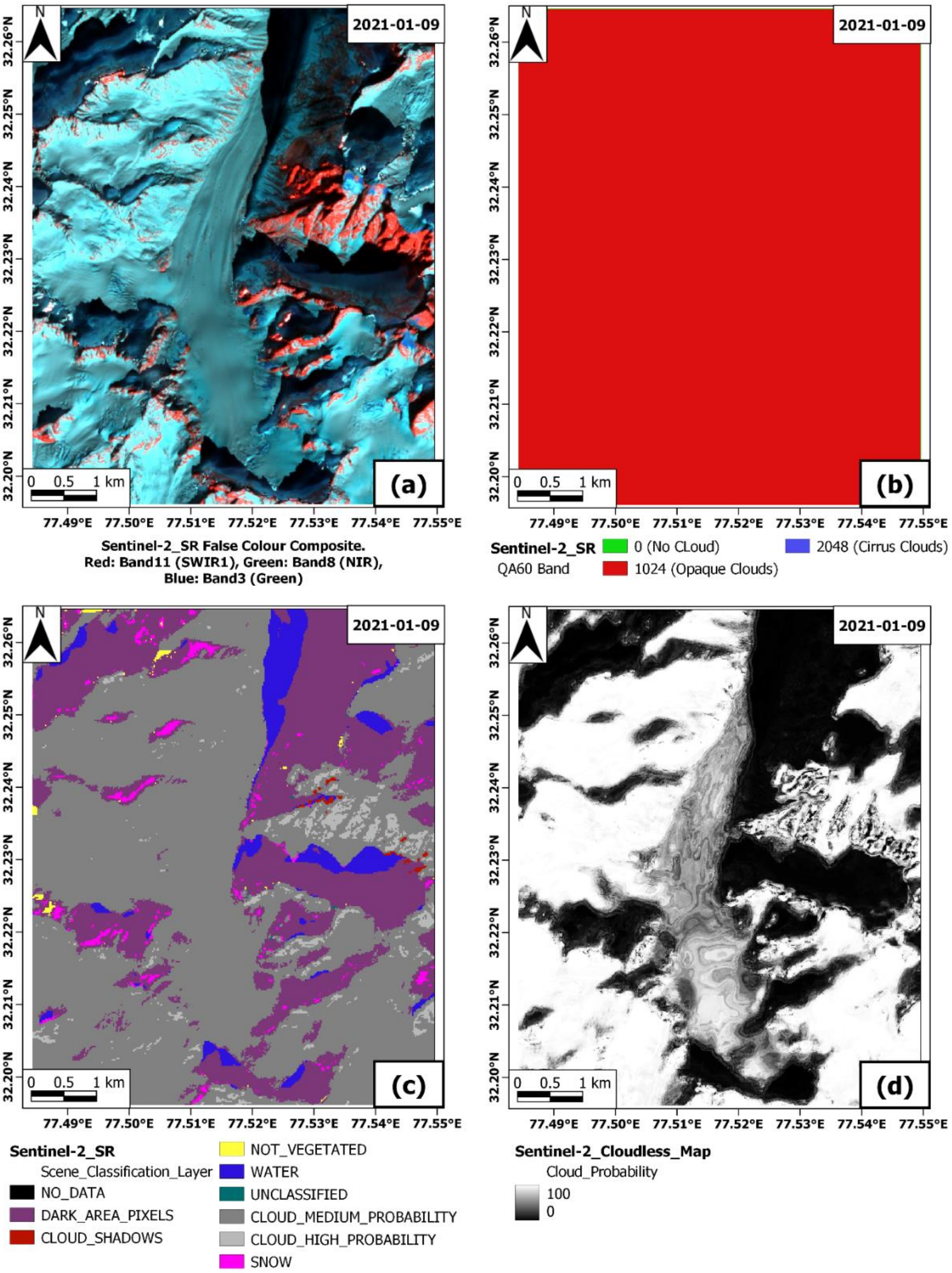


Figure 3.4 Cloud detection by standard cloud products in a winter cloud-free scene. True cloud cover is shown by a polygon. Figure (a) shows the false-colour-composite. Figure (b), (c) and (d) shows the image in QA60 band, SCL and S2cloudless, respectively.

3.3.2. Time-series behaviour of standard cloud products

Each polygon, glacial and non-glacial, covers 100 pixels of 10m spatial resolution each. Therefore, one image contains 8 polygons. Each of the 8 polygons will retrieve a mean value from an image band (the image band refers to the standard cloud product). Thus, for one image, there are 8 mean image band values. For the entire image collection, with 253 images, there are 8x253 mean image band values. These values, along with their standard deviation, are plotted as time-series for the three standard cloud products (Figure 3.5, 3.6 and 3.7).

Since the QA60 band can only have no-clouds (0 band value), opaque clouds (1024 band value) and cirrus clouds (2048 band value), the presence of intermediate band values indicate the presence of non-homogenous polygons. Polygons containing both clouds and no-clouds pixels in different proportions lead to this non-homogenous nature.

The time-series, of QA60 band, shows peculiar behaviour of no-clouds (0 band value), opaque clouds (1024 band value) and cirrus clouds (2024 band value). The majority of representative polygon values are classified as opaque clouds followed by no-clouds and cirrus clouds (Figure 3.5). Almost all the no-clouds values lie in the summer months of June to October. Interestingly, after April 2021, the proportions of no-cloud values increase relative to opaque cloud values. However, no polygon values are classified as either opaque or cirrus clouds after February 2022.

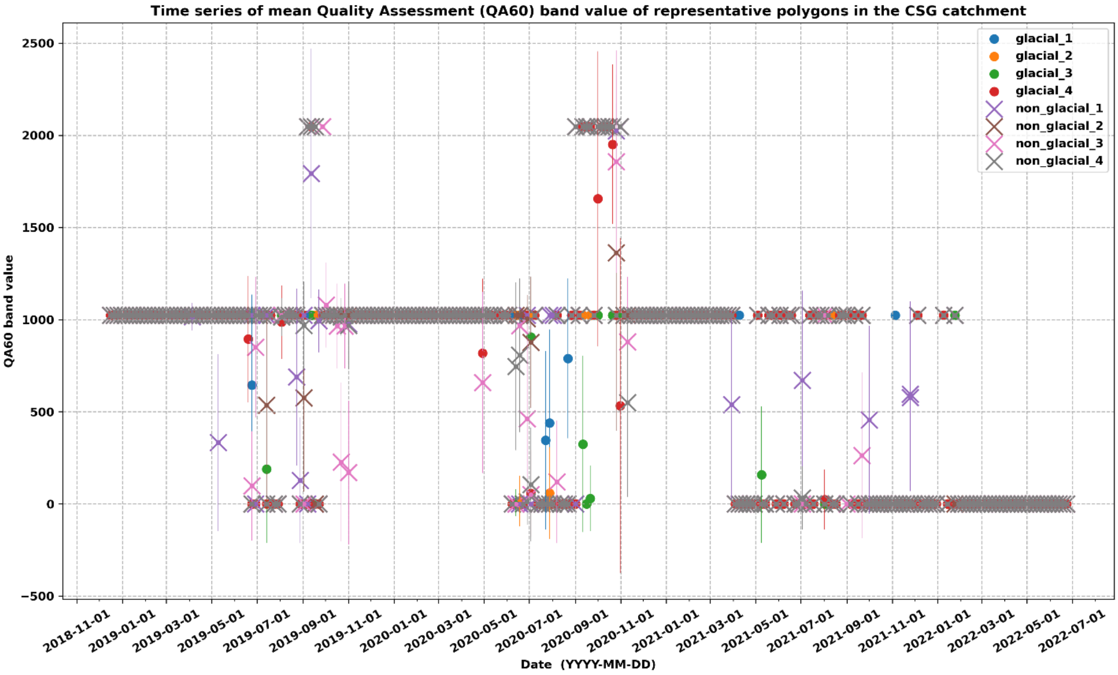


Figure 3.5 Time-series of QA60 band value for all the representative polygons in the study area.

The time series of mean SCL band classes for all representative polygons is shown in Figure 3.6. The figure shows that most of the representative polygons are classified as the class ‘CLOUD_MEDIUM_PROBABILITY’. The majority of ‘SNOW’ class is present in the summer months and dominated by glacial polygons. The summer months also show the presence of the class ‘CLOUD_HIGH_PROBABILITY’ and only one polygon has a mean band value belonging to the class of ‘CLOUD_SHADOWS’. Notably, some of the polygon have mean band values which are interpreted as the class ‘WATER’ during the summer months.

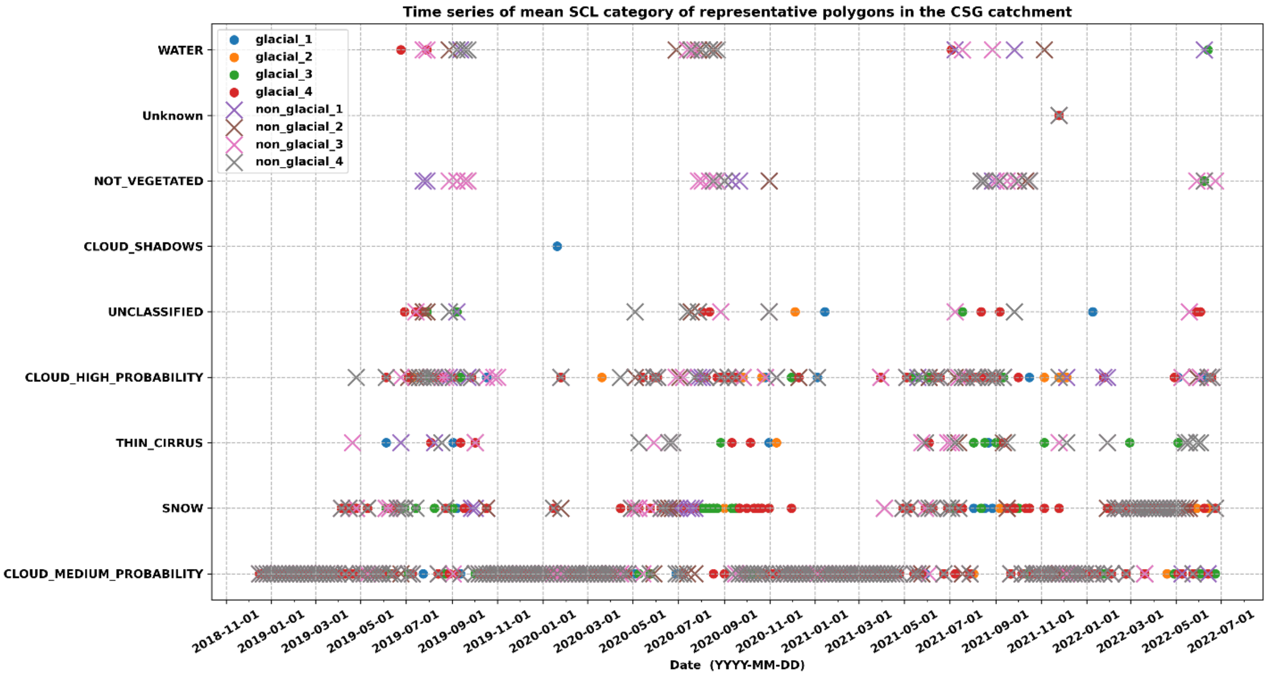


Figure 3.6 Time-series of SCL for all the representative polygons in the study area.

The time series of mean S2cloudless band value for all representative polygons along with their standard deviation is shown in Figure 3.7. The series show a recognisable pattern between winter and summer months. The winter months have above 90% cloud probability and the values are almost mutually exclusive with summer months. All intermediate values between 0-100% have high standard deviation, representing non-homogenous polygon values.

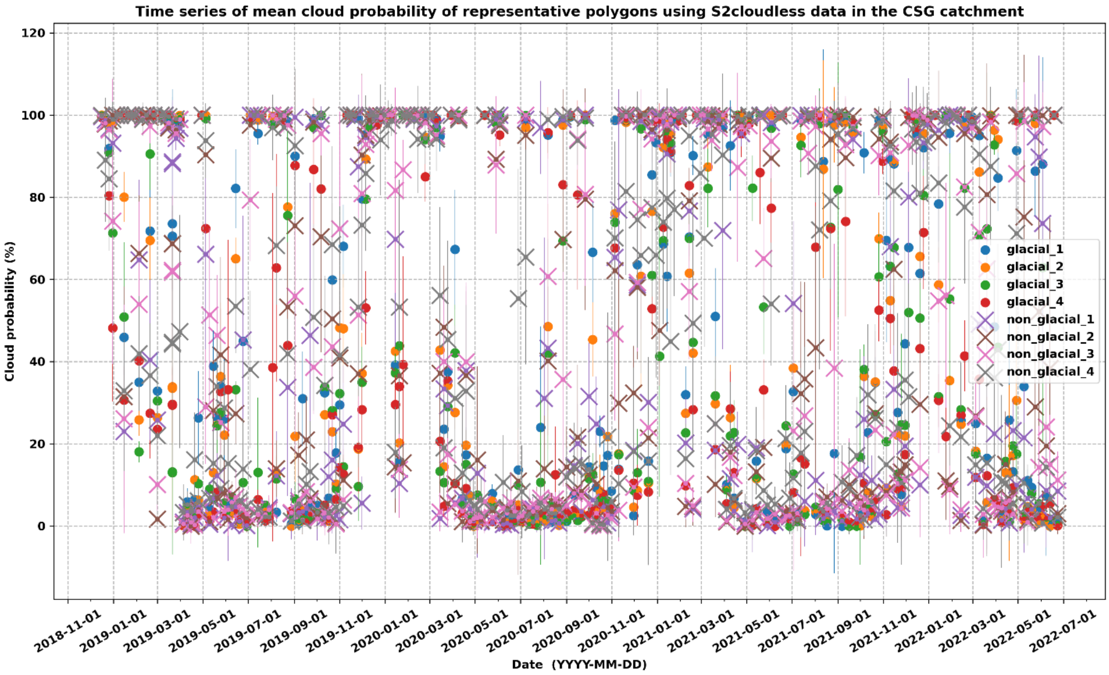


Figure 3.7 Time-series of S2cloudless for all the representative polygons in the study area.

The true-colour-composites of a cloud-free image for every month of 2019 is shown in Figure 3.8. This implies that there is at-least one image in each month where the standard cloud products should not show any cloud presence for the year 2019. Figure 3.8 also shows the colour change of bright snow cover to displaying a brownish tinge during summer months (Figure 3.8 (f, g, h and i)).

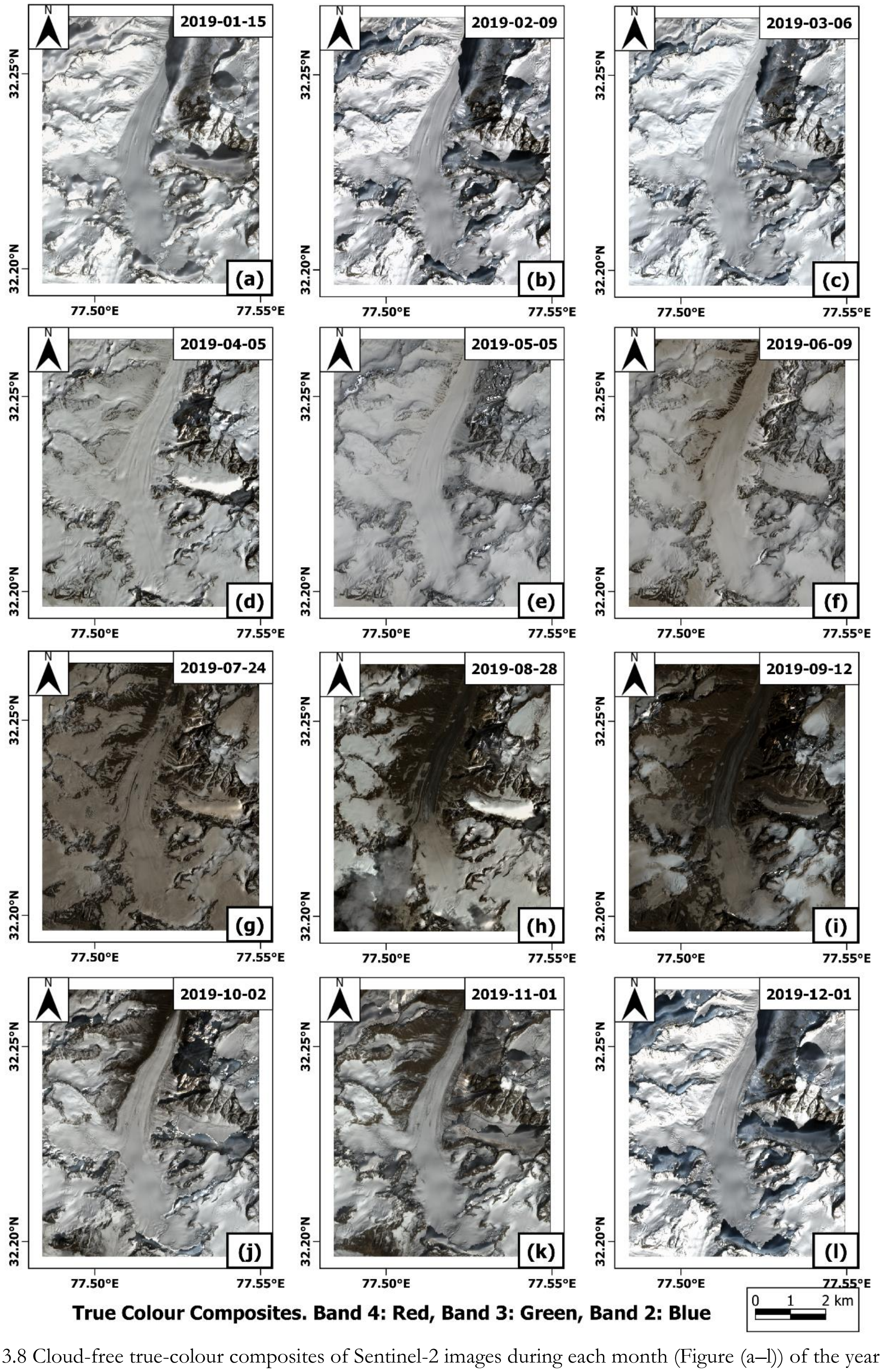


Figure 3.8 Cloud-free true-colour composites of Sentinel-2 images during each month (Figure (a–l)) of the year 2019.

