Mapping Flood Waters with Satellite Imagery and Topography using Convolutional Neural Networks

University College London

Department of Civil, Environmental and Geomatic Engineering

A thesis submitted in fulfilment of the requirements for the degree

of

Master of Science Geospatial Sciences (GIS & Computing)

of

University College London

Author: Christopher Koido-Bunt

Supervisor: Dr Aldo Lipani (University College London)

Subsidiary Supervisor: Dr Calogero Schillaci (JRC European Commission)

Submitted: 11th September 2021

Word Count: 11,310

Abstract

Using satellite imagery to automatically map flood extents in near-real-time is crucial for effective disaster response. Convolutional neural networks (CNNs) perform well in segmenting flood waters (FWs) at the global scale in medium-resolution (10m) multispectral imagery. Methods to improve FW mapping by incorporating ancillary data sources exist but can often only be implemented with hindsight or if the flooded area is known a priori. One ancillary data source which constrains the extent of FW is topographic data. However, its utility is limited to wetland studies at high spatial resolutions with little variety in flood events or topographic complexity.

Given the unpredictable locations of flood events, this study seeks to determine if the 90m resolution MERIT global topographic dataset can improve FW mapping when incorporated into the DeepLabV3+ CNN as an extra imagery channel. Additionally, this study will answer whether incorporating topographic data is more effective as a digital elevation model (DEM) or as the DEM-derived height above nearest drainage (HAND) terrain index (TI).

Sentinel-2 images from the World Floods FW-mapping benchmark dataset have been matched with DEM and HAND data from the MERIT topographic product. Matched scenes have been resampled to make two coregistered datasets at 90m and 10m resolutions. For each dataset a baseline CNN has been trained on (red, narrow near-infrared, short-wave infrared) tri-band imagery with other models additively including either DEM or HAND data.

Among the 90m resolution models, incorporating topography was found to have a negligible or even detrimental effect as the baseline model reported the highest mIOU (21.47%). However, 90m resolution model are thought to suffer from DeepLabV3+'s stepped dilation rate intervals of its atrous convolutions which incorporates redundant information. Results from the 10m resolution models suggest that the MERIT topographic data product can improve global FW mapping and that the HAND TI is the most useful (highest mIOU, 15.94%).

Acknowledgements

I would like to thank to Dr. Aldo Lipani and Dr Calogero Schillaci for their input and expert guidance. The pragmatic advice I received has kept my head above water during the research project and this study would not have been possible without them.

Thank you to my parents for their support and understanding under a particularly trying set of circumstances and accommodating my needs. I would like to thank Rachel, George, Gee and GEE for your comments, counsel, and cloud computing services. And a thank you to Kathryn for encouraging me to do what I wanted to do.

Last but not least, I give thanks to all the researchers who have completed the works my study uses for its foundations as I am but stood on the shoulders of giants.

Table of Contents

List of Figures

List of Tables

List of Equations

Mapping Flood Waters with Satellite Imagery and Topography using Convolutional Neural Networks

1 Introduction

1.1 Motivation

Flood events are ubiquitous global phenomena that can cause widespread damage to property and human life. Under a changing climate, flood risk in many flood-prone areas is projected to worsen. This is both in terms of intensity (Hirabayashi et al., 2013; Tabari, 2020) and in terms of exposed people and assets (Hirabayashi, Tanoue, et al., 2021). To inform stakeholders including disaster responders and reinsurers, disaster management organisations use satellite imagery to map large flood events in near real-time (NRT).

Medium-resolution multispectral (MS) image data from the Sentinel-2 satellite mission is of particular interest in mapping flood water (FW). This is due to its high spatial resolution for an open data source (~10m) coupled with its high revisit times when compared to often-proprietary radar imaging systems. With sun-synchronous continuous collection resulting in high volumes of remotely sensed (RS) data, automatable and scalable machine learning (ML) methods have grown in popularity over manual human-analyst approaches (Sun & Scanlon, 2019).

Accompanied by improvements in the availability of computing power, Convolutional Neural Networks (CNNs) have become highly effective in 'big data' Earth observation (EO) image tasks such as land use mapping (Hoeser & Kuenzer, 2020) and specifically FW mapping (e.g., Mateo-Garcia et al., 2021). Use of CNNs in EO is currently in an 'advanced transition phase' (Hoeser et al., 2020) with global benchmark datasets for FW mapping becoming available (e.g., Bonafilia et al., 2020; Mateo-Garcia et al., 2021). The creation of such datasets has invited different approaches to improve FW mapping; one approach has been fusing RS imagery with ancillary topographic data. Incorporation of digital elevation models (DEMs) (e.g., Muñoz et al., 2021) and DEM-derived terrain indices (TIs) (e.g., Du et al., 2020) utilise the constraining effect topography has on the movement of water (Cohen et al., 2019).

Previous studies incorporating topography report high overall classification accuracies of 95%-97% (Du et al., 2020; Muñoz et al., 2021, respectively) that improve upon using RS imagery alone. However, these studies are limited by the narrow variety of study area types and flood events considered, so the transferability of their findings have yet to be assessed on a global scale.

Whilst open MS imagery datasets have improved in accessibility and spatial and temporal resolutions, the same cannot be said for open global topographic datasets (Schumann & Bates, 2018, 2020) which are limited to coarse (30m-90m) resolutions. One open data product, the MERIT DEM (~90m resolution), is noted to perform similarly to the premium TanDEM-X DEM (~8m resolution) in flood modelling studies (Archer et al., 2018; McClean et al., 2020). Considering the MERIT DEM's ability to perform similarly to higher resolution topographic products, and the lack of advancement in open products, questions are raised about the utility of using existing topographic datasets with 10m resolution MS imagery to map FWs.

Therefore, this study aims to determine how effective incorporating open topographic data is on a global scale. This study will also assess whether the MERIT DEM can be used in applications at a finer scale than the original product and whether TIs or DEMs are more useful.

1.2 Aims

- To determine on a global scale whether incorporating topographic data as an additional channel of RS imagery improves FW segmentation in CNNs.
- To determine if a performance difference is observed when using 'raw' DEM elevation data or the 'derived' height above nearest drainage (HAND) TI.
- To determine whether coarse resolution topographic data is beneficial in FW segmentation at a finer resolution (10m) than that of the original product (90m).

1.3 Objectives

- Building a coregistered MS image and topographic dataset by obtaining MERIT DEM and HAND data from Google Earth Engine to match Sentinel-2 images in the World Floods dataset (Mateo-Garcia et al., 2021).
- The DeepLabV3+ CNN will be assessed in a multi-class segmentation task. Training will be undertaken on three-channel images with terrain models incorporating topographic data in an additional imagery channel. The models are as follows;

    - Baseline model: red (R), narrow near infrared (nNIR), short-wave infrared 1 (SWIR1).
    - DEM model: R, nNIR, SWIR1, DEM
    - TI model: R, nNIR, SWIR1, HAND
- The first set of models will be trained with 10m resolution imagery to assess the utility of the MERIT topographic product at spatial scales finer than its intended use.
- The second set of models will be trained with 90m resolution imagery to assess whether the effects of topography are scale dependent.
- The relative utility of terrain data in 'raw' (DEM) or 'derived' (TI) forms will be assessed by comparing model performance through a series of computer-vision metrics.

1.4 Overview

Chapter 2 will review the data sources and approaches commonly adopted in the literature in mapping FW. Chapter 3 outlines the details of the datasets, their pre-processing steps and model experiments. Chapter 4 reports model predictions on a test set of flood events. Chapter 5 discusses the results in context of wider literature. Finally, Chapter 6 provides a conclusion summarising the main findings of this study.

2 Literature Review

This chapter will introduce the concept of remote sensing and the data sources used to map flood water. Additionally, the mechanics of Convolutional Neural Networks and their contributions to mapping SW will be covered. This chapter ends with an assessment of the shortcomings of the literature about mapping FWs when using topography.

2.1 Introduction to Remote Sensing for Flood Water Mapping

Remote Sensing (RS), otherwise known as Earth Observation (EO) when at the planetary scale, is the field of study concerned with observing the surface of the Earth to obtain information about its characteristics, such as land cover (Hoeser & Kuenzer, 2020). Presently, a wide array of satellite missions and sensors exist which use different RS technologies to retrieve different forms of information about the Earth's surface. All RS platforms (airborne or spaceborne systems) utilise the principles of surface reflection of waves in the electro-magnetic (EM) spectrum. However, they can be broadly divided into two groups by their reflective mechanisms: optical and backscattering.

The optical group of RS technologies work on the principle of different objects reflecting varying intensities of light in the visible and infrared wavelengths of the EM spectrum. The reflection intensity across different wavelengths is recorded by the sensor to obtain a 'spectral signature' (Huete, 2004) of an object's reflectance to infer object properties. These RS technologies are considered 'passive' systems as they do not produce their own energy source and instead use that of the Sun.

The second group of RS technologies work on the analogous principle of interference and backscatter. These RS systems are 'active' systems that transmit their own EM waves using an onboard energy source. The emitted EM wave bounces off a target in a manner determined by an object's geometric or textural properties and the returning wave's intensity is recorded to infer object properties.

Both groups of RS technologies (summarised in Table 1) are used in mapping water (Bijeesh & Narasimhamurthy, 2020). Commonly used RS imagery and ancillary data sources (C. Huang et al., 2018; Musa et al., 2015) are expanded upon below.

Table 1 - Remote sensing (RS) technologies used in mapping water

RS Mechanism	RS Technology	System Energy Source	Predominant Use
Optical Reflectance	Visible (Red, Green, Blue)	Passive	LULC
	Multispectral	Passive	LULC
	Hyperspectral	Passive	LULC /Mineral classification
Backscattering	SAR	Active	LULC
	InSAR	Active	DEM development
	LiDAR	Active	DEM development

LULC–Land Use Land Cover classification, (In)SAR–(Interferometric) Synthetic Aperture Radar, LiDAR–Light Detection and Ranging

2.1.1 Identifying Methods for Mapping Surface Water and Flood Water

This study notes a subtle distinction exists between the problem tasks of mapping surface water (SW) and mapping flood water (FW). SWs are defined as normally inundated areas such as rivers, lakes and wetlands. Therefore, they can vary in spectral and temporal characteristics from FWs (Jain et al., 2020; Mateo-Garcia et al., 2021; Rambour et al., 2020), which are defined as inundation in normally dry areas.

Whilst there is an overlap in the techniques applicable to both problem tasks, due to scale and unpredictability of flood events, FW mapping studies must often rely solely on single-acquisition imagery from RS systems with wide fields of view. This constraint was used to focus the literature review.

A systematic literature review was carried out using the SCOPUS records of journal articles in Harzing's Publish or Perish search software using combinations of the query terms: 'earth observation', 'remote sensing', 'flood*', 'water', 'map*'. After an initial review of the methods used in FW mapping, the terms 'machine learning', 'CNN', 'DEM' and 'data fusion' were queried. The top 200 query results, ordered by annual citation count**,** were filtered based on title and abstract relevance. Exclusion criteria included journal abstracts which did not have clear research aims, objectives and findings (as defined by PRISMA guidelines) or those that were not relevant to near-real-time (NRT) mapping of FW with single-acquisition imagery.

2.2 Remote Sensing Data Sources for Flood Mapping

2.2.1 Multispectral Imagery and Spectral Indices

Multi-Spectral (MS) sensors measure reflection within the range of the EM spectrum that is visible to the human eye, and in the non-visible near infrared (NIR) and short-wave infrared (SWIR) wavelengths. The wider group of optical imagery includes; single channel panchromatic (black and white) imagery, three channel 'RGB' images (comprised of 'red', 'green', and 'blue' spectral bands), and multi-channel multispectral (MS) and hyperspectral (HS) imagery. Data from MS and HS imagery is similarly arranged to RGB imagery but across ~13 or several-hundred spectral bands, respectively.

Pixel based methods for water detection, including spectral indices (SI), have been developed by leveraging reflective information within the spectral signature of water (e.g., Equation 1). Water and non-water pixels can be classified by setting a site-specific class membership threshold (Sezgin & Sankur, 2004) on the spectral index. A comprehensive list of SI for mapping water is available in Bijeesh & Narasimhamurthy (2020, Table1, p3-4).

$$NDWI = (Green - NIR)/(Green + NIR)$$

Equation 1 - Normalised Differential Water Index,(McFeeters, 1996)

Optical RS technologies, however, suffer from occlusions such as cloud or dense vegetation (forests and bush) that can cause misclassifications. In addition, misclassification may occur because of the many different 'appearances' of water due to; sun-angle, shadow, submerged vegetation (Bijeesh and Narasimhamurthy, 2020), and particularly in dynamic FWs (Gómez-Palacios et al., 2017), turbidity.

2.2.2 Synthetic Aperture Radar Imagery

Synthetic aperture radar (SAR) is a microwave radar RS system that is highly effective in mapping water (Schumann & Moller, 2015). SAR is an active system which is not subject to atmospheric lighting conditions and its microwave frequencies can penetrate cloud, which allow constant observation of ephemeral FWs (Schumann & Moller, 2015).

In SAR images water can be identified by its 'smooth' appearance with little backscatter. However, because information recorded by SAR sensors relates to object structural and surface roughness properties, misclassifications can occur without post-processing (Schumann & Moller, 2015). Misclassification of 'smooth' non-inundated areas (e.g., roads, Rambour et al., 2020) and conversely 'rough' inundated areas (areas affected by wind-driven waves, Mason et al., 2018; Vickers et al., 2019) are both possible.

Like optical imagery, SAR imagery suffers from interference from vegetation and pixel-wise classification is usually undertaken by setting a reflectance threshold (Manavalan, 2017). As SAR data is not the main data source in this study, the reader is referred to Manavalan (2017) for a detailed overview of methods used to map waterbodies in SAR images.

2.2.3 Topographic Data: Digital Elevation Models and Terrain Indices

As RS imagery suffers from occlusions, a wide body of literature has sought to incorporate topographic data to better map water (section 2.6.1). The importance of topographic data is supported by the field of numerical flood modelling which has established the driving forces affecting the distribution of FWs are inertia (surface friction) and topography-directed gravity (Cohen et al., 2018, 2019). RS technologies such as LiDAR (Gomes Pereira & Wicherson, 1999) or interferometric SAR (InSAR) (Faherty et al., 2020) can generate three-dimensional representations of the Earth's topography in the form of digital elevation models (DEMs) or their derived terrain indices (TIs).

DEMs represent absolute elevation values and include both 'as is' digital surface models (DSMs) and 'bare earth' digital terrain models (DTMs). DSMs contain surface features such as buildings and vegetation while DTMs have had these features removed. TIs developed for hydrology provide specific subsets of topographical characteristics such as gravity-based drainage direction maps (e.g., Yamazaki et al., 2019) or relative height above the nearest drainage network (Rennó et al., 2008).

Some topographic data sources commonly used in mapping water with RS imagery are provided in Table 2.

Table 2 - Common sources of terrain data used in mapping water with RS imagery

Topographic Information	Description
Digital Elevation Model (DEM)	Elevation profile expressed as height above reference datum. Includes DSMs and DTMs.
Topographic Wetness Index (TWI, Quinn et al., 1991)	TWI is a TI for soil moisture mapping that describes the tendency of a pixel to accumulate water. TWI uses pixel-specific (up)stream catchment area (SCA) and slope angle (φ) under the principle of mass-balance. $$TWI = \ln\left(\frac{SCA}{tan\ \varphi}\right)$$ *Equation 2 – Topographic Wetness Index formula*
Multi-Resolution Valley Bottom Flatness (MrVBF, Gallant & Dowling, 2003)	MrVBF is a TI measuring topographic flatness on valley bottoms at multiple scales for mapping hydrogeomorphological deposits.
Height Above Nearest Drainage (HAND, Rennó et al., 2008)	Values of the HAND TI represent the relative height of each pixel above the elevation of its nearest downstream drainage pixel. This requires hydrologically adjusting elevation to remove sinks.

