University of
BRISTOL

DEPARTMENT OF ENGINEERING MATHEMATICS

Evaluation of the usefulness of Digital Elevation Models for Floodwater segmentation using Optical Imagery and Satellite Aperture Radar

Alex Beavil

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 4$^{\text{th}}$ December, 2023

Supervisor: Mohammad Golbabaee

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Alex Beavil, Monday 4$^{\text{th}}$ December, 2023

Contents

List of Figures

Abstract

Identifying land that is inundated during a flood event has many potential uses, from understanding the effects of climate change to training models to better predict flood inundation in the future. In this study we use the hand labelled portion of the Sen1Floods11 dataset to try and identify features that can be added to water segmentation methods that can improve their performance. We focus on the addition of elevation data in the form of features derived from a Digital Elevation Model as well as smaller combinations of composite optical features than tested by other authors. Following a methodology similar to other authors using the Sen1Floods11 dataset, we found similar results with total and mean IOU values of 0.8427 and 0.6246 respectively on a large feature space, compared to 0.8751 and 0.5873 from previous attempts. We note far higher performance on a region outside of the training set than other studies, with a far lower drop in performance between in-domain and out-of-domain image sets. Similarly, a feature space using only optical imagery was able to perform nearly as well as the mixed SAR-optical spaces used by other authors, with a total IOU score of 0.8319. Single source feature spaces have much potential as they are less dependent on coincident imagery from multiple satellites.

The research hypothesis for this project is that the addition of a digital elevation model (DEM) to the feature spaces used by Iselborn, et al. is likely to improve the classification performance on the Sen1Floods11 dataset. We were not able to confirm this with the DEM tested due to data quality issues although some derived features demonstrated modest gains of around 5%.

As part of this project, I:

- Wrote a new implementation of the model framework used by Iselborn, et al.

- Downloaded and co-registered a DEM with the Sen1Floods11 dataset

- Tested all feature spaces used by Iselborn, et al., finding similar results

- Experimented with extra feature spaces that combined DEM derived features with optical and SAR

The code and usage instructions are contained within a GitHub repository, which can be found at:

- https://github.com/aeroniemi/floodwater-classification

Supporting Technologies

In completing this project, a number of third party resources were used. These are detailed in the following subsections. Unless otherwise stated, these are python libraries.

0.0.1 Geographical

A number of libraries were used to import and process the Sen1Floods11 dataset other associated geotiff files, including:

- Rasterio

- GDAL

- GeoWombat

The QGIS GIS platform forms the core of the geographical analysis.

0.0.2 Digital Elevation Models

- GeoWombat was used to coregister the SRTM DEM with the Sen1Floods11 chips through its integration with AROSICS.

- Richdem was used to derive additional features from the DEM, including slope and water flow characteristics.

0.0.3 Models

- The majority of ML models used are sourced from scikit-learn.

- GBDT models from LightGBM

- The Optuna experiment framework ensures easy analysis and repeatability.

0.0.4 General

- TDQM, colored and alive-progress were used to monitor the progress of code.

- Scikit-image for analysing and rescaling images

- Visual Studio Code IDE and the Python extension set

- Latex, Overleaf and MarkText for formatting the thesis.

Notation and Acronyms

(SAR) VV / VH	: Polarisations used by SAR
ACU	: Water ACcUmulation feature
ASTER	: Advanced Spaceborne Thermal Emission and Reflection
	: Radiometer (a satellite / DEM)
AWEI	: Automatic Water Extraction Index
AWEISH	: Automatic Water Extraction Index (Shadowed)
cAWEI	: Combined feature space containing AWEI and AWEISH
cNDWI	: Combined feature space containing MNDWI and NDWI
DEM	: Digital Elevation Model
DP	: Binary feature identifying depressions
GBDT	: Gradient Boosted Decision Trees
HSV	: Hue, Saturation, Value (colour space)
IB	: Experiments performed by Iselborn, et al.
IOU	: Intersection-Over-Union
IR	: Infrared
JRC	: Joint Research Council, references a permanent water dataset
LDA	: Linear Discriminant Analysis
LDEM	: Locally scaled DEM
LiDAR	: Light Detection and Ranging
LX	: Experiments performed as part of this study
ML	: Machine Learning
MNDWI	: Modified Normalised Difference Water Index
NB	: Naïve Bayes
NDVI	: Normalised Difference Vegetation Index
NDWI	: Normalised Difference Water Index
NIR	: Near Infrared
O3	: Colour space containing
QDA	: Quadratic Discriminant Analysis
RGB	: Red, Green, Blue (colour space)
SAR	: Satellite Aperture Radar
SAVI	: Soil Adjusted Vegetation Index
SRTM	: Shuttle Radar Topography Mission (a DEM)
SWIR1/SWIR2	: Short Wave Infrared bands
WRI	: Water Resource Index

Acknowledgements

Chapter 1

Introduction

1.1 Textual

Flooding is the most frequent form of natural disaster, accounting for more than 40% of all events, and affecting more than 120 million people per year globally over the period 1993-2013 [7]. Storm surge and riverine flooding combined caused over 1.6 trillion USD in damages over the same period globally. Understanding areas at risk of flooding is therefore important to allow the governments, developers and residents to apply appropriate mitigation to reduce the impact of flooding. This will be essential in the future, as rising sea levels and increases in extreme weather events increase the frequency and severity of flood events. It thus stands true that predicting flooding has long been a goal of science. This roughly falls into two categories: determining the magnitude and likelihood of events that cause flooding, and identifying areas that are at risk of these floods. The latter task is the focus of this project. Numerical models are used in both tasks, and allow for a variety of different scenarios to assist the development of mitigation measures. This can provide an accurate assessment of the flood hazard - essential for planners and insurers, as well as to the design and operation of essential infrastructure such as roads, railways, electrical grids and hospitals, where the potential consequences of even a small amount of flooding are very high.

In order for an inundation classifier to be used globally, it is essential that it depends primarily on remote sensing data. Collecting data in the field is both challenging and expensive, especially during natural disasters when access may be very limited. A wide array of remote sensing methods exist, including aerial and satellite systems; the latter provides far better coverage on a regular basis, but may be lower resolution and more affected by atmospheric conditions than aerial systems, which are generally flown on-demand and can often travel below cloud layers. Optical and Synthetic Aperture Radar (SAR) are the two most commonly used technologies in this field [25, 28, 21, 29]. Optical imagery, collected by satellites such as ESA's Sentinel-2 collect visual images in a variety of bands, including visual RGB, near infrared and shortwave infrared. These bands can be used individually, or combined into composite layers such as AWEI [11] which have been developed to better highlight water. Optical imagery in all bands is affected by clouds and cloud shadows [5], which block the view of the ground. This is especially problematic for flood classification as floods tend to occur during storms, when cloud cover is high. SAR is an active technology that emits radar pulses that are reflected by different land covers in different amounts. Flat water tends to reflect more of the incident beam than rougher land, making inundated land stand out from the background. SAR is not affected by cloud cover, and can (often) penetrate vegetation cover due to its longer (3.8-7.8cm) wavelength [6]. It is however affected by speckle - random interference caused by the interaction of the radar beams - which can make images appear fuzzy. Methods that combine SAR and optical imagery require both images to have been taken as close in time as possible to ensure that the ground state that they are imaging has not changed too much between them [29]. This is especially important for floods as a difference of 12 hours may lead to the river being in a very different state (e.g. having breached its banks). Historically this has been challenging to achieve globally due to long satellite return times, although upcoming satellite constellations are likely to greatly reduce this. Given the demand for a global flood water classification method and the wide availability of freely accessible remote sensing data it is something of a surprise that limited progress has been made over the last decade. A number of studies have addressed the issue in part, usually focusing on a single flood event or location [18, 19, 30, 28] where they often achieve high metric scores, but few have attempted to create a global classifier. One reason for this is that there have not been any datasets that combine remote sensing

data with a hand labelled layer, meaning that each study has to create its own dataset. Hand labelling is resource intensive, and thus most authors choose to focus on just one flood event. Bonafilla, et al. [5] have tackled this task though, creating a global optical+SAR+labels dataset called Sen1Floods11. First published in 2020 and since updated to improve accessibility, they aimed to create a basis for expansion by future studies and a dataset to allow for the comparison of different approaches to flood classification, including traditional machine learning (ML), neural networks and other statistical methods.

One of the first studies to utilise this dataset is Iselborn, et al. [16]They compared a number of machine learning techniques, including discriminant analysis, gradient boosted decision trees (GBDT) and Gaussian naïve Bayes (NB). In particular they focused on the selection of features, which they identified to be "the single most important parameter to tune". Composite features such as cAWEI and cNDWI generated using multiple bands of optical imagery were especially important to this performance; models using these always outperformed raw image bands. Unsurprisingly, they found that models using a combination of SAR and Optical performed best, with Intersection-Over-Union (IOU) scores of up to 0.875 and F1 scores exceeding 0.9. Although they test SAR-only and optical-only feature sets, they receive little mention in the results and discussion despite the far higher potential for operationalisation of single-source methods.

This study builds on the work of Iselborn et al. [16] but aims to answer slightly different questions. We test the addition of another feature commonly used for water segmentation - elevation. "Live" satellite images are difficult to come by, but digital elevation models (DEMs) are both easy to find for most of the planet and relatively unchanging over time. It makes logical sense that an elevation model could be useful for identifying flooded land. For example, the peak of a mountain or strongly sloped land is far less likely to be inundated than low-lying coastal or river valleys. Much of Europe and the US has LiDAR DEMs available with resolution up to 25cm, with the data for many floodplains being updated every few years. Global data is available freely at 30m resolutions (SRTM), with higher resolutions hopefully becoming available in the coming years. Finding a way to use these to augment the expensive satellite imagery with a DEM could improve accuracy considerably. Iselborn, et al. [16] focus lots of attention on the metric scores which are impressively high, but this study seeks to take a more geographical approach. Metric performance is important, but it is equally important to ensure that the models provide outputs that are usable in the real world. We look at areas surrounding rivers and floodplains to provide a qualitative assessment of their usefulness for planning and calibrating river basin hydraulic models, and seek to address the question of whether these techniques would be operationally useful to flood forecasters.

1.2 Summary

- An accurate methodology for identifying inundated land that can be applied outside of the training domain has many potential uses, including live risk assessment during flood events and better long term climate and flood forecasting.

- This must use remote sensing to keep costs manageable, as collecting on site data is often expensive or impossible. A SAR-only approach would be ideal, although a mixed SAR+Optical method would also be fine.

- Iselborn, et al. [16] used a variety of traditional ML approaches to test which features would give the highest classification accuracy on the Sen1Floods11 dataset, which covers 11 flood events globally with SAR, optical imagery and hand labelled imagery. They found that a combined SAR+Optical composite feature space gave the highest metric scores.

- We were able to reproduce most of the findings of Iselborn, et al. [16] using a slightly different methodology, and identified other optical-only feature spaces that performed nearly as well as SAR+Optical.

- We also tested a variety of features derived from a coregistered DEM, but did not find any of these features provided a significant improvement in model performance.

- A number of issues with the Sen1Floods11 dataset that may affect performance are highlighted, including regional imbalances, time gaps between SAR and optical images and issues with the hand labelled layer.