3.3.3. Spectral behaviour of standard cloud products

Two polygons were selected from the 8 representative polygons namely, glacial_1 and non_glacial_1. The glacial_1 polygon was drawn over an area of permanent snow cover. On the other hand, non_glacial_1 was drawn over seasonal bare rock. The reflectance values in the Green and SWIR wavelengths were extracted from the image collection for these two polygons and graphically represented as a scatterplot in Figure 3.9. The reflectance values are colour coded according to the classification by standard cloud products.

The glacial_1 polygon is expected to have either snow or cloud spectral signals in the scatterplot (Figure 3.9(a, c, e)), whereas the non-glacial_1 polygon is expected to transition between rocks and snow with possible occurrence of clouds Figure 3.9(b, d, f). Snow has low reflectance in SWIR wavelength where clouds have high reflectance (Figure 1.2). Therefore, this distinction should be reflected in the standard cloud products. For instance, the QA60 band should distinguish cloud values (1024 and 2048) from no-cloud value (0) in the SWIR wavelength. However, the cloud label 1024 (black colour in Figure 3.9 (a)) covers the entire range of reflectance values in the SWIR wavelength. The no-cloud label 0 (blue colour in Figure 3.9 (a)) also has a widespread in the reflectance values of SWIR wavelength. This behaviour is also observed in the SCL and S2cloudless cloud classes (Figure 3.9 (c, e)).

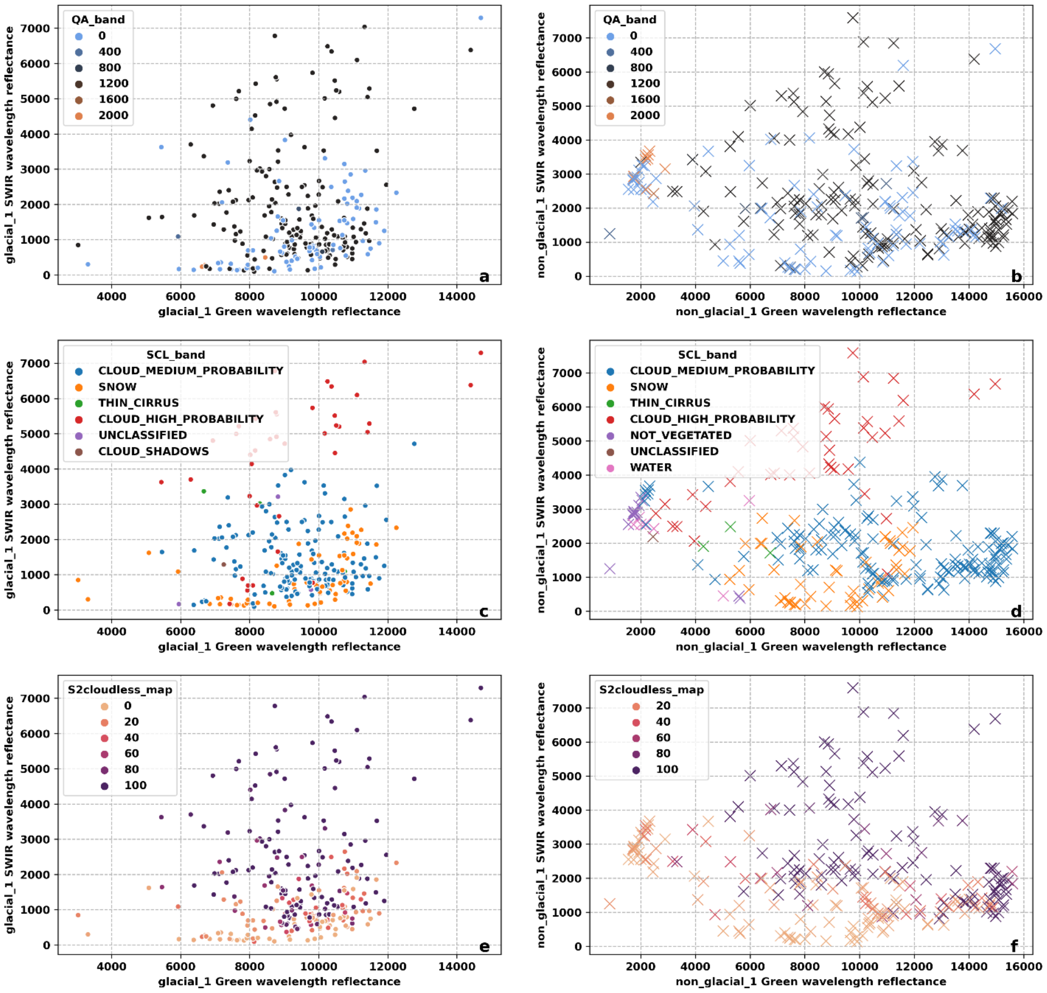


Figure 3.9 Scatterplot between reflectance values of Green and SWIR wavelengths for glacial_1 (a, c, e) and non_glacial_1 (b, d, f) polygon. The reflectance values are colour coded by the standard cloud products. QA60 band (a, b); SCL (c, d); S2cloudless (e, f).

3.4. Discussion

Research Question: 1.1: How does the visual interpretation differ for cloud masking by standard cloud products in a cloudy summer and winter image?

The QA60 band visually interprets major cloud cover correctly for the summer image. However, it is not able to delineate a clear boundary for the cloud shape as compared to the SCL and S2cloudless which match very well with the true cloud cover (manually digitised clouds) as seen in Figure 3.1.

The SCL and S2cloudless are able to detect true cloud covered areas, both in summer and winter images (Figure 3.1, 3.2). However, they overestimate cloud cover for majority of the study area in the winter image (Figure 3.2). The false colour composite shows the true cloud areas as reddish (of high SWIR reflectance) and the non-cloudy areas in hues of green and blue (of high Green and NIR but low SWIR), indicative of snow cover (Dong, 2018). Therefore, the SCL and S2cloudless estimate both clouds and snow as cloud cover in the winter image. This implies that the SCL and S2cloudless fail to spectrally separate cloud and snow signals in winter. Since the study area receives maximum precipitation as snowfall in winter (Azam et al., 2014), the freshly fallen snow in winter may have higher reflectance than summer. This is also visible in Figure 3.8 where summer snow appears brownish (may be contaminated) and winter snow is bright white. This can be explained by the errors of nearly all popular cloud masking algorithms including SCL and S2cloudless over bright objects like snow (Skakun et al., 2022).

Research Question: 1.2: How does the visual interpretation differ for cloud masking by standard cloud products in a cloud-free summer and winter image?

The cloud-free winter and summer images are classified with significant cloud presence across all three standard cloud products. The SCL shows cloud presence in areas of high reflectance in SWIR wavelength (Figure 3.3(a, c)) in summer cloud-free image. The similarity between bare rocks and clouds in the SWIR wavelength (high reflectance) (Dong, 2018) might explain the inability of SCL in correctly discriminating the two.

The QA60 band classifies the entire winter scene as cloud, independent of the actual presence of cloud. The S2cloudless identifies majority of the snow covered area as cloud for the winter image as discussed in previous research question 1.1.

Research Question 1.3. What is the time-series behaviour of standard cloud products for glacial and non-glacial regions in the study area?

The observations made in mono-temporal cloudy and cloud-free images during winter (research question 1.1., 1.2) showed that the standard cloud products detect snow as cloud. These observations are in-line with the time series behaviour of glacial and non-glacial polygons in winter. Therefore, during winter the standard cloud products identify snow as cloud with varying proportions. The visual inspection revealed that every month has at-least one cloud-free image for the year 2019 (Figure 3.8). The presence of cloud-free images solidifies the support for the empirical evidence that snow and bare rock areas are clearly misclassified as clouds.

Research Question 1.4. How does the standard cloud products discriminate spectral signals of permanent snow cover and bare rock areas in the reflectance values of Green and SWIR wavelengths?

The standard cloud products are based on the multi-spectral band values of Sentinel-2. Therefore, they should be able to show clear demarcation in the scatterplot of spectral band values. However, such results are not seen in this study. The standard cloud products fail to separate cloudy and non-cloudy values along the Green and SWIR wavelengths. The primary division is loosely along the SWIR wavelength. However, the boundaries are fuzzy and show a large overlap of cloudy and non-cloudy values. The distinction is poorest in the QA60 band, intermediate in SCL and relatively better among the three in S2cloudless. Interestingly, the bare rock areas having low Green and high SWIR wavelength reflectance cluster are only classified as such in S2cloudless map. SCL and QA60 band show a range of classes and values for these clusters. One primary reasons for such poor cloud detection might be the elevated reflectance values of the study area. In general, the Green wavelength reflectance ranges till 16,000 (unscaled 1.6) for non_glacial_1 and till 12,000 (unscaled 1.2) for glacial_1. Such high reflectance values (above 10,000 unscaled 1.0) shows the presence of directed radiance towards the satellite sensor (Dozier and Painter, 2004). The spectral thresholds used in the standard cloud product algorithms might not be tuned for such high reflectance surfaces.

Scientific studies on Sentinel-2 images for terrestrial monitoring services heavily use cloud masks provided in the satellite products. However, the current cloud masks available for Sentinel-2 fail to separate cloud, snow and bare rock areas in the high mountain regions. Among these cloud masks, S2cloudless performs relatively better in detecting and demarcating cloud areas. However none of the cloud masks area suitable for winter months. For landcover studies, it is imperative that alternate methods of cloud masking should be performed. Otherwise, the omission errors by these cloud masks are too large and a large data would be lost for scientific analysis (Kokhanovsky et al., 2019; Main-Knorn et al., 2017; Zupanc, 2017).

4. SEMI-SUPERVISED IMAGE CLASSIFICATION FOR CLOUD MASKING OVER SNOW COVER

4.1. Background

The standard cloud products, although derived from the spectral bands of Sentinel-2, have poor abilities to distinguish clouds and landcover features in the study area. The spectral behaviour of standard cloud products show no clear-cut distinction between clouds, snow and bare rocks in the reflectance values of Green and SWIR wavelengths. However, the manual cloud identification was possible for both summer and winter images using spectral band combinations to create false colour composites. The Green and SWIR bands showed the most differences in landcover features in the visual interpretation of false colour composites in Chapter 3. In this chapter an attempt is made to classify the Sentinel-2 images using spectral classes identified in the feature space made by the Green and SWIR wavelength. These wavelengths show the maximum distinction for snow and clouds (Figure 1.2) and bare rocks can also be identified using these bands (as seen in Chapter 3).

The identification and grouping of digital image pixels into valid categories in satellite imagery is called classification (Arthur and Vassilvitskii, 2007). The classification process assigns labels to each pixel based on the multi-spectral behaviour of the image. The multi-spectral information of a pixel can be effectively represented in a graph or a pattern space, where the axes (dimensions) are the multi-spectral bands (Richards and Jia, 2005). Thus, the multi-spectral values of a pixel can be represented by a vector. The graph or pattern space is also known as multi-spectral vector space or 'feature space' (Arthur and Vassilvitskii, 2007). An example of a feature space is shown in Figure 4.1. The multi-spectral signature of all the pixels of an image class/object should ideally show similar spectral behaviour. The groupings of such pixels in the feature space is known as 'information class'. However, there might also exist separate classes within the same information class due to intra-variabilities. For example, a bare rock area under snow will exhibit different spectral behaviour. Such classes are referred to as a spectral class and collectively they make up the information class (Figure 4.1).

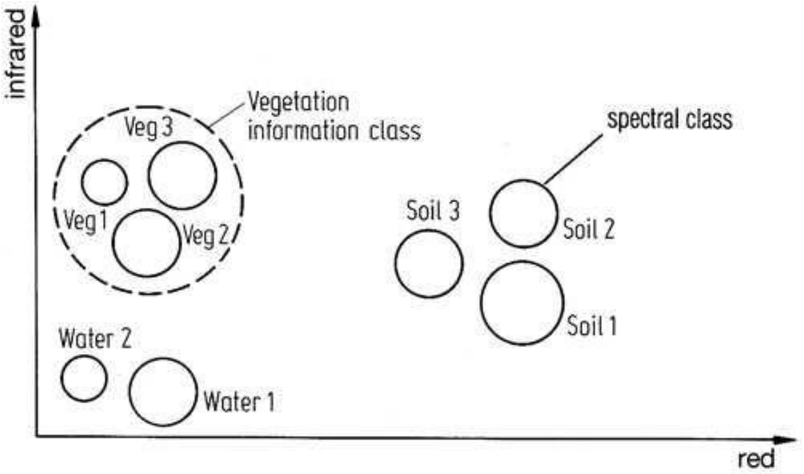


Figure 4.1 Example of a feature space with two spectral dimensions (red and infrared wavelength) showing spectral and information classes. Source: (Richards and Jia, 2005).

The spectral classes occur as cluster of pixels in the feature space. The clusters are defined as "continuous regions of this space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points" (Koutroumbas and Theodoridis, 2008). In un-supervised methods, the clusters in a feature space are quantified by either similar or dissimilar measures. For example, the Euclidean distance is a dissimilarity measure and useful for compact clusters. The type of clustering method used depends on the distribution of the pixels in the feature space (Koutroumbas and Theodoridis, 2008) Figure 4.3 shows the different types of clusters. The most popular clustering methods generally use distance between data points in the feature space. However, the dis-similarity criteria 'distance' can also vary, for example, Euclidean or Manhattan (Richards and Jia, 2005), and even the number of clusters can be assigned beforehand based on different statistical methods (Arthur and Vassilvitskii, 2007). Hence, the final clustering results depend very much on the initial conditions of clustering and feature space properties.

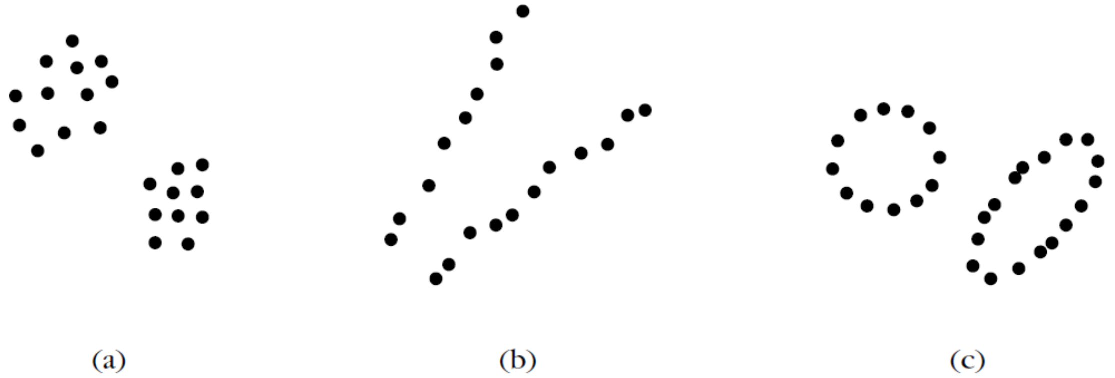


Figure 4.2 Types of clusters (a) compact, (b) elongate and, (c) spherical or ellipsoidal clusters. Source: (Koutroumbas and Theodoridis, 2008).