While the use of DEMs allows satellite water altimetry measurements to be used, they have been omitted from this study as their coincidence with flood events are rare (Musa et al., 2015) and because in the absence of numerical flood models altimetry data is only useful in 'bathtub approaches' (section 2.6) to mapping FW.

2.3 Machine Learning: Artificial Neural Networks

As RS data is spatially extensive and nearly continuously collected, the field of EO has moved from traditional manual approaches towards using automatable and scalable tools such as machine learning (ML) models (Sun & Scanlon, 2019; X. X. Zhu et al., 2017). ML models can be summarised as algorithms which learn a target function (the problem to solve) from experience. In supervised learning approaches the relationships of an input's features (e.g., pixel attributes) to its label (classification) are learned from a training dataset with performance evaluated on an unseen test dataset. While varied in internal mechanisms, ML models can be split into two groups: weak learners like decision trees (DTs) or support vector machines (SVMs), and deep learners like artificial neural networks (ANNs).

2.3.1 Network Structure and Learning

ANNs emulate the structure of the mammalian brain and are comprised of networks of interconnected neurons (Figure 1b) (Rosenblatt, 1958). Based on the perceptron model (Rosenblatt, 1958), each neuron (Figure 1a) intakes a vector of feature values (x) with multiplicative weights (w) and an additive bias term (b).

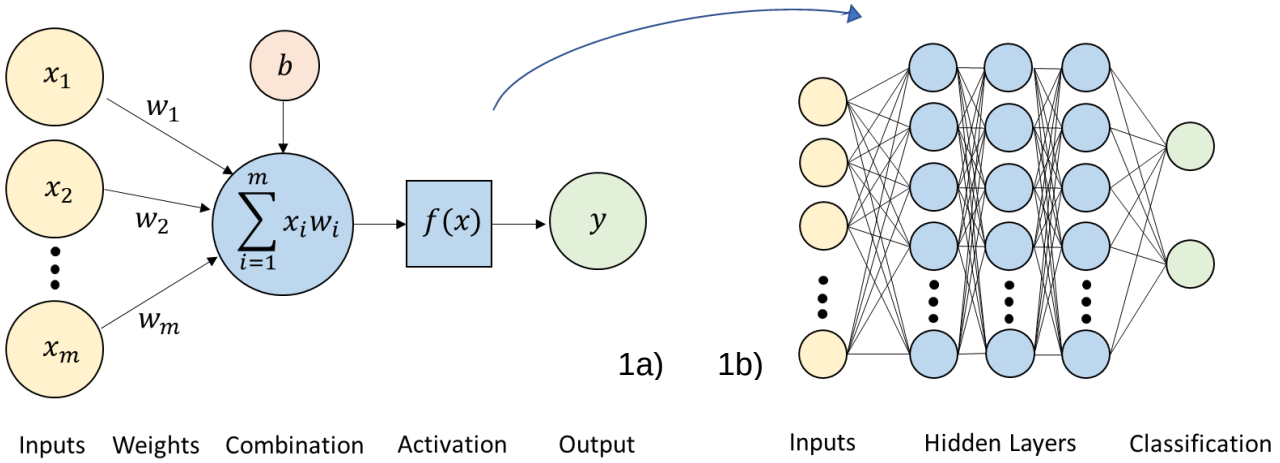


Figure 1 – Individual neuron structure (1a) and neuron arrangement in a network (1b)

For each input instance, features are 'fed forward' through a neuron which sums weights and adds bias before calculating output through an activation function. This output (e.g., a classification prediction) is then compared to the target's label in the loss function to calculate the size of the loss (error).

During training, loss is 'backpropagated' through each neuron in the network to adjust the weights proportionally. Therefore, given the same input features on the next iteration, the network will output a prediction with a smaller loss.

2.3.2 Hyperparameters in Model Training

Hyperparameters are a range of model parameters which control the learning process and affect the model's predictive performance.

Activation Functions determine neuron output. Two common activation functions are the rectified linear unit (ReLU) and sigmoid functions. The ReLU function (Equation 3) is popular due to its computational simplicity and its prevention of vanishing gradient descents; both important considerations in deep networks (Rasamoelina et al., 2020).

$$r(x) = max(0, x)$$

Equation 3 – Rectified Linear Unit (ReLU) activation function

The sigmoid function is slower to implement within the network (Han Jun and Moraga, 1995), it is often used in classification output to represent class membership probability as it returns a value between 0 and 1. In multi-class problems (where the labels are mutually exclusive) the generalised form of the sigmoid function, the softmax function, ensures that probabilistic output for each class sums to 1.

$$\sigma(x) = \frac{1}{1 + e^{-x_i}} \qquad s(x) = \frac{e^{x_i}}{\sum_{j=0}^{k} e^{x_j}}$$

Equation 2 – Sigmoid activation function *Equation 3 – Softmax activation function*
Where x is the weighted attribute, k is the number of classes.

All three activation functions allow non-linearity to be added to the model's outputs. Non-linearity is fundamental for solving complex ML problems which do not have linearly separable classification boundaries (Mitchell, 1997).

Loss Functions describe the measure of fit of a model to the data. While different problem tasks will achieve optimum results with different types of loss functions, a shared characteristic of loss functions is that they must be convex and therefore differentiable.

Gradient Descent is the iterative optimisation process used in reaching a minima of a loss function (Polyak, 1987). In each iteration, the loss function's gradient is calculated through differentiation and the weights of each feature are updated so the next prediction will take a 'step' in the direction of the steepest gradient. Depending on the complexity of the loss surface (Figure 2), the random initialisation of weights during training can cause some models to become stuck in local minima rather than reaching the global minima (the optimum solution).

Optimisers - As minimising the loss is mathematically an optimisation problem, different 'optimiser' algorithms can be applied. The Adaptive Moment Estimation (Adam) algorithm is highly suitable for complex loss surfaces as it is less sensitive to noisy or shallow gradients and so less likely to get stuck on saddle points (Figure 2) (Kingma and Ba, 2014).

Learning Rates and Momentum - The learning rate controls the size of the step (Figure 2) taken towards the minima of the loss function. The momentum can increase the step size (Figure 2) based on the gradients of past steps (Polyak, 1987). Large learning rates and momentum parameters result in spiky learning where model predictions overshoot the minima, as seen in training loss curves (section 4).

Epochs and Batchsize - Batch learning calculates average loss over a batch of instances before taking a step (Mitchell, 1997). This results in smoother training curves which are less likely to get stuck in local minima as models learn relationships applicable to more instances rather than instance-specific relationships (Mitchell, 1997). Training a model for multiple epochs (a single cycle through the training dataset) allows for instances to be reused and relationships to be better learned.

2.3.3 Training considerations: Under- & Over-fitting and Data Leakage

A model with a poor fit to the training data is said to 'underfit' and training for multiple epochs can improve model fit. As instances are re-used during training it is important to stop training before the model 'overfits' to the training set. Overfitting models fit the training data well but 'generalise' poorly on unseen data as they go beyond learning general relationships within the data and have since learned non-general relationships specific to that particular training set. Therefore, an independent validation set is used to monitor overfitting and to obtain a model with a good trade-off between overfitting and underfitting (Geman et al., 1992). Training can be ended early should losses on the validation set increase (section 4).

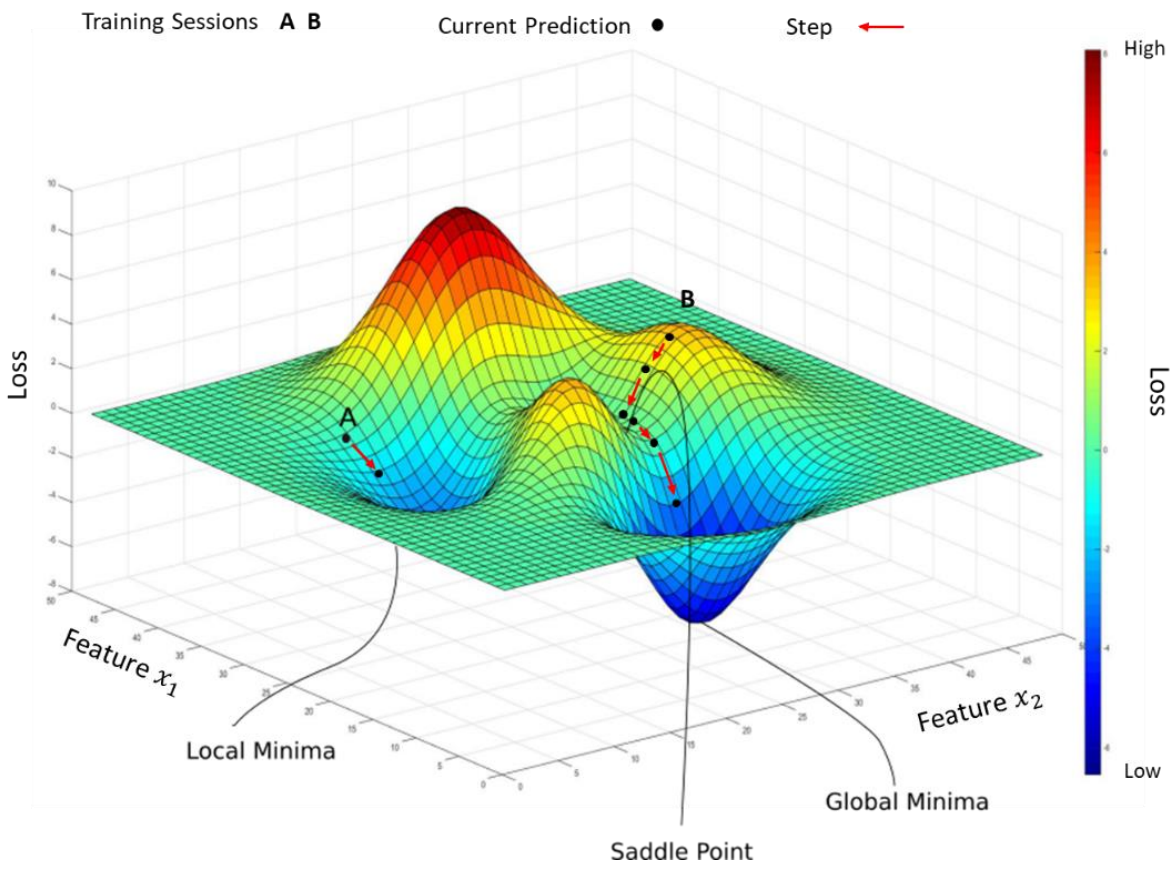


Figure 2 - Visualisation of Gradient Descent across an arbitrary loss surface for two training sessions. Adapted from Kathuria 2018 blogpost

Noise is anything that obscures the relationship between an instance's feature and its label (Frénay & Verleysen, 2014) and can be applied to the feature or label. While excessive noise increases error (Frénay & Verleysen, 2014), introduction of limited noise can act to regulate the network and reduce overfitting (Bishop, 1995) by learning general relationships which should be nearly correct for a label instance with typical pixel distributions (Angluin & Laird, 1988; Valiant, 1984).

Data augmentation is another common method in deep learning to reduce overfitting (Kukačka et al., 2017). Augmentation can yield better performances (Yu et al., 2017) than non-augmented datasets or achieve similar results with less training data (Angluin & Laird, 1988). In computer vision tasks augmentation can be achieved through random (noisy) geometric or colour alterations to features (e.g., Shorten & Khoshgoftaar, 2019; Yu et al., 2017).

Data leakage is an unintended transfer of information between independent datasets which violates the principle of a fair 'blind' test (Larsen & Becker, 2018).

2.3.4 Image Segmentation and Performance Metrics

The task of mapping water in terms of computer vision tasks is an image segmentation problem. As illustrated in Figure 3, this problem involves assigning each pixel in the image a class, as opposed to identifying; a label for the image (image recognition), a bounding box of flooded areas (object detection) or segmenting individual water bodies (instance segmentation).

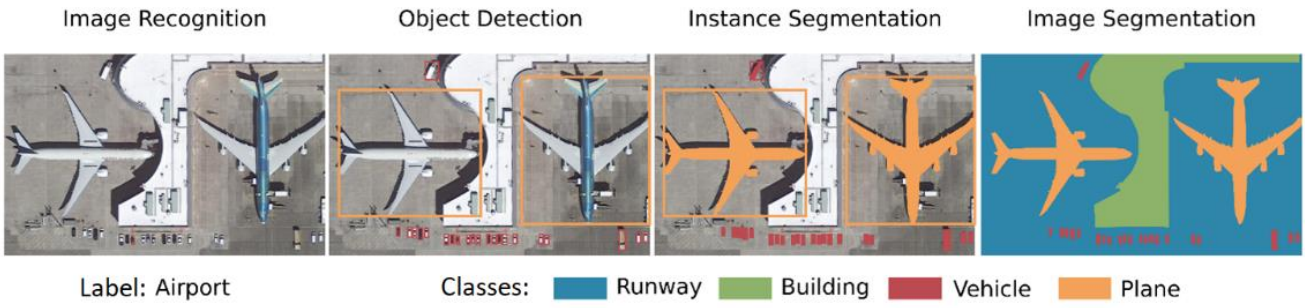


Figure 3 – Examples of computer vision classification tasks. Adapted from Hoeser and Kuenzer 2020, Figure 3. The example image is from the DOTA dataset (Xia et al., 2018)

Two common performance metrics in segmentation tasks are F1-scores and intersection over union (IoU). These metrics have binary and multi-class versions with the former listed in Table 3. For an extension to multi-class metrics see Grandini et al., (2020).

Table 3 - classification metrics used in this study.

Metric	Description	Formula
Number of classes	Total number of classes	k
True Positive	The sum of pixels correctly classified as water	TP
True Negative	The sum of pixels correctly classified as non-water	TN
False Positive	The sum of pixels incorrectly classified as water	FP
False Negative	The sum of pixels incorrectly classified as non-water	FN
Precision	The proportion of all pixels predicted as water that were correctly classified	$\frac{TP}{TP + FP}$
Recall	The proportion of all 'real' water pixels that were correctly classified	$\frac{TP}{TP + FN}$
Accuracy	The proportion of correctly classified instances	$\frac{TP + TN}{TP + TN + FP + FN}$
F1 Score	The harmonic mean of the precision and recall	$2 \times \frac{precision \times recall}{precision + recall}$
Intersection over Union	The intersection over union of a class	$\frac{TP}{TP + FP + FN}$
Mean Intersection over Union (mIOU)	The mean intersection over union of k classes	$\frac{1}{k}\sum_{k=1}^{k}\frac{Area_{k\ pred} \cap Area_{k\ label}}{Area_{k\ pred} \cup Area_{k\ label}}$

2.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) (Hubel & Wiesel, 1962) are a type of ANN designed to handle imagery data arranged in multiple arrays (LeCun et al., 2015). They can achieve high performance in RS classification tasks such as land use land cover (LULC) mapping (Carranza-García et al., 2019; Y. Li et al., 2018). In the task of segmenting water CNNs can outperform weak-leaners (Rezaee et al., 2018; Sarker et al., 2019) and even traditional SIs (James et al., 2021; Mateo-Garcia et al., 2021) as CNNs learn multi-level abstract representations (i.e., edges, object parts, and patterns) in the image data (LeCun et al., 2015).

CNN architectures vary among design 'families' but all utilise convolutional and pooling layers in a 'convolutional backbone'. The layer types used within the CNN implemented in this study are outlined below.

2.4.1 Convolution Layer

In a convolution layer a sliding window (kernel) with learnable multiplicative weights moves across the image (Figure 4). By traversing across height and width dimensions, convolutional layers act as feature extractors to generate stacks of feature maps through element-wise matrix multiplication. Kernel parameters affect the size of the output and include its size (field of view), stride (traversal interval), and padding. Padding controls how the kernel treats image borders where without padding kernels crop images to produce a feature map that is smaller than the input, termed down sampling.