1.3 Aims and Objectives

This study aims to assess the performance of traditional machine learning techniques at identifying inundated land from multispectral optical and radar imagery from satellites, and the effect, if any, of including a digital elevation model. In order to achieve this, it will:

- Attempt to replicate the results of Iselborn, et al. [16] using the Sen1Floods11 dataset.

- Co-register a DEM with the Sen1Floods11 dataset, and train the models used by Iselborn, et al. [16] on this expanded source.

- Test a wide variety of feature spaces, and compare their performance to those favoured by Iselborn, et al. [16].

- Create an easily reproducible method of running Iselborn, et al.'s [16] models that can be published publicly.

- Evaluate the usefulness of Sen1Floods11 for various tasks including flood model calibration.

Chapter 2

Background

2.1 Remote Sensing

The original method for identifying flooded land was to send human observers to the flooded area with a map and pencil. This is obviously not ideal for many reasons:

- Regions where flooding is occurring are often dangerous and difficult to access

- Sending human observers is expensive

- A single human can only survey a limited area at any one time

- It is often difficult to identify flooded land from ground level.

In more recent times aircraft have been used - these can take photos which can be classified by a human observer at a later date. Imaging capacity still remains an issue though as one aircraft can only survey a certain amount of land, and their services may be in high demand for other purposes (e.g. evacuations).

Remote sending is the practice of measuring attributes of a place from a long distance away. In general, this refers to space based satellite solutions, although this is not always the case (e.g. aircraft based LiDAR mapping for DEMs [1]). A number of different satellite platforms have been used for flood classification in the past, usually using SAR or Optical imaging.

2.1.1 Satellite Aperture Radar (SAR)

SAR is a active imaging technology that generates its own electromagnetic waves, and records the response. It operates in the microwave range, with wavelengths between 0.8-100cm [8]. Longer wavelengths require larger antenna to image at a given resolution, meaning that they can become impractically large very quickly. The 'synthetic' component of SAR is a work around to this issue, using the movement of the satellite (or aircraft) to artificially increase the size of the antenna [6]. As a result, resolution becomes less dependent on the size of the antenna, and more dependent on the frequency that the satellite can record the responses at. Due to the long wavelengths SAR can penetrate clouds and vegetation, giving an image of the bare ground. This is ideal for identifying flooded land, and gives the possibility to identify water in vegetated areas, as well as when cloud cover is high. Historically, access to SAR imagery has been limited as the technology was viewed as having military uses, although missions such as Sentinel-1 have made the technlogy more accessible to the public in recent years.

2.1.2 Optical Imagery

Imaging using visible light was one of the first uses for satellites in the earth observation space. Cameras mounted on satellites can image the earth in the way one would expect, and are the source of much of the imagery that people think of as "satellite imagery". However, the technology is not only limited to visible light - sensors can be setup to record any wavelength that is of interest. For water segmentation, the infrared space provides far better results than visible light, as it has a more distinctive signature. All wavelengths of optical imagery experience issues due to cloud cover as they cannot image through clouds. A wide array of optical imagery sources are available; some of the most commonly used in academia include Sentinel-2 and LANDSAT.

(a) Optical (True colour RGB) (b) Labels (Blue = Wet; White = Dry; Black = unclassified) (c) SAR (synthetic colour where VV is Red and VH is Green)
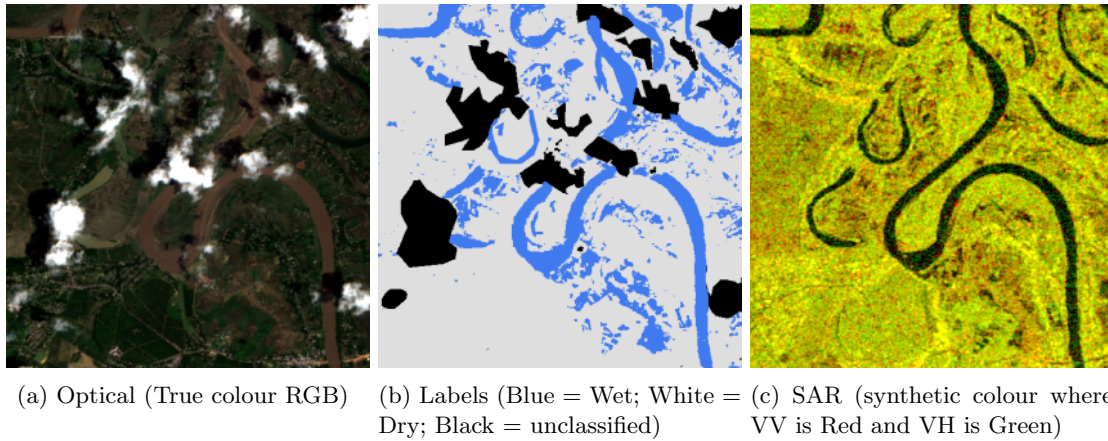
Figure 2.1: Example chip from Sen1Floods11, showing the three types of data included. Image is from Paraguay. Cloud cover is clearly visible in (a); this area is masked in (b)

2.2 Sen1Floods 11

Many different techniques for extracting flood extent maps from remote sensing data have been developed over the past 30 years using a variety of data sources. Most studies [18, 19, 30, 28, 13] tend to focus on one region, and one or more repeated flood events in this region. The reasons for this are obvious - it is much easier to acquire high quality data for one location than it is to produce comparable layers for many different locations in different environments. Techniques that use both optical and SAR imagery are particularly reliant on finding areas where both images are available at the same moment in time (else each image would show a different phase of the flood). Bonafilla, et al. [5] developed the Sen1Floods11 dataset to allow researchers to compare the results of different modelling approaches across different biomes. It contains 446 hand labelled 512*512 pixel chips that have both optical and SAR imagery, sourced from the Sentinel-2 and Sentinel-1 satellites respectively An example of each image is shown in figure 2.1, with visible RGB, hand labels and a SAR image for the same chip. These were split into four subsets by the original authors using a random sampling approach, with all Bolivia images being allocated to a separate out-of-domain set to test generalisation performance:

- Training - 250 images (60%)

- Validation - 86 images (20%)

- Test - 86 images (20%)

- Bolivia - 14 images

A small number of images (10) were removed from the dataset as they either contained no classified pixels, significant numbers of NAN pixels or other issues that prevented the automatic testing system working correctly. A full list of each image used in each split can be found in the GitHub repository.

2.3 Labels and masking

The Sen1Floods11 hand labels layer contains images classified with the following values:

- 1 - Wet

- 0 - Dry

- -1 - Unknown

The unknown category contains pixels that are covered by clouds, do not have imagery for both SAR and Optical (e.g. at the edge of a coverage pass) or were not able to be classified by the trained remote sensing analyst. These pixels were masked from the dataset for training along with any pixels that displayed NAN in any band. The various features used during this study have magnitudes varying between 0 and 20 000 with distributions equally as variable. Some models used, such as those trained using Stochastic

Gradient Descent (SGD) benefit from all features having the same range. In order to achieve this, all features not marked as unscaled were standardised by removing the mean and scaling to unit variance using a sklearn StandardScaler trained on the masked training set images. Some features are min-max scaled or scaled on a per-image basis; details are described in their feature description below.
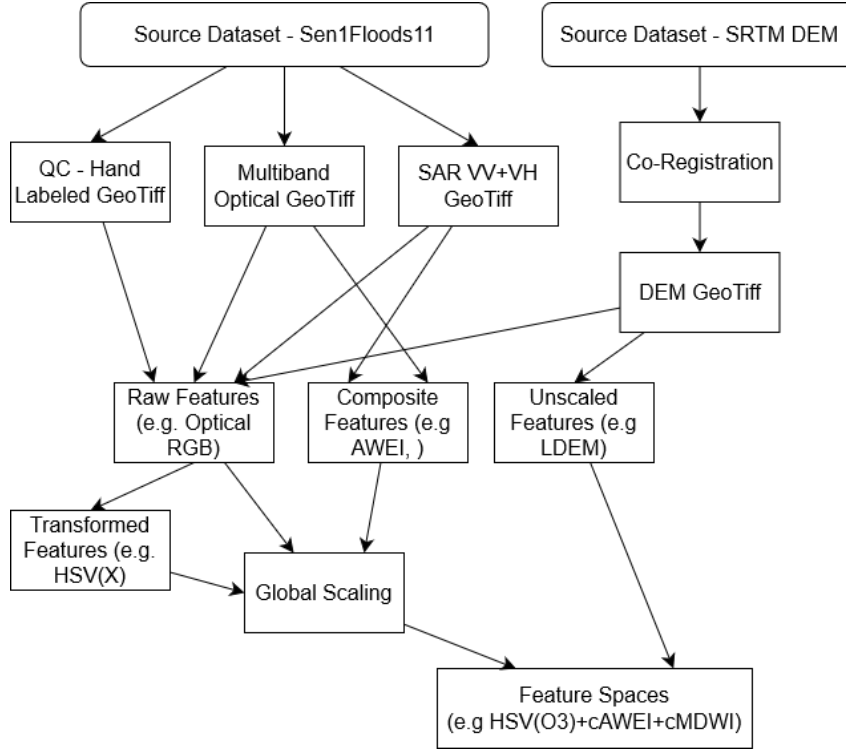
2.4   Feature spaces

Figure 2.2: Flow chart showing the steps involved in generating feature spaces

Figure 2.2 shows the process used to generate individual features and feature spaces. Each step is discussed in more detail in the following sections. When training ML models, the selection of features is important. Groups of multiple features are combined together to create feature spaces. Following the schema used by Iselborn, et al. [16] feature spaces are named using a string of format DEM_SAR_OPTICAL with individual features of each type being denoted with a +. Transformations applied to feature spaces are denoted as TRANSFORMATION(SPACE). For example, LDEM_SAR_HSV(O3)+cNDWI denotes a space with the following features:

- LDEM - Locally scaled DEM

- SAR - SAR VV+VH

- HSV(O3)+cNDWI - HSV transformed O3 space with cNDWI

Where possible feature spaces used in this study match those used in Iselborn, et al., allowing for easy comparison.

2.5   Features

2.5.1   Satellite Aperture Radar (SAR)

SAR data used in Sen1Floods11 dataset is derived from Sentinel-1. Launched by the European Space Agency in the early 2010s, Sentinel-1 is a two satellite constellation with C and S band SAR antenna operating in near polar sun synchronous orbits [12]. This orbit means that each satellite returns to a specific point once every 12 days, giving global coverage with a return period under 6 days [9]. Sentinel-1

Band	Wavelength	Resolution
B1 - Coastal	443nm	60m
B2 - Blue	490nm	10m
B3 - Green	560nm	10m
B4 - Red	665nm	10m
B5 - RedEdge-1	705nm	20m
B6 - RedEdge2	740nm	20m
B7 - RedEdge3	783nm	20m
B8 - Near Infared (NIR)	842nm	10m
B8A - Narrow Near Infared (NNIR)	865nm	20m
B9 - Water Vapour	945nm	60m
B10 - Cirrus	1375nm	60m
B11 - SWIR1	1610nm	20m
B12 - SWIR2	2190nm	20m

Table 2.1: Optical imagery bands found in Sentinel-2 imagery in the Sen1Floods11 Dataset

SAR imagery contains two bands which correspond to the two polarisations used (VV and VH). The following features are derived:

- SARVH - VH band only

- SARVV - VV band only

- SAR - VV + VH bands

SAR imagery is noisy, and particularly affected by speckle. This is a often severe form of noise which occurs when the radar returns from different cells interfere destructively and constructively, resulting in semi-random brightening and darkening of pixels, and can make analysis difficult [21]. A number of filtering techniques are used to reduce the impact of speckle on the output imagery of SAR; some of the most common include mean/median and lee filters. Both of these methods were tested, but did not provide significantly better results, likely due to the use of pixel wise classifiers.