In this study, the spectral signals of the study area are analysed to find the spectral class of clouds in the Green and SWIR feature space. The separation of values in the feature space is done using convex (spherical) and non-convex (non-spherical) clustering (unsupervised) methods. These clusters are then labelled and used as training data for the nearest-neighbour classifier used for image classification. The classified images are then analysed to develop a cloud mask using cluster labels and their accuracy is assessed to find the best cloud mask in the study area.

4.2. Methods

This chapter fulfils the second sub-objective of the thesis to develop a cloud mask over glacial snow cover by applying semi-supervised image classification techniques using Green and SWIR wavelength bands of Sentinel-2. The broad outline of the steps involved in this methodology are:

1. Semi-supervised image classification
    a. Classifying unlabelled data in the Green and SWIR wavelength feature space
        i. Theoretical interpretation of spectral classes in the feature space – to address Research Questions 2.1 & 2.2.
        ii. Clustering methods to find data groups in the feature space – to address Research Questions 2.3 & 2.4.
    b. Nearest-Neighbour image classification using cluster labels – to address Research Question 2.5
        i. Cloud-free images for summer and winter months
        ii. Cloud-covered images for summer and winter months

2. Accuracy assessment of cloud labels for different clustering methods – to address Research Question 2.6.

The detailed description of these steps are give below:

4.2.1. Semi-supervised image classification

The method is called semi-supervised because the class labels are generated through unsupervised methods and the image classification is done using supervised classification.

- At first, the representative polygons made for Chapter 3 (Appendix B) are used to extract the mean and standard deviation of the spectral values of the Green and SWIR wavelength bands for the entire image collection using Google Earth Engine.
- The Green and SWIR wavelengths show the most distinction between clouds and snow in the optical range of wavelengths (Figure 1.2). These can also identifying bare rock areas distinct from clouds and snow as they have low and high reflectance in the Green and SWIR wavelengths respectively (Figure 1.2).
- The occurrence of cloud cover over these polygons will register the spectral signature of clouds at the sensor, instead of the underlying landcover. The spectral signals of cloud occurrences will therefore, register as an anomalous value deviating form the natural spectral behaviour of the underlying landcover.

a. Classifying unlabelled data in the Green and SWIR wavelength feature space.

i. Theoretical interpretation of spectral classes in the feature space

The extracted reflectance values of Green and SWIR wavelengths for the glacial and non-glacial polygons are plotted as a time-series and scatterplot.

- The variation of reflectance values will be visually examined for patterns to understand the seasonal evolution of the glacial and non-glacial areas.
- The feature space is the scatterplot between the Green and SWIR wavelength values. The combined spectral property of polygon values will be visually examined to find clusters of data points in the feature space.

ii. Clustering methods to find data groups in the feature space

The clustering methods are based on statistical association and not on physical relationship between data points. The following pre-processing steps are done before applying clustering methods.

- Removing non-homogenous representative polygon values

The wavelength values for glacial and non-glacial polygons are assumed to be representing either cloud or landcover classes. However, these polygons might contain mixture of cloud and/or different landcover pixels. This might result in misclassification of data points in the feature space. To eliminate this error, the band values of representative polygons with high standard deviation are removed. The standard deviation values are scaled by taking a standard deviation of the standard deviation values. The scaled values between -1 to +1 are used. The results are shown in Appendix C.

- Standardizing the Green and SWIR band values using z-score

The clustering methods use vector distance between data points in a feature space as a measure of similarity (Arthur, 2006). However, the scale of an axes depends on the range of its values. Therefore, unit distances will vary along different axes direction in a feature space. This means that the band with larger spread of values will be favoured to define clusters. To bring the reflectance values of the Green and SWIR wavelength on the same scale, z-score is used to standardize the reflectance data. z-score standardises the data based on mean and standard deviation. One unit axis represent values separated by one standard deviation from the mean, which is at zero.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where, z is the standardized value, x is the reflectance value, μ is the mean and σ is the standard deviation of the data.

- Applying clustering algorithm

Three different clustering methods are used in this study. K-Means, Mean-Shift and Spectral. All these methods require some parameter to estimate the number of cluster in the feature space. Therefore, the optimal number of clusters for each method is calculated followed by clustering. Three different clustering methods, two convex and one non-convex method is used in this study.

- K-Means:

The value of 'K' in K-Means is the expected data centers in the feature space. It is provided beforehand for clustering. The clustering begins by assigning arbitrary 'k' data centers in the feature space. The nearest data points are assigned to each of these 'centers'. The total squared distance is calculated between each data point and its assigned center. New arbitrary 'k' data centers are assigned and the process is repeated until the total squared distances are minimized for a particular distribution of 'k' data centers (Arthur and Vassilvitskii, 2007). The optimal number of clusters is found by plotting the sum of square errors from the cluster center. The cluster number where the sum of square errors saturates is selected.

- Mean-Shift

Mean-Shift clustering also known as kernel density estimation is a type of density estimation. This method does not require prior information about the number of clusters. Instead, clustering is dependent on the kernel size. A bandwidth parameter dictates the size of processing window (kernel) in the feature space.

The method begins at each data point by calculating the mean value within a kernel. The centroid of this kernel is then shifted to the calculated mean value in the feature space. A new mean value is calculated and the kernel centroid is shifted to this value. The process continues until the centroids converge (Comaniciu and Meer, 2002). The optimum number of clusters are identified by the stabilization of bandwidth values for the number of clusters.

- Spectral

Spectral clustering is based on graph theory. In simple terms, the feature space is converted into a graph where data points are nodes and the distance between connected nodes are the edge weights. The algorithm then cuts this graph along the low weight edges to form clusters (Yu and Shi, 2003). The optimum number of clusters are identified by the number of lowest eigenvalues. The process consist of 3 steps:

- Forming an adjacency matrix

The feature space is converted to a graph by nearest neighbour method. 5 data points close to each other are connected to form a graph. Then an adjacency matrix of size NxN is made, where 'N' is the number of nodes (data points). The entries of this matrix represent the presence or absence of an edge between the nodes.

- Finding the eigen vectors and eigen values (Eigenvalue decomposition)

The adjacency matrix is converted to a graph Laplacian. Graph Laplacian is another matrix where, the diagonal elements are the degree (connectedness) of nodes and the off-diagonal values are the negative of edges weight.

The eigenvalues of this graph Laplacian is calculated. The number of zero eigenvalues corresponds to the number of independently connected clusters. However, eigenvalues closer to zero means that there is an almost separation of clusters. The eigenvectors associated with these small eigenvalues are selected.

- K-Means clustering on Eigenvectors

Each eigenvector for these near-zero eigenvalues gives information on how to make cuts to cluster the graph. K-Means clustering of the entries in the eigenvector matrix will assign cluster labels to the graph.

b. Nearest-Neighbour image classification using cluster labels

The kNearest classification algorithm on GEE is used in classification mode (https://developers.google.com/earth-engine/apidocs/ee-classifier-minimumdistance, last access 15-08-2022). It assigns labels to unlabelled pixels based on the distance of the closest class/label of the training set to that pixel. The distance metric used is Euclidean.

i. Cloud-free images for summer and winter months

The pixel count percent of class labels are calculated for each clustering method in the cloud-free images of summer and winter month. The cloud class labels are expected to have 0 or negligible pixel count percent in the image. These classes of negligible occurrences are identified for the next step.

ii. Cloud-covered images for summer and winter months

The pixel count percent of class labels, not present in cloud-free images of the previous step, is calculated for each clustering method in the cloud-covered images of summer and winter month. If present, the count values are checked (not negligible) and the location is visually checked to match the cloud cover.

4.2.2. Accuracy assessment of cloud labels for different clustering methods

The manually digitized cloud cover polygon is overlayed over the classification image for each clustering method for summer and winter month. The percent of pixels belonging to different classes is evaluated for the cloud cover polygon . An accuracy assessment is made to identify the clustering labels with highest cloud cover percent.

4.3. Results

Sub-objective 2: To develop a cloud mask over glacial snow cover by applying semi-supervised image classification techniques using Green and SWIR wavelength bands of Sentinel-2.

4.3.1. Time-series behaviour of glacial and non-glacial areas in Green and SWIR wavelengths

The time-series of reflectance values in the Green wavelength for the glacial and non-glacial polygons is shown in Figure 4.3. The first striking observation made from the time-series plot is the large range of reflectance values with a seasonal fluctuation by the non-glacial polygons. The non-glacial polygons occupy the peak and troughs of the reflectance values. These polygons reach peak reflectance of 15,000 (approx.) during January-February every year. Even though the reflectance values are scaled by 10,000, the non-glacial polygons show peak reflectance values greater than 1 (approx.. 1.5). The minimum reflectance values are shown in September-October every year. The fall in reflectance values from January-February to September-October is continuous and steep. The reflectance values also abruptly rises in the month of November-December every year.

The glacial polygons, on the other hand, show a rather vague seasonal pattern in the reflectance values of Green wavelength (Figure 4.3). The peak values are achieved somewhat around April-May every year. The maximum reflectance values attained are around 12,000 (unscaled reflectance approx. 1.2) which is still higher than 10,000 (unscaled 1). The lowest reflectance values vary among the glacial polygons. The glacial_4 (and to a lesser extent glacial_3) shows minimum reflectance values similar to non-glacial polygons. On the other hand, the glacial_1 polygon never shows reflectance values below 5,000 (approx.). The fall in reflectance values is also continuous and gradual like non-glacial polygons during April-May to September-October. In the month of September during 2021, both glacial and non-glacial polygons show a sudden rise in reflectance values.

The time-series of reflectance values in the SWIR wavelength for glacial and non-glacial polygons is shown in Figure 4.4. There exists a seasonal fluctuation with a small amplitude (0-2,000) for glacial polygons and with a slightly higher amplitude (0-3,000) for non-glacial polygons. The highest SWIR wavelength values appear discontinuous to the cycles of glacial and non-glacial polygons.

The non-glacial polygons show peak values in summer-autumn (July-November). Interestingly, the glacial and non-glacial polygons show similar reflectance during winter and the reflectance values continue to drop for all polygons. However, ever year starting from August (approx.), the glacial polygons reflectance values stoops down to near zero values while the non-glacial ones reach their maximum (Figure 4.3).

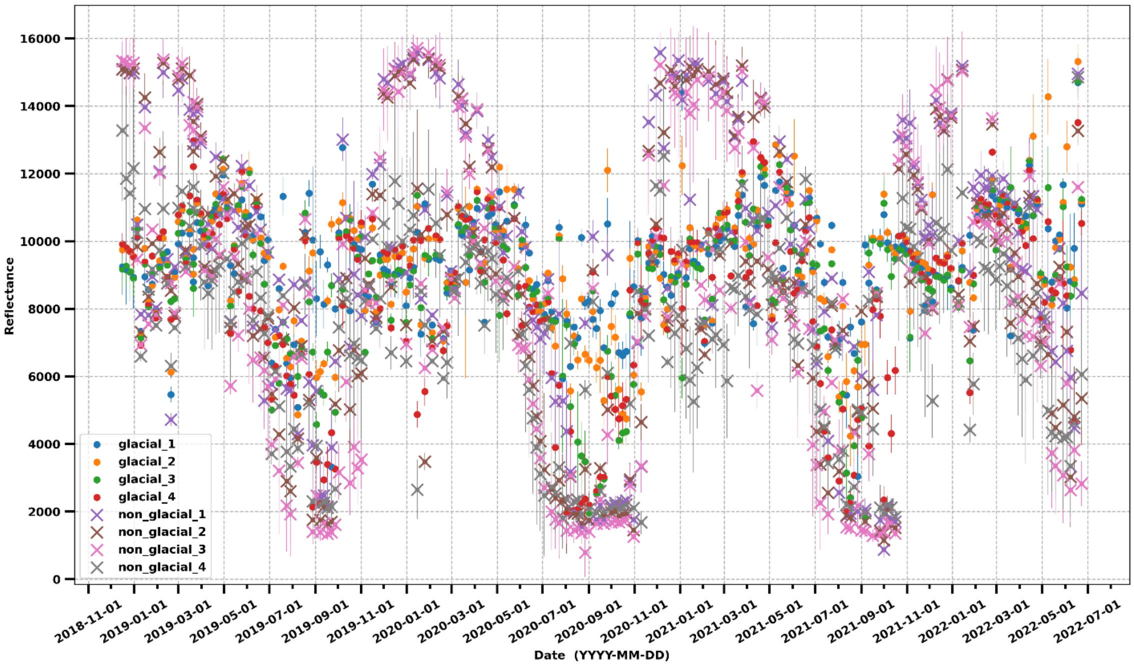


Figure 4.3 Time series of mean reflectance value of Green wavelength for representative polygons. The error bars are standard deviation of mean polygon reflectance values.

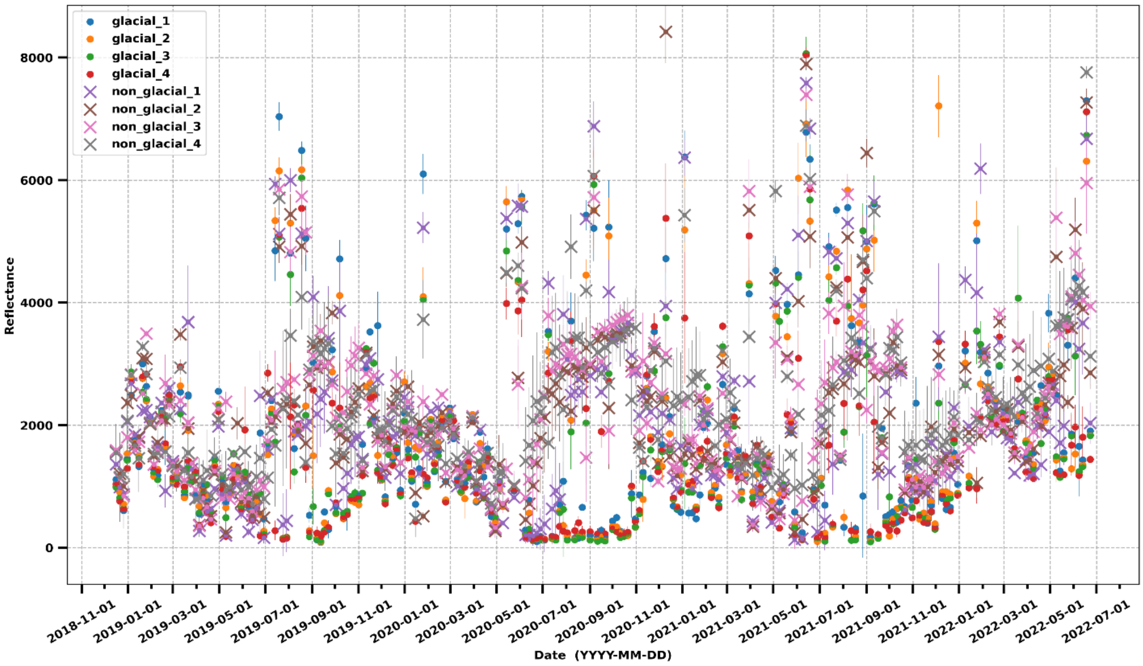


Figure 4.4 Time series of mean reflectance value of SWIR wavelength for representative polygons. The error bars are standard deviation of mean polygon reflectance values.

4.3.2. Spectral signals of cloud and landcover in the feature space

The scatterplot between reflectance values of Green and SWIR wavelength for glacial and non-glacial polygons is shown in Figure 4.5. This scatterplot is also termed as a feature space, where the combined spectral signals of clouds and landcover is interpretated. Shapes are drawn on the feature space to highlight the different spectral signals visually observed. Although the boundaries between groups of data points is not distinct, the feature space still shows some clusters.

The non-glacial polygons occupy two ends of the feature space (red circle in Figure 4.5). This high and low reflectance value was also observed in the Green wavelength time series (Figure 4.3). However, the high values in Green wavelength has low values in SWIR wavelength. Similarly, low values in Green wavelength is associated with high values in SWIR wavelengths (red circles in Figure 4.5). Most of the data is clustered around the black circle which shows a mixture of glacial and non-glacial polygons showing almost 1 reflectance (unscaled) (Figure 4.5). The ellipsoidal blue shape shows the cluster of glacial polygons (mostly glacial_4 and glacial_3) showing low Green and SWIR wavelength values (Figure 4.5). The most unusual distribution of data points is in the green shape (Figure 4.5). The data points are not close together to form a cluster yet not too far apart to be called sparse. It has a mixture of glacial and non-glacial polygon values and a broad range of Green and SWIR wavelength. The visual clusters identified are not spherical in their shape.