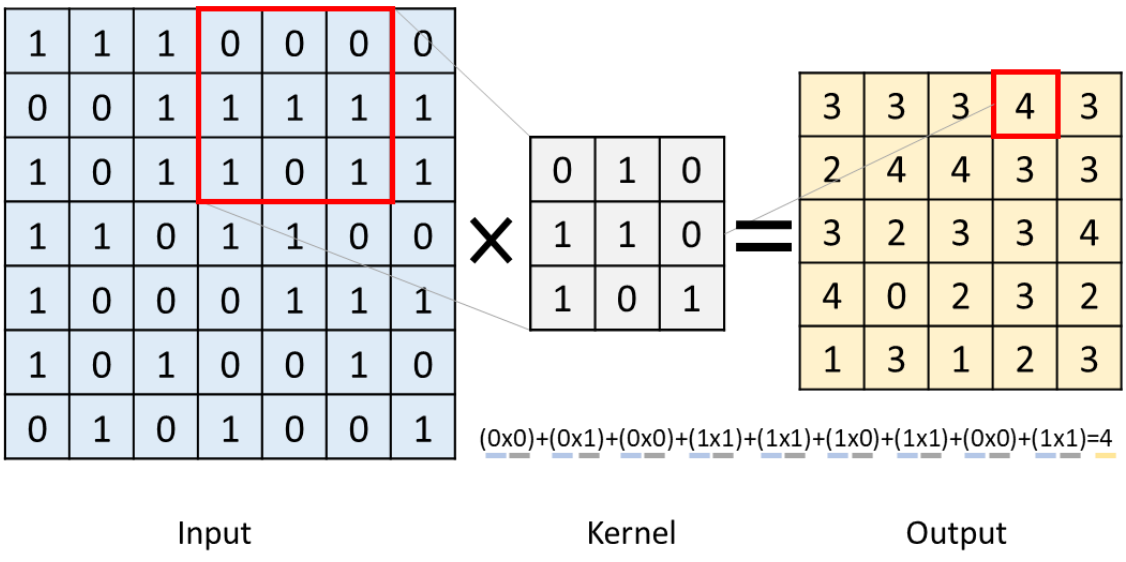


Figure 4 - Illustration of a 3x3 kernel convolving across a 7x7 input to produce a down sampled 5x5 output (padding=0)

Convolutions can be depthwise (applying the same operation to each channel, Figure 4) and pointwise (Figure 5) where the latter can reduce or expand the number of channels for each element in the image data.

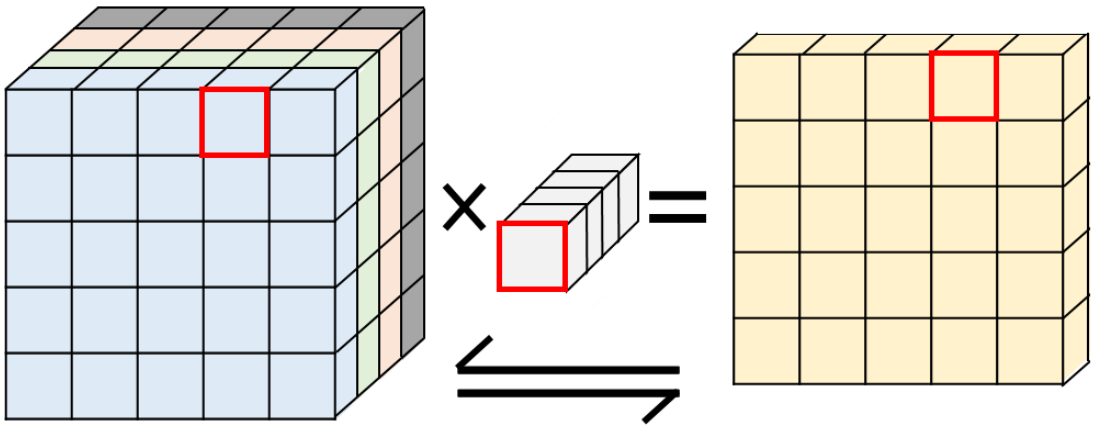


Figure 5 - Pointwise convolution producing a single channel feature map from four channels

2.4.2 Pooling Layer

Pooling layers summarise information across a pool of elements according to a set kernel size. These layers reduce the number of learnable parameters, and subsequently computational cost, while only keeping useful information (Gholamalinezhad & Khosravi, 2020). After multiple convolutions the original location of the summarised information is not retained which allows the model to identify features regardless of their location within the image (termed spatial invariance) (Gholamalinezhad & Khosravi, 2020). Summary can be undertaken through maximum or average pooling (Figure 6) among other forms.

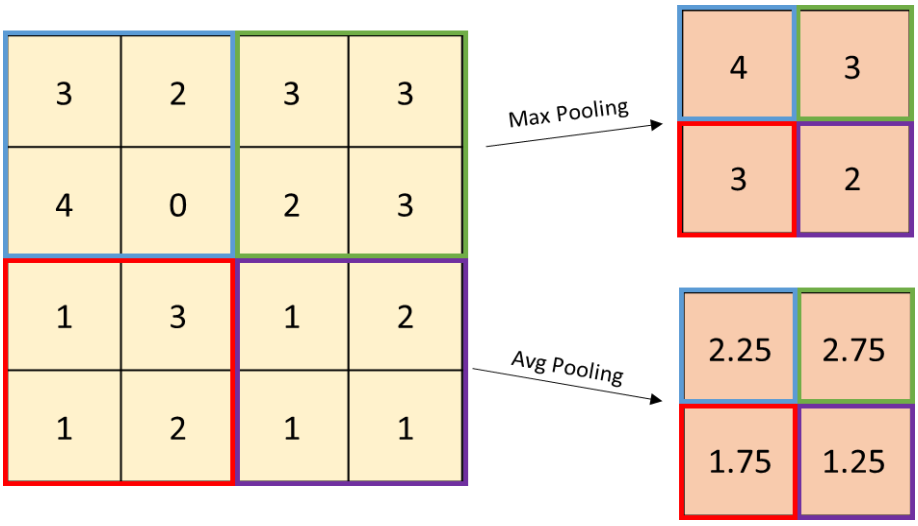


Figure 6 - Example of maximum and average pooling with a 2x2 kernel

2.4.3 Atrous Spatial Pyramidal Pooling

Atrous spatial pyramidal pooling (ASPP) employs a series of dilated kernels in atrous convolutions (Figure 7). By using kernels 'with holes' (*'à trous' - French*), combination (concatenation) of the produced feature maps allow incorporation of semantic information at a range of spatial scales without excessively increasing computational burden (Chen et al., 2017).

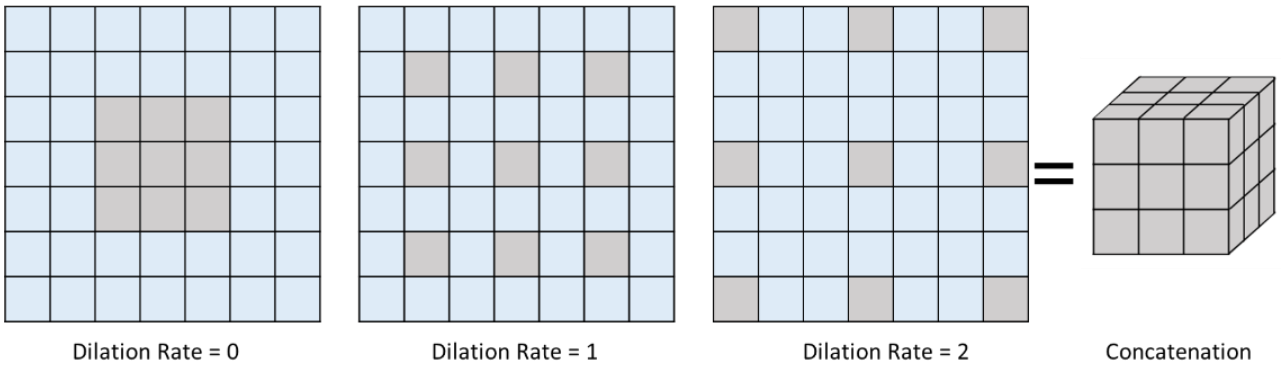


Figure 7 - Atrous Spatial Pyramidal Pooling for three 3x3 kernels with illustrative dilation rates

2.4.4 Fully Connected Layer and Output layer

After the convolutional backbone, the fully connected layer in a 'vintage' (Hoeser & Kuenzer, 2020) CNN feeds information through 'hidden' layers of neurons of arbitrary width (number of neurons in a layer) and depth (number of hidden layers in the network) to an output layer forming the classification (Figure 1).

2.4.5 Convolutional Backbones

Various CNN layer types are combined to form convolutional backbones, or feature encoders. Only the approaches used in the study model are outlined but for a coverage of CNN architectures used in EO studies see Hoeser & Kuenzer (2020).

Encoders in the ResNet family use 'residual blocks' during convolution that use residual (or 'skip') connections that retain information from feature maps at multiple levels (Figure 8a). The ResNet-101 encoder used by Chen et al., (2017, 2018) uses batch normalisation after each convolution and before the activation function. By normalising the mean and variance of each instance in the batch, numerical values within the batch are kept stable while only the relationships between instances and instance values change (Ioffe & Szegedy, 2015). This allows for higher learning rates to be implemented resulting in faster training times and also higher model accuracies (e.g., He et al., 2015; Ioffe & Szegedy, 2015).

The Xception encoder uses parallel kernels of different sizes and depthwise separable convolutions (a depthwise convolution followed by a pointwise convolution, Figure 8b) to extract features (Chollet, 2016). Depthwise separable convolutions allow cross channel correlations and spatial correlations to be mapped separately to learn channel-specific patterns (Chollet, 2016). The Xception encoder is arranged into staged entry, middle, and exit blocks that, while varying in channel number and stride, utilise ResNet 'residual blocks' (Figure 8c) and can similarly implement batch normalisation (BN) (Chen, Papandreou, et al., 2018).

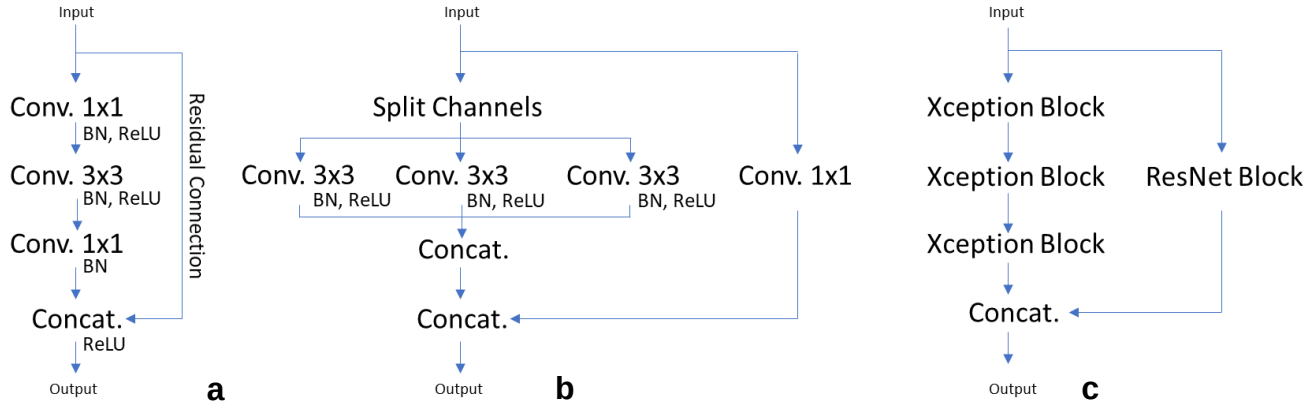


Figure 8 - a) a ResNet block, b) an Xception block and their arrangement within c) a generic staged block in the Xception backbone., Concat. - concatenation

2.5 CNN Architectures for Semantic Segmentation

Model architectures for semantic segmentation can be broadly split into: Naïve Decoders and Encoder-Decoders.

Naïve Decoders use simple bilinear interpolation on the output of the encoder to restore the resolution of the prediction to that of the original image (termed upsampling). Naïve Decoders, however, struggle to produce fine-grained segmentation predictions as the encoder module reduces the input resolution significantly (Long et al., 2014).

Encoder-Decoder architectures, like the seminal U-Net model, add a symmetrical decoding (upsampling) path to mirror the encoding (downsampling) path. The decoder gradually restores resolution through deconvolutions while using skip connections to transfer information from feature maps at multiple depth levels to obtain fine-grained output predictions (Ronneberger et al., 2015).

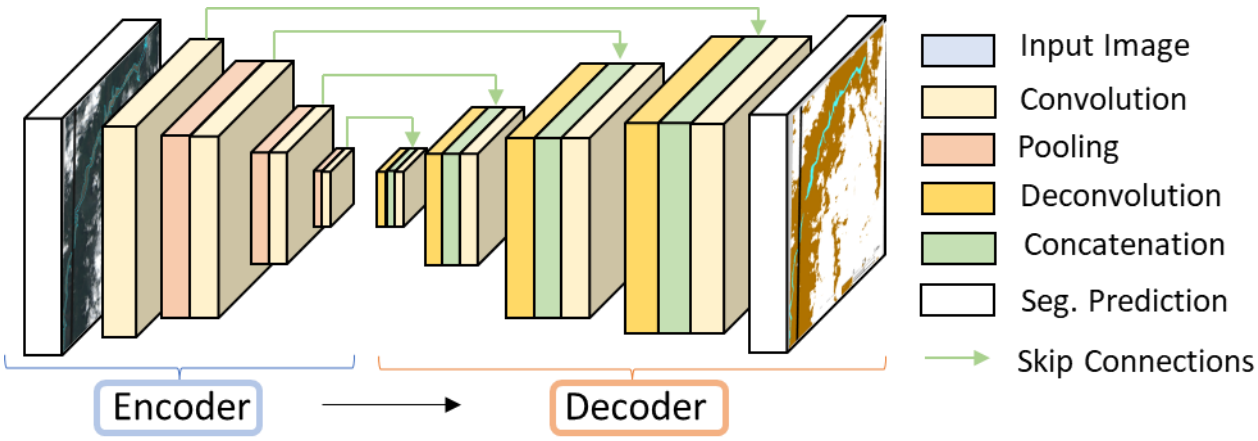


Figure 9 - Illustrative example of encoder-decoder model that is four convolutional layers deep. Seg - segmentation

2.6 Mapping Water: Incorporating Topographic Data

Previous work that has incorporated topographic data has done so in limited ways using non-ML or weak-ML approaches. Non-ML approaches typically use topography as a post-processing step to improve classifications derived solely from RS imagery. 'Mask' approaches filter pixels by absolute or relative elevations which are unlikely to be water (commission errors) (Chow et al., 2016; Cohen et al., 2019; C. Huang et al., 2017; Martinis et al., 2015; Song et al., 2007; Thissen, 2019). 'Bathtub' approaches use simple assumptions of planar water heights and open hydraulic connectivity to 'grow' water classified regions in occluded areas (omission errors) from visible flooded pixels that are known to be at equal or lower elevations (Cohen et al., 2018; Hashemi-Beni & Gebrehiwot, 2021; C. Huang et al., 2014; Wang et al., 2002).

Weak-ML approaches can directly incorporate topographic information into the classification (e.g., Aristizabal et al., 2020; Hird et al., 2017; Irwin et al., 2017) and have reported classification improvements over using RS imagery alone (+36%, HAND, Aristizabal et al., 2020). However, these approaches treat each pixel in isolation and are known to be outperformed by context-aware deep learning approaches (Jiang, 2020; Rezaee et al., 2018; Sarker et al., 2019).

While CNNs have been used to successfully segment water with MS (James et al., 2021; Z. Li et al., 2019; Mateo-Garcia et al., 2021; Sarker et al., 2019) or SAR (Bonafilia et al., 2020; Kang et al., 2018; Nemni et al., 2020) imagery, only the works of Hosseiny et al., (2019), Du et al., (2020), Muñoz et al., (2021) have tried to incorporate topographic data. As the field is so small, conclusions are limited by the disparate nature of the methodologies of these studies.

2.6.1 Gap Analysis

While Hosseiny et al., (2019) directly input the DEM into CNNs, the authors apply change detection approaches to SAR imagery, so their work differs from the focus of this study.