Optical imagery used in the Sen1Floods11 dataset is derived from Sentinel-2 imagery. Sentinel-2 optical imagery contains 13 bands each described in table 2.1. Bands 1, 5-7, 9 and 10 are not used for this study but are primarily used for cloud and vegetation detection. Bands 2-5 form the visible RGB component; bands 8, 8A, 11 and 12 are different wavelengths of infrared light with increasing wavelength used for identifying clouds and water vapour in the atmosphere.

2.5.2 Colour spaces

RGB

This space forms the visible light true colour component, in the RGB colour space. Intensity values for Red, Green and Blue form separate bands, with no separation between colour and brightness.

O3

O3 is a false RGB colour set that is formed from SWIR2, Near Infrared and Red. Some studies [26, 16] have suggested that this colour space performs well for water identification, especially when transformed to HSV.

HSV(X)

Pekel, et al. [26] suggested that it may be useful to transform RGB and O3 to follow the HSV colour model. This decouples chromaticity - the shade of colour - from brightness. Changing observation conditions tend to affect the brightness before the shade of colour, which makes the H and S components of HSV more stable than any individual band in the RGB/O3 colour spaces. Iselborn, et al. tested HSV transformations on both RGB and O3, finding that they improved performance; HSV(O3) outperformed HSV(RGB) in all cases. The skimage rgb2hsv function is used for this transformation.

2.5.3 Colour indices

Different surfaces are more reflective to different wavelengths of light. For example, yellow objects tend to appear vibrant under blue light, making them stand out well from the background. A similar approach can be taken with satellite imagery to highlight features of interest. The following sections note some composite indices designed to identify water and vegetation.

Normalised Difference Water Index (NDWI)/ Modified Normalised Difference Water Index (MNDWI)

McFeeters [22] proposed a Normalised Difference Water Index (NDWI) as a way of more easily identifying water from early Landsat imagery, which was one of the first optical satellites to include near infrared bands. Near infrared (NIR) is strongly absorbed by water but reflected well by vegetation, and allows for a better distinction between the two land cover types. NDWI is defined as:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR}$$

and gives a value in the range -1 to +1, with water tending towards positive values while bare earth and vegetation give zero or negative values. Xu [32] noted that whilst generally effective in rural areas, NDWI tended to highlight built up areas as wet due to noise and surface types. They proposed a modified index (Modified NDWI, MNDWI):

$$MNDWI = \frac{GREEN - SWIR1}{GREEN + SWIR1}$$

which uses a middle infrared band (corresponding to SWIR1 in Sentinel-2 imagery). Whilst the same noise issues also exist in this band, reflectivity tends to be lower which gives higher contrast values, thus rendering the noise a smaller issue. NDWI and MNDWI are combined together by Iselborn, et al. to form a two band feature, cNDWI.

Automatic Water Extraction Index (AWEI)

In order to address some of the issues identified with NDWI in urban areas, various other authors have proposed alternative indices. Using imagery from the same satellite constellation, Feyisa, et al. [11] developed the Automatic Water Extraction Index (AWEI), specifically to better classify shadows and low albedo surfaces such as asphalt. There are two version of this index:

$$AWEI = 4 * (GREEN - SWIR1) - \frac{1}{4} * (NIR + 11 * SWIR2)$$

$$AWEISH = BLUE + \frac{5}{2} * GREEN - \frac{3}{2} * (NIR + SWIR1) - \frac{SWIR2}{4}$$

AWEI is designed for use in scenarios where shadows are not an issue, whilst AWEISH is intended for shadowed scenarios such as where tall mountains or buildings are present. AWEI and AWEISH are combined together by Iselborn, et al. [16] to form a two band feature, cAWEI.

Water Resource Index (WRI)

The Water Ratio Index is another method that "considers the dominant spectral reflectance of water"[19]. It is defined as:

$$WRI = \frac{GREEN + RED}{NIR + SWIR1}$$

The WRI value for water bodies tends to be above 1, whilst wetlands tend to cover the range of 0.7-1.2. This index was not used by Iselborn, et al. [16] but showed promising results during testing, and so is included for completeness.

Soil Adjusted Vegetation Index (SAVI)

In many cases, flooded vegetation is the most difficult land to correctly classify as wet or dry [19]. Vegetation is often taller than the water level meaning that it can change the absorbance spectrum significantly. This is especially true with SAR, which can generally see through vegetation due to its longer wavelength but is sensitive to noise introduced by interactions with taller vegetation. Thus, it makes sense that the inclusion of an index to identify vegetation may be useful to the overall performance of the models. Another potential benefit is that a vegetation+SAR feature space is far less dependent on the two images being taken at the same time, as overall vegetation levels tend to vary on longer timescales. The Soil Adjusted Vegetation Index is a commonly used method that highlights high-chlorophyll areas whilst accounting for changes in soil brightness associated with moisture variations [14]. The latter is important when moisture levels are likely to change a lot - such as during a flood event - as wet soil tends to be darker than dry soil. It is defined as:

$$SAVI = \frac{NIR - RED}{NIR + RED + 0.5 * (1 + L)}$$

Where L is a variable between 0 and 1 that should correspond to the inverse of expected vegetation densities. A value of 0.5 has been suggested as a good balance for most scenarios, and is used here.

2.5.4   Digital Elevation Model (DEM)

A digital elevation model is a "representation of the bare earth topographic surface" [31] in a computer readable format. It is used to add terrain data, both for description and modelling purposes. A wide array of sources are available that cover countries or regions, but very few global models exist due to the cost associated with their development. For this study, SRTMv3 DEM was selected. The Shuttle Radar Topography Mission (SRTM) launched in 2000, and used C-band and X-band SAR to create a global DEM at 90m resolution [24]. This has long been the "default" DEM selected by researchers due to its global coverage, acceptable resolution and nature as a freely available product maintained by NASA. SRTM v3, released in 2011, combines the raw SRTM DEM with data from ASTER v1 and a variety of other localised datasets to fill gaps and give an overall resolution of 30m over land between 60°N/30°S [10]. In particular, a void-filling approach was used to reduce errors and improve accuracy over steep terrain, where SRTM struggled due to the slow capture rate and high incident angle of the orbit. ASTER GDEM v3, released in 2019, is both more recent and has higher recorded resolution (30m) but the web portal/api was unavailable due to maintenance when selecting my data. Unlike SRTM, ASTER uses an optical sensor to take multi spectral images, which are then processed to create a DEM [24]. One particular flaw with this approach is that it cannot penetrate vegetation and trees, and requires extensive post-processing to remove these features and create a "ground" elevation model. As such, it is likely to perform worse for the inundation classification in this study, where ground level is important. The SRTM v3 DEM was downloaded from the NASA repository [23] in 1x1 degree files. These were then converted to GeoTIFF and co-registered with the Sen1Floods11 chips using the AROSICS [27] co-register function in Python. Where a chip spanned two files, these were joined together using GeoWombat before co-registration. The inclusion of elevation data into the water classifiers is based on the concept that some areas are inherently more prone to flooding than others due to factors that can be determined before the event. This is known as flood susceptibility mapping. Avand, et al. [2] reviewed a number of flood influencing factors that can be derived from a DEM; these features are discussed in the following subsections.

Altitude (DEM and LDEM)

Avand, et al. identified altitude to be one of the most critical factors to flood susceptibility. As altitude decreases, the amount of water runoff possible increases as there is more land above that point. Surface runoff is a major way in which water makes its way to rivers (or congregates into pools). The raw elevation data in metres is included as the feature DEM.

Across the range of locations in the hand labelled Sen1Floods11 dataset, elevation varies between 3m and 3000m, but these elevations are not evenly distributed. The elevation range in each region is far smaller. Through experimentation, it was found that applying a min-max scaler (sklearn MinMaxScaler) to each image individually improved performance somewhat. This feature is named LDEM.

Slope

Another factor identified by Avand, et al. is slope. Slope affects infiltration rates; higher slopes tend to lead to more runoff. Additionally, water pooling is less likely on sloped terrain than flat plains. The Python RichDem library was used to derive a slope layer from the the unscaled DEM; the results were then scaled globally using the standard technique.

Distance to river

Terrain closest to a river or lake is the first to become inundated during a flood event. Global river layers do exist, but they experience the same issues as regular water classifiers as they use the same remote sensing techniques. An alternative approach to this are flow accumulation algorithms. These use the DEM to count the number of cells that flow into an individual cell (generally the number of cells higher than each point). This works well for areas where elevation changes, but is less effective over wide rivers and lakes due to the flat surface. Flat area can be filled using a depression filling technique, which ensures that every point away from the edge of the image flows into at least one other point. The RichDem D8 method with epsilon depression filling was used here, and the resultant feature is known as ACU.

Depressions

DEMs often contain sinks - areas where there is no clear water output. In the real world, drainage in these areas tends to happen through small surface processes, infiltration or evaporation [20]. Often, these sinks do not actually exist in the real world - they are already water bodies or artifacts of the way in which the DEM was produced. When using a DEM for hydrology it is common to fill these sinks using a depression filling algorithm, to ensure that drainage is always possible. As they identify flat areas, these algorithms may also be useful for identifying permanent water. The Epsilon algorithm [3] fills depressed cells to a elevation slightly higher than the next closes pixel to the nearest outlet. As such, it means that every pixel in the depression can drain linearly to one of its neighbours. For the purposes of this study we are not interested in using the model to calculate flow. Instead, we are interested in the areas that are identified as depressions.

The RichDem epsilon depression fill method [4] is used to create a filled DEM. A binary layer is created by the difference between the original and filled DEM, marking any cells where the elevation has changed as filled as 1 and others as 0. The resultant feature is known as DP.

2.6 Evaluation Methodology

Many different metrics exist for assessing the performance of classification models. The Sen1Floods11 dataset shows a large class imbalance, with less than 10% of cells being classified as wet, and some images having no water at all. As such, it is important to choose metrics that can account for this imbalance. A wide array of metrics are reported in the results section for completeness, although IOU is the primary method selected for hyperparameter optimization. Intersection-Over-Union (IOU)

$$IOU = \frac{TP}{TP + FP + FN}$$

Both Bonafilla, et al. and Iselborn, et al. use Intersection Over Union as their main metric for use on the Sen1Floods11 dataset. IOU "describes the percentage of the area that is correctly identified as inundated (the intersection) out of the area that is actually water or has incorrectly been recognized as water (the union)" (cite). Thus, it is less sensitive to the class imbalance shown between wet and dry land in the training set. This metric can be calculated on a per-image and full dataset scale. The former is more sensitive to performance on individual images, in particular those that show very little water.