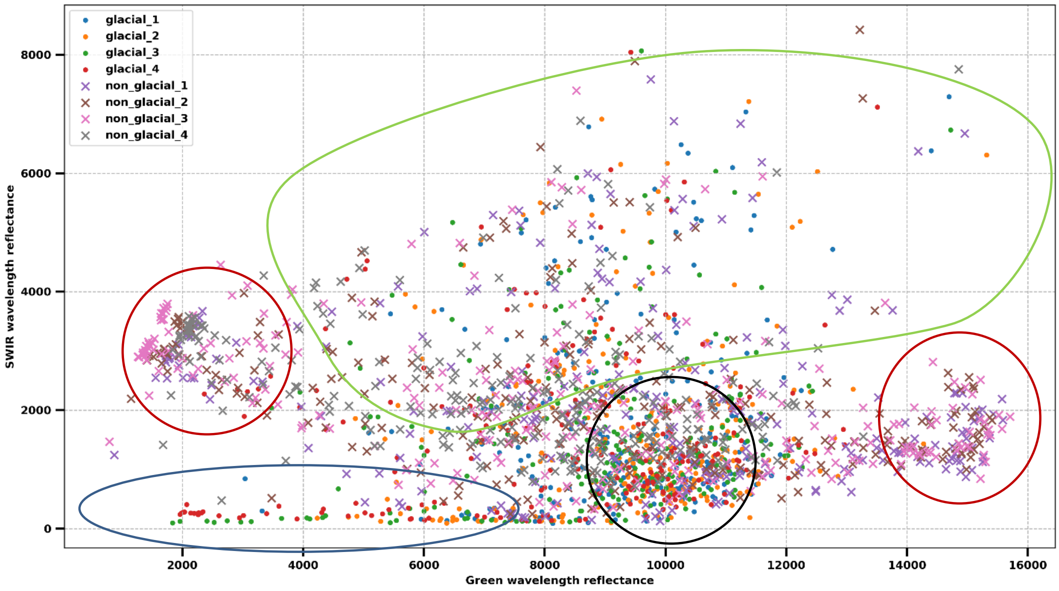


Figure 4.5 Scatter plot between mean reflectance values of Green and SWIR wavelengths for glacial and non-glacial polygons. The data points and their cluster patterns are highlight with different shapes.

4.3.3. Standardized reflectance of Green and SWIR wavelengths

Figure 4.9 shows the scatterplot of standardized reflectance values for Green and SWIR wavelengths after removing non-homogenous polygon values (Appendix C). In this scaled feature space, the Green wavelength reflectance lies approximately between two standard deviations from the zero mean. The data points also show a pronounced separation from the zero mean in SWIR wavelength reflectance. The range of values in this new feature space lies approximately between -2 to +2 and -1 to +5 for Green and SWIR wavelength reflectance respectively. This implies that a unit change in SWIR is equal to 1.5 times the unit change in Green wavelengths. Therefore, the new units have higher weights in the SWIR than in Green wavelength and are opposite to the non-standardized feature space (Figure 4.6).

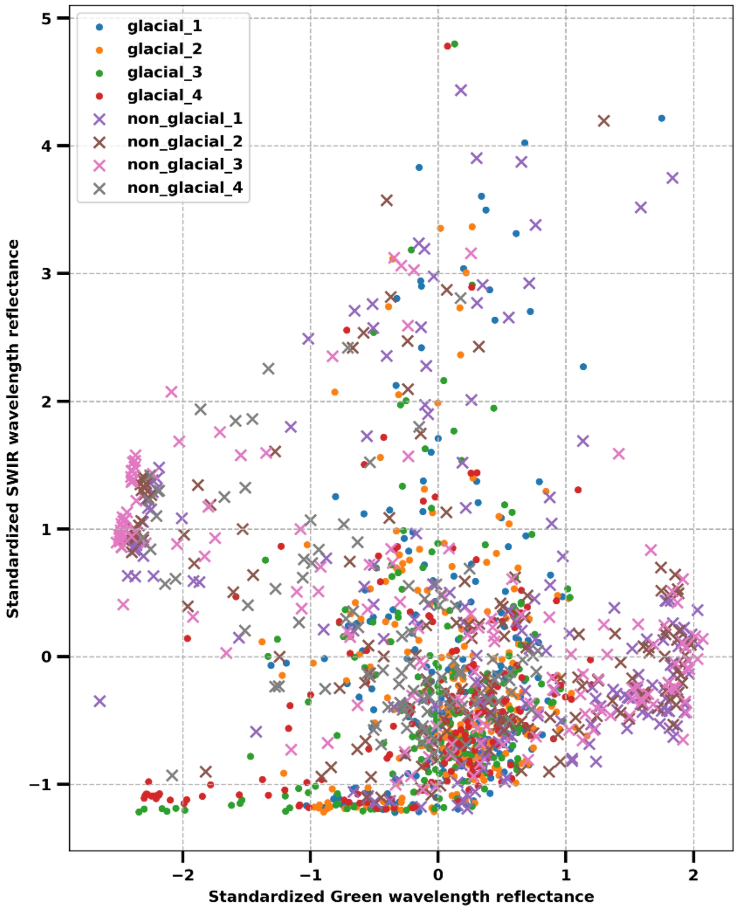


Figure 4.6 Scatter plot between standardized mean reflectance values of Green and SWIR wavelength for representative polygons.

4.3.4. Optimum clusters

1. K-Means clustering

The variation of inertia with the number of clusters for the scaled feature space (Figure 4.6) is shown in Figure 4.7(a). Inertia is a measure of goodness of K-Means clustering (Arthur and Vassilvitskii, 2007). It is defined as the squared sum of distances of samples to the nearest cluster center. Minimization of inertia is a criteria for choosing the optimal number of clusters in K-Means and it is also know as elbow method (Arthur and Vassilvitskii, 2007). Figure 4.7(a) shows that the lower inertia values are associated with higher number of clusters (>10). However, after six number of clusters, the inertia values saturate and the decrease become linear. Six clusters, relatively minimizes the inertia and does not produce too many clusters. Therefore, the optimal number of clusters for K-Means clustering is chosen to be six for this study. The K-Means clustering of scaled feature space using 6 number of clusters is shown in Figure 4.7(b).

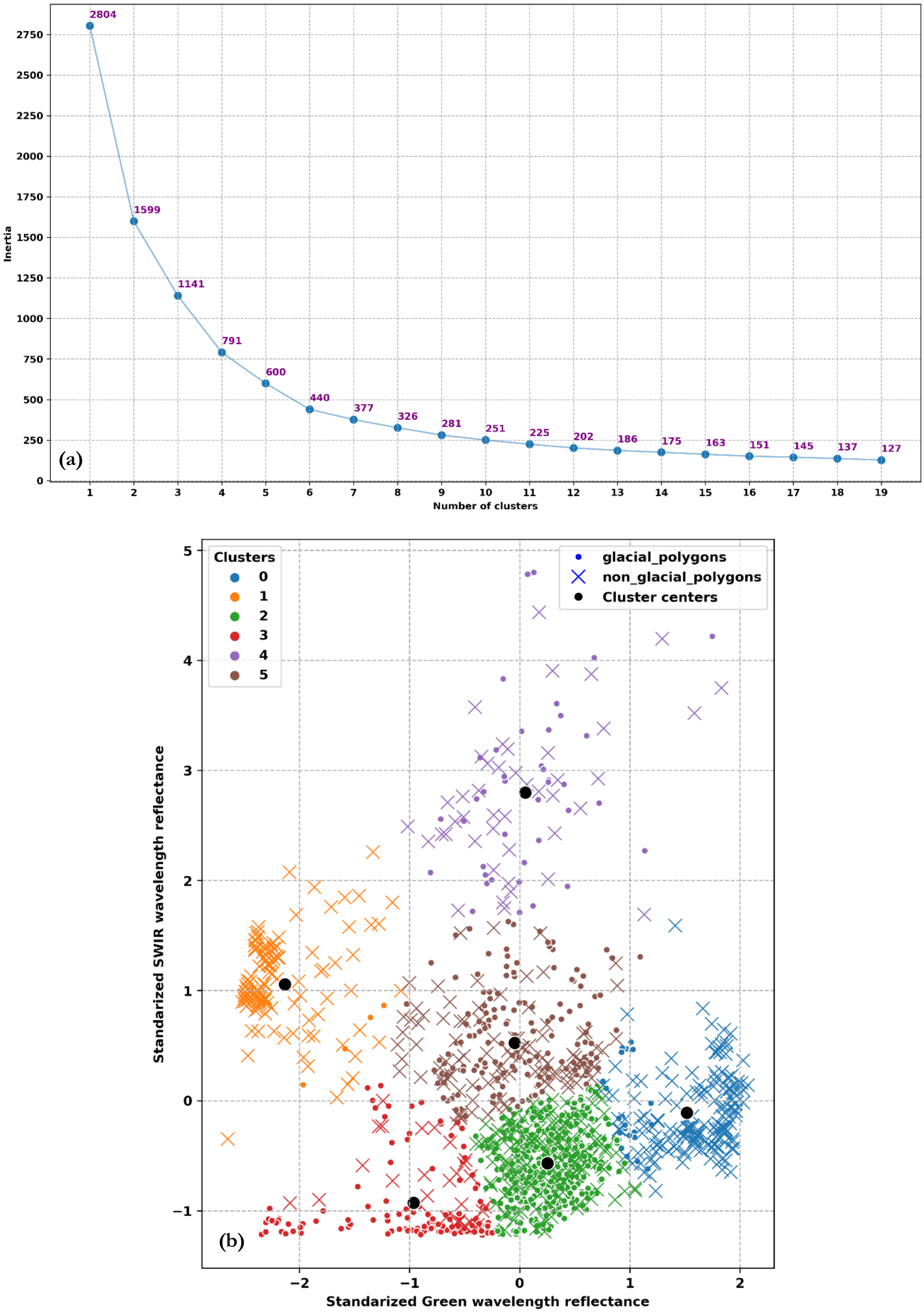


Figure 4.7 The clustering of scaled feature space using optimum number of K-Means clusters. Figure (a) shows the variation of inertia and number of clusters used in K-Means clustering. The rounded inertia values (in purple) are also plotted at each data point in the graph. Figure (b) shows the partitioning of scaled feature space into clusters. The cluster centers are shown in black dots in (b).

2. Mean-Shift clustering

The number of clusters in the Mean-Shift clustering is dependent on the bandwidth (kernel size) parameter. Figure 4.8(a) shows the relationship between the number of clusters in the scaled feature space (Figure 4.6) for the estimated bandwidth values. The whole feature space is classified as a single cluster for bandwidth values greater than 1.36 (Figure 4.6 (a)). The Mean-Shift algorithm separates the feature space into 2, 3, 4 and 10 number of clusters over a broad range of bandwidth values (Figure 4.6(a)). The cluster centers for these bandwidth values are thus considered stable in the sense that a small change in bandwidth does not have a significant effect on the partitioning of the feature space. Also, the minimum number of clusters visually observed in the feature space is 5 (Figure 4.5). The only value fulfilling this criteria among the number of clusters is 10.

The optimum number of clusters can also be evaluated by looking at the kernel density estimation (KDE) plots for the bandwidth values. KDE smoothens (sharpens) as the bandwidth values increases (decreases) in Figure 4.9(a-f). For relatively low bandwidth values (Figure 4.9(f)), it is able to highlight small dense regions within the feature space, which are otherwise smoothened out for relatively high bandwidth values (Figure 4.9(a)). However, all the plots show 1 central and 4 surrounding regions of high density (blue coloured regions in Figure 4.9).

The Mean-Shift algorithm is run for the 6 bandwidth values used to make these 6 KDE plots in Figure 4.9. The cluster centers are calculated for each of the six bandwidth values. These cluster center are plotted on top of the KDE plot in Figure 4.9. and indicated as red dots. The bandwidth value where the cluster centers can identify all the dense regions are noted (Figure c, d, e, f). The number of clusters for these bandwidth values are 8, 10, 14 and 20. Among these, the feature space is over-classified by bandwidths associated with 14 and 20 number of clusters (Figure 4.9(e-f)).

Together with the results of Figure 4.8, the 4.9, the 10 number of clusters seems appropriate for clustering the feature space. The result of Mean-Shift clustering using the bandwidth value 0.66 for clustering the feature space into 10 classes is shown in Figure 4.9(b).

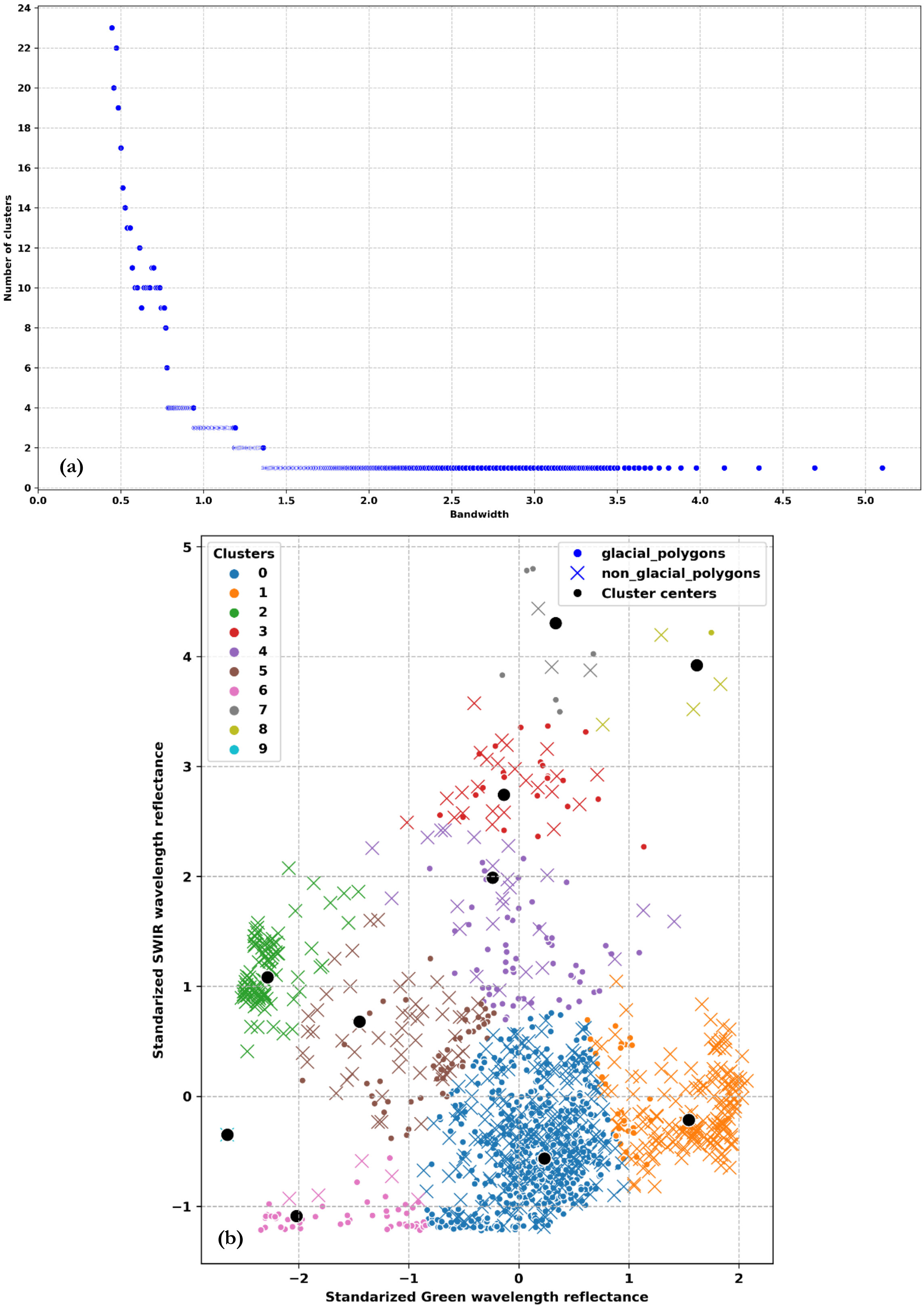


Figure 4.8 The clustering of scaled feature space using optimum number of Mean-Shift clusters. Figure (a) shows the number of clusters in the feature space for different Mean-Shift bandwidth values. Figure (b) shows the partitioning of scaled feature space into clusters. The cluster centers are shown in black dots in (b).