Muñoz et al., (2021) incorporate DEMs into their CNNs when using LandSat MS images to map compound flooding at 30m resolution in coastal wetlands of the Atlantic US. Their comprehensive ablation tests show that incorporating a DEM does improve overall accuracy by 2-4 percentage points when using either, all 13 Landsat bands or just the NIR, SWIR1 and SWRI2 bands, respectively.

Similarly at a finer resolution of 2m, Du et al., (2020) show use of a LiDAR DEM with WorldView-3 MS imagery improves overall accuracy from 92% to 95% in a single-catchment wetland area. Du et al., (2020), go further and assess the effects of the TWI TI in ablation studies. Substitution of the DEM for TWI is found not to improve model overall accuracy past 95%, but slightly improves median IoU scores when compared to using the DEM alone, or the DEM in conjunction with TWI (Table 4). This suggests the TI better represents information useful for the problem task than the noisy raw DEM.

While the beneficial effects of topography are supported by both studies, they each have their limitations that limit the transferability of their findings to other settings. In the Muñoz et al., (2021) study there is possible data leakage between features and labels. In addition to inputting a DEM as a feature, their flood extent labels are generated by a hydrodynamic model which uses a similar LiDAR-derived DEM (NOAA's CUDEM, Muñoz et al., 2021). Additionally, unlike Du et al.'s, (2020) study which uses a U-Net segmentation model, Muñoz et al's., (2021) model architecture uses a pixel-wise classification approach with separate convolutional arms for MS imagery and topographic data sources. Therefore, it is unclear if results obtained in both studies are specific to the model architectures in the study areas selected.

Finally, both studies are limited to high resolution imagery in flat wetlands with little variety in the flood events or topographic settings. Therefore, it is unknown if including DEMs, or other TIs, can improve FW mapping on a global scale across a range of study areas in medium resolution imagery.

Table 4 - A selection of model tests that use topographic data in CNNs to segment water. -- indicates decimals not reported

Author	k	Model	Overall Accuracy (%)	F1-Score (macro) (%)	Median IoU (%)
Muñoz et al., 2021b	8	MS (13 bands)	95.73	95.57	-
		MS (13 bands)+DEM	97.09	96.97	-
		MS (nNIR,SWIR1,SWIR2)	92.37	91.82	-
		MS (nNIR,SWIR1,SWIR2)+DEM	96.62	96.49	-
Du et al., 2020	2	MS (8 bands)	92.--	91.--	66.--
		MS+DEM	95.--	95.--	69.--
		MS+TWI	95.--	95.--	70.--
		MS+TWI+DEM	95.--	94.--	68.--

3 Data and Methodology

This section details the datasets used and their pre-processing steps. The intercomparison experimental setup is outlined along with the training regimes implemented. This includes the hyperparameters and data splits used in each model resolution group. The evaluation methods on the test flood events are described. The overall methodology workflow can be summarised in Figure 10.

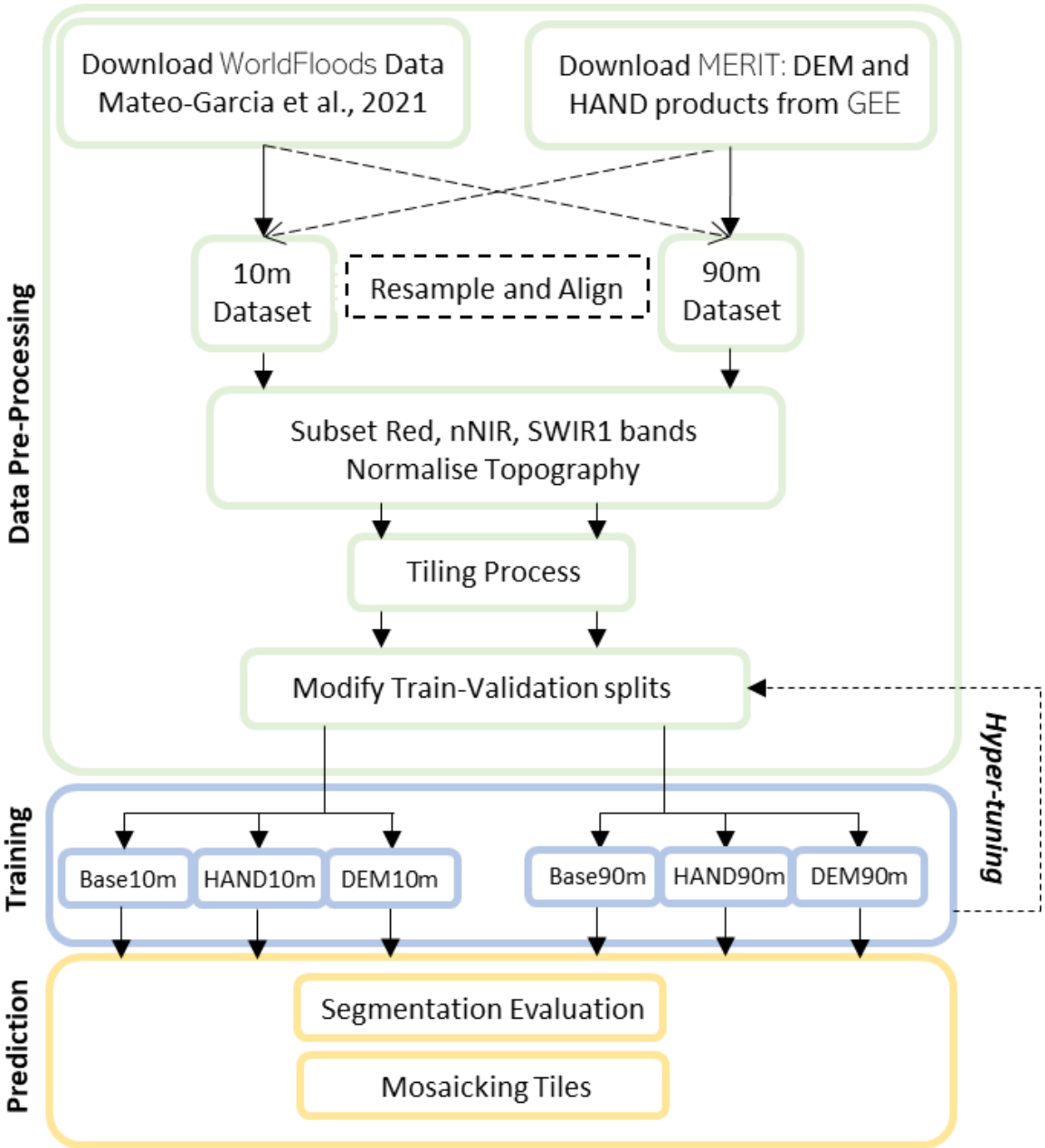


Figure 10 – Methodology Workflow

3.1 Datasets

The RS imagery is a curated selection of Sentinel-2 MS imagery from the World Floods dataset (Mateo-Garcia et al., 2020). The HAND TI (Yamazaki, et al., 2019) is derived from the MERIT DEM product (Yamazaki, et al., 2017) and both terrain datasets were downloaded via the Google Earth Engine (GEE, Gorelick, et al., 2017) data catalogue. The details of the datasets are provided below.

3.1.1 MERIT Digital Elevation Model: DEM and HAND

MERIT: DEM

The Multi-Error-Removed Improved Terrain (MERIT) DEM is an open-access bare-earth DTM with global coverage (Yamazaki et al., 2017). This DTM is a composite data product comprised of SRTMv2.1 and ALOS AW3D30m DSM information and supplemented with a coarse Viewfinder Panoramas DSM to fill voids (0.5% contribution). This dataset has higher absolute and relative accuracies than its components by removing the main sources of error in interferometry derived DEMs (absolute bias, stripe noise, speckle noise and tree-heights) (Yamazaki et al., 2017).

A drawback is that after error-removal, the final product is a DTM at 90m resolution while its DSM inputs were at 30m resolution. However, this data product has been selected as post-processed DTMs have been shown to be more suitable than unprocessed DSMs for hydrological modelling applications at global scales (Jarihani et al., 2015). This is as the latter include surface artefacts which obstruct flow.

Furthermore, two separate DEM intercomparison studies demonstrated that the MERIT product is found to be better at modelling flood extents in urban (Carlisle, UK, McClean et al., 2020) and rural (Ba catchment, Fiji, Archer et al., 2018) study areas (when comparing against TanDEM-X, AW3D30m, ASTER, SRTM DEMs).

MERIT: HAND

From the MERIT DEM, Yamazaki et al. (2019) have released a global hydrography dataset, MERIT Hydro, which includes the HAND TI. The HAND TI is produced by ensuring artificial enforcement of sub-pixel scale drainage networks in the coarse global product. This layer is an improvement on previous global hydrography datasets such as HydroSHEDS (Lehner et al., 2008, derived from the SRTMv3 DEM) in terms of coverage and representation of small streams (Yamazaki et al., 2019).

Of the available products in MERIT Hydro, such as flow direction maps, only the HAND TI has been used. This is because flow direction maps are more suited for running process-based flood models across entire catchments (e.g., Hirabayashi et al., 2021). In Sentinel-2 images of incomplete river catchments and without access to process-based methods, the author of this study theorises that height in relation to the river, is the most important MERIT Hydro layer in determining if the pixel is flooded.

The HAND TI has been selected over other TIs like TWI and MrVBF. While informative in wetland studies (Du et al., 2020; Hird et al., 2017), TWI only considers only upstream contributing areas and has even been shown to be unfit for purpose in flat areas (Western et al., 1999). MrVBF, which is also useful in SW (Yang et al., 2017) and FW (Aristizabal et al., 2020) mapping, has not been used as HAND is found to outperform MrVBF across a variety of study areas as a post-processing mask, (C. Huang et al., 2017).

Finally, by using a TI derived from the same DEM, no uncertainties in coregistration, absolute or relative error are introduced by using different data products

3.1.2 Sentinel-2 Multispectral Imagery (WorldFloods)

The World Floods dataset of Mateo-Garcia et al. (2021) comprises of processed Sentinel-2 imagery and georeferenced labels covering 119 globally verified flood events between November 2015 and March 2019. These events were sourced from international disaster management agencies; Copernicus Emergency Management Service, United Nations Satellite Centre flood portal and the Global Flood Inundation Map Repository. This provides a globally varied dataset observing flood events in a range of landscapes.

Features are taken from Sentinel-2 imagery which is found to be superior to similar MS sensors (Table 5) as Sentinel-2 possesses the highest; repeat time, spatial and spectral resolutions, and largest swath width (field of view). This allows for a wider range of flood events to be captured and more information gleaned from Sentinel-2 images than other sensors.

Table 5 - Comparison of Sentinel-2 with important heritage missions

Attribute	LANDSAT 1-7, 8	SPOT 7	SENTINEL-2
Access	Open	Usually Premium	Open
Repeat cycle (days)	16	26	5*
Swath width (km)	185	120 (2 x 60)	290
Spectral bands	7, 11	4 (VIS, NIR)	13
Spatial resolution (metres)	30, 60	1.5**, 5, 20	10, 20, 60
Source	USGS (n.d.)	ESA (2021)	ESA (2015, Table1, p35)

**at the equator, when using both Sentinel vehicles in cloud-free conditions*

***attained from two 5m panchromatic sensors and so not useful in this study*

Labels have been produced using manual and semi-automatic photo interpretation utilising official flood maps. These labels have mostly been generated by radar-based satellites (e.g., SAR, Mateo-Garcia et al., 2021, Figure 5a, p5). Given the suitability of SAR in mapping flood water due to its cloud (Schumann & Moller, 2015), and even vegetation (Hess et al., 2003), penetrating ability, SAR can be taken as the best available data source to provide labels at the required scale.

Considering both features and labels are provided from RS data, with reliance on unspecified 'automatic' methods, the term 'ground truth' is explicitly avoided by Mateo-Garcia et al. (2021). This introduces two forms of label noise. Firstly, through temporal misalignment between acquisition of feature imagery and label imagery. Secondly through 'weakly' labelled data using unspecified automated approaches.

However, the temporal misalignment is experimentally validated by Mateo-Garcia et al. (2021) to be beneficial as worse performance is achieved with a model trained only on the fewer temporally aligned samples. Additionally, a similar study segmenting FWs (Bonafilia et al., 2020) finds that, given enough training instances, assigning 'weak' automatic labels to Sentinel-1 or Sentinel-2 imagery results in better performance than when training on few hand-labelled images alone. The World Floods test and validation sets include flood events with no temporal misalignment between the feature generating Sentinel-2 imagery and the label generating SAR imagery (Mateo-Garcia et al., 2021). This has been undertaken to assess model performance under conditions of NRT applications.

Because MS imagery suffers from occlusions this dataset has implemented the official Sentinel Hub's (2019) automatic cloud detector to simplify labels into; 0: No data, 1: Land, 2: Flood, 3: Cloud.

3.2 Band selection and input to CNN channels

To avoid issues of data redundancy in high-dimensional data and to speed up training times, the bands used in this study were reduced to those found by Jain et al. (2020) to be the most effective at classifying water. Jain et al. (2020) iteratively tested 120 tri-band combinations of Sentinel-2 data in mapping FW using a CNN with a ResNet-18 backbone. The authors found that the two best-performing band combinations; 1) Red, nNIR, SWIR1 and 2) Red, SWIR1, Blue, both produced F1-scores of 96%. Despite similar performance, the former was selected as its component bands have been used to successfully map FWs in studies using CNNs (Muñoz et al., 2021) and in SI (Bijeesh & Narasimhamurthy, 2020).

The 20m resolution nNIR and SWIR1 bands distinguish water well from other classes but provide less spatial information than other bands. Therefore, similar to the concept of pan-sharpening (Ranchin & Wald, 2000), incorporation of the 10m resolution Red band is reported to aid boundary delineation when mapping water (Yang et al., 2017).

Table 6 summarises the data sources and bands input into the CNNs.

Table 6 - The data sources and bands input into CNN models

Data source	Input channel into CNNs	Band width (nm)	Original resolution
Sentinel-2 Imagery (Mateo-Garcia et al.,, 2021 via WorldFloods)	Red	30	10m
	nNIR	115	20m
	SWIR1	90	20m
MERIT DEM (Yamazaki et al., 2017, Yamazaki et al., 2019, via GEE)	DTM	-	90m
	HAND	-	90m

3.3 Experimental Setup

To evaluate the relative contribution of topographic data to semantic segmentation of FW, two sets of baseline and additive experiments were designed (Table 7). Muñoz et al. (2021) and Du et al. (2020) suggest the beneficial effects of topography are present at spatial resolutions of 30m and 2m, respectively. Mateo-Garcia et al. (2021, Table2) note that resampling MS imagery to 80m resolution only causes a several percent decrease in

performance across all models. Therefore, the experiments are repeated at scales native to both data sources (10m and 90m) to assess the effects of topography at the original scale of the MERIT product and at a finer scale than its intended use. This setup will confirm if observed effects are scale-dependant.

Table 7 - Summary table of modelled experiments

Input Channels	Model Name	
	10m Group	90m Group
Red, nNIR, SWIR1	BASE10m	BASE90m
Red, nNIR, SWIR1, DEM	DEM10m	DEM90m
Red, nNIR, SWIR1, HAND	HAND10m	HAND90m

3.4 Data Pre-Processing

Firstly, the complementary MERIT DEM data was downloaded from GEE using the spatial extent of the Sentinel-2 images. Care was undertaken to include a small buffer (at least one DEM-cell-wide) when obtaining MERIT data. This was done to reduce edge effects when resampling the two data sources in both scale and alignment to produce two coregistered datasets at 10m and 90m resolution. Potential loss of information through alignment resampling has been minimised by only modifying the dataset which was not native to the scale of analysis (Figure 10, rescaling and aligning 10m Sentinel-2 imagery to match 90m MERIT data).

Resampling of all features of data products was undertaken via bilinear interpolation. However, care was taken to preserve the location of the drainage network in the 10m HAND product (cells where HAND=0) as resampling was found to excessively smooth out the fine drainage network. Similarly, labels were resampled using nearest neighbour approaches to avoid smoothing categorical data.