2.6.1 Models

Model hyperparameters were tuned using Optuna, following a random search pattern (RandomSampler). Hyperparameters to tune were chosen to closely match those selected by Iselborn, et al.; the specific hyperparameters tuned for each model are detailed in their model description below.

Bayesian Classifiers (naïve Bayes, LDA & QDA)

Bayesian Classifiers are simple probabilistic classifiers that apply Bayes' theorem to estimate the probability of class membership. All three models used here assume class members follow a Gaussian distribution. The difference between the models is in the assumptions that are made. LDA forms linear boundaries between classes in the search space, whilst QDA forms quadratic boundaries. naïve Bayes is a subset of these models, with the added assumption that individual features are independent of each other.

- naïve Bayes models have no hyperparameters.

- LDA:

  - Shrinkage: $[0, 0.1, 0.2, 0.3...0.9, 1.0]$

- QDA:

  - Covariance Regularisation: $[0.0, 0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1]$

This is a smaller range than tested by Iselborn, et al. [16] who also included $[2, 4, 8, 10]$. The model used by both parties (sklearn QuadraticDiscriminantAnalysis) only support values in the range 0-1, meaning that it is not possible to use these extra values.

Gradient Boosted Decision Trees (GBDT)

Decision trees are one of the simplest yet most effective machine learning algorithms. The LightGBM [17] implementation is used here, which creates trees that grow leaf wise. This reduces overall tree size by only splitting leaves with high numbers of samples. Rather than using each sample to create the trees, data is bucketed into bins using a histogram of its distribution. This is significantly quicker and more efficient than using individual bins, allowing high numbers of trees and iterations to be calculated.

LightGBM uses Gradient-based One-Side Sampling (GOSS) to sample the dataset, weighing points where the gradient is highest the most. This ensures that points that are far from the current model are weighted higher and included in the binned set.

2.7 On The Importance of Feature Representation for Flood Mapping using Classical Machine Learning Approaches

It is often assumed that neural network approaches will outperform traditional ML algorithms in every task, but currently this does not appear to be true for water segmentation. Iselborn, et al. tested a variety of traditional ML models on the Sen1Floods11 dataset, focusing on hyperparameter tuning and feature selection. Their best model outperformed a similar NN study on the same dataset by more than 19% across multiple metrics. In their paper, Iselborn, et al. [16] highlighted the importance and usefulness of flood inundation maps for many of the same uses as described above. They noted the promising results that remote sensing data has given in other similar fields, notably land use classification, as well as the issues that have been highlighted by other researchers associated with automatic label generation. The paper explores the potential usefulness of traditional ML algorithms trained on the small high quality hand labelled segment of Sen1Floods11. As part of their study the following models were tested:

- Gaussian naïve Bayes

- Linear Discriminant Analysis

- Quadratic Discriminant Analysis

- Gradient Boosted Decision Trees (GBDT)

- Linear SGD

These models were tested on a selection of 23 feature spaces incorporating SAR-only, optical-only and mixed groups of the following features:

- SAR (VV+VH)

- RGB

- NIR

- cNDWI (NDWI+MNDWI)

- cAWEI (AWEI+AWEISH)

- O3 (SWIR2+NIR+Optical Red)

- HSV transformed O3 and RGB

A wide array of metrics are calculated and provided, including mean and total IOU, F1 scores, accuracy, recall and precision. The selection of features was found to be the single most important parameter to tune, with a combination of SAR, HSV(O3), cAWEI and cNDWI giving the highest performance across multiple metrics and models. Combinations of raw satellite imagery bands were found to provide relatively poor results compared to the inclusion of transformed and composite features. Perhaps most importantly, they note that none of the SAR or Optical-only feature space/model combinations give satisfactory performance for the use cases previously identified, although combinations of both do approach this level. The metric scores noted are impressive, and are on par with or exceed other studies, including those using neural networks. When searching for a basis for this project, Iselborn, et al. [16] stood out as a good place to start. A major contributing factor to its selection was the inclusion of a GitHub repository [15] which contains the code used in their study, as well as instructions on how to interpret and use the dataset. Unfortunately, whilst extensive, the code contained within their repository does not function in the way that it is described to, and it was not possible to easily reproduce the findings of their study. As a result, a larger part of this project focused on reproducing their findings using a new methodology before extending their work to include a co-registered digital elevation model, and testing other models and feature spaces that appear to be overlooked in the original study.

Chapter 3

Quantitative Analysis

| Split	Run	Model	Feature Space	Mean IoU Flooded (std)	Total IoU Flooded	Mean Accuracy (std)	Total Accuracy	Total Precision Flooded	Total Recall Flooded	Total Recall Dry	Total F₁ Score
Test	LX	GBDT	SAR_HSV(O3)+cAWEI+cNDWI	0.6246 (±0.3105)	**0.8427**	**0.9750 (±0.0452)**	0.9782	0.8991	0.9307	0.9850	0.9147
	LX	Naïve Bayes	SAR_HSV(O3)+cAWEI+cNDWI	0.5675 (±0.3015)	0.7738	0.9610 (±0.0574)	0.9652	0.8074	**0.9490**	0.9675	0.8725
	IB	GBDT	SAR_HSV(O3)+cAWEI+cNDWI	**0.7031 (±0.2984)**	0.8751	0.9718 (±0.1176)	**0.9838**	**0.9577**	0.9103	**0.9943**	**0.9334**
	IB	Naïve Bayes	SAR_HSV(O3)+cAWEI+cNDWI	0.5187 (±0.3168)	0.7399	0.9522 (±0.0860)	0.9586	0.7784	0.9378	0.9616	0.8503
Bolivia	LX	GBDT	SAR_HSV(O3)+cAWEI+cNDWI	0.5686 (±0.3379)	0.8287	0.9682 (±0.0298)	0.9706	0.9254	0.8880	0.9864	0.9063
	LX	Naïve Bayes	SAR_HSV(O3)+cAWEI+cNDWI	0.5131 (±0.3169)	0.7447	0.9434 (±0.0562)	0.9482	0.7788	**0.9445**	0.9489	0.8537
	IB	GBDT	SAR_HSV(O3)+cAWEI+cNDWI	**0.6070 (±0.3542)**	**0.8357**	**0.9705 (±0.0293)**	**0.9730**	**0.9684**	0.8671	**0.9929**	**0.9105**
	IB	Naïve Bayes	SAR_HSV(O3)+cAWEI+cNDWI	0.4402 (±0.2889)	0.6160	0.8984 (±0.0776)	0.9033	0.6268	0.9728	0.8902	0.7620

Table 3.1: Comparison of our findings (marked as LX) with the headline results of Iselborn, et al. [16, Table 3] (marked as IB). Highest scores for each metric on each data split are shown in bold. Reported standard deviation refers to the variation across images in that set, rather than between runs.

A comparison of our attempts can be seen in table 3.1 where best scores for each split are highlighted in bold. In the following sections, results recorded by Iselborn, et al. [16] are referenced as IB, while results from our methods are referenced as LX.

- In most cases it was not possible to replicate the very high metric scores recorded by IB, particularly with the GBDT model. For example, their GBDT test set mean IOU value of 0.7031±0.2984 was nearly 10% higher than our best attempt value of 0.6246±0.3105.

- A number of our metric scores are higher than or comparable to IB. For example, the mean accuracy recorded by our GBDT method is within 2% of their method with similar or lower standard deviation. The performance gap is higher on the test set than Bolivia, although the gap is still significant.

- Our naïve Bayes method outperformed IB by a wide margin, with both splits showing a 0.04 point improvement in mean IOU and a nearly 0.013 point increase in Bolivia total IOU. This is unexpectedly high, and brings our Bolivia performance on-par with IB's test set naïve Bayes results. naïve Bayes models do not have any hyperparameters, suggesting that differences in pre-processing may affect the results here.

- Both IB and TS show higher performance across most metrics on the test split than the out-of-domain Bolivia data with the exception of total precision and recall. The difference is often small though.
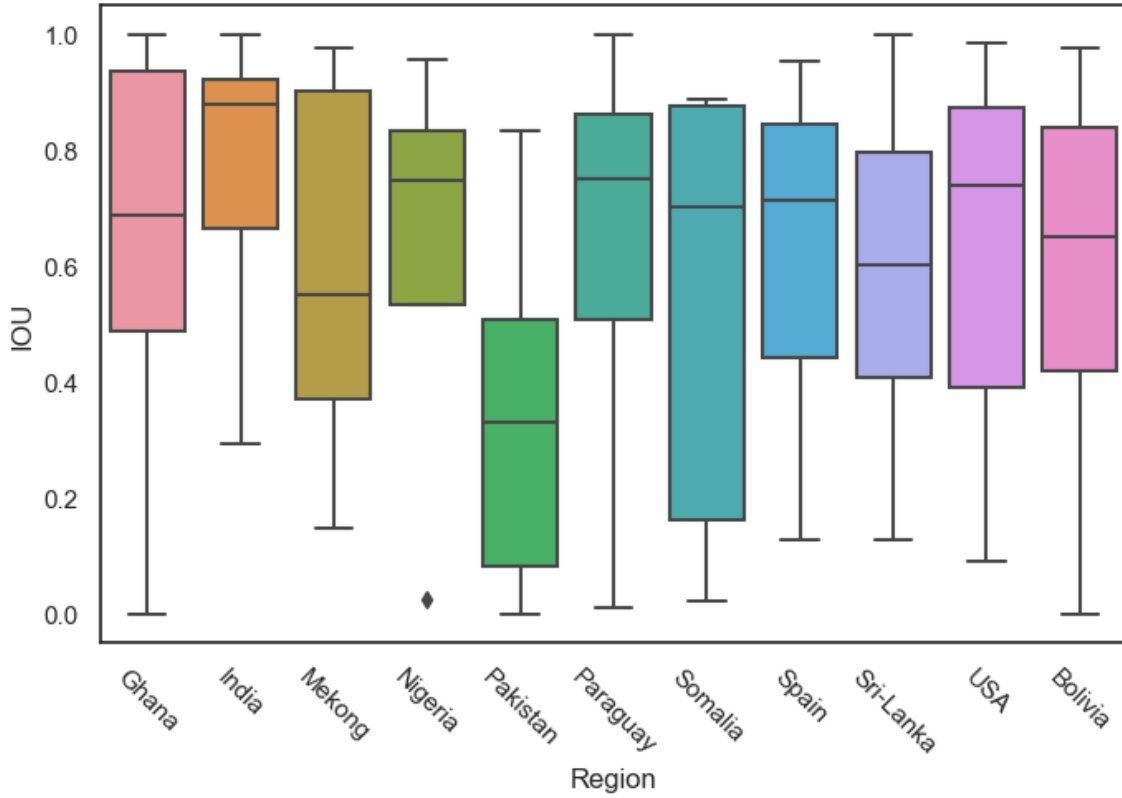
(a) SAR_HSV(O3)+cAWEI+cNDWI feature space
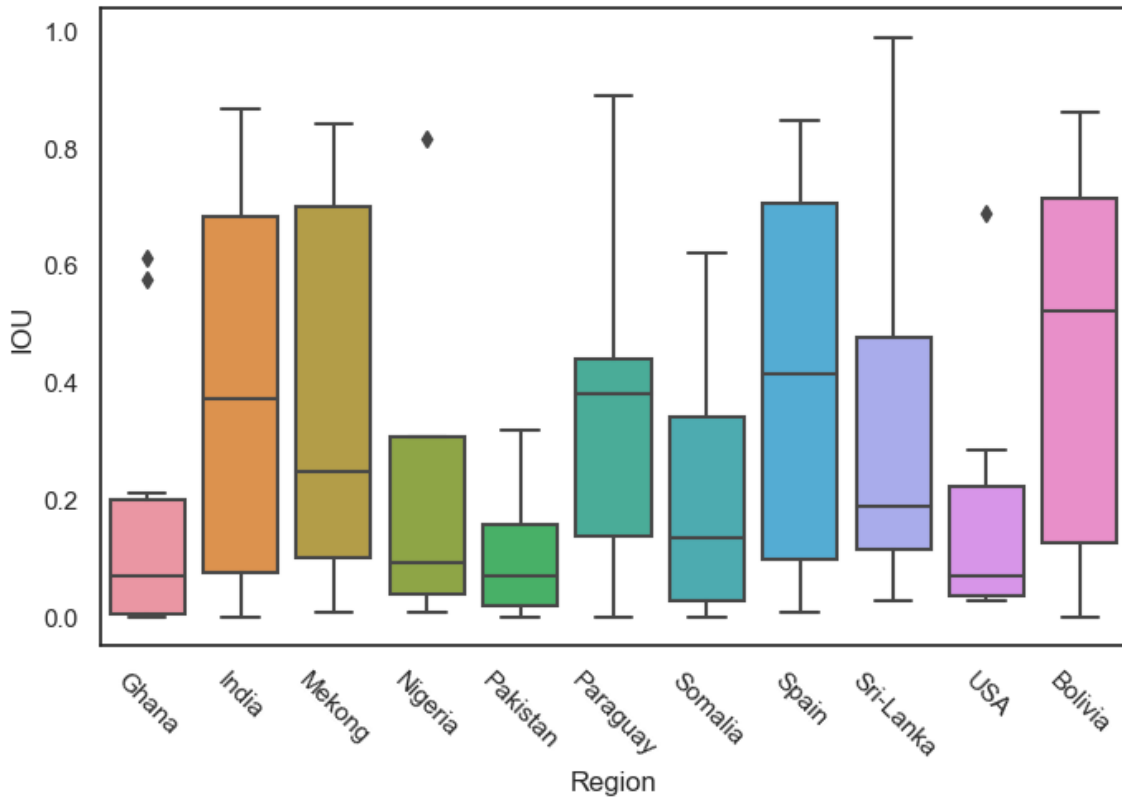
(b) SAR feature space