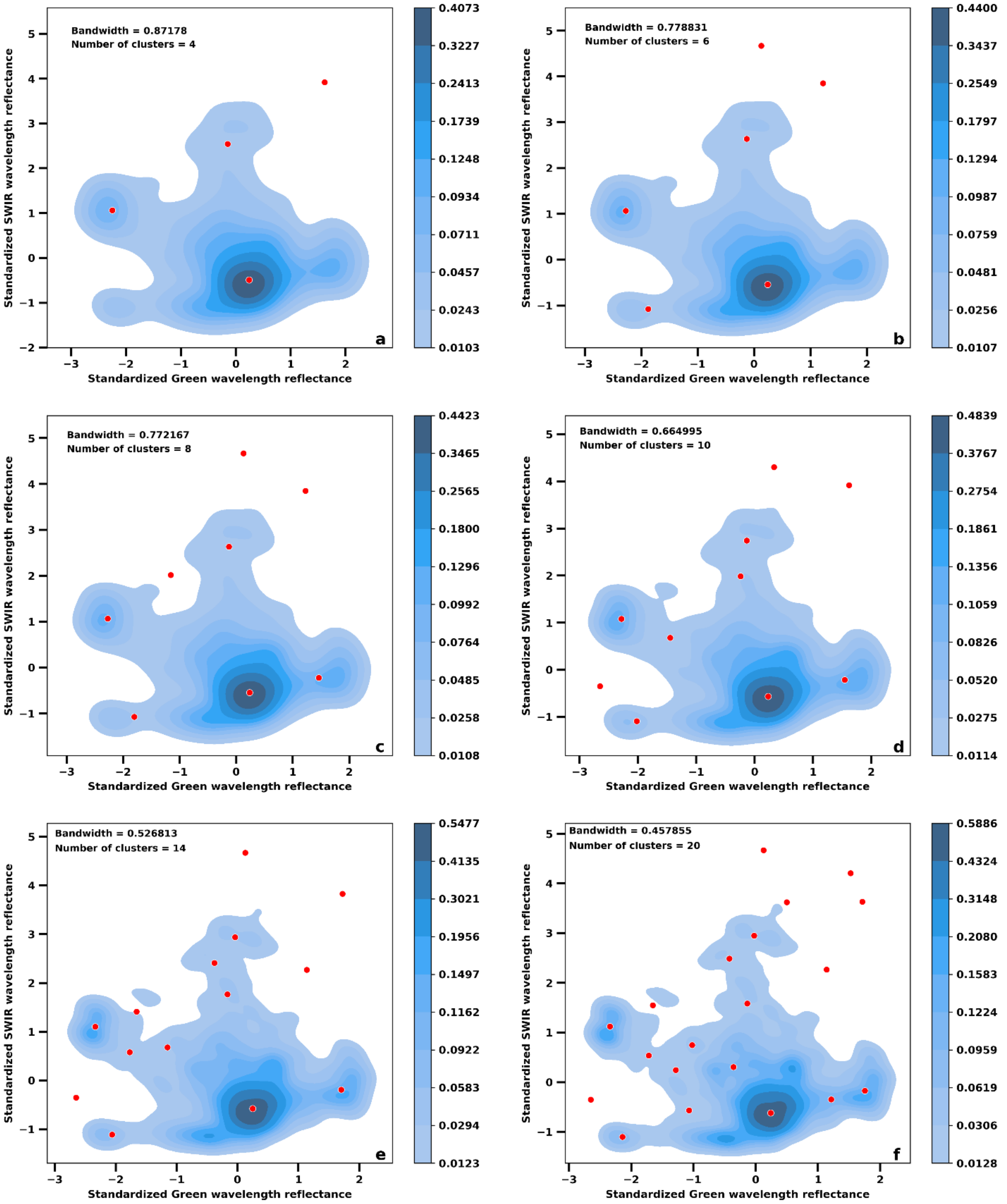


Figure 4.9 Bi-variate kernel density estimation plot between standardized Green and SWIR wavelength reflectance values for different bandwidth values (a-f). The bandwidth values for estimating the kernel density and calculating the Mean-Shift clusters is shown in each figure along with number of clusters. The location of cluster centers in the feature space is shown by red circles. The contour lines are iso-proportion to the kernel density values and the colour bar shows the range of kernel density values.

3. Spectral clustering

The eigenvalues (of the eigenvectors) of the graph Laplacian for the feature space (Figure 4.6) are sorted in the increasing order and plotted against their respective index (Figure 4.10(a)). The figure only show the 30 lowest eigenvalues (among the 1402 eigenvalues) for their respective eigenvectors.

Figure 4.10(a) depicts that from index 11 to 12, the eigenvalues show sudden rise. Before index 11, there is no significant difference among the eigenvalues. Therefore, these 11 eigenvectors are assumed to separate the data into the optimal number of graphs (clusters). The K-Means algorithm is used to cluster the data points in this reduced dimensional data whose axes are the 11 eigenvectors corresponding to the first 11 eigenvalues (Figure 4.10 (a)). Therefore, there is no presence of cluster center in the resulting clustered feature space in Figure 4.10(b).

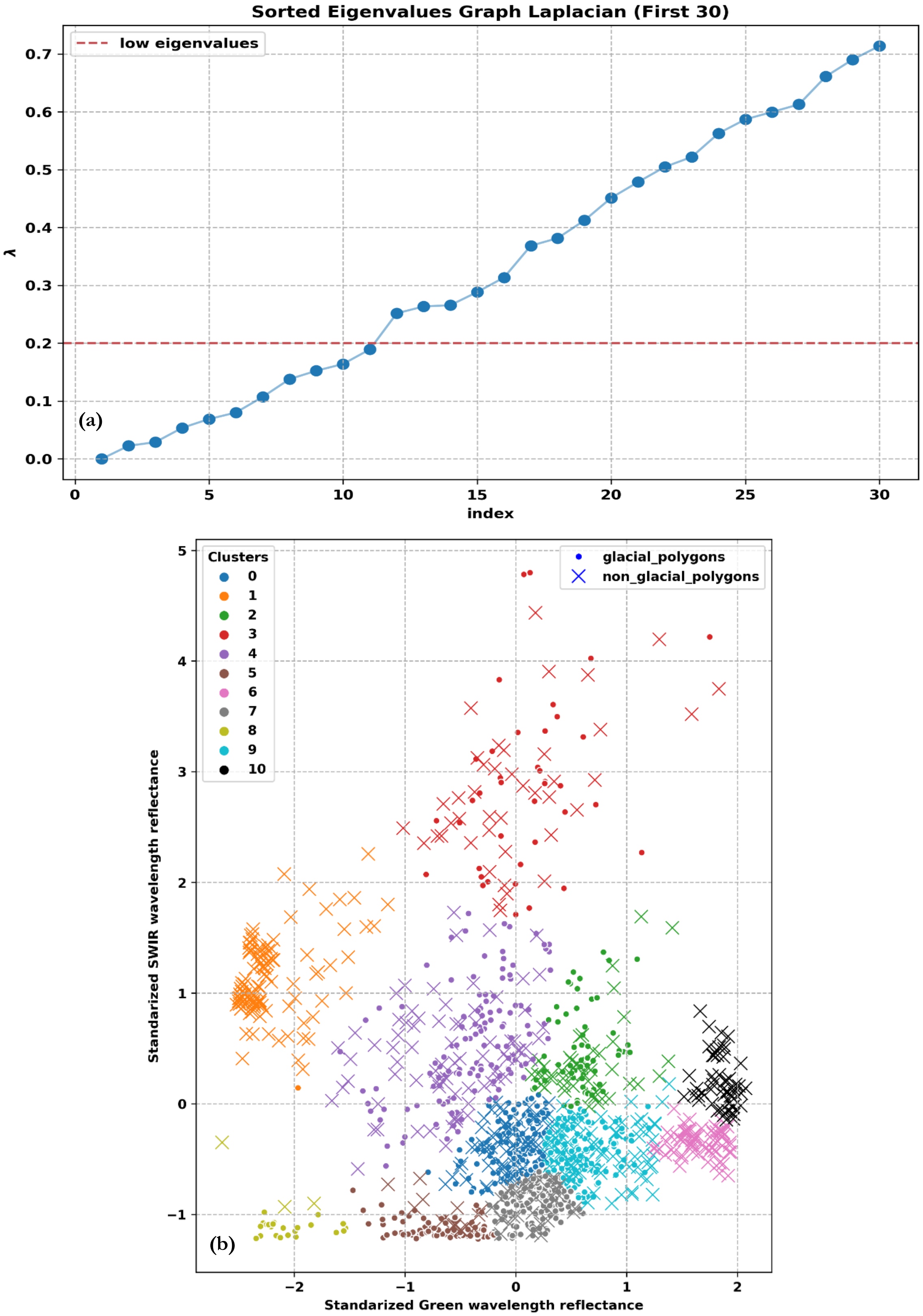


Figure 4.10 The clustering of scaled feature space using optimum number of Spectral clusters. Figure (a) shows the variation of eigenvalues of the graph Laplacian in increasing order of magnitude. Figure only show eigen values up to the index 30. The red line indicates the low eigenvalues used for clustering as explained in Section 4.3.4 on Spectral clustering. Figure (b) shows the partitioning of scaled feature space into clusters.

4.3.5. Nearest-Neighbour image classification using cluster labels

Cloud-free images: For K-Means, the cluster labels with the lowest pixel percent in both summer and winter image are 4 and 5 (Table 4.1). For Mean-Shift, the cluster labels with the lowest pixel percent in both summer and winter image are 3, 4, 7 and 8 (Table 4.1). For Spectral, the cluster labels with the lowest pixel percent in both summer and winter image are 2, 3 and 4 (Table 4.1). These cluster labels are assumed to be representing clouds.

Table 4.1 The percentage of pixels in the classified images of 4.12 (cloud-free images) belonging to the label classes of different clustering methods. The values (colour coded in green) show less than 5% values only if found in both summer and winter images.

K-Means			**Mean-Shift**			**Spectral**		
	Summer	**Winter**		**Summer**	**Winter**		**Summer**	**Winter**
Cluster labels	Pixel count (%)	Pixel count (%)	Cluster labels	Pixel count (%)	Pixel count (%)	Cluster labels	Pixel count (%)	Pixel count (%)
0	1.0	25.8	0	8.9	29.1	0	2.1	12.1
1	61.1	19.6	1	0.7	26.6	1	38.1	2.4
2	5.0	24.9	2	33.2	1.5	2	**0.0**	**2.7**
3	32.2	24.3	3	**0.3**	**0.3**	3	**0.5**	**0.6**
4	**0.5**	**0.9**	4	**0.0**	**2.0**	4	**0.5**	**3.1**
5	**0.2**	**4.6**	5	10.1	10.1	5	22.3	15.7
			6	26.7	18.1	6	0.3	7.5
			7	**0.3**	**0.0**	7	2.1	6.2
			8	**0.0**	**0.1**	8	33.1	24.9
			9	19.7	12.2	9	0.7	11.6
						10	0.3	13.2

Cloud-cover images: The cluster labels assumed to be clouds (in cloud free images above) are checked for their combined pixel percent of cluster labels for cloudy images. For K-Means, the combined pixel percent of cluster label 4 and 5 in summer is 15.8% and in winter is 16.2%. For Mean-Shift, the combined pixel percent of cluster label 3, 4, 7 and 8 in summer is 11.4% and in winter is 8.5%. For Spectral, the combined pixel percent of cluster label 2, 3 and 4 in summer is 19.3% and in winter is 24.5%.

Table 4.2 The percentage of pixels in the classified images of 4.13 (cloud-covered images) belonging to the label classes of different clustering methods. The values (colour coded in blue) shows values for the classes (<5%) identified in Table 4.1.

K-Means			**Mean-Shift**			**Spectral**		
	Summer	**Winter**		**Summer**	**Winter**		**Summer**	**Winter**
Cluster labels	Pixel count (%)	Pixel count (%)	Cluster labels	Pixel count (%)	Pixel count (%)	Cluster labels	Pixel count (%)	Pixel count (%)
0	3.3	28.7	0	19.8	26.2	0	3.0	13.2
1	45.6	13.0	1	4.1	28.3	1	26.2	0.7
2	18.3	18.9	2	21.2	0.4	2	**0.5**	**15.4**
3	16.9	23.1	3	**7.3**	**1.0**	3	**9.0**	**2.8**
4	**10.9**	**3.7**	4	**2.5**	**5.2**	4	**9.8**	**6.3**
5	**4.9**	**12.5**	5	13.4	12.1	5	8.6	14.4
			6	13.0	17.2	6	1.3	4.3

				7	**1.1**	**0.3**		7	7.9	1.3
				8	**0.5**	**2.0**		8	24.1	19.4
				9	17.1	7.4		9	9.3	4.0
								10	0.2	18.0

K-Means: The cluster label 4 and 5 cannot be seen in the classified cloud-free images of summer and winter (Figure 4.11 (a, b)). The cluster label 4 and 5 can be seen in the true cloud cover in summer and winter (Figure 4.12(a, b)). The cluster 5, in the cloud cover winter classified image (Figure 4.12(b), appears over glacial snow (Figure 3.2(a)).

Mean-Shift: The cluster label 3, 4, 7 and 8 cannot be seen in the classified cloud-free images of summer and winter (Figure 4.11 (c, d)). The cluster label 3, and 4 can be seen in the true cloud cover in summer and winter (Figure 4.12(c, d)).

Spectral: The cluster label 2, 3 and 4 cannot be seen in the classified cloud-free images of summer and winter (Figure 4.11 (e, f)). The cluster label 2, 3 and 4 can be seen in the true cloud cover in summer and winter (Figure 4.12(e, f)). The cluster 2, in the cloud cover winter classified image (Figure 4.12(f), appears over glacial snow (Figure 3.2(a)).

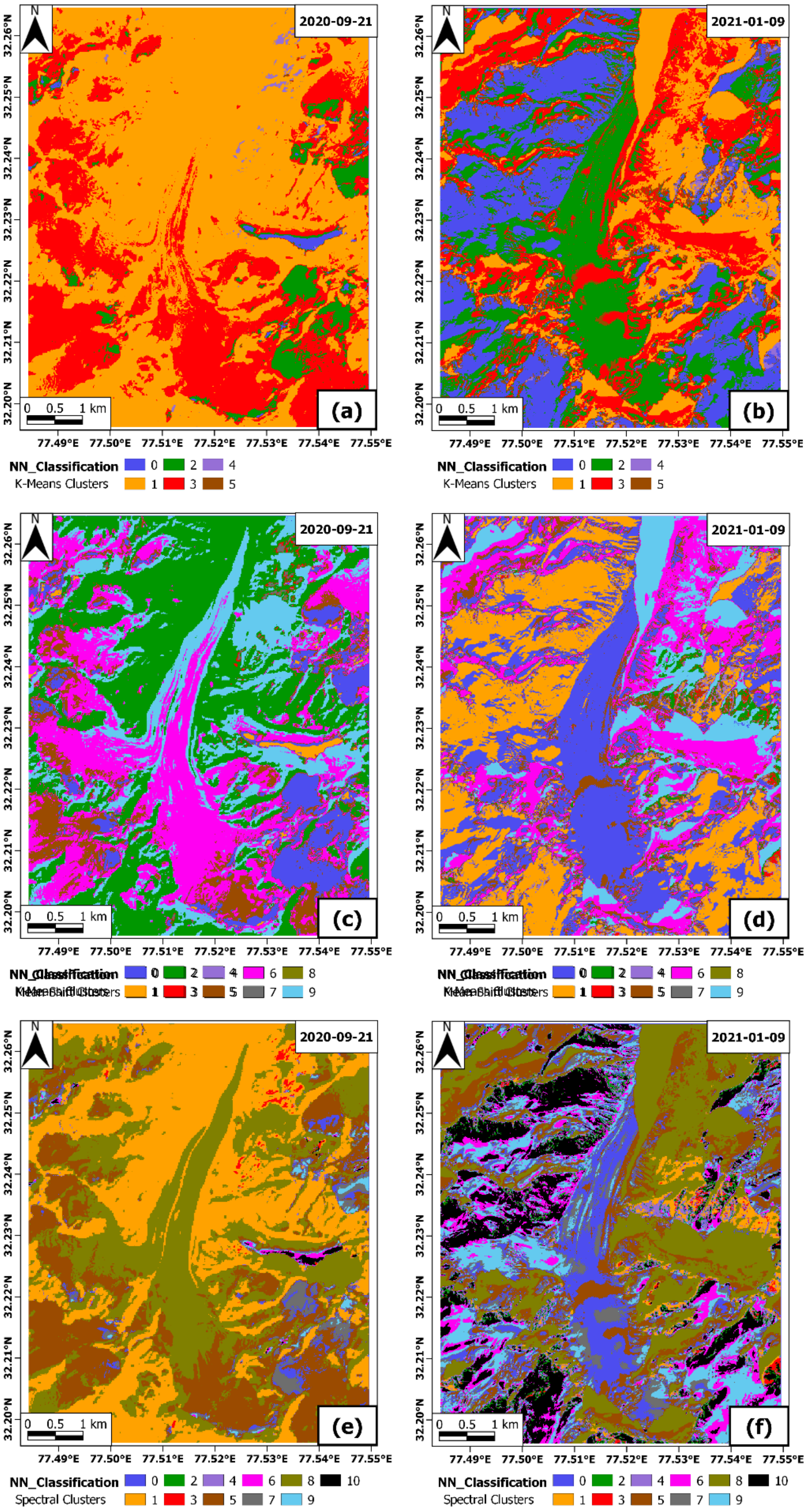

2020-09-21
2021-01-09
(a)
(b)
(c)
(d)
(e)
(f)
N
0 0.5 1 km
32.26°N
32.25°N
32.24°N
32.23°N
32.22°N
32.21°N
32.20°N
77.49°E 77.50°E 77.51°E 77.52°E 77.53°E 77.54°E 77.55°E
NN_Classification 0 2 4
K-Means Clusters 1 3 5
NN_Classification 0 2 4 6 8
Mean-Shift Clusters 1 3 5 7 9
NN_Classification 0 2 4 6 8 10
Spectral Clusters 1 3 5 7 9

Figure 4.11 The Nearest-Neighbour image classification result of cloud free images using different clustering labels. Figure (a), (c) and (e) shows summer classified images using K-Means, Mean-Shift and Spectral cluster labels, respectively for the summer image (dated: 01-10-2021). Similarly, Figure (b), (d) and (f) shows classified images using K-Means, Mean-Shift and Spectral cluster labels, respectively for the winter image (dated: 05-12-2020).