Sentinel-2 and MERIT 'images' underwent a tiling process as CNNs are designed to intake RGB images of regular dimensions where smaller image tiles with fewer training parameters resulting in faster training times (B. Huang et al., 2018).

MS imagery values were unaltered while topographic values were scaled to the same 0-65536 range. Integer values were used as the augmentation package implemented was designed for natural images in integer (.jpg) form. Real elevation values in the DEM that

were negative were translated above zero before scaling values by the new dataset minimum and maximum. The same scaling was applied to the HAND TI, except between 0-65535. HAND values were then shifted by +1 before enforcing ‘NoData’ values as 0. The shift was done to distinguish the real original ‘0’ values of the HAND data from the new ‘0’ values representing ‘NoData’.

To allow comparison of the results to those of Mateo-Garcia et al. (2021) the models were trained on the same 256x256 pixels tile size and were padded with ‘NoData’ values to ensure consistent dimensions. The lower resolution models have the same tile size as the higher resolution models and the effects of this are discussed in section 5.

3.5 Dataset Splits

The table below shows the total number of tiles used in the training, testing, and validation sets. Figures reported in this study differ slightly from those of Mateo-Garcia et al. (2021). This is because corrupt or non-downloadable files (n=13/422) were omitted, and the original data was subset in two ways.

Firstly, a new validation set has been created by adding files from the original training set until a 90:10 split was achieved. This was undertaken as overfitting behaviour was difficult to spot due to ‘spiky’ validation loss curves when using the original validation set. This behaviour was suspected to be due to the small number of events in the original validation set (n=6) not adequately representing the underlying distributions of the training data (n=422). Entire flood event files were randomly added rather than random tiles to ensure no data leakage between sets. Therefore, rather than by file count, splits were balanced by label file size as a simple proxy for pixel count.

Secondly, to speed up training times to practical durations the data was dramatically subset. Even when training on tiles that only contained water (~1/3 or 1/2 of the original data in 10m and 90m sets, respectively), training times were too long to run six models. Therefore, training sets were further subset to a seeded-random sample of 1000 tiles. Across the 1000 tiles, their different resolutions create an imbalance in terms of the proportion of the training data fed to the 10m and 90m models (2.3% and 56.9%). The effects of this imbalance are discussed in section 5.

The original test set was retained to allow results to be comparable to those of Mateo-Garcia et al. (2021).

Table 8 - Tiles present in the datasets. Bold text denotes data splits used to train the models. Italics represent counts of tiles that only contain water

Data Set & Split		Flood events	Tiles	Water pixels (%)	Land pixels (%)	Cloud Pixels (%)	Invalid Pixels (%)
Original 10m	Train	422	182,413	2.70	43.24	50.25	3.81
	Val.	6	1,132	8.33	76.72	13.27	1.68
	Test	11	2,029	21.39	59.05	16.21	3.34
Study 10m	Train	*377*	*43,748*	*25.00*	*25.00*	*25.00*	*24.99*
		-	***1000***	***25.48***	***25.47***	***25.48***	***23.58***
	Val.	51	1,284	25.08	24.98	24.97	24.97
	Test	11	2,270	"	"	"	"
Study 90m	Train	*377*	*1,755*	*24.92*	*25.00*	*25.06*	*25.02*
		-	***1000***	***24.81***	***24.81***	***24.96***	***25.21***
	Val.	51	210	24.65	25.95	24.41	24.99
	Test	11	44	"	"	"	"

3.6 CNN Implementation and Training Regimes

The DeepLab group of CNNs have performed well in semantically segmenting water in RS tasks (mIOU, 0.986 James et al., 2021, 0.936, Li et al., 2019 in DeepLabV3,) and are implemented in this study. The latest model in the series, DeepLabV3+, is an encoder-decoder model which uses an Xception encoder and implements atrous, depthwise-separable, and their combined atrous-separable, convolutions in the ASPP and encoder (Chen, Zhu, et al., 2018). DeepLabV3+'s staged decoder, upsamples by a factor of four while incorporating skip connections from the encoder. DeepLabV3+ builds upon DeepLabV3 to achieve the highest mIoU score (89.0%) on the PASCAL-VOC 2012 benchmark dataset, while also remaining computational lighter than its predecessors (Chen, Zhu, et al., 2018). An overview diagram can be found below.

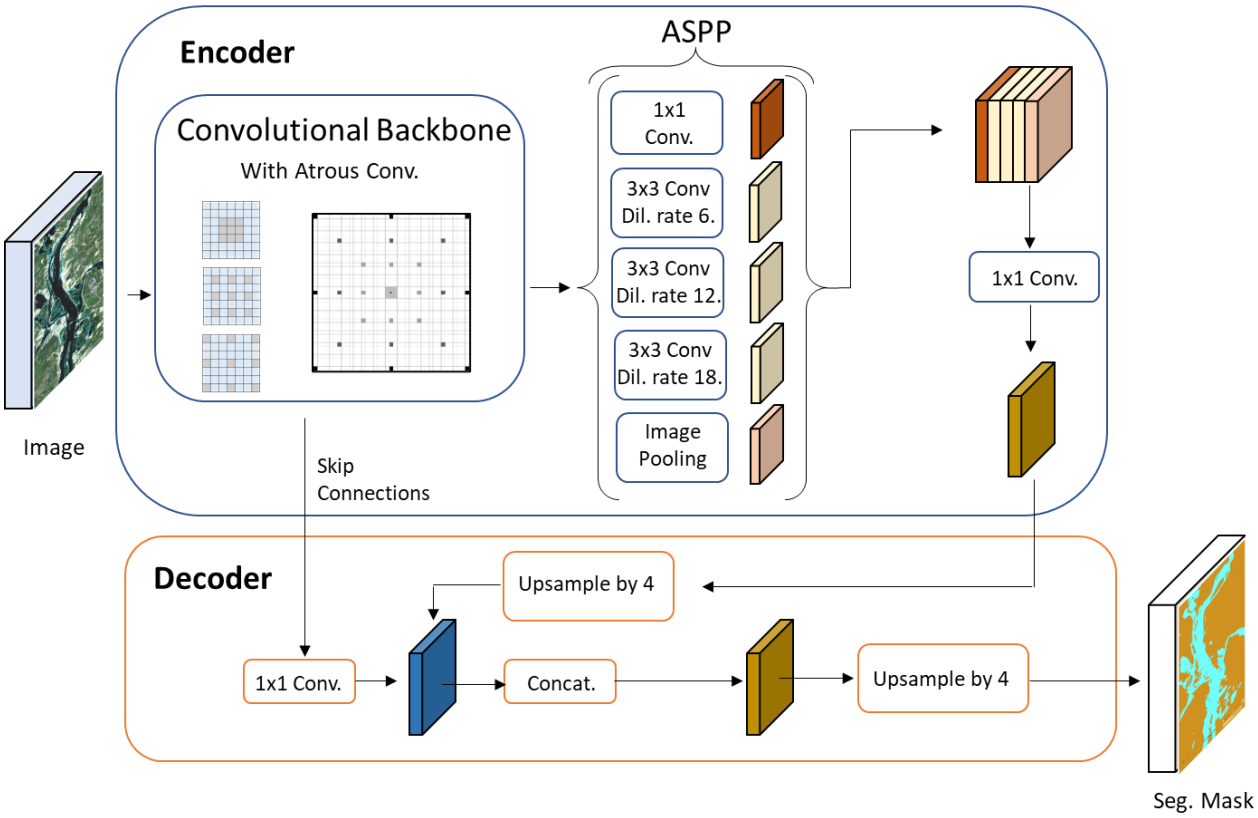


Figure 11 - DeepLabV3+ model implementation. Dil. – dilation,

DeepLabV3+ was implemented in the Pytorch Segmentation Models API (Yakubovskiy, 2019) with a ResNet-50 backbone. This API did not allow implementation of the Xception backbone so a light ResNet backbone with relatively few layers was chosen as a stand in due to training time constraints.

A generalised Dice Loss has been used as the loss function (Equation 4). The Dice score is a lightweight loss function designed to measure overlap in computer vision tasks (Milletari et al., 2016) and has been implemented due to being particularly robust to class imbalance (Sudre et al., 2017).

$$L_{Dice}(X,Y) = -\frac{2\sum_{m=1}^{M} P(X_m)P(Y_m) + \varepsilon}{\sum_{m=1}^{M} P(X_m) + \sum_{m=1}^{M} P(Y_m) + \varepsilon}$$

Equation 4 - Dice Loss. X-feature, Y-label, M-batchsize

Recommended data augmentations such as contrast/brightness adjustments, flips and rotations, and Poisson (shot) noise were added (B. Huang et al., 2018; Yu et al., 2017). Batch sizes were set to those used by Mateo-Garcia et al. (2021, GitLab). Pixel values were not mean centred as learning was observed with non-centred data and the data type was constrained to integers by the augmentation package used. Weights were randomly initialised for all channels, as while pretrained Imagenet weights were usable for the R, nNIR, SWIR1 channels, use of pretrained weights would not allow for fair comparison to models with a topography channel.

Models were each trained for 4000 epochs, or until overfitting was observed on the validation sets. Training was undertaken on a Linux x86_64 operating system using a NVIDIA TITAN RTX (24GB) GPU.

Table 9 shows the hyperparameters adjusted during training experiments with Table 10 showing the parameters used in the final six models.

Table 9 - Hyperparameters explored during training

Hyperparameter	Values Tested
Learning rate	0.0001, 0.0002, 0.0005, 0.0010
Batch size	32, 110
Augmentations	With, without
Loss Ignore Channels in Label	Ignoring No Data, Including No Data
DeepLab encoder depth	3, 5
Backbone	ResNet-50, ResNet-101

Table 10 - Final Model Parameters

Optimiser	Learning Rate	Loss Function	Batch Size	Weights	Backbone	Augmentation
Adam	0.0001	Dice Loss, Including no data	32	Randomly initialised	ResNet-50	contrast/brightness adjustments, flips and rotations, Poisson (shot) noise

3.7 Model Evaluation

Considering the number of models ran and test sites, models have been assessed quantitatively through segmentation performance metrics (section 4). Tile-wise performance has been assessed via box plots to examine the distribution of predictions among model resolution groups.

Qualitative assessment throughout training has been undertaken across 3 of the 11 test sites. These flood events were selected to include a range of land covers and topographic ranges. Events include; an upland forested area of Ituango, Colombia, a meandering agricultural scene in Ylitornio, Finland, and a scrubland estuary scene of Northmanton, Australia. For ease of visual communication and to focus on the main objective of the study, the four-class maps have been reduced to binary classification maps of water and non-water. Misclassifications among non-water classes are not represented in the images and so performance metrics are not informative and have not been calculated.

3.8 Assumptions

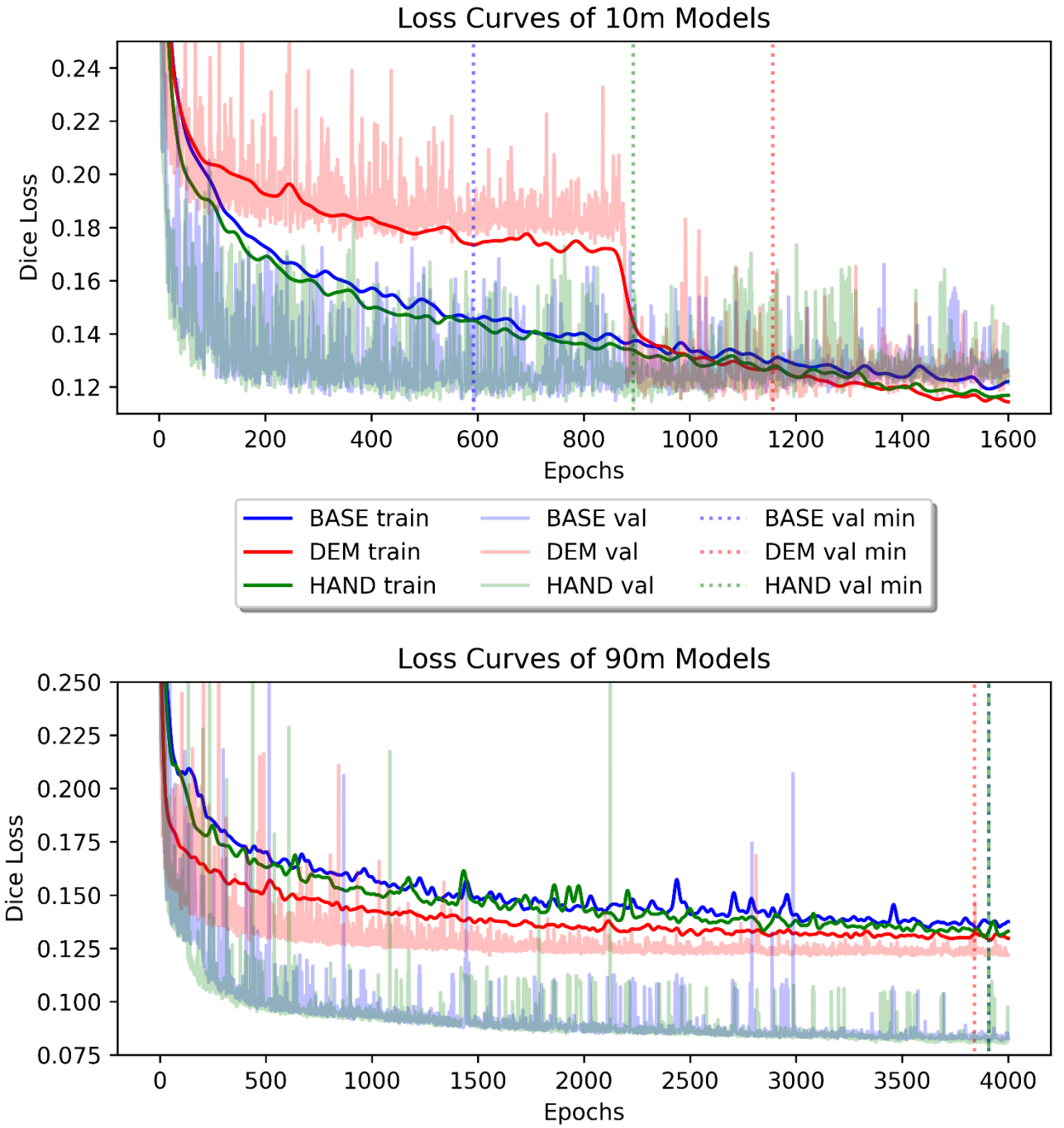
Several key assumptions have been made within the study:

1. The label noise in the 1000 image subsets allows the correct feature-label relationships to be learned.
2. Resampling of the features and labels via different methods (bilinear and nearest neighbour) will not detrimentally increase label noise.
3. It is appropriate to infer topographic sub-pixel information by simple bilinear interpolation. Artificially creating sub-pixel information is thought to be acceptable as topography is only an additional data source in all models and never replaces any spectral bands, the main source of information.
4. Topographic information collected from a range of different times than the MS imagery (SRTM, 2000, Farr & Kobrick, 2000, AW3D30m, 2006-2011, Tadono et al., 2014) will still be useful.
5. The target function can still be learned adequately with the stand-in ResNet-50 backbone. This is as Xception backbones significantly outperform ResNet backbones in natural image segmentation tasks using DeepLabV3+ (Kamann & Rother, 2019).

4 Results

The loss curves from training are presented in Figure 12. The model weights with the lowest validation loss were taken for the final model.

Figure 12 - Model loss curves for 10m (12a) and 90m models (12b). Training set curves have been smoothed using a gaussian filter, sigma=10



4.1 10m Resolution Models

4.1.1 Quantitative Assessment

Across water-specific performance metrics (Table 11), the HAND10m model is noted to perform the best, followed by the DEM10m and BASE10m models. This behaviour also holds for all-class overall accuracy (OA) but not for the F1-score.

Table 11 - Performance metrics of the 10m models. Bold font indicates the best model.