Figure 3.1: Box plots showing variable performance across different regions for two feature spaces. IOU is calculated across individual images. Outliers, marked by diamonds, are defined as more than 1.5 times the inadequately range whiskers is 1.5 times the inadequately range away from the median.

Looking at all of the results included in table 3.1 it is clear that the predictive performance is highly variable on a image-to-image basis, demonstrated by the high standard deviation and differences between mean and total metric scores. Figure 3.1a shows the regional distribution of IOU scores calculated on a per-image basis for our best GBDT model on the SAR_HSV(O3)+cAWEI+cNDWI feature space. Performance across many regions (Spain, Sri Lanka, USA, Bolivia, Ghana, Nigeria) is generally good, with high median scores and reasonably high lower quartiles. Pakistan stands out as an outlier, with scores far lower than all other regions. A number of regions show reasonably high numbers of points with scores close to zero and equal to 1. A likely explanation for this is that these images include very few inundated pixels; false negatives and even a few false positives would sway the score in these cases. The 15 Bolivia images were checked for this using QGIS; 3 were found to have less than 5 inundated pixels and 2 had no inundated pixels.

Figure 3.1b shows a similar boxplot for the GBDT SAR-only feature space. Performance across the board is significantly lower, with Ghana, Pakistan and the USA standing out as having very low values. The USA result is especially surprising as this region has the highest number of images (40) in the training set. On the other hand, the SAR-only model performs best in Bolivia with an IOU of 0.5181 ± 0.3406, which is outside of the training domain. This result is around 15% worse than the full feature space.
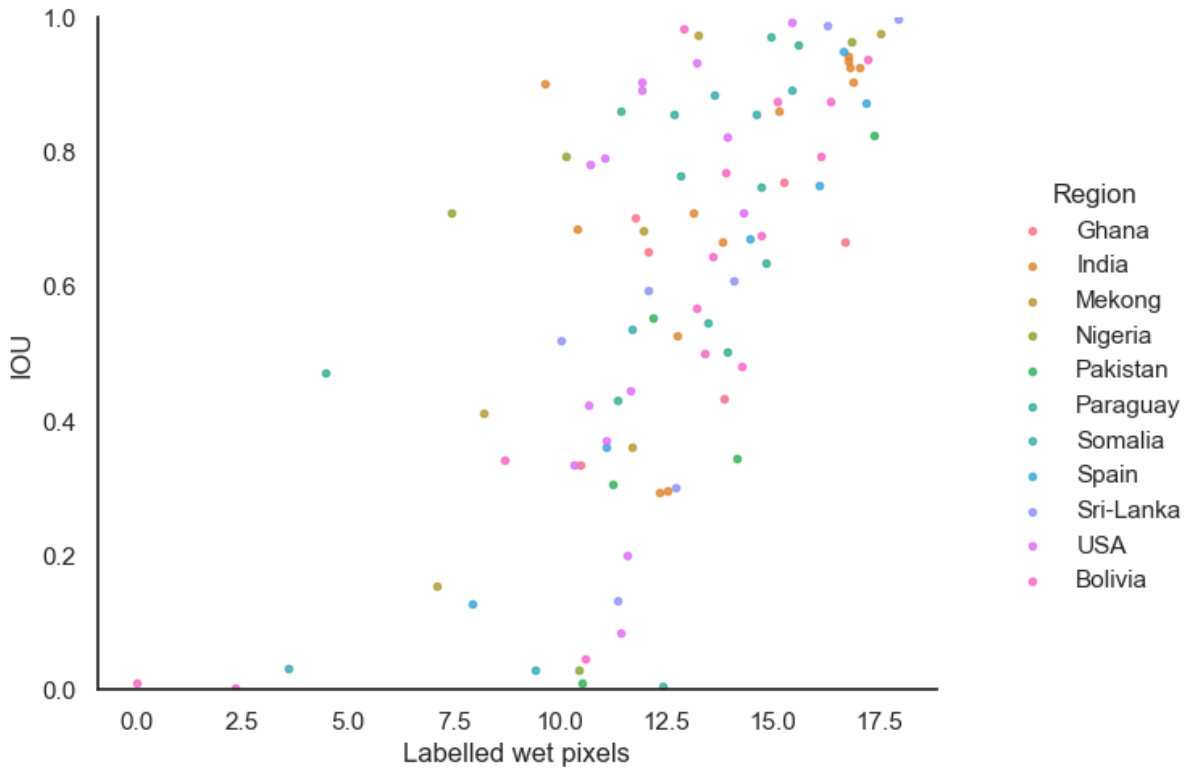

Figure 3.2: Distribution of IOU by number of wet pixels in the hand labeled layer across the test and Bolivia sets. IOU shown is from the GBDT model using the SAR_HSV(O3)+cAWEI+cNDWI feature space, points split by region. Number of pixels is transformed by $log_2(n)$ to make correlation clearer.

Even within a region, performance varies a lot. For example, on the full feature space the interquartile range for Somalia is 0.67 - whilst the median performance is on par with other regions, the lowest quartile value of 0.18 is far worse than all other regions except Pakistan. One possible explanation for this is shown in figure 3.2 - there is a clear link between the IOU recorded for an individual image and the number of wet pixels present in its label. This appears to be true of all regions, with no obvious differences in the distribution shown. This chart also gives a better look at the pixels where IOU is equal to 1 or 0. There seem to be a number of pixels where IOU = 0 and the number of wet pixels is 0, but this is not the case for all recorded cases, with some having 1024 or more wet pixels. Similarly, all of the cases where IOU=1 have many wet pixels - up to 185000 (70%). Most images (92%) in the set have more than 180 wet pixels.

Alongside those highlighted by IB, a wide array of other feature space were tested, with features derived from the DEM added. The test split performance of the 10 best models from a random sample of 40 model-feature space combinations are shown in table 3.2. The best performing model by most metrics was

Split	Model	Feature Space	Mean IOU Flooded (std)	Total IOU Flooded	Total Precision Flooded	Total Recall Flooded	Total F₁ Score
	GBDT	SAR_cAWEI	**0.6692(±0.3007)**	**0.8743**	**0.9418**	0.9242	0.9329
	GBDT	LDEM_SAR_HSV(O3)+cAWEI+cNDWI	0.6061(±0.3147)	0.8466	0.9238	0.9102	**0.9390**
	GBDT	SAR_HSV(O3)+cAWEI+cNDWI	0.6246(±0.3105)	0.8427	0.8991	0.9307	0.9147
	GBDT	cAWEI	0.6235(±0.3072)	0.8349	0.8884	**0.9327**	0.9100
Test	LDA	cAWEI	0.5982(±0.3143)	0.8262	0.8932	0.9168	0.9049
	QDA	DEM_SAR_HSV(O3)+cAWEI+cNDWI	0.5604(±0.3022)	0.7991	0.8830	0.8937	0.8883
	LDA	SAR_HSV(O3)+cAWEI+cNDWI	0.5635(±0.2998)	0.7966	0.8718	0.9023	0.8868
	NaiveBayes	SAR_HSV(O3)+cAWEI+cNDWI	0.5635(±0.2998)	0.7966	0.8718	0.9023	0.8868
	NaiveBayes	SAR_cAWEI+cNDWI	0.5730(±0.2998)	0.7922	0.8746	0.8936	0.8840
	LDA	SAR_cAWEI+cNDWI	0.5730(±0.2998)	0.7922	0.8746	0.8936	0.8840

Table 3.2: Top 10 best performing models from a randomly selected space of 50 feature space - model combinations, sorted by total IOU flooded. Metric values shown correspond to the in-domain test set; highest scores for each metric are shown in bold. Reported standard deviation refers to the variation across images in that set, rather than between runs.

a GBDT trained on SAR_cAWEI, giving a total IOU score of 0.8743 - the highest recorded by any of the LX models, and nearly as high as IB's best performing GBDT (trained on SAR_HSV(O3)+cAWEI+cNDWI, total IOU of 0.8751).

The standard deviation for both IOU and accuracy is very similar for all model runs, rounding to 0.3 and 0.05 respectively in all cases. This suggests that performance on the outlier locations is not changing based on the model and feature space used. The same effect can be seen in table 3.1 and IB's original optical and mixed space results, although the variation there is slightly higher. Far lower variability is seen in their SAR-only results, although overall mean performance is lower.

The bottom four models shown in figure 3.2 display very similar results. When shrinkage is set to zero, the LDA model performs the same as a Gaussian naïve Bayes model. This appears to be the optimum hyperparameter configuration in these cases.