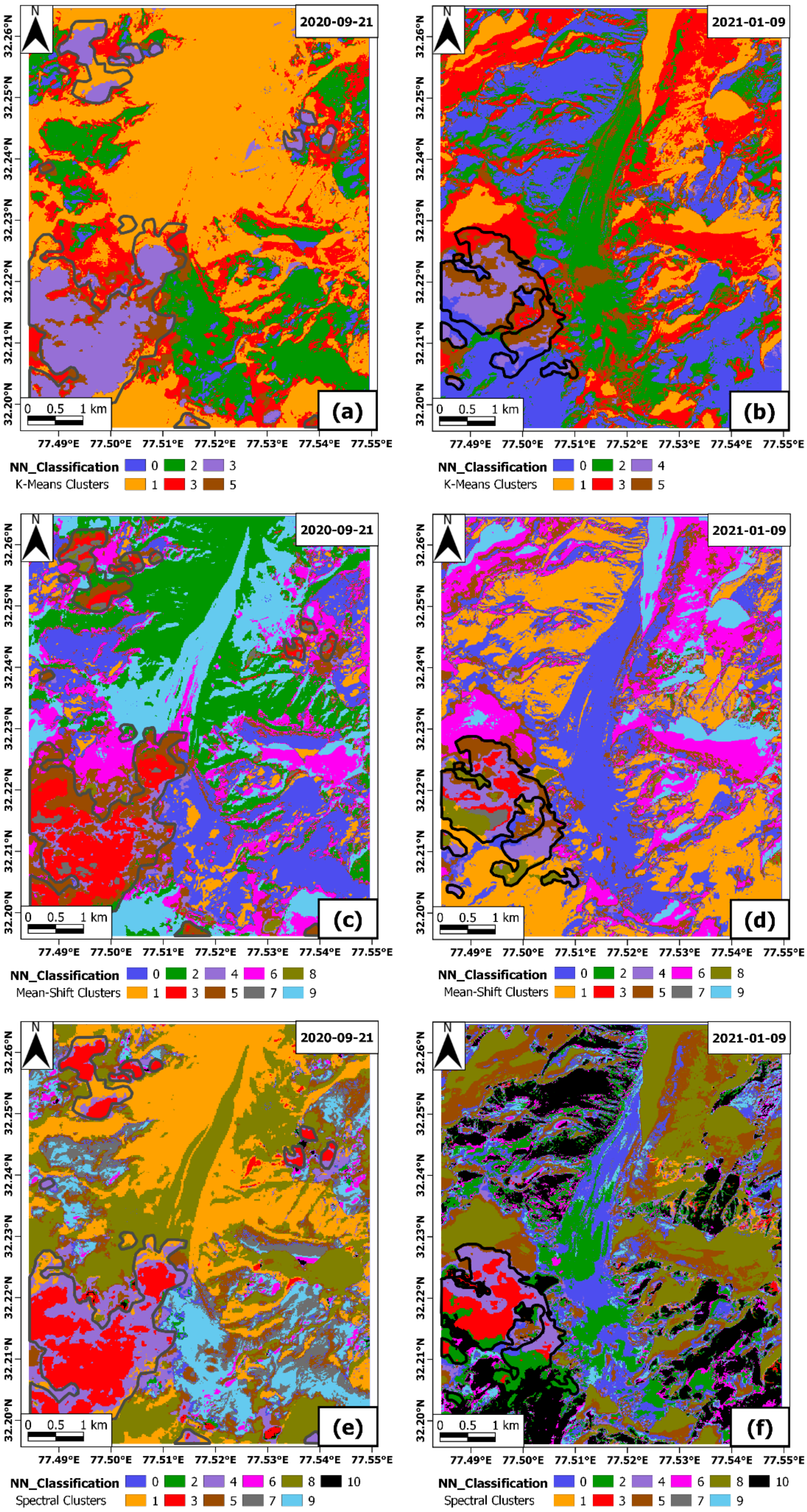


Figure 4.12 The Nearest-Neighbour image classification result of cloud covered images using different clustering labels. Figure (a), (c) and (e) shows summer classified images using K-Means, Mean-Shift and Spectral cluster labels, respectively for the summer image(dated: 21-09-2020). Similarly, Figure (b), (d) and (f) shows classified images using K-Means, Mean-Shift and Spectral cluster labels, respectively for the winter image (dated: 09-01-2021).

Research Question 2.6: Which clustering method performs better at cloud masking over snow cover in the study area?

The percentage of labelled pixels within the manually digitized cloud polygon for different clustering methods is shown in Table 4.1. The labels identified as non-occurring in cloud-free images and appearing in appreciable amount in cloud-covered images are highlighted. The results show that the Spectral cluster labels detect the most 'true positives', approx.. 85% in summer and approx.. 89% in winter cloudy images, followed by K-Means cluster labels (approx.. 75% in summer and winter cloudy images). Mean-Shift cluster labels perform the worst cloud detection among the three clustering methods.

Table 4.3 The percentage of pixels belonging to different cluster labels, lying within the manually digitized cloud polygon for summer and winter images for different clustering methods.. The rows (colour coded in green) show values in either summer and winter images for the <5% classes identified in table 4.1. The total percent covered by these identified classes within the manually digitized polygon 'true positive' is also shown along with 'false negative'.

K-Means			**Mean-Shift**			**Spectral**		
	Summer	**Winter**		**Summer**	**Winter**		**Summer**	**Winter**
Cluster labels	**Pixel count (%)**	**Pixel count (%)**	**Cluster labels**	**Pixel count (%)**	**Pixel count (%)**	**Cluster labels**	**Pixel count (%)**	**Pixel count (%)**
0	0	14.5	0	0.1	1.5	0	0.1	1.5
1	17	0	1	0	6.6	1	12.2	0
2	0.1	0.5	2	4.7	0	2	**0.2**	**24.1**
3	7.3	8.8	3	**44**	**12.1**	3	**51.1**	**32.1**
4	**62.2**	**40.7**	4	**8.9**	**38.8**	4	**34.8**	**33.3**
5	**13.4**	**35.6**	5	33.1	19.4	5	0.1	2
			6	2.1	0.7	6	0	0.1
			7	**5.5**	**4.2**	7	0	0
			8	**1.1**	**16.8**	8	1.4	0.3
			9	0.5	0	9	0	0
						10	0	6.6
True positives	**75.6**	**75.13**	**True positives**	**58.15**	**70.19**	**True positives**	**85.11**	**89.5**
False negatives	**24.4**	**14.87**	**False negatives**	**41.85**	**29.81**	**False negatives**	**14.89**	**10.5**

The cluster labels identified in previous steps for each clustering method, is checked for its pixel count in the classified cloud-covered images. The pixel count of manual cloud covered polygon is also checked for these cluster labels in the classified cloud-covered images. The percent of pixels occupied by these labels in manual cloud covered polygon to the classified cloud covered images is calculated and shown in Table 4.2. The low percent value are identified and shows the 'percent of true positives within each label'. The K-Means label '5' for winter shows only 19% pixels are actually in manual cloud covered polygon. The Spectral label '2' for both summer and winter shows very low coverage of cloud cover in this label. Mean-shift, however, have >50% coverage for all the identified labels.

Table 4.4 Percent of manual cloud covered pixel count to the total pixel count by the label in the image. The low percent values are highlighted.

K-Means			**Mean-Shift**			**Spectral**		
Cluster labels	**Summer (%)**	**Winter (%)**	**Cluster labels**	**Summer (%)**	**Winter (%)**	**Cluster labels**	**Summer (%)**	**Winter (%)**
4	90.9	75.4	3	95.2	82.7	2	**7.5**	**10.7**
5	**42.9**	**19.6**	4	**56.5**	**51.6**	3	90.1	78.5
			7	80.2	92	4	56.2	36.4
			8	36.5	58.2			

4.4. Discussion

Research Question 2.1: What is the spectral behaviour of glacial and non-glacial areas in the Green and SWIR wavelengths?

The non-glacial polygons show seasonal snow cover during winter and bare rock during summer. The winter snow cover on bare rocks show unusually high peak reflectance (approx. 1.5). This is due to the approximate measurement of bi-directional reflectance value by the sensors of passive satellite missions (Schaepman-Strub et al., 2006). The approximate measurement is due to the large solid angle used by satellite sensors, in their instantaneous field of view, for integrating the radiance/reflectance values (Schaepman-Strub et al., 2006). This is not in agreement with the infinitesimal requirement (theoretical) of solid angle to measure bi-directional reflectance (Schaepman-Strub et al., 2006). The natural reflected radiance from surface features also contain diffuse (scattered) component, in addition to the direct component. This diffuse component is affected by the atmosphere, topography and nature of topographic surface (Schaepman-Strub et al., 2006). Therefore, bright surfaces like snow, which also possess forward reflecting surfaces, may show reflectance values greater than 1 (Dozier and Painter, 2004).

Sentinel-2 surface reflectance images is able to capture the seasonal transitioning of snow cover associated with summer and winter months in the study area. The winter images are clearly snow covered due to general high reflectance in the Green wavelength. The glacial areas have a very low range of reflectance values compared with non-glacial areas. The lowest spread in reflectance values is shown by the high elevation areas (accumulation zone) having permanent snow cover throughout the year. The transition from bare rocks to snow in the annual cycle leads to the large spread of reflectance values observed in the non-glacial areas. The non-glacial areas show unusually high reflectance in the Green wavelength during winter months. This could be attributed to either the snow is more pure (uncontaminated) on the non-glacial mountain regions than glacial snow or, the reflectance values measured at the satellite sensor are affected by some other factor. Some of the non-glacial areas lie on the mountain slopes. Hence, geometric factors cannot be ruled out for such high reflectance values.

Research Question 2.2: How does clouds exhibit their spectral signals in the Green and SWIR wavelength reflectance of glacial and non-glacial areas?

The presence of clouds definitely impart anomalous reflectance values to the otherwise cyclic time-series of Green wavelength reflectance. However, they cannot be separated using spectral thresholds. On the other hand, the time-series of SWIR wavelength reflectance clearly show elevated cloud signals. The time-series is filled with unusually high reflectance values. However, bare rocks can also show seasonal high reflectance in SWIR wavelength and makes it difficult to use a threshold to separate clouds from landcover.

Except the cloud signals, the snow and bare rocks show a clear seasonal cycles of low and high reflectance in both Green and SWIR wavelengths. The snow and bare rocks reflectance in both Green and SWIR wavelength is always at odds with each other. They show clear spectral separation in summer months. Although the distinction is more pronounced in the SWIR wavelength. The maximum separation is observed in May-June of each year, especially in the year 2020. This might be explained by the melting of winter snow cover and exposure of rocks in non-glacial areas.

The combined Green and SWIR wavelength behaviour of clouds and landcover classes creates unique clusters in the feature space. The snow pixels have the densest cluster. The non-glacial areas show two end members along the Green wavelength associated with snow covered bare rocks and snow-free bare rocks. Only the low elevation glacial areas show large spread in Green wavelength due to transition from snow to glacial ice (Vincent et al., 2013).

Research Question 2.3: How does the standardization of reflectance data affect the clustering process?

The vector distance between data points in the standardized feature space (Figure 4.6) assigns more weights along the SWIR wavelength compared to the SWIR wavelength in the unscaled feature space (4.5). The presence of cloud over a pixel raises the SWIR wavelength reflectance relative to the underlying landcover. However, the amount of rise in this reflectance is dependent on the cloud physical and optical properties. The aim of this study to successfully detect such anomalous rise in SWIR wavelength reflectance and classify it as clouds. Therefore it was important to change the original scale of reflectance values. z-score brings the Green and SWIR wavelength reflectance on the same scale and in doing so assigns more weight to the SWIR axis compared to the SWIR axis in unscaled feature space. This change in axis weights meant that the clusters are much more separated along the SWIR axis than the Green axis in the scaled feature space compared to the unscaled feature space.

Research Question 2.4: How does the optimum number of clusters vary for different clustering methods?

The optimum clusters calculated for K-Means, Mean-Shift and Spectral clustering are 6, 10 and 11 respectively. The Spectral clustering gives non-spherical clusters in the feature space (Figure 4.10(b)). For example, the cluster label 3 and 10 shows very different cluster shapes. Cluster label 10 is compact while 3 has sparsely populated data points. Although the cluster number of Mean-Shift and Spectral clustering show near similar values, the clustering of feature space by Mean-Shift algorithm shows more similarity to the K-Means clustering (Figure 4.7(b) and Figure 4.8(b)). Some of the cluster centers in K-Means and Mean-Shift clustering have near identical values. For example, the clusters of non-glacial polygon representing bare rock and snow covered bare rock (Figure 4.7(b), 4.8(b)). The number of clusters in K-Means clustering is easily calculated by minimizing the inertia value. However, the cluster formed are spherical in nature irrespective of the cluster type.

Some of the clusters formed by Mean-Shift algorithm have very few data point and are part of nearby clusters in other clustering results (Figure 4.7(b), 4.8(b) and 4.10(b)). For example, in the Mean-Shift clustering the cluster label 9 has only one data point. These isolated clusters are formed due to the nature of Mean-Shift algorithm (Comaniciu and Meer, 2002). The algorithm tries to find the mode of values within a region defined by the bandwidth parameter. The sparsely located points will act as local maxima and for specific bandwidth parameter, these will be assigned a cluster. This is also seen in the KDE plot of the feature space (Figure 4.9). These small sized sparse clusters (7, 8 and 9) and their cluster centers calculated by the Mean-Shift algorithm have too low densities to be represented on the KDE plot. However, trying to remove these clusters by choosing higher bandwidth values ends up removing important clusters too (Figure 4.9).

Research Question 2.5: What are the cloud classes identified using different cluster labels in nearest-neighbour image classification?

The K-Means clustering shows high likelihood of two clusters, labelled 4 and 5 to be the cloud classes. These cloud classes account for <1% of total pixels in the cloud-free images, except class 5 which shows slightly higher pixel count (approx. 5%) in the winter cloud-free image (Figure 4.11). As expected, the pixel count increases for both cloud classes in the cloud-covered images but still remain less than 13% of the total pixels (Figure 4.11). The combined pixel count of these cloud classes for summer and winter cloud covered images are 15.8% and 16.2% respectively (Figure 4.11). The true pixel count for the summer and winter cloud covered images are 16% and & 7%, respectively (Appendix A). The winter cloud cover is over estimated by twice as much by these cloud classes 4 and 5. The probable source of this over estimation is the classification of snow pixels as clouds by cloud class 5. The classified winter image (Figure 4.13(b)) shows the cloud class 5 is assigned to both true cloud pixels and glacial snow pixels. However, the cluster label 2 is assigned to snow cover pixels in the classified images of the study area (Figure 4.12(a, b), Figure 2.13(a, b)). In the K-Means clustered feature space, the cluster label 2 and 5 lie adjacent to each other (Figure 4.7(b)). Therefore, the data points in cluster 5 are not correctly partitioned by the K-Means clustering algorithm and contain spectral signals of snow cover.