Model	OA (%)	F1 (%)	mIOU Water (%)	Recall Water (%)	Precision Water (%)	Epochs (n)
BASE10m	79.24	36.86	12.96	13.36	24.69	592
DEM10m	80.53	36.22	14.02	16.06	24.82	1156
HAND10m	**88.03**	**36.72**	**15.94**	**18.28**	**26.99**	**893**

Examining the interquartile range across all classes (Figure 13a) shows, that with the exception of precision, the HAND10m model performs the most consistently and the BASE10m model has the most variable performance. Conversely, when examining water-only performance in Figure 13b, the HAND10m model is the most variable model and has a greater proportion of tiles in which more water pixels are identified. In all models, water-specific classification metrics of each tile are noted to be heavily skewed towards zero.

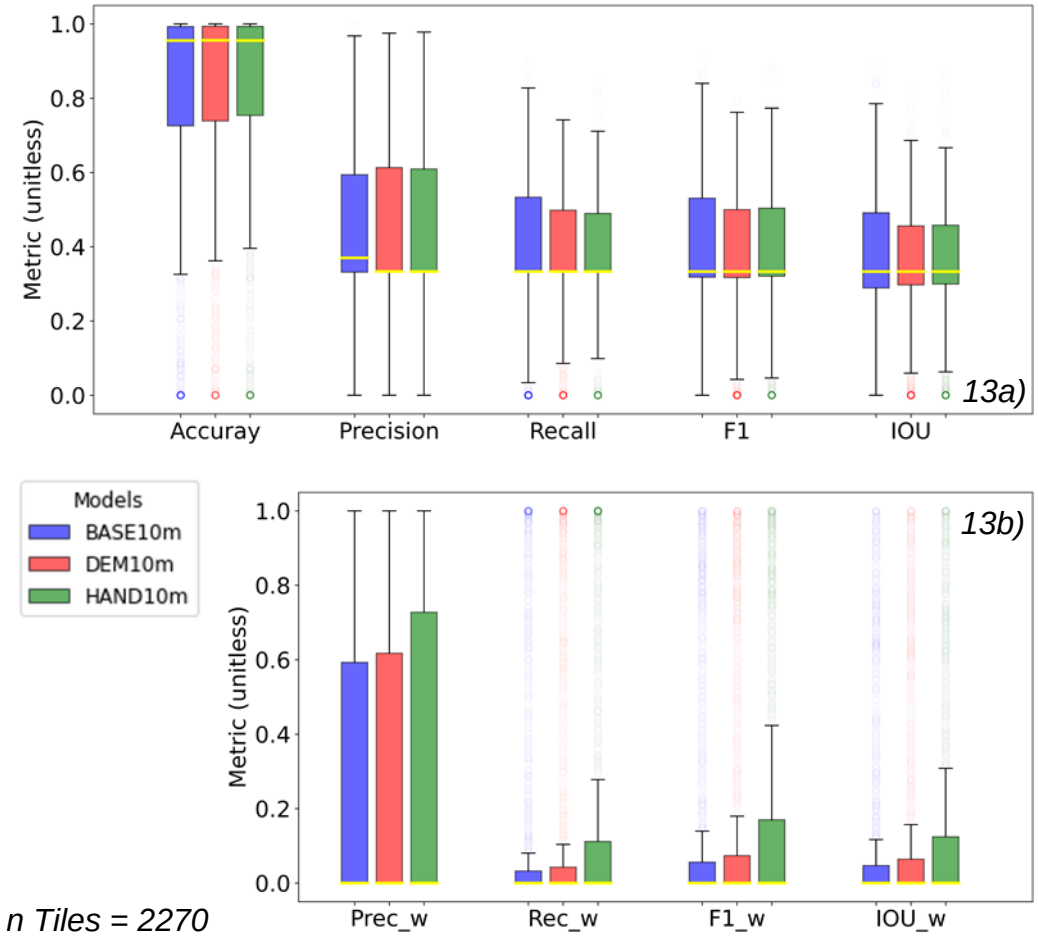


Figure 13 - Box plots displaying performance of the 10m models across 13a) all-classes and the 13b) water-class

4.1.2 Qualitative Assessment

Figure 15 shows all models performed similarly well in the Ituango scene where the main body of the river was clearly identified with a high TP count. Only two areas of water to the southeast and northeast were undetected. In this scene the BASE10m has the highest TP count, followed by the DEM10m and HAND10m models. All models tended to underpredict water areas with differences observed in FN count along boundaries

In the Ylitornio scene all models slightly overpredicted the presence of water along boundaries as seen by the high FP count. A notable difference was the HAND10m model has the lowest FN count and was the only model to not dramatically misclassify water in tiles in the centre of the image. All models fail to identify narrow waterbodies.

In the Northmanton scene the DEM10m model reports the highest TP count, followed by the HAND10m model and lastly the BASE10m model. The behaviour of failing to predict water (Figure 13b zero-skew) can be clearly seen in the FN counts of the BASE10m prediction. In the Northmanton scene misclassification across entire tiles is suspected to be

due to a high sensitivity to atmospheric water vapour causing misclassification of water as cloud. Underprediction along water boundaries in this scene was observed across all models.

4.2 90m resolution models

4.2.1 Quantitative Assessment

With the exception of water-specific precision, the BASE90m model performed best across all performance metrics (Table 12). The HAND90m performs similarly or slightly worse than the BASE90m model while the DEM90m model has the lowest performance due to not being able to predict the water class.

Table 12 - Performance metrics of the 90m models. Bold font indicates the best model.

Model	OA (%)	F1 (%)	mIOU Water (%)	Recall Water (%)	Precision Water (%)	Epochs (n)
BASE90m	**87.55**	**46.14**	**21.47**	**18.17**	**50.98**	**3907**
DEM90m	81.41	35.91	0.00	0.00	0.00	3840
HAND90m	87.24	45.18	20.77	17.90	52.40	3908

The box plots in Figure 14a show the HAND90m model yields slightly more variable results in both the all-class and water-specific performance metrics. In the water-specific metrics (Figure 14b) the HAND90m model is noted to have lower median scores in each metric than BASE90m model. This is with the exception of water-specific precision in which the HAND90m model has a slightly higher mean (Table 12) and median (Figure 14b) value than the BASE90m model.

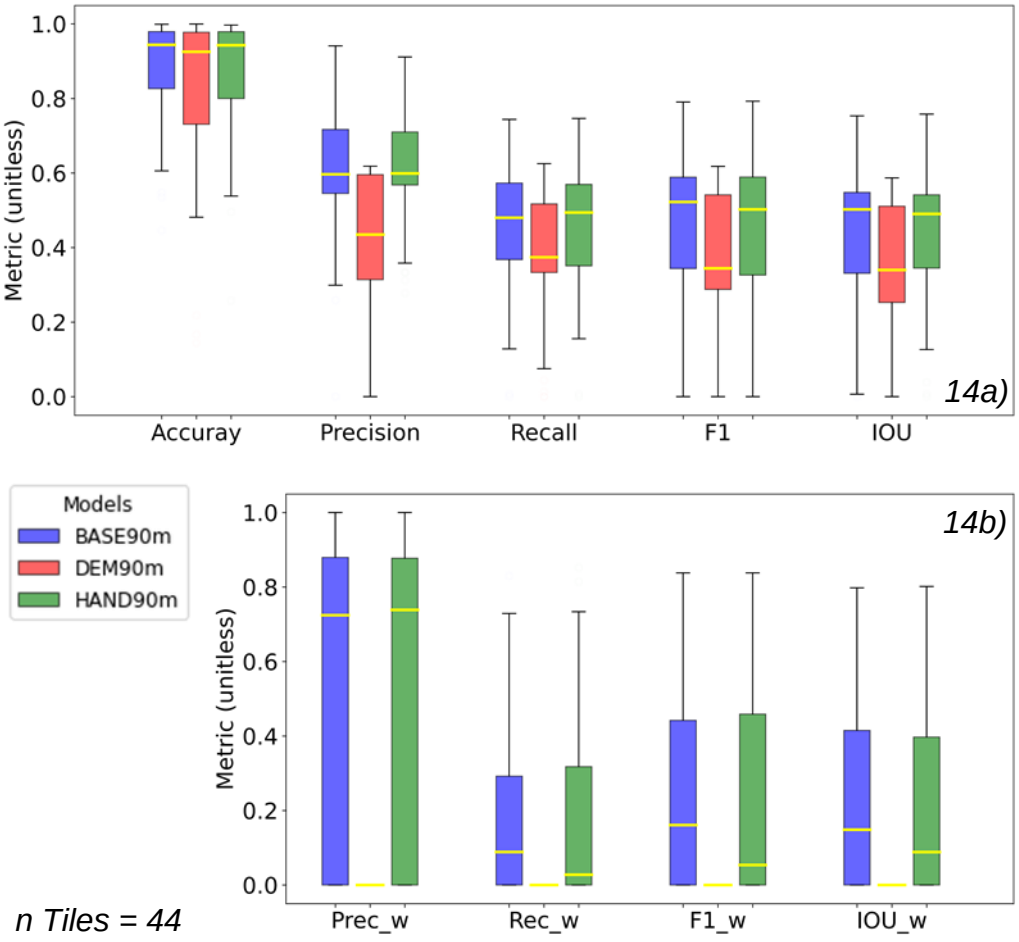


Figure 14 - Box plots of performance of the 90m models across 14a) all-classes and the 14b) water-class

4.2.2 Qualitative Assessment

The DEM90m model is left out of the qualitative assessment as the model did not learn to predict the presence of water pixels.

With the exception of the DEM90m model, the overall trends observed in the 90m models (Figure 15) are similar to those observed in the 10m models (Figure 16). In the Ituango scene models underpredict near water boundaries and struggle to predict the presence of water in the southwest and northeast of the scene. In the Ylitornio scene the BASE90 and HAND90 models overpredict water boundaries slightly. Unlike the 10m model, no misclassification of the river centre occurs. In the Northmanton scene the HAND90m model is noted to have a higher TP count that the BASE90m model which underpredicts the water class. Both models performed poorly in the presence of cloud.

A summary of the results of models of both resolution groups is found in Table 13.

Table 13 - Performance metrics of the 10m and 90m models. Bold font indicates the best model.

Model	OA (%)	F1 (%)	mIOU Water (%)	Recall Water (%)	Precision Water (%)	Epochs (n)
BASE10m	79.24	36.86	12.96	13.36	24.69	592
DEM10m	80.53	36.22	14.02	16.06	24.82	1156
HAND10m	**88.03**	**36.72**	**15.94**	**18.28**	**26.99**	**893**
BASE90m	**87.55**	**46.14**	**21.47**	**18.17**	**50.98**	**3907**
DEM90m	81.41	35.91	0.00	0.00	0.00	3840
HAND90m	87.24	45.18	20.77	17.90	52.40	3908

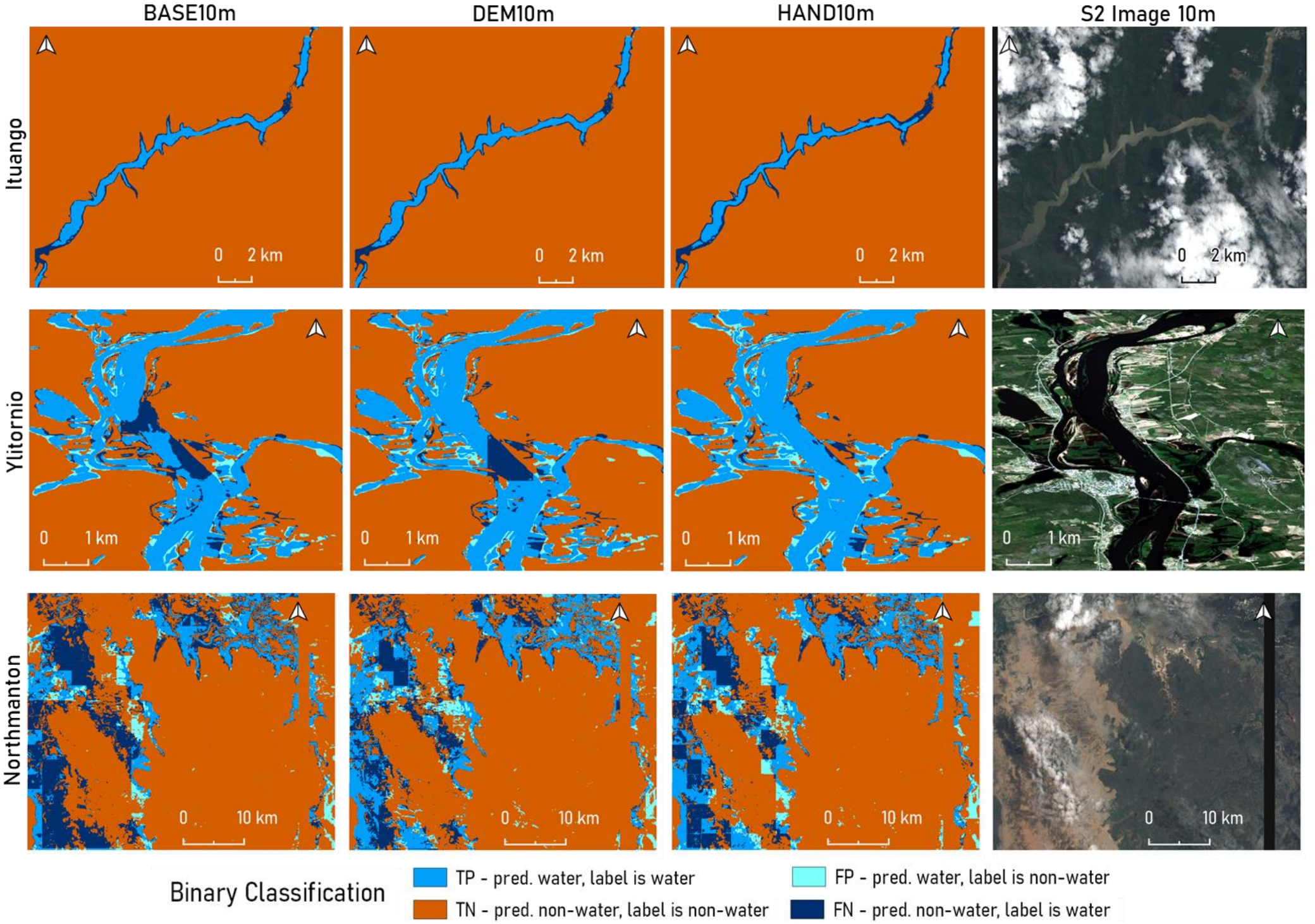


Figure 15 - Predictions of 10m models on three flood events, simplified to a binary classification problem