		Test		Bolivia	
Model	Feature Space	Mean IOU Flooded (std)	Total IOU Flooded	Mean IOU Flooded (std)	Total IOU Flooded
Naïve Bayes	SAR_cAWEI+cNDWI	0.5693(±0.3039)	0.7812	0.5212(±0.3250)	0.7607
Naïve Bayes	LDEM_SAR_HSV(O3)+cAWEI+cNDWI	0.5718(±0.3006)	0.7755	0.5116(±0.3165)	0.7441
Naïve Bayes	SAR_HSV(O3)+cAWEI+cNDWI	0.5675(±0.3015)	0.7738	0.5131(±0.3169)	0.7447
Naïve Bayes	DEM_SAR_HSV(O3)+cAWEI+cNDWI	0.5646(±0.3031)	0.7732	0.5148(±0.3189)	0.7508
Naïve Bayes	DP_SAR	0.3061(±0.2889)	0.5580	0.4191(±0.3109)	0.6391
Naïve Bayes	SAR	0.2983(±0.2899)	0.5333	0.4071(±0.3069)	0.6329
Naïve Bayes	ACU+SDEM+LDEM_SAR	0.2983(±0.2899)	0.5333	0.4071(±0.3069)	0.6329
GBDT	DP_SAR	0.2845(±0.2908)	0.5121	0.4300(±0.3093)	0.6559

Table 3.3: Selection of feature spaces using DEM derived features, sorted descending by test split total IOU flooded. Example models from other non-DEM feature spaces included for comparison.

In general, the DEM derived features did not improve performance enough to consider them to be worth including, and in some cases actively made metric scores worse. naïve Bayes tended to be more resilient than GBDT in dealing with DEM features. The instances where DEM features made the biggest improvement are shown in table 3.3. The biggest positive difference was with DP_SAR, where DP is a binary layer denoting whether the pixel is within a depression. Here, a marginal increase of 5% to 0.5580 is seen. The feature did not improve performance by more than 1% on any other feature spaces.

LDEM_SAR_HSV(O3)+cAWEI+cNDWI is the highest performing space that includes DEM features, although the value is only 0.002 points higher than the space without the locally scaled DEM. Similarly, the unscaled DEM makes this model worse.

Chapter 4

Qualitative Analysis

In this chapter, we review the original aims of the project, their success (or lack thereof), and recommendations for how our findings could be used and fit into the wider context.

4.1 Replicating Iselborn, et al.'s findings

The project started out by aiming to reproduce the findings presented in Iselborn, et al. [16]. This proved more challenging than expected as the code released in their GitHub repository is incomplete and not inherently reproducible. As a result, a new methodology (detailed in our own GitHub repository) was made. Changing the methodology significantly is always undesirable as it introduces uncertainty in how comparable the results will be, and makes it more difficult to question results generated using the original methods. We were not able to directly replicate the highest performance figures presented by IB, although overall our performance is close enough to their values to suggest that the methodologies are broadly comparable. naïve Bayes models were used internally to check the general method performance as they have no hyperparameters to tune, and thus should be the easiest to reproduce. We were successful in this respect - our best attempt naïve Bayes models outperform IB's significantly. On the full feature space it gave total IOU values of 0.7738 compared to IB's equivalent value of 0.7399 The gap is especially highlighted on the Bolivia split, where our total IOU value of 0.7447 is over 10 points higher than IB's value of 0.6160, and on-par with their models' scores on the in-domain test split. Across all of the models highlighted in table 3.1 we show a far lower performance drop than IB between the test and Bolivia sets, even when using the same hyperparameter values. One possible explanation for this is the standardisation applied as part of our pre-processing. IB do not mention standardisation as a step in their pre-processing pipeline, although their code suggests that it was at least tested. It is unclear whether their implementation is functional, although its existence was the reason for us to include it here. Standardisation helps in dealing with out-of-domain data by limiting the effect that values outside of the range seen by the model can have on the results.

There is a big difference between mean and total IOU across all models. This is also true of IB's reported results, and shows the importance of specifically defining how each metric will be calculated and used. Total IOU gives higher numbers, but is perhaps less useful for understanding where the models are performing well as it obscures the variation shown across the dataset. In a similar vein, mean IOU is also strongly affected by images that have either very high or very low (including zero) levels of inundation. For example, on an image with only 5 pixels defined as flooded, the IOU is being calculated solely on the performance of these 5 pixels. This can be seen looking through the individual image metrics, where many images have an IOU value of 1 or 0. It may be useful to create a metric that accounts for this, that weights individual pixel performance equally across the image, but we could not come up with any designs that especially helped with this issue.

4.2 Alternative Feature Spaces

The four features used here are far fewer than then 9 required for the full feature space, suggesting that these extra features may not be adding as much as suggested by IB. They do not report findings for SAR_cAWEI in the main body or appendix (which contains tables of other tested features), suggesting that this space may have been overlooked, or else not given similarly high performance.

4.3 Digital Elevation Models

Somewhat surprisingly the DEM derived features do not improve model performance overall by much, although the locally scaled DEM (LDEM) and depression (DP) layers do make marginal improvements in some cases. One potential reason for this can be seen in the bottom left of the DEM in figure 4.1c, - the river has moved since the DEM was produced. This isn't uncommon - meandering rivers move on a constant basis due to erosional processes. One way to improve this could be to use the ASTER DEM instead of SRTM. Whilst the first version of ASTER v3 was released in the early 2010s, it is regularly updated with the most recent major version being released in 2019. Continuous iteration is a strong reason to use a product that depends on current remote sensing technology rather than data that was recorded in a single mission more than 20 years ago, and will not be updated again.

Inclusion of the slope layer decreased performance in every case. The theoretical reasoning for including slope is clear - flooding is less likely on a slope than flat land - but the DEM is not of a high enough quality to allow the signal to be seen through the noise. One reason is that although SRTM v3 is released at a 30m resolution, some of the original data was recorded at a 90m resolution due to equipment limitations. The difference in resolution can be seen when comparing Pakistan (figure 4.1c) to the USA (figure 4.1b) - the Pakistan DEM image looks much smoother with less hard edges, whilst it is possible to see clear elevation differences in the USA image. Again, the difference between the 30m resolution of the DEM products and the 10m optical image above is clear. In order to calculate slope, multiple pixels need to be analysed, thus the output layer is of lower detail than the inputs even if the raw resolution is the same. As a result, it is preferable that the DEM resolution is 2 or 4 times higher than the analysis resolution, which is not the case here. At best the DEM resolution is 1/3rd as high as the analysis resolution, and at worst 1/9th. As a result, it is not surprising that this layer does not provide useful results. Higher resolution DEMs with global coverage do not currently exist in the public domain, although commercial ones do exist that may be usable in a more operational implementation of a flood classifier. The most effective DEM derived layer tested was a binary representation of areas that were classed as depressions by the depression fill algorithm. The cases where this is likely to help are where noise makes identifying open water difficult, such as lakes and ponds. These are areas usually classified as depressions because they do not have a clear exit for water entering them. During a flood event these areas are very likely to be the first to flood, especially in rural areas, and so this acts like thresholding on SAR imagery. These findings suggest that the inclusion of a DEM shows some potential although its usefulness is hampered by the low quality of data used. A higher resolution, more recent DEM may show much higher performance.

4.4 Sen1Floods11

Across all models used in this study, Iselborn, et al. [16] and Bonafila, et al. [5] there appears to be an upper limit on model performance by total IOU, in the range of 0.83-0.87. This suggests that the models are making errors when classifying the dataset, which is not especially surprising due to the random noise present, as well as varying conditions. However, this is true even if training and fitting against the same images and region, suggesting that there may be issues with the original hand classified labels layer used. The reference layer was generated using thresholding on MNDWI and NDVI, followed by cloud shadow removal, and is described in more detail in Bonafilla, et al. [5]. From this, trained analysts were given two composite RGB images and the SAR VH band to make changes to the reference layer for use as labels. 4.1 shows three separate chips, from Bolivia, the USA and Pakistan. Analysts were given access to the middle two images alongside the automatically generated labels and another false colour image.

Looking at the Bolivia image, it appears that there are a number of locations where inundated land is misclassified as dry. In particular, the bottom left hand corner shows a meandering river, which in reality must be continuous, especially during a flood event. The false colour (third from top) image clearly shows the river to be larger than the labelled inundated area, but this area is less than half covered by labelled inundated land. The SAR image for this area appears to be offset by 1-2 pixels, and shows a much smaller river. In the top left quarter of the image is an area where the river has breeched its banks and is covering the floodplain. Here, the SAR image appears to match the labels well, although again there are areas where it appears land is inundated but not labelled as such. The DEM layer does not really assist with classification here - the resolution and height accuracy is too low to easily identify the river boundaries, and the overall height variation is very low. It also appears that the river has moved slightly since the DEM was recorded in 2000, which also raises issues. Looking at the optical composite images it appears to be much more difficult to classify this area, as there are no clear colour boundaries
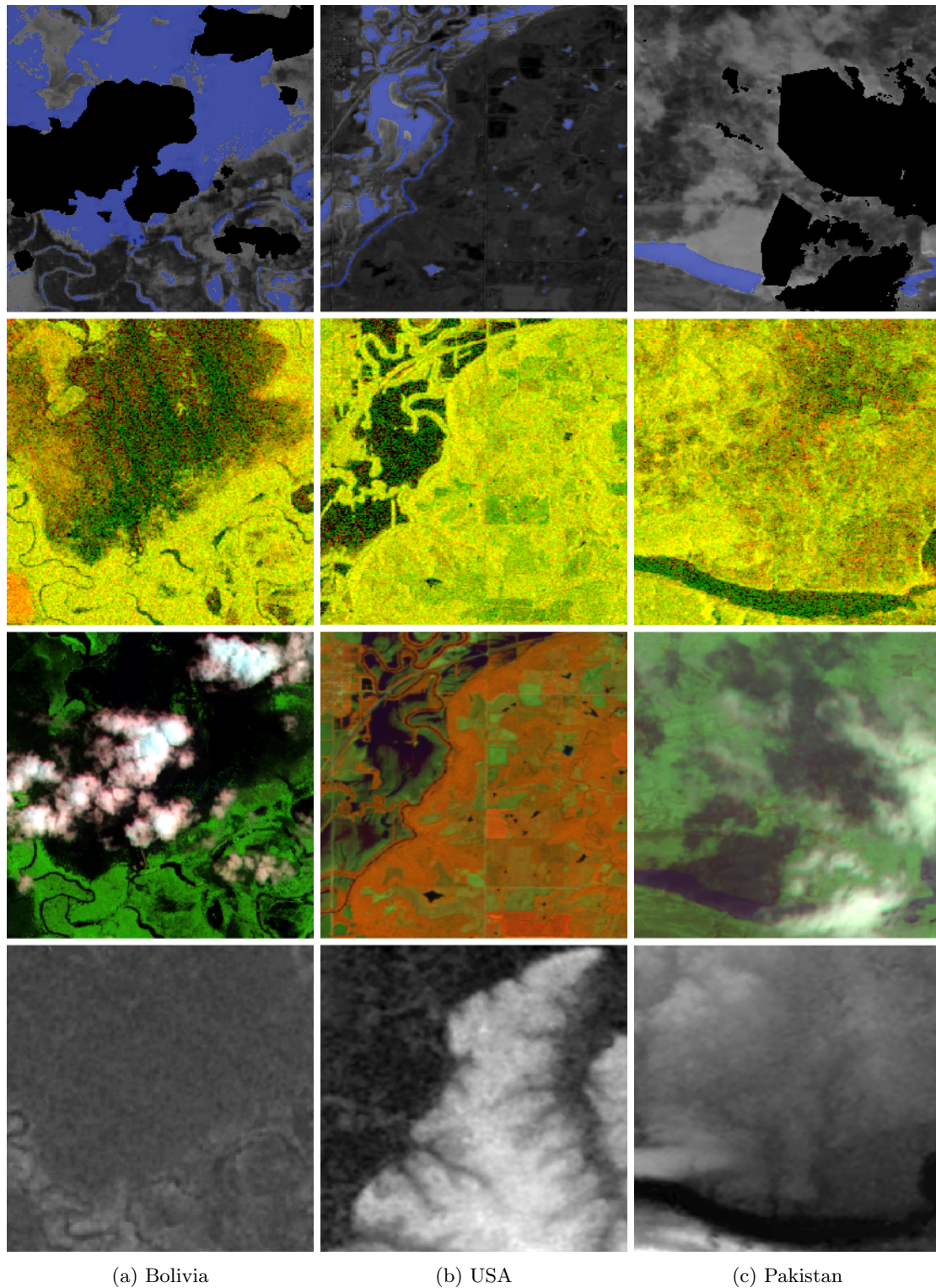
(a) Bolivia   (b) USA   (c) Pakistan