The Mean-Shift clustering shows high likelihood of four clusters, labelled 3, 4, 7 and 8 to be the cloud classes. These cloud classes account for not more than 2% of pixels in cloud-free images. The cloud classes 7 and 8 account even less (<0.3%) to the total pixels in the cloud-free images and only increased slightly to not more than 2% of pixel in the cloud-covered images. The cloud classes 3 and 4 show higher pixel count in cloud covered image approx. 7% to the total pixels in the image. The Spectral clustering shows high likelihood of three clusters, labelled 2, 3, and 4 to be the cloud classes.

Research Question 2.6: Which clustering method performs better at cloud masking over glacial snow cover?

The cloud classes identified among the labels of Spectral clustering are '2', '3' and '4'. These cloud classes together detect more cloud pixels (>85% for both summer and winter images) than all cloud classes in K-Means and Mean-Shift clustering. However, in Spectral clustering a cloud class ('2' – green colour) shows very high error in cloud pixel classification (Table 4.2). In the winter image, it also classifies glacial snow in the aforementioned label (Figure 4.13(f)). This cloud class could thus be a mix of snow and cloud spectral properties. In the clustered feature space (Figure 4.10), this cloud class is located at high Green and intermediate SWIR wavelength reflectance. Notably, corresponding cluster of cloud class 2 is surrounded by cloud or snow classes. This may explain why that specific cloud class contains mixed spectral properties (belonging to clouds and snow). However, the removal of cloud class 2 would not significantly affect the cloud detection capability of summer image (with 0.2% contribution as shown in Table 4.2). This might be due to few snow pixels of such high reflectance in Green wavelength.

This mix of snow and cloud classification by cloud class '2' (green) might also be due to the nature of classifying algorithm used. The Nearest-Neighbour algorithm assigns labels to image pixels based on the distance to the nearest available class label in the feature space and therefore, does not depend on the shape or density of the cluster (Boiman et al., 2008). This implies that data points in a cluster located far from the cluster centers can assign class labels to the nearest unlabelled image pixel. This image pixel might have no similarity to the cluster itself. Looking closely at the spread of data values in cloud class '2' (green), it appears that more data points are clustered at lower SWIR wavelength values and few data points extend along the SWIR axis. At higher SWIR wavelength the data points are labelled '3' (red) and '4' (purple) belonging to the cloud classes. Therefore, snow pixels and cloud pixels of similar reflectance values in the Green wavelength but different in SWIR wavelength is assigned the same class '2' (green).

5. CONCLUSION AND RECOMMENDATIONS

To conclude this study, there are some main takeaways to be briefly discussed. The standard cloud masks did not prove to be very useful for cloud masking in the glaciers terrain as they are not able to discriminate spectral signals of snow and clouds. Though, S2cloudless performs better in the summer month when the overall reflectance values are lower for all landcover types, but fails in winter. All the standard clouds failed to discriminate spectral signals of snow and cloud in the Green and SWIR wavelength feature space.

Notably, the Sentinel-2 spectral bands are capable of identifying spectral properties of snow, bare rocks and cloud. The clouds are visible in the SWIR wavelength and its band combinations with visible bands. In this case, the standardization of data is important for applying clustering methods. However, it also depends on the data type and purpose of clustering. The statistical clustering methods were able to detect >50% clouds for all the methods used in this study. Spectral clustering performed the best detecting (over 85%) of clouds in summer and winter.

The lack of reference images limited the validation of classification results. Moreover, the study is based on spectral signals from 4 glacial and 4 non-glacial areas. This was not able to capture the topographic and cloud shadows. The high reflectance values for non-glacial polygons is probably due to slope effects which makes it difficult to compare them to snow cover on relatively flat surfaces

The cloud masking over glacial snow cover is a real issue. Without cloud masks the glacier studies cannot be conducted. The current cloud masking methods are too complicated to implement in a small study area. To modify this method used in this study, a bigger study can be used. The feature space can be made from all the pixel values to account for intra-class variabilities. More spectral bands could be used, especially Cirrus (Band 10) of Sentinel-2 which detects high altitude clouds. Sentinel-2 lacks a thermal band which can easily identify cloud tops as they are colder compared to the surrounding landcover pixels in a satellite image. In the absence of referenced images, cloud cover detected from Landsat missions can be used for validation in further studies.

LIST OF REFERENCES

Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding, in: Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, New Orleans, Lousiana, pp. 1027–1035.

Azam, M.F., Wagnon, P., Vincent, C., Ramanathan, A.L., Favier, V., Mandal, A., Pottakkal, J.G., 2014. Processes governing the mass balance of Chhota Shigri Glacier (western Himalaya, India) assessed by point-scale surface energy balance measurements. The Cryosphere 8, 2195–2217. https://doi.org/10.5194/tc-8-2195-2014

Azam, M.F., Wagnon, P., Vincent, C., Ramanathan, A.L., Kumar, N., Srivastava, S., Pottakkal, J.G., Chevallier, P., 2019. Snow and ice melt contributions in a highly glacierized catchment of Chhota Shigri Glacier (India)over the last five decades. Journal of Hydrology 574, 760–773. https://doi.org/10.1016/j.jhydrol.2019.04.075

Biemans, H., Siderius, C., Lutz, A.F., Nepal, S., Ahmad, B., Hassan, T., von Bloh, W., Wijngaard, R.R., Wester, P., Shrestha, A.B., Immerzeel, W.W., 2019. Importance of snow and glacier meltwater for agriculture on the Indo-Gangetic Plain. Nature Sustainability 2, 594–601. https://doi.org/10.1038/s41893-019-0305-3

Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification, in: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR. https://doi.org/10.1109/CVPR.2008.4587598

Bookhagen, B., Burbank, D.W., 2010. Toward a complete Himalayan hydrological budget: Spatiotemporal distribution of snowmelt and rainfall and their impact on river discharge. Journal of Geophysical Research: Earth Surface 115, 1–25. https://doi.org/10.1029/2009JF001426

Boucher, O., D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao, U. Lohmann, P. Rasch, S. K. Satheesh, S. Sherwood, B. Stevens, X. Y. Zhang, 2013. Clouds and aerosols, in: Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley (Eds.), Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 571–658. https://doi.org/10.1017/CBO9781107415324.016

Coluzzi, R., Imbrenda, V., Lanfredi, M., Simoniello, T., 2018. A first assessment of the Sentinel-2 Level 1-C cloud mask product to support informed surface analyses. Remote Sensing of Environment 217, 426–443. https://doi.org/10.1016/j.rse.2018.08.009

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619. https://doi.org/10.1109/34.1000236

Cuffey, K.M., Paterson, W.S.B., 2010. The Physics of Glaciers, 4th ed. Elsevier, Inc.

Dong, C., 2018. Remote sensing, hydrological modeling and in situ observations in snow cover research: A review. Journal of Hydrology 561, 573–583. https://doi.org/10.1016/j.jhydrol.2018.04.027

Dozier, J., Painter, T.H., 2004. Multispectral and hyperspectral remote sensing of alpine snow properties. Annual Review of Earth and Planetary Sciences 32, 465–494. https://doi.org/10.1146/annurev.earth.32.101802.120404

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Frey, H., Paul, F., Strozzi, T., 2012. 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

Goodwin, N.R., Collett, L.J., Denham, R.J., Flood, N., Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: An automated method for Landsat TM/ETM + time series. Remote Sensing of Environment 134, 50–65. https://doi.org/10.1016/J.RSE.2013.02.019

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

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sensing of Environment 114, 1747–1755. https://doi.org/10.1016/J.RSE.2010.03.002

Hollstein, A., Segl, K., Guanter, L., Brell, M., Enesco, M., 2016. Ready-to-Use Methods for the Detection of Clouds, Cirrus, Snow, Shadow, Water and Clear Sky Pixels in Sentinel-2 MSI Images. Remote Sensing 8. https://doi.org/10.3390/RS8080666

Immerzeel, W.W., Beek, L.P.H. Van, Bierkens, M.F.P., 2010. Climate Change Will Affect the Asian Water Towers, American Association for the Advnacement of Science.

Kokhanovsky, A., Lamare, M., Danne, O., Brockmann, C., Dumont, M., Picard, G., Arnaud, L., Favier, V., Jourdain, B., Meur, E. le, di Mauro, B., Aoki, T., Niwano, M., Rozanov, V., Korkin, S., Kipfstuhl, S., Freitag, J., Hoerhold, M., Zuhr, A., Vladimirova, D., Faber, A.K., Steen-Larsen, H.C., Wahl, S., Andersen, J.K., Vandecrux, B., van As, D., Mankoff, K.D., Kern, M., Zege, E., Box, J.E., 2019. Retrieval of snow properties from the Sentinel-3 Ocean and Land Colour Instrument. Remote Sensing 11. https://doi.org/10.3390/rs11192280

Koutroumbas, K., Theodoridis, S., 2008. Pattern Recognition . Academic Press.

López-Puigdollers, D., Mateo-García, G., Gómez-Chova, L., 2021. Benchmarking deep learning models for cloud detection in landsat-8 and sentinel-2 images. Remote Sensing 13. https://doi.org/10.3390/rs13050992

Luo, Y., Trishchenko, A.P., Khlopenkov, K. v., 2008. Developing clear-sky, cloud and cloud shadow mask for producing clear-sky composites at 250-meter spatial resolution for the seven MODIS land bands over Canada and North America. Remote Sensing of Environment 112, 4167–4185. https://doi.org/10.1016/J.RSE.2008.06.010

Lutgens, F.K., Tarbuck, E.J., Tusa, D., 2010. The Atmosphere, 11th ed, Pearson. Pearson Education.

Lutz, A.F., Immerzeel, W.W., Shrestha, A.B., Bierkens, M.F.P., 2014. Consistent increase in High Asia's runoff due to increasing glacier melt and precipitation. Nature Climate Change 4, 587–592. https://doi.org/10.1038/nclimate2237

Lyapustin, A.I., Wang, Y., Frey, R., 2008. An automatic cloud mask algorithm based on time series of MODIS measurements. Journal of Geophysical Research Atmospheres 113. https://doi.org/10.1029/2007JD009641

Mahajan, S., Fataniya, B., 2020. Cloud detection methodologies: variants and development—a review. Complex & Intelligent Systems 6, 251–261. https://doi.org/10.1007/s40747-019-00128-0

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2Cor for Sentinel-2, in: Proceedings of SPIE. Warsaw, Poland. https://doi.org/10.1117/12.2278218

Pedregosa, F., Varoquaux, G., Gramfort, A., Michael, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournpeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Qiu, S., Zhu, Z., He, B., 2019. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. Remote Sensing of Environment 231. https://doi.org/10.1016/j.rse.2019.05.024

Richards, J.A., Jia, X., 2005. Remote Sensing Digital Image Analysis, 4th ed, Remote Sensing Digital Image Analysis. https://doi.org/10.1007/978-3-662-03978-6

Sarangi, C., Qian, Y., Rittger, K., Ruby Leung, L., Chand, D., Bormann, K.J., Painter, T.H., 2020. Dust dominates high-altitude snow darkening and melt over high-mountain Asia. Nature Climate Change 10, 1045–1051. https://doi.org/10.1038/s41558-020-00909-3

Schaepman-Strub, G., Schaepman, M.E., Painter, T.H., Dangel, S., Martonchik, J. v., 2006. Reflectance quantities in optical remote sensing—definitions and case studies. Remote Sensing of Environment 103, 27–42. https://doi.org/10.1016/J.RSE.2006.03.002

Scherler, D., Bookhagen, B., Strecker, M.R., 2011. Spatially variable response of Himalayan glaciers to climate change affected by debris cover. Nature Geoscience 23. https://doi.org/10.1038/NGEO1068

Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O., López-Puigdollers, D., Louis, J., Lubej, M., Mateo-García, G., Osman, J., Peressutti, D., Pflug, B., Puc, J., Richter, R., Roger, J.C., Scaramuzza, P., Vermote, E., Vesel, N., Zupanc, A., Žust, L., 2022. Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. Remote Sensing of Environment 274. https://doi.org/10.1016/j.rse.2022.112990

Vincent, C., Ramanathan, Al., Wagnon, P., Dobhal, D.P., Linda, A., Berthier, E., Sharma, P., Arnaud, Y., Azam, M.F., Jose, P.G., Gardelle, J., 2013. Balanced conditions or slight mass gain of glaciers in the Lahaul and Spiti region (northern India, Himalaya) during the nineties preceded recent mass loss. The Cryosphere 7, 569–582. https://doi.org/10.5194/tc-7-569-2013

Wagnon, P., Linda, A., Arnaud, Y., Kumar, R., Sharma, P., Vincent, C., Pottakkal, J.G., Berthier, E., Ramanathan, A., Hasnain, S.I., Chevallier, P., 2007. Four years of mass balance on Chhota Shigri Glacier, Himachal Pradesh, India, a new benchmark glacier in the western Himalaya. Journal of Glaciology 53, 603–611. https://doi.org/10.3189/002214307784409306

Wang, B., Ono, A., Muramatsu, K., Fujiwarattt, N., 1999. Automated detection and removal of clouds and their shadows from landsat TM images. IEICE Transactions on Information and Systems E82-D, 453–460.

Warren, S.G., 1982. Optical Properties of Snow. Reviews of Geophysics 20, 67–89. https://doi.org/10.1029/RG020i001p00067

Winsvold, S.H., Kääb, A., Nuth, C., 2016. Regional Glacier Mapping Using Optical Satellite Data Time Series. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, 3698–3711. https://doi.org/10.1109/JSTARS.2016.2527063

Wu, Q., 2020. geemap: A Python package for interactive mapping with Google Earth Engine. The Journal of Open Source Software 5. https://doi.org/10.3390/rs10050691

Yu, S.X., Shi, J., 2003. Multiclass spectral clustering, in: Proceedings of the IEEE International Conference on Computer Vision. Nice, France, pp. 313–319. https://doi.org/10.1109/iccv.2003.1238361

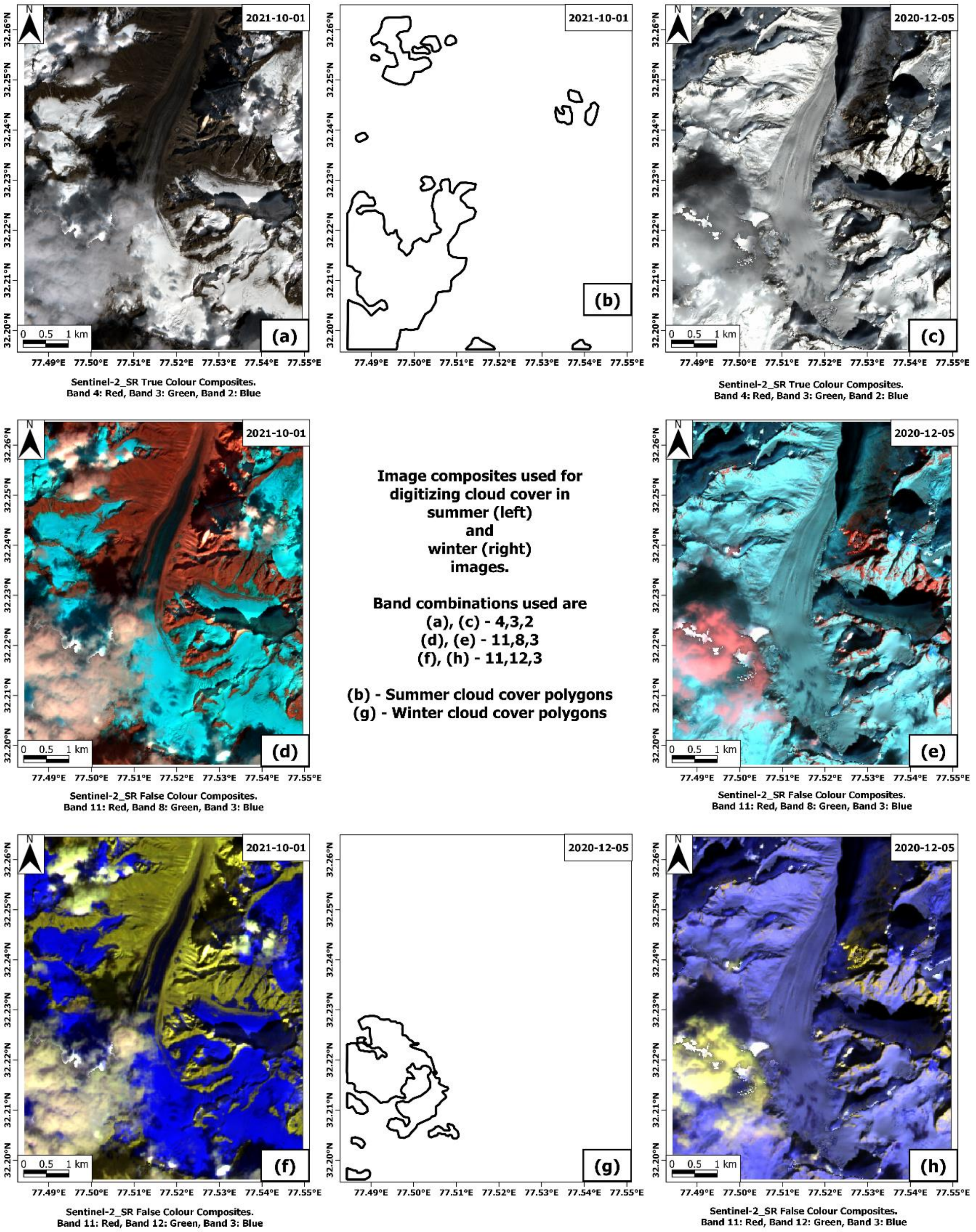
Zhu, X., Helmer, E.H., 2018. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. Remote Sensing of Environment 214, 135–153. https://doi.org/10.1016/j.rse.2018.05.024

Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. Remote Sensing of Environment 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014

Zhu, Z., Woodcock, C.E., 2014. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. Remote Sensing of Environment 152, 217–234. https://doi.org/10.1016/j.rse.2014.06.012

Zupanc, A., 2017. Improving Cloud Detection with Machine Learning [WWW Document]. URL https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13 (accessed 8.17.22).