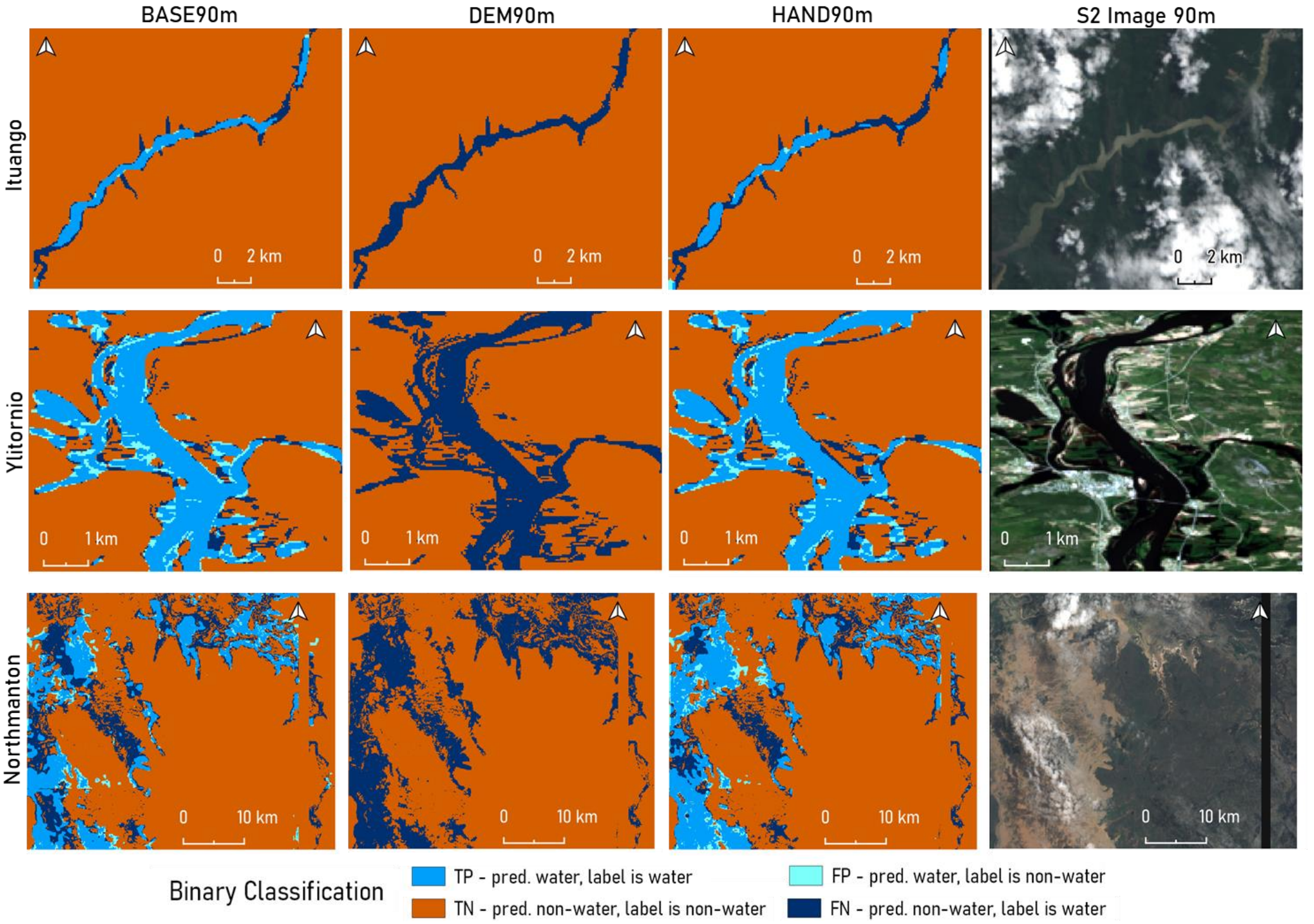


Figure 16- Predictions of 90m models on three flood events, simplified to a binary classification problem

5 Discussion

While including topography, particularly the HAND TI, was beneficial in mapping FW at the 10m scale, at the 90m scale topography has a benign or even detrimental effect. This chapter reviews the conflicting behaviours observed between the model resolution groups and explores landscape- and scale dependent characteristics which are thought to affect results. The findings are discussed in context of the wider literature and the limitations of the study are explored.

5.1 Effects of Topography

In the 10m model group, the addition of topography was noted to have a consistent beneficial effect over the baseline model with the HAND10m model performing the best overall. Unlike the other models the HAND10m model does not underpredict water pixels around the river centreline in the Ylitornio scene. Correct classification is suspected to be due to explicit enforcement of the drainage network in the HAND layer, while the DEM only implicitly contains this information.

Whilst the HAND10m model performs well in the topographically simple Ylitornio scene, the DEM10m model performs best in the topographically complex Northmanton scene. The flood mechanism in the relatively flat (48m inundated range across 4km^2) Northmanton scene is from two major rivers meeting the ocean and backing up across minor rivers and sub-catchments. The DEM10m model is thought to outperform the HAND10m model in this scene because absolute elevation information is thought to be more informative than relative elevation in flatter or multi-catchment scenes. As the HAND TI sacrifices absolute elevation values for relative values, the HAND TI assumes flooding is controlled by the closest drainage network. Without absolute elevation information the HAND TI cannot distinguish between confluencing catchments which are at different minimum elevations (an extreme example being a glaciated hanging valley landform). The lack of this absolute elevation information is suspected to be why the HAND10m model performs worse than the DEM10m model in the Northmanton scene, where multi-catchment flooding is determined by absolute elevation rather than by relative. Additionally, Yamazaki et al., (2019) note the HAND layer may be unreliable in flat estuarine areas. However, robust conclusions of the effects of topography are undermined by the confounding behaviours observed in the 90m models.

The beneficial effects of topography in the 10m resolution model group were not observed in the 90m resolution model group. Adding topography was found to yield similar

(HAND90m) or even worse (DEM90m) performance than the baseline model. This behaviour is surprising, especially considering the richer training data provided to the 90m resolution models in terms of training set proportion and fields of view.

Comparisons of the DEM10m and DEM90m models relative to other models within their resolution group suggest similar loss trajectories but with a notable difference (Figure 12). Initially both DEM models have markedly higher losses than the HAND and BASE models of the same resolution. Around 850 epochs, the DEM10m model exhibits a stepped-decrease in both training and validation loss. Examination of model checkpoints either side of the step show this decrease in loss relates to the DEM10m model learning to predict the presence of the water class for the first time. The lack of a similar step in the 90m models suggest that the DEM90m model is stuck in a local minima of the loss surface.

The cause of the poor performance of the DEM models is suspected to be the intervals of the dilation rates of the atrous convolutions. The atrous convolutions undertake stratified sampling of the surrounding cells at dilation rates of 6, 12, and 18 (Figure 11). This is fixed for both 10m and 90m models and so provides fields of view at different scales.

In a large single catchment scene with smooth topography, the intervals between dilation rates allow values from both sides of a river valley to be sampled and feature maps can identify water-filled local depressions. However, in complex multi-catchment scenes, or at the 90m scale, the dilation rate intervals are unlikely to capture important local variations in topography such as catchment boundaries. In addition to sampling with too large intervals, the wide field of view of the atrous convolutions are mostly sampling topographic values from adjacent catchments which do not affect the presence of water at the pixel within the convolution centre. Incorporation of this irrelevant information in feature maps is thought to explain why the beneficial effects of topography were not observed in any of the 90m models despite the greater proportion of the training data used.

5.2 Comparison to the Wider Literature

5.2.1 Band selection

Despite the differences in the effects of adding topography, some similarities were observed between models in the qualitative assessments. With the exception of the DEM90m model, all models; struggled to identify narrower water bodies in all scenes, were very sensitive to misclassifications in the presence of cloud in the Northmanton scene, and overpredicted water boundaries in the Ylitornio scene.

Under detection of narrow water bodies is noted in similar studies that use solely spectral information (James et al., 2021; Mateo-Garcia et al., 2021). In studies using SIs, this behaviour is noted to be due to shallow waters containing mixed spectral profiles that include a strong signal return of the waterbed (H. Jiang et al., 2014; Tulbure et al., 2016). Misclassifications due to depth would explain the underprediction of water around boundaries by all models, most clearly seen in the Ituango scene. However, Mateo-Garcia point towards temporal misalignment as the reason for poor performance along water boundaries which allows water extents to grow (Ylitornio) or shrink (Ituango) in MS imagery since label acquisition

A notable difference from the model output of Mateo-Garcia et al., (2021) (Figure 9, p11) are that the models of this study misclassify entire tiles, particularly in the presence of cloud and cloud shadow. The implemented tri-band combination is suspected to be the cause as other studies which report the predictive power of this tri-band combination in mapping FWs have done so in images relatively free of cloud (Jain et al., 2020; Muñoz et al., 2021). Mateo-Garcia et al., (2021) still report some misclassification issues around cloud boundaries but their study benefits from utilising richer spectral information provided across all 13 Sentinel-2 bands. In scenes with cloud occlusion, it is likely that better classification of water-pixels can be achieved when using more bands, including the cirrus SWIR band (2202.4nm ESA, 2015) that is particularly useful for discriminating cloud even from similar reflective surfaces such as snow (Z. Zhu et al., 2015).

5.2.2 Incorporating Topography

Inclusion of topography in the 10m models was found to increase model performance by several percentage points, as also found by Du et al., (2020) and Muñoz et al., (2021). Furthermore, like Du et al.'s (2020) study, the inclusion of a TI was found to yield performance increases of a similar magnitude over using raw elevation values alone. However, these behaviours were not observed in the 90m models.

In models using absolute elevation of the DEM, the target function seems harder to learn than when using relative elevations of the HAND TI, at both 10m and 90m resolutions. In addition to the dilation intervals of the atrous convolutions, high variability in DEM values and the model architecture may limit the learnability of the target function.

This global study differs from the single-site or neighbouring-site studies of Du et al., (2020) and Muñoz et al., (2021), respectively, which are characterised by low variance in elevation values in the wetland (a range of 45m, Du et al., 2020) and coastal regions (Muñoz et al., 2021) examined. The low variance in study area elevation implies a low variance in elevation values of inundated pixels. Conversely the WorldFloods dataset examines floods in a range of lowland and mountainous scenes and the elevations of inundated pixels vary widely. This contrasts to the values in the Red, nNIR, SWIR1 and HAND channels where inundated pixels have relatively stable values. Therefore, when including absolute elevation, the target function relies little on the absolute values of inundated cells and relies mostly on patterns of height of inundated cells relative to their surroundings, in conjunction with the imagery channels.

Furthermore, the CNN architecture used in this study to incorporate topography is not as advanced as methods found in analogous high resolution LULC classification studies. Pan et al., (2018) note that when incorporating LiDAR DSMs with tri-band imagery, best results are obtained when extracting feature maps of each channel separately. The ResNet-50 backbone implemented as a stand-in for the Xception backbone, is not designed to separately learn inter- and intra- channel patterns and could limit learnability of the target function.

Other studies like those of Muñoz et al., (2021) undertake data fusion through separate convolutional branches for imagery and topographic data sources (e.g., Xu et al., 2018; H. Zhu et al., 2020), or multiple branches for the same data source (e.g., Liu et al., 2021). Explicitly extracting spectral and spatial information through separate branches is found to be successful in LULC classification studies fusing HS imagery and LiDAR data (Xu et al.,

2018) and Panchromatic and MS imagery (H. Zhu et al., 2020). While these works relate to pixel-based classifiers, encoder-decoder models can include multiple convolutional arms to focus channel attention differently (Hu et al., 2017) between arms extracting spectral and spatial features (Liu et al., 2021). This approach is noted to outperform DeepLabV3+ on the Vaihingen (Cramer, 2010) LULC classification benchmark dataset (Liu et al., 2021).

5.3 Implications for the Field

This study has sought to be a first step in including the effects of topography both on a global scale and at different scales of resolution. Model performance is theorised to be sensitive to the dilation rate intervals in the atrous convolutions and varies with catchment size and with pixel scale. This suggests dilation rates and their intervals should be tailored to the scale of the causal features of interest (catchments).

Given the beneficial effects of adding topography in the 10m models, it seems that the MERIT global open dataset is useful at smaller scales than its intended use. Wider work could examine new uses of existing coarse-resolution data sources as ancillary information to improve LULC classification EO studies.

6 Conclusions

6.1 Study Findings

This study has shown that, under the correct conditions, incorporating topography into CNNs can improve FW segmentation at a global scale.

In the 10m models, topography was noted to have a beneficial effect on segmentation with the HAND TI yielding a slightly higher performance than the DEM. This suggests that the MERIT topographic data product can be used at finer scales than its intended use when used as an ancillary data source.

In the 90m models, the effects of topography were negligible or even detrimental to segmentation. This behaviour is suspected to be due to misspecification of the dilation rate intervals causing models to be provided with irrelevant topographic information from neighbouring catchments.

Performance differences between 10m and 90m models could not be fairly made due to the different proportions of the training data used between model groups.

6.2 Limitations and Future Work

All models struggled to identify water pixels along water-cloud boundaries. Future work including a greater number of spectral bands may aid segmentation performance around cloud-boundaries.

The findings of this study suggest that treating topography as an extra channel in CNNs can improve model performance. However future work should assess whether improvements can be made through, altering the dilation rate intervals of the atrous convolutions, implementing an Xception backbone, or by using multiple convolutional arms.

Sub-setting the data to reduce model training times introduced a data imbalance. Due to the imbalance, inter-resolution group comparisons could not be directly made, and direct comparisons are limited to each resolution group. Future work should assess if the behaviours observed between models are found to hold when using the full training data.

Future work should seek to differentiate between classification of permanent SW and ephemeral FWs. Due to differences in their appearance and the damaging effects of the latter the differentiation is noted to be highly important (Bonafilia et al., 2020; Mateo-Garcia et al., 2021).

7 Bibliography

Angluin, D., & Laird, P. (1988). Learning From Noisy Examples. Machine Learning, 2(4), 343–370. https://doi.org/10.1023/A:1022873112823

Archer, L., Neal, J. C., Bates, P. D., & House, J. I. (2018). Comparing TanDEM-X Data With Frequently Used DEMs for Flood Inundation Modeling. Water Resources Research, 54(12), 10, 205–210, 222. https://doi.org/https://doi.org/10.1029/2018WR023688

Aristizabal, F., Judge, J., & Monsivais-Huertero, A. (2020). High-Resolution Inundation Mapping for Heterogeneous Land Covers with Synthetic Aperture Radar and Terrain Data. Remote Sensing, 12(6). https://doi.org/10.3390/rs12060900

Bijeesh, T. v, & Narasimhamurthy, K. N. (2020). Surface water detection and delineation using remote sensing images: a review of methods and algorithms. Sustainable Water Resources Management, 6(4), 68. https://doi.org/10.1007/s40899-020-00425-4

Bishop, C. M. (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), 108–116. https://doi.org/10.1162/neco.1995.7.1.108

Bonafilia, D., Tellman, B., Anderson, T., & Issenberg, E. (2020, June). Sen1Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Carranza-García, M., García-Gutiérrez, J., & Riquelme, J. C. (2019). A Framework for Evaluating Land Use and Land Cover Classification Using Convolutional Neural Networks. Remote Sensing, 11(3). https://doi.org/10.3390/rs11030274

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. http://arxiv.org/abs/1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. https://arxiv.org/abs/1802.02611

Chollet, F. (2016). Xception: Deep Learning with Depthwise Separable Convolutions. https://arxiv.org/abs/1610.02357

Chow, C., Twele, A., & Martinis, S. (2016). An assessment of the Height Above Nearest Drainage terrain descriptor for the thematic enhancement of automatic SAR-based flood monitoring services. Proc.SPIE, 9998. https://doi.org/10.1117/12.2240766

Cohen, S., Brakenridge, G. R., Kettner, A., Bates, B., Nelson, J. M., McDonald, R. R., Huang, Y.-F., Munasinghe, D., & Zhang, J. (2018). Estimating floodwater depths from flood inundation maps and topography. Journal of the American Water Resources Association, 54(4), 847–858. https://doi.org/10.1111/1752-1688.12609

Cohen, S., Raney, A., Munasinghe, D., Loftis, J. D., Molthan, A., Bell, J., Rogers, L., Galantowicz, J., Brakenridge, G. R., Kettner, A. J., Huang, Y.-F., & Tsang, Y.-P. (2019). The Floodwater Depth Estimation Tool (FwDET v2.0) for improved remote sensing analysis of coastal flooding. Natural Hazards and Earth System Sciences, 19(9), 2053–2065. https://doi.org/10.5194/nhess-19-2053-2019

Cramer, M. (2010). The DGPF-Test on Digital Airborne Camera Evaluation - Over- view and Test Design. Photogrammetrie Fernerkundung Geoinformation, 2010, 73–82.

Du, L., Zhang, X., Lang, M., Vanderhoof, M., Li, X., Huang, C., Lee, S., & Zou, Z. (2020). remote sensing Mapping Forested Wetland Inundation in the Delmarva Peninsula, USA Using Deep Convolutional Neural Networks. Remote Sensing, 12. https://doi.org/10.3390/rs12040644

ESA, (2021). SPOT 6 and 7 ESA archive, available from https://earth.esa.int/eogateway/catalog/spot-6-and-7-esa-archive, [Accessed 24th July 2021]

ESA, (2015). S2 user handbook, Available from https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook, [Accessed 24th July 2021]

Faherty, D., Schumann, G. J.-P., & Moller, D. K. (2020). Bare Earth DEM Generation for Large Floodplains Using Image Classification in High-Resolution Single-Pass InSAR. Frontiers in Earth Science, 8, 27. https://doi.org/10.3389/feart.2020.00027

Farr, T. G., & Kobrick, M. (2000). Shuttle radar topography mission produces a wealth of data. Eos, Transactions American Geophysical Union, 81(48), 583–585. https://doi.org/https://doi.org/10.1029/EO081i048p00583

Frénay, B., & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. Neural Networks and Learning Systems, IEEE Transactions On, 25, 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Gallant, J. C., & Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39(12). https://doi.org/https://doi.org/10.1029/2002WR001426

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. Neural Computation, 4, 1–58.