Figure 4.1: A selection of three images from across Sen1Floods11, highlighting the variation shown between regions. From top to bottom: Hand labels overlaid on AWEI (blue = wet, black = unknown, greyscale = AWEI); SAR (red = VV, green = VH), False colour (red = SWIR1, green = near infrared, blue = red); DEM (white = highest elevation; black = lowest elevation). All three locations have the same scale per image for comparison purposes.

shown, instead they tend to blend. It is clear at least some of this land is inundated but how much is difficult to say using these layers, meaning that the labels are probably not an especially good fit for the reality at the time that the images were recorded. Whilst the optical and SAR images were taken on the same day, it appears that the flooding in this pixel is quite different between the images, with the SAR image showing less flowing water, although similar levels on the floodplain.

Looking at the USA images, the background of the false colour image stands out as very different from the other two images. This is due to the orange land being largely bare with no vegetation. A much clearer height difference can be seen on the DEM, as much of the area to the east of the river is higher up than the river and corresponding floodplain (top left). The labels here appear to be more accurate than in Bolivia, but still miss a number of small water bodies that are evident on the AWEI/false colour image. These are likely small irrigation ponds used by the agricultural activity evident across the centre of the image. On the floodplain, the SAR image suggests there are very clear boundaries between the edge of the water and the dry land, but the same boundaries appear to be in a different location on the false colour image - suggesting the water has receded in the gap. It is unclear how long the gap was between the two images, although they were stated to have been taken on the same date (22-05-2019, [5, Figure 3])

The Pakistan images show many of the same issues identified in the other two chips. Cloud cover is much more of an issue here as there is a low density of cloud across much of the wet area of the image making it appear out of focus in some areas. The cloud masking used in the original dataset is less effective here, and has clearly been edited by the trained analyst (due to the very straight edges shown on the mask, which are characteristic of a polygon based approach). The SAR image appears to show a river or perhaps reservoir at the bottom of the image with little flooding in the rest of the image, whilst the optical images suggest far more water is present. the DEM is helpful here - it is possible to see elevation markings suggesting that there are rivers or streams feeding into the main river at the bottom of the image from higher elevations. This image is perhaps the most difficult to accurately classify by hand although it seems reasonably clear that more of the image is wet than the labels suggest.

These images highlight the difficulty of water classification, especially during flood events. Across the light and radar spectra different surfaces display different reflective characteristics which are affected by factors as diverse as soil moisture, plant type, incident angle and season even without considering inundation. Additionally, floods are fast-moving events - during the space of 24 hours water levels can go from normal to peak flood and back down to lower levels. Any gap between images can make classification more challenging, but this becomes especially true as that period increases.

In creating Sen1Floods11, Bonafilla, et al. [5] aimed to provide a first-of-its-kind global dataset that would simplify the process of creating and validating modelling approaches to classify flooded land across multiple biomes. To some extent, it does meet this goal - both this study and Iselborn, et al. [16] were able to achieve good metric scores even on out-of-band regions. However, this dataset appears to not be especially useful for the goals set out by this study - in particular, it does not appear that the dataset labels classify water (and particularly flooded land/vegetation) to a high enough accuracy to allow models to achieve the required accuracy needed for use as calibration layers for hydraulic models. This is a task where a remote sensing solution has been "coming soon" for well over two decades, although unfortunately it does not appear to have arrived yet.

4.5 Links to wider goals

For a classification model to be used to calibrate hydrodynamic flood models, it needs to show performance as least as high as the hydrodynamic model, which likely means consistent IoU scores about 0.90. It is quite clear that none of the models used by us or Iselborn, et al. [16] perform at this level across all regions, and thus do not constitute a reliable and usable methodology for identifying floods globally. Despite this, the best techniques give insight into ways in which the features involved could be used to produce a model that could achieve those goals. Additionally, the current models may be effective enough for some uses, such as identifying permanent water and providing a general overview of the level of inundation that is occuring in a region, rather than identifying specific locations that are wet. The GBDT model trained on SAR_HSV(O3)+cAWEI+cNDWI or even just cAWEI alone does a good job of meeting these needs; the latter is more ideal as it only requires one data source, and thus doesn't suffer from issues with acquiring coincident imagery from two separate satellite platforms. Combining SAR with Optical imagery remains challenging for many of the same reasons it has for the past few decades: lack of data, resolution and cost.

Chapter 5

Future Work

The creation of better labelled flood event datasets is the most obvious place where future work would be useful, although achieving that goal is rather challenging. Sen1Floods11 is useful for comparing the performance of different modelling techniques, but is not ideal for training models for use in the real world, at least not where local performance is just as important as total metrics. Many uses do not need a model that can predict well in every biome, just the one which is of interest. If the biome in question is similar to the area of Bolivia, the USA or India which is included in this dataset you may be in luck, but if it is not then performance is likely sub-par. Hand labelling large numbers of images is very expensive, which is the main reason that there are so few datasets of this type. An alternative approach that is more feasible would be to expand the models to use the weakly labelled section of Sen1Floods11, which is nearly 10 times larger than the hand labelled subset. In this instance we did not do this as the use of a much larger dataset would have limited the amount of different experiments that could have been performed due to limitations on compute resources. Currently most of the models used trained on the largest feature space take between 5 and 30 minutes to run a single iteration and struggle to fit the entire dataset within the 16 gigabytes of memory which the device used has. The code used is rather inefficient in this, and could be optimised to reduce memory usage and speed up each individual model run. One way to do this would be to switch to partial fitting, where the full set of samples are split into blocks, and the model is trained iteratively on each block. This allows for the use of far larger datasets than can be fit within device memory, and is supported by many of the models used here (although not GBDT)

By design, all of the models used in this study are pixel-wise, which means that they ignore the surrounding context by assuming each pixel is independent. This assumption is obviously not true in the real world - if a pixel is wet and the elevation has not changed significantly, it is likely that its neighbours are also wet. This leads to errors where e.g. a pixel in the middle of a river is marked as dry due to an anomaly in the data (e.g. a small cloud, floating debris or just noise). This has been an issue since the early days of hydrodynamic modelling, where human observers would draw lines on paper aerial photos. The DEM features, particularly the depression filled binary layer help with this as they do a good job of identifying areas of flat ground that should almost certainly be wet during a flood (i.e. lakes and rivers). However, it does not improve overall performance much, suggesting that these kind of errors are not very common. Indeed, across the board the DEM did not improve performance much or at all. It appears that the DEM resolution, vertical accuracy and age makes it rather unsuitable for this task. Newer DEMs exist that provide global coverage (such as ASTER) although they too face similar quality issues. A better approach for testing the usefulness of the DEM as a feature may be to select one country to focus on. Much of Europe, the UK and USA have very high quality LiDAR DEMs updated on a frequent basis at 2m resolution or better. These may allow for better derivative features to be generated that may give them more of a chance to shine.

The obvious extension from a pixel-wise model where independence between pixels is not established is to instead use patch processing. With this approach, the model looks at a patch of pixels (e.g. a 3x3 grid around the selected pixel) to make its classification. This is likely to help in cases where large bodies of contiguous water are present, but may pose issues for small water bodies that are only a few pixels in size. The use of data science models that treat water segmentation as a classification task rather than a statistical edge finding task is relatively recent, so this field may have a lot of potential when used with hand crafted feature spaces. This may also be more resilient to mis-registered images, where different sources do not line up exactly. In working on this study, it has become apparent that the goal of making a "global" model that can accurately identify flooding from satellite imagery for

hydraulic model validation is not currently feasible, at least not using these techniques. However, there are some applications where these techniques may be useful, for example permanent and seasonal water classification. This information is used in multiple disciplines - for land use classification, water resource management and long term climate research. Current datasets such as the JRC water dataset, tend to use thresholding approaches as these have a very low computational complexity. Simple models used in this study, including LDA and QDA trained on small datasets using cAWEI do not add a very large computational cost over thresholding, but may perform significantly better. This is a case where the techniques used here may be easily operationalised, so may be worthwhile exploring.

Chapter 6

Conclusions

As part of this project we have expanded on the work of Iselborn, et al. [16], using additional feature spaces and features to provide more resilient and useful results on the same dataset.

When comparing our results to Iselborn, et al. [16] we were unable to achieve the highest scores that they noted with the GBDT model, but found similar scores with other model/feature combinations. Remarkably we demonstrated a far smaller drop in performance across multiple models between the in domain test set and the out-of-domain Bolivia images. Whilst the reasoning for this is unclear, it may be related our differing approach to feature scaling. This is nevertheless a useful improvement if it can be maintained on other images/regions.

Through testing more feature spaces than Iselborn, et al. we were able to identify spaces other than SAR_HSV(O3)+cAWEI+cNDWI that performed as well The SAR_cAWEI gave metric scores slightly higher than we were able to produce with larger feature space with total IOU value of 0.8743 - with 1% of the value recorded by Iselborn, et al. We also identified that cAWEI alone gave total IOU valus of 0.8319. This is perhaps a more useful finding, as it opens the door to methods that only use optical imagery, and thus do not suffer from as many data acquisition issues as finding coincident optical and SAR imagery.

Although theoretically sound we found that the use of DEM derived features such as slope and flow accumulation did not make a material improvement to classification performance by any of the metrics used. They showed marginal gains of less than 5% at best. Whilst disappointing, there are clear reasons why this approach did not have much potential in this case. The quality of the DEM is the most obvious, as its resolution and vertical accuracy is not good enough for this kind of analysis. Its age is also problematic as the ground state has changed since it was recorded in a number of instances. A better quality DEM may make these features more useful, but such a DEM is not currently available in the public domain on a global scale.