APPENDIX A: CLOUD COVER DIGITIZATION



Sentinel-2_SR True Colour Composites.
Band 4: Red, Band 3: Green, Band 2: Blue

Sentinel-2_SR True Colour Composites.
Band 4: Red, Band 3: Green, Band 2: Blue

Image composites used for digitizing cloud cover in summer (left) and winter (right) images.

Band combinations used are
(a), (c) - 4,3,2
(d), (e) - 11,8,3
(f), (h) - 11,12,3

(b) - Summer cloud cover polygons
(g) - Winter cloud cover polygons

Sentinel-2_SR False Colour Composites.
Band 11: Red, Band 8: Green, Band 3: Blue

Sentinel-2_SR False Colour Composites.
Band 11: Red, Band 8: Green, Band 3: Blue

Sentinel-2_SR False Colour Composites.
Band 11: Red, Band 12: Green, Band 3: Blue

Sentinel-2_SR False Colour Composites.
Band 11: Red, Band 12: Green, Band 3: Blue

Figure 5.1 Band combinations used for digitizing cloud cover in summer and winter image.

The cloud cover area in summer is ~16% and ~7% in winter image.

APPENDIX B: REPRESENTATIVE POLYGONS

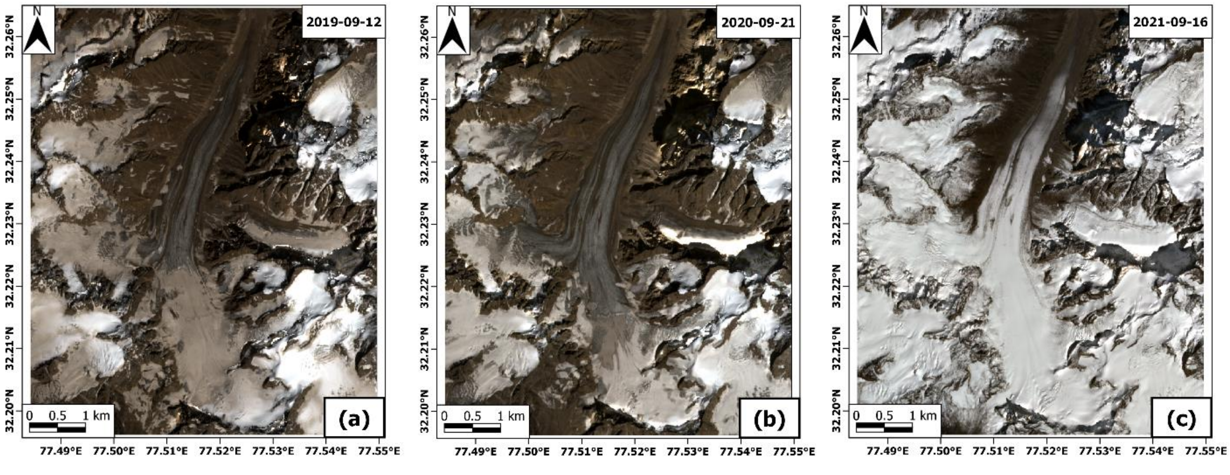
The time period of the image collection (Section 2.2) covers three periods for the month of September. The true-colour-composites of the cloud-free images closest to the end of September in shown in Figure 5.1. In the year 2021, the September month has only one cloud-free image. It shows complete snow cover over the glacial area (with no glacial ice) and more snow coverage for the month of September than in 2019 and 2020 (Figure 5.1(a, b)).

The permanent snow cover region (accumulation zone) of high elevation can be seen in the lower middle of the image. The glacial snow has a brownish tinge and becomes progressively white/bright moving up the accumulation zone for the year 2019 and 2020 (Figure 5.1(a, b)). It is completely white/bright for the year 2021 (Figure 5.1(c)).

The September image of 2020 shows the least snow cover and maximum exposure of the catchment rocks (Figure 5.1(b)). The image shows no presence of water bodies or any vegetated areas. The only landcover classes seen are bare rock/soil from catchment rocks and glacial moraines, glacial ice, and glacial and non-glacial snow. These inferences are presented in Table 3.2.

Table 5.1 The observed landcover classes in the study area.

Landcover	Surface type		Landcover extent
Snow	Glacial	Non-glacial	All year, maximum during winter
Ice	Glacial	-	Maximum during end of summer
Rocks	Glacial (Moraines)	Non-glacial	Maximum during end of summer



True Colour Composites. Band 4: Red, Band 3: Green, Band 2: Blue

Figure 5.2 Cloud free true-colour-composite images of the end-of-summer (September) month of the study area for the year (a) 2019, (b) 2020 and, (c) 2021.

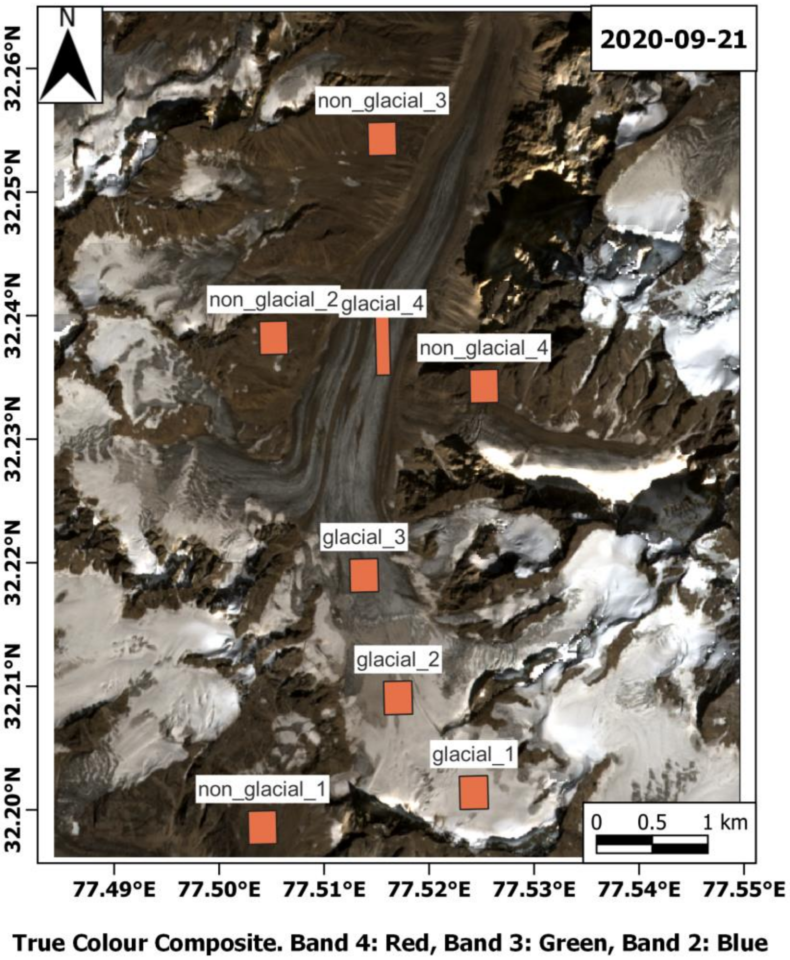


Figure 5.3 Cloud free true-colour-composite image of the end-of-summer (September) month for the year 2020. The figure shows the location and distribution of representative polygons in the study area.

The September 2020 image is used to select representative polygons for the above mentioned landcover classes in the study area. The four glacial and four non-glacial representative polygons cover the spatial extent of the study area (Figure 5.2). The non-glacial polygons are underlain by catchment rocks and surround the glacier. The glacial polygons follow the glacial central flow from South to North (accumulation to ablation zone). The glacial_1 polygon is underlain by permanent snow cover and 'glacial_4' is underlain by seasonal exposure of glacial ice. Polygons 'glacial_2' and 'glacial_3' show intermediate surface behaviour between 'glacial_1' and 'glacial_4'.

APPENDIX C: HOMOGENOUS AND NON-HOMOGENOUS REPRESENTATIVE POLYGONS

The mean values of Green and SWIR wavelength reflectance for representative polygons also contain values with large standard deviation. The polygons cover 100 pixels of approximately 10,000 sq. meters area. A large standard deviation in the mean reflectance value is indicative of the presence of different cloud and/or landcover classes within the polygon. This can be seen in the scatterplot of the standard deviation values in Figure 5.3, where the data points show a large spread in both axes and are densely clustered around low standard deviation values. The threshold of the standard deviation value for defining homogenous and non-homogenous polygons is difficult to make from the absolute standard deviation values.

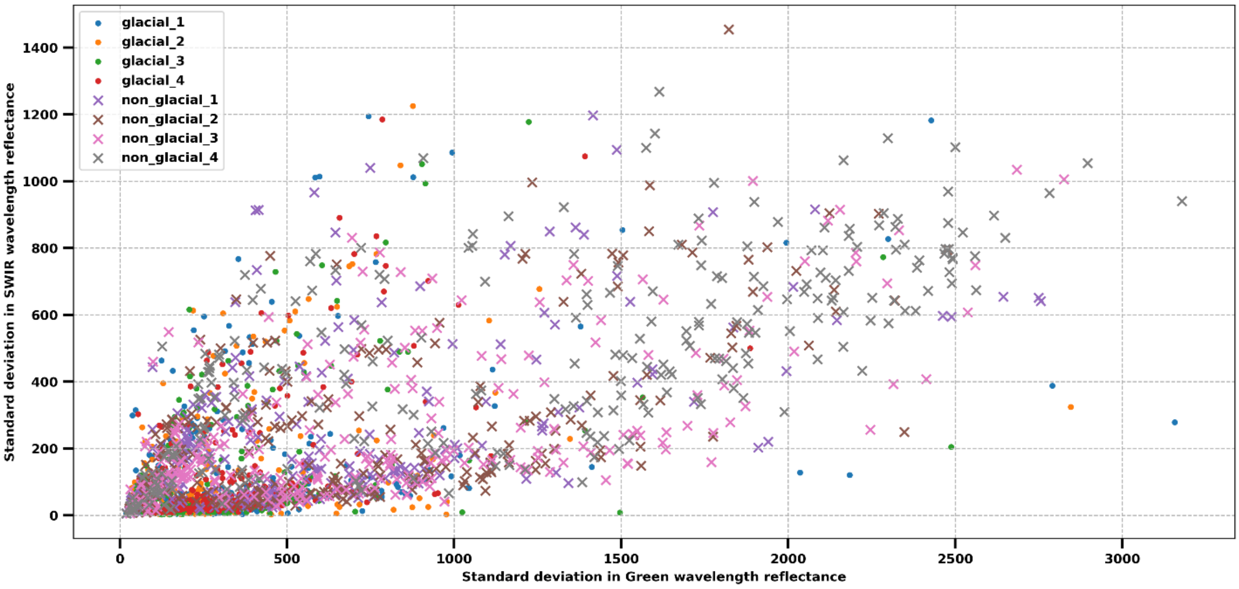


Figure 5.4 Scatter plot between the standard deviation in mean reflectance values of Green and SWIR wavelengths for the representative polygons.

Figure 5.4 shows the scatterplot of the normalized standard deviation (standard deviation of the standard deviation values) and mean reflectance values of Green and SWIR wavelengths for all the representative polygons, respectively. The red dashed lines indicate all the data points between -1 to +1 normalized standard deviation values. High normalized standard deviation values indicate the non-homogenous polygon values. These values are eliminated and the data points within the red dashed lines (-1 to +1) indicative of homogenous polygons are used for further clustering analysis.

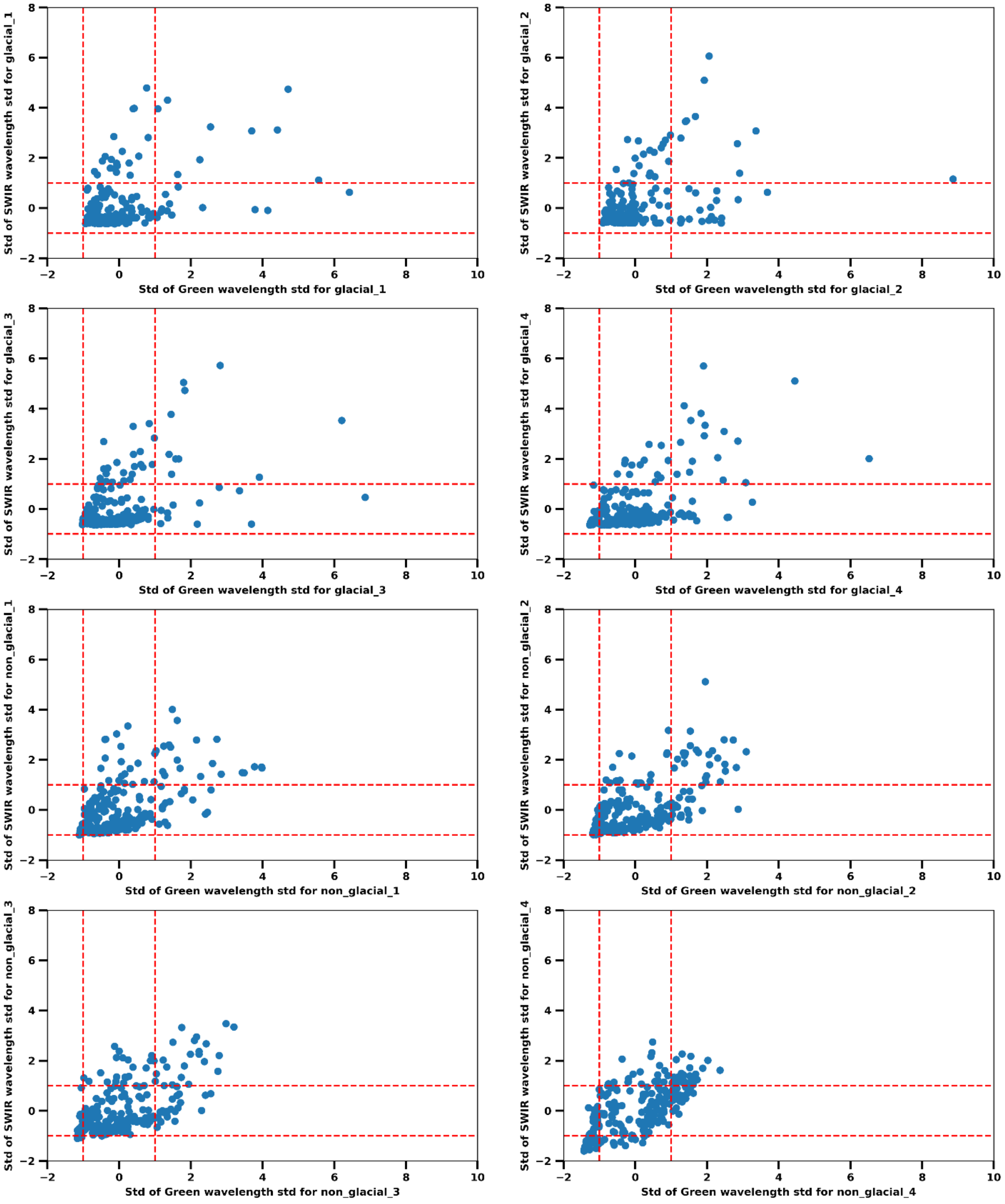


Figure 5.5 Scatter plot between the standard deviation of standard deviation values in the Green wavelength and the standard deviation of standard deviation values in the SWIR wavelength for respective representative polygons. The red dashed lines denote standard deviation values between -1 to +1.