The Sen1Floods11 dataset provides a good starting point to compare the performance of different modelling techniques at identifying flooded land, but its usefulness as a calibration layer for hydrodynamic models is unclear. The inclusion of a wide array of biomes across many different countries is useful, however it seems to result in models that perform acceptably across the board but do not excel in any specific region - they are the jack of all trades, but master of none. Depending on the task that it is used for this may be optimum - such as if you're trying to classify large numbers of images globally to identify changes in the frequency and magnitude of flooding due to climate change. But for many tasks where this kind of water segmentation would be very helpful - calibrating hydrodynamic models, planning for and responding to natural disasters - the predictive accuracy is not there yet.

The usefulness of a global model that can accurately identify flooded land to a good enough level for its outputs to be used to calibrate high resolution hydrodynamic models on a rapid basis is undeniable, but unfortunately this remains out of reach. Sen1Floods11 and the research done on it is however worthwhile as it allows for the comparison of methods, to ascertain where future research should focus. It is not suitable for training models to be used for these purposes in the real world, at least not using the techniques presented in this study, and likely not by only using the hand labelled data. We presented different feature spaces that may be as effective as those used by others whilst requiring less training features. In this case, less does appear to be more.

Bibliography

[1] Eric S. Anderson et al. "Horizontal resolution and data density effects on remotely sensed LIDAR-based DEM". In: *Geoderma* 132.3 (June 1, 2006), pp. 406–415. ISSN: 0016-7061. DOI: 10.1016/j.geoderma.2005.06.004. URL: https://www.sciencedirect.com/science/article/pii/S001670610500193X.

[2] Mohammadtaghi Avand et al. "DEM resolution effects on machine learning performance for flood probability mapping". In: *Journal of Hydro-environment Research* 40 (Jan. 1, 2022), pp. 1–16. ISSN: 1570-6443. DOI: 10.1016/j.jher.2021.10.002. URL: https://www.sciencedirect.com/science/article/pii/S1570644321000769 (visited on 08/29/2023).

[3] Richard Barnes, Clarence Lehman, and David Mulla. "An Efficient Assignment of Drainage Direction Over Flat Surfaces in Raster Digital Elevation Models". In: *Computers & Geosciences* 62 (Jan. 2014), pp. 128–135. ISSN: 00983004. DOI: 10.1016/j.cageo.2013.01.009. arXiv: 1511.04433[cs]. URL: http://arxiv.org/abs/1511.04433 (visited on 08/30/2023).

[4] Richard Barnes, Clarence Lehman, and David Mulla. "Priority-flood: An optimal depression-filling and watershed-labeling algorithm for digital elevation models". In: *Computers & Geosciences* 62 (Jan. 1, 2014), pp. 117–127. ISSN: 0098-3004. DOI: 10.1016/j.cageo.2013.04.024. URL: https://www.sciencedirect.com/science/article/pii/S0098300413001337 (visited on 08/30/2023).

[5] Derrick Bonafilia et al. "Sen1Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020, pp. 210–211. (Visited on 06/22/2023).

[6] Jason Brown. *SAR 101: An Introduction to Synthetic Aperture Radar*. Capella Space. Feb. 10, 2020. URL: https://www.capellaspace.com/sar-101-an-introduction-to-synthetic-aperture-radar/ (visited on 08/20/2023).

[7] Centre for Research on the Epidemiology of Disasters. *The human cost of natural disasters 2015: a global perspective — English*. June 7, 2016. URL: https://climate-adapt.eea.europa.eu/en/metadata/publications/the-human-cost-of-natural-disasters-2015-a-global-perspective (visited on 06/27/2023).

[8] NASA Earth Science Data Systems. *What is Synthetic Aperture Radar?* Earthdata. Publisher: Earth Science Data Systems, NASA. Apr. 10, 2020. URL: https://www.earthdata.nasa.gov/learn/backgrounders/what-is-sar (visited on 08/20/2023).

[9] European Space Agency. *Acquisition Segments - Sentinel-1 - Sentinel Online*. Sentinel Online. URL: https://copernicus.eu/missions/sentinel-1/observation-scenario/acquisition-segments (visited on 06/30/2023).

[10] Tom G. Farr et al. "The Shuttle Radar Topography Mission". In: *Reviews of Geophysics* 45.2 (2007). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005RG000183. ISSN: 1944-9208. DOI: 10.1029/2005RG000183. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183 (visited on 08/30/2023).

[11] Gudina L. Feyisa et al. "Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery". In: *Remote Sensing of Environment* 140 (Jan. 1, 2014), pp. 23–35. ISSN: 0034-4257. DOI: 10.1016/j.rse.2013.08.029. URL: https://www.sciencedirect.com/science/article/pii/S0034425713002873 (visited on 08/27/2023).

[12] Dirk Geudtner et al. "Sentinel-1 System capabilities and applications". In: *2014 IEEE Geoscience and Remote Sensing Symposium*. 2014 IEEE Geoscience and Remote Sensing Symposium. ISSN: 2153-7003. July 2014, pp. 1457–1460. DOI: 10.1109/IGARSS.2014.6946711.

[13] M. S. Horritt, D. C. Mason, and A. J. Luckman. "Flood boundary delineation from Synthetic Aperture Radar imagery using a statistical active contour model". In: *International Journal of Remote Sensing* 22.13 (Jan. 1, 2001). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431160116902, pp. 2489–2507. ISSN: 0143-1161. DOI: 10.1080/01431160116902. URL: https://doi.org/10.1080/01431160116902 (visited on 07/02/2023).

[14] A.R Huete. "A soil-adjusted vegetation index (SAVI)". In: *Remote Sensing of Environment* 25.3 (Aug. 1, 1988), pp. 295–309. ISSN: 0034-4257. DOI: 10.1016/0034-4257(88)90106-X. URL: https://www.sciencedirect.com/science/article/pii/003442578890106X.

[15] Kevin Iselborn et al. *DFKI-Earth-And-Space-Applications/Flood_Mapping_Feature_Space_Importance*. Mar. 1, 2023. URL: https://github.com/DFKI-Earth-And-Space-Applications/Flood_Mapping_Feature_Space_Importance (visited on 06/22/2023).

[16] Kevin Iselborn et al. *On the Importance of Feature Representation for Flood Mapping using Classical Machine Learning Approaches*. Mar. 1, 2023. DOI: 10.48550/arXiv.2303.00691. arXiv: 2303.00691[physics]. URL: http://arxiv.org/abs/2303.00691 (visited on 06/22/2023).

[17] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (visited on 08/29/2023).

[18] Zachary Kreiser, Brian Killough, and Syed R Rizvi. "Water Across Synthetic Aperture Radar Data (WASARD): SAR Water Body Classification for the Open Data Cube". In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia: IEEE, July 2018, pp. 437–440. ISBN: 978-1-5386-7150-4. DOI: 10.1109/IGARSS.2018.8517447. URL: https://ieeexplore.ieee.org/document/8517447/ (visited on 07/02/2023).

[19] Jeerapong Laonamsai et al. "Utilizing NDWI, MNDWI, SAVI, WRI, and AWEI for Estimating Erosion and Deposition in Ping River in Thailand". In: *Hydrology* 10.3 (Mar. 2023). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 70. ISSN: 2306-5338. DOI: 10.3390/hydrology10030070. URL: https://www.mdpi.com/2306-5338/10/3/70 (visited on 08/27/2023).

[20] John B. Lindsay. "Efficient hybrid breaching-filling sink removal methods for flow path enforcement in digital elevation models". In: *Hydrological Processes* 30.6 (2016). _eprint: https://onlinelibrary.wiley.com/doi/pdf/ pp. 846–857. ISSN: 1099-1085. DOI: 10.1002/hyp.10648. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10648 (visited on 08/30/2023).

[21] R. Manavalan. "SAR image analysis techniques for flood area mapping - literature survey". In: *Earth Science Informatics* 10.1 (Mar. 1, 2017), pp. 1–14. ISSN: 1865-0481. DOI: 10.1007/s12145-016-0274-2. URL: https://doi.org/10.1007/s12145-016-0274-2 (visited on 03/31/2023).

[22] S K McFeeters. "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features". In: *International Journal of Remote Sensing* 17.7 (May 1, 1996). Publisher: Taylor & Francis, pp. 1425–1432. ISSN: 0143-1161. DOI: 10.1080/01431169608948714. URL: https://doi.org/10.1080/01431169608948714.

[23] NASA Jet Propulsion Laboratory. *SRTM DEM*. Version 3. Oct. 2015. URL: https://lpdaac.usgs.gov/documents/179/SRTM_User_Guide_V3.pdf.

[24] K. G. Nikolakopoulos, E. K. Kamaratakis, and N. Chrysoulakis. "SRTM vs ASTER elevation products. Comparison for two regions in Crete, Greece". In: *International Journal of Remote Sensing* 27.21 (Nov. 1, 2006). Publisher: Taylor & Francis, pp. 4819–4838. ISSN: 0143-1161. DOI: 10.1080/01431160600835853. URL: https://doi.org/10.1080/01431160600835853.

[25] Mahdi Panahi et al. "Large-scale dynamic flood monitoring in an arid-zone floodplain using SAR data and hybrid machine-learning models". In: *Journal of Hydrology* 611 (Aug. 1, 2022), p. 128001. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2022.128001. URL: https://www.sciencedirect.com/science/article/pii/S0022169422005765 (visited on 07/02/2023).

[26] J. -F. Pekel et al. "A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data". In: *Remote Sensing of Environment* 140 (Jan. 1, 2014), pp. 704–716. ISSN: 0034-4257. DOI: 10.1016/j.rse.2013.10.008. URL: https://www.sciencedirect.com/science/article/pii/S0034425713003763 (visited on 08/29/2023).

[27]   Daniel Scheffler et al. "AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data". In: *Remote Sensing* 9.7 (2017). ISSN: 2072-4292. DOI: 10.3390/rs9070676. URL: https://www.mdpi.com/2072-4292/9/7/676.

[28]   Ahad Hasan Tanim et al. "Flood Detection in Urban Areas Using Satellite Imagery and Machine Learning". In: *Water* 14.7 (Jan. 2022). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 1140. ISSN: 2073-4441. DOI: 10.3390/w14071140. URL: https://www.mdpi.com/2073-4441/14/7/1140 (visited on 07/02/2023).

[29]   J. Teng et al. "Flood inundation modelling: A review of methods, recent advances and uncertainty analysis". In: *Environmental Modelling & Software* 90 (Apr. 1, 2017), pp. 201–216. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2017.01.006. URL: https://www.sciencedirect.com/science/article/pii/S1364815216310040 (visited on 07/02/2023).

[30]   Uddin, Matin, and Meyer. "Operational Flood Mapping Using Multi-Temporal Sentinel-1 SAR Images: A Case Study from Bangladesh". In: *Remote Sensing* 11.13 (July 3, 2019), p. 1581. ISSN: 2072-4292. DOI: 10.3390/rs11131581. URL: https://www.mdpi.com/2072-4292/11/13/1581 (visited on 06/30/2023).

[31]   United States Geological Survey. *What is a digital elevation model (DEM)?* URL: https://www.usgs.gov/faqs/what-digital-elevation-model-dem (visited on 08/30/2023).

[32]   Hanqiu Xu. "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery". In: *International Journal of Remote Sensing* 27.14 (July 20, 2006), pp. 3025–3033. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431160600589179. URL: https://www.tandfonline.com/doi/full/10.1080/01431160600589179 (visited on 08/27/2023).