Gholamalinezhad, H., & Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, a Review. https://arxiv.org/abs/2009.07485

Gomes Pereira, L. M., & Wicherson, R. J. (1999). Suitability of laser data for deriving geographical information: A case study in the context of management of fluvial zones. ISPRS Journal of Photogrammetry and Remote Sensing, 54(2), 105–114. https://doi.org/https://doi.org/10.1016/S0924-2716(99)00007-6

Gómez-Palacios, D., Torres, M. A., & Reinoso, E. (2017). Flood mapping through principal component analysis of multitemporal satellite imagery considering the alteration of water spectral properties due to turbidity conditions. Geomatics, Natural Hazards and Risk, 8(2), 607–623. https://doi.org/10.1080/19475705.2016.1250115

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. https://arxiv.org/abs/2008.05756

Han Jun and Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In F. Mira José and Sandoval (Ed.), From Natural to Artificial Neural Computation (pp. 195–201). Springer Berlin Heidelberg.

Hashemi-Beni, L., & Gebrehiwot, A. A. (2021). Flood Extent Mapping: An Integrated Method Using Deep Learning and Region Growing Using UAV Optical Data. IEEE Journal of

Selected Topics in Applied Earth Observations and Remote Sensing, 14, 2127–2135. https://doi.org/10.1109/JSTARS.2021.3051873

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. https://arxiv.org/abs/1512.03385

Hess, L. L., Melack, J. M., Novo, E. M. L. M., Barbosa, C. C. F., & Gastil, M. (2003). Dual-season mapping of wetland inundation and vegetation for the central Amazon basin. Remote Sensing of Environment, 87(4), 404–428. https://doi.org/https://doi.org/10.1016/j.rse.2003.04.001

Hirabayashi, Y., Alifu, H., Yamazaki, D., Donchyts, Gennadii, & Kimura, Y. (2021). Detectability of variation in river flood from satellite images. Hydrological Research Letters, 15(2), 37–43. https://doi.org/10.3178/hrl.15.37

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., & Kanae, S. (2013). Global flood risk under climate change. Nature Climate Change, 3(9), 816–821. https://doi.org/10.1038/nclimate1911

Hirabayashi, Y., Tanoue, M., Sasaki, O., Zhou, X., & Yamazaki, D. (2021). Global exposure to flooding from the new CMIP6 climate model projections. Scientific Reports, 11(1), 3740. https://doi.org/10.1038/s41598-021-83279-w

Hird, J. N., DeLancey, E. R., McDermid, G. J., & Kariyeva, J. (2017). Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping. Remote Sensing, 9(12). https://doi.org/10.3390/rs9121315

Hoeser, T., Bachofer, F., & Kuenzer, C. (2020). Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review—Part II: Applications. Remote Sensing, 12(18). https://doi.org/10.3390/rs12183053

Hoeser, T., & Kuenzer, C. (2020). Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. Remote Sensing, 12(10). https://doi.org/10.3390/rs12101667

Hosseiny, B., Ghasemian, N., & Amini, J. (2019). A CONVOLUTIONAL NEURAL NETWORK FOR FLOOD MAPPING USING SENTINEL-1 AND SRTM DEM DATA: CASE STUDY IN POLDOKHTAR-IRAN. ISPRS - International Archives of the

Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W18, 527–533. https://doi.org/10.5194/isprs-archives-XLII-4-W18-527-2019

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. http://arxiv.org/abs/1709.01507

Huang, B., Reichman, D., Collins, L. M., Bradbury, K., & Malof, J. M. (2018). Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. https://arxiv.org/abs/1805.12219

Huang, C., Chen, Y., Wu, J., Chen, Z., Li, L., Liu, R., & Yu, J. (2014). Integration of remotely sensed inundation extent and high-precision topographic data for mapping inundation depth. 2014 The Third International Conference on Agro-Geoinformatics, 1–4. https://doi.org/10.1109/Agro-Geoinformatics.2014.6910580

Huang, C., Chen, Y., Zhang, S., & Wu, J. (2018). Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review. Reviews of Geophysics, 56(2), 333–360. https://doi.org/https://doi.org/10.1029/2018RG000598

Huang, C., Nguyen, B. D., Zhang, S., Cao, S., & Wagner, W. (2017). A Comparison of Terrain Indices toward Their Ability in Assisting Surface Water Mapping from Sentinel-1 Data. ISPRS International Journal of Geo-Information, 6(5). https://doi.org/10.3390/ijgi6050140

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huete, A. R. (2004). 11 - REMOTE SENSING FOR ENVIRONMENTAL MONITORING. In J. F. Artiola, I. L. Pepper, & M. L. Brusseau (Eds.), Environmental Monitoring and Characterization (pp. 183–206). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012064477-3/50013-8

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. http://arxiv.org/abs/1502.03167

Irwin, K., Beaulne, D., Braun, A., & Fotopoulos, G. (2017). Fusion of SAR, Optical Imagery and Airborne LiDAR for Surface Water Detection. Remote Sensing, 9(9). https://doi.org/10.3390/rs9090890

Jain, P., Schoen-Phelan, B., & Ross, R. (2020). Tri-Band Assessment of Multi-Spectral Satellite Data for Flood Detection.

James, T., Schillaci, C., & Lipani, A. (2021). Convolutional neural networks for water segmentation using sentinel-2 red, green, blue (RGB) composites and derived spectral indices. International Journal of Remote Sensing, 42(14), 5338–5365. https://doi.org/10.1080/01431161.2021.1913298

Jarihani, A. A., Callow, J. N., McVicar, T. R., van Niel, T. G., & Larsen, J. R. (2015). Satellite-derived Digital Elevation Model (DEM) selection, preparation and correction for hydrodynamic modelling in large, low-gradient and data-sparse catchments. Journal of Hydrology, 524, 489–506. https://doi.org/10.1016/j.jhydrol.2015.02.049

Jiang, H., Feng, M., Zhu, Y., Lu, N., Huang, J., & Xiao, T. (2014). An Automated Method for Extracting Rivers and Lakes from Landsat Imagery. Remote Sensing, 6(6), 5067–5089. https://doi.org/10.3390/rs6065067

Jiang, Z. (2020). Spatial Structured Prediction Models: Applications, Challenges, and Techniques. IEEE Access, 8, 38714–38727. https://doi.org/10.1109/ACCESS.2020.2975584

Kamann, C., & Rother, C. (2019). Benchmarking the Robustness of Semantic Segmentation Models. https://doi.org/10.1109/cvpr42600.2020.00885

Kang, W., Xiang, Y., Wang, F., Wan, L., & You, H. (2018). Flood Detection in Gaofen-3 SAR Images via Fully Convolutional Networks. Sensors, 18(9). https://doi.org/10.3390/s18092915

Kathuria, A., (2018). Intro to optimization in deep learning: Gradient Descent, https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Kukačka, J., Golkov, V., & Cremers, D. (2017). Regularization for Deep Learning: A Taxonomy. https://arxiv.org/abs/1710.10686

Larsen, K., & Becker, D. (2018). Chapter 24. Seven Types of Target Leakage in Machine Learning and an Exercise.

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

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data. Eos, Transactions American Geophysical Union, 89(10), 93–94. https://doi.org/https://doi.org/10.1029/2008EO100001

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. WIREs Data Mining and Knowledge Discovery , 8(6), e1264. https://doi.org/https://doi.org/10.1002/widm.1264

Li, Z., Wang, R., Zhang, W., Hu, F., & Meng, L. (2019). Multiscale Features Supported DeepLabV3+ Optimization Scheme for Accurate Water Semantic Segmentation. IEEE Access, 7, 155787–155804. https://doi.org/10.1109/ACCESS.2019.2949635

Liu, Y., Zhu, Q., Cao, F., Chen, J., & Lu, G. (2021). High-Resolution Remote Sensing Image Segmentation Framework Based on Attention Mechanism and Adaptive Weighting. ISPRS International Journal of Geo-Information, 10(4). https://doi.org/10.3390/ijgi10040241

Long, J., Shelhamer, E., & Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation. http://arxiv.org/abs/1411.4038

Manavalan, R. (2017). SAR image analysis techniques for flood area mapping - literature survey. Earth Science Informatics, 10(1), 1–14. https://doi.org/10.1007/s12145-016-0274-2

Martinis, S., Kersten, J., & Twele, A. (2015). A fully automated TerraSAR-X based flood service. ISPRS Journal of Photogrammetry and Remote Sensing, 104, 203–212. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2014.07.014

Mason, D. C., Dance, S. L., Vetra-Carvalho, S., & Cloke, H. L. (2018). Robust algorithm for detecting floodwater in urban areas using synthetic aperture radar images. Journal of Applied Remote Sensing, 12(4), 1–20. https://doi.org/10.1117/1.JRS.12.045011

Mateo-Garcia, G., Veitch-Michaelis, J., Smith, L., Oprea, S. V., Schumann, G., Gal, Y., Baydin, A. G., & Backes, D. (2021). Towards global flood mapping onboard low cost satellites with machine learning. Scientific Reports, 11(1), 7249. https://doi.org/10.1038/s41598-021-86650-z, GitLab repository available from https://gitlab.com/frontierdevelopmentlab/disaster-prevention/cubesatfloods, [Accessed July 2nd 2021]

McClean, F., Dawson, R., & Kilsby, C. (2020). Implications of Using Global Digital Elevation Models for Flood Risk Analysis in Cities. Water Resources Research, 56(10), e2020WR028241. https://doi.org/https://doi.org/10.1029/2020WR028241

McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7), 1425–1432. https://doi.org/10.1080/01431169608948714

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. http://arxiv.org/abs/1606.04797

Mitchell, T. M. (1997). Machine Learning (1st ed.). McGraw-Hill, Inc.

Muñoz, D. F., Muñoz, P., Moftakhari, H., & Moradkhani, H. (2021). From local to regional compound flood mapping with deep learning and data fusion techniques. Science of The Total Environment, 782, 146927. https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.146927

Musa, Z. N., Popescu, I., & Mynett, A. (2015). A review of applications of satellite SAR, optical, altimetry and DEM data for surface water modelling, mapping and parameter estimation. Hydrology and Earth System Sciences, 19(9), 3755–3769. https://doi.org/10.5194/hess-19-3755-2015

Nemni, E., Bullock, J., Belabbes, S., & Bromley, L. (2020). Fully Convolutional Neural Network for Rapid Flood Segmentation in Synthetic Aperture Radar Imagery. Remote Sensing, 12(16). https://doi.org/10.3390/rs12162532

Pan, X., Gao, L., Zhang, B., Yang, F., & Liao, W. (2018). High-Resolution Aerial Imagery Semantic Labeling with Dense Pyramid Network. Sensors, 18(11). https://doi.org/10.3390/s18113774

Polyak, B. (1987). Introduction to Optimization, Reprint. New York, Optimization Software, Inc. ~ Publications Division (2010)

Quinn, P., Beven, K., Chevallier, P., & Planchon, O. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes, 5(1), 59–79. https://doi.org/https://doi.org/10.1002/hyp.3360050106

Rambour, C., Audebert, N., Koeniguer, E., le Saux, B., Crucianu, M., & Datcu, M. (2020). FLOOD DETECTION IN TIME SERIES OF OPTICAL AND SAR IMAGES. The International Archives of the Photogrammetry, Remote Sensing and Spatial

Information Sciences, XLIII-B2-2020, 1343–1346. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1343-2020

Ranchin, T., & Wald, L. (2000). Fusion of high spatial and spectral resolution images: The arsis concept and its implementation. Photogrammetric Engineering and Remote Sensing.

Rasamoelina, A. D., Adjailia, F., & Sinčák, P. (2020). A Review of Activation Function for Artificial Neural Network. 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 281–286. https://doi.org/10.1109/SAMI48414.2020.9108717

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., & Waterloo, M. J. (2008). HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. Remote Sensing of Environment, 112(9), 3469–3481. https://doi.org/https://doi.org/10.1016/j.rse.2008.03.018

Rezaee, M., Mahdianpari, M., Zhang, Y., & Salehi, B. (2018). Deep Convolutional Neural Network for Complex Wetland Classification Using Optical Remote Sensing Imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(9), 3030–3039. https://doi.org/10.1109/JSTARS.2018.2846178

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. http://arxiv.org/abs/1505.04597

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

Sarker, C., Mejias, L., Maire, F., & Woodley, A. (2019). Flood Mapping with Convolutional Neural Networks Using Spatio-Contextual Pixel Information. Remote Sensing, 11(19). https://doi.org/10.3390/rs11192331

Schumann, G. J.-P., & Bates, P. D. (2018). The Need for a High-Accuracy, Open-Access Global DEM. Frontiers in Earth Science, 6, 225. https://www.frontiersin.org/article/10.3389/feart.2018.00225

Schumann, G. J.-P., & Bates, P. D. (2020). Editorial: The Need for a High-Accuracy, Open-Access Global Digital Elevation Model. Frontiers in Earth Science, 8, 544. https://www.frontiersin.org/article/10.3389/feart.2020.618194

Schumann, G. J.-P., & Moller, D. K. (2015). Microwave remote sensing of flood inundation. Physics and Chemistry of the Earth, Parts A/B/C, 83–84, 84–95. https://doi.org/https://doi.org/10.1016/j.pce.2015.05.002

Sentinel Hub (2019) s2cloudless:Sentinel Hub's cloud detector for Sentinel-2 imagery, Available from https://github.com/sentinel-hub/sentinel2-cloud-detector

Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging, 13(1), 146–165. https://doi.org/10.1117/1.1631315

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Song, Y.-S., Sohn, H.-G., & Park, C. (2007). Efficient Water Area Classification Using Radarsat-1 SAR Imagery in a High Relief Mountainous Environment. Photogrammetric Engineering & Remote Sensing, 73, 285–296. https://doi.org/10.14358/PERS.73.3.285

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. https://doi.org/10.1007/978-3-319-67558-9_28

Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7), 073001. https://doi.org/10.1088/1748-9326/ab1b7d

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability. Scientific Reports, 10(1), 13768. https://doi.org/10.1038/s41598-020-70816-2

Tadono, T., Ishida, H., Oda, F., Naito, S., Minakawa, K., & Iwamoto, H. (2014). Precise Global DEM Generation by ALOS PRISM. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II–4, 71–76. https://doi.org/10.5194/isprsannals-II-4-71-2014

Thissen, J. J. M. (2019). Automating surface water detection for rivers : the estimation of the geometry of rivers based on optical earth observation sensors. http://essay.utwente.nl/77537/

Tulbure, M. G., Broich, M., Stehman, S. v, & Kommareddy, A. (2016). Surface water extent dynamics from three decades of seasonally continuous Landsat time series at subcontinental scale in a semi-arid region. Remote Sensing of Environment, 178, 142–157. https://doi.org/https://doi.org/10.1016/j.rse.2016.02.034

USGS, (n.d.). What are the band designations for the Landsat satellites?, Available from, https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?qt-news_science_products=0#qt-news_science_products, [Accessed 25th July 2021]

Valiant, L. G. (1984). A theory of the learnable. Communications of the ACM, 27(11), 1134–1142.

Vickers, H., Malnes, E., & Høgda, K.-A. (2019). Long-Term Water Surface Area Monitoring and Derived Water Level Using Synthetic Aperture Radar (SAR) at Altevatn, a Medium-Sized Arctic Lake. Remote Sensing, 11(23). https://doi.org/10.3390/rs11232780

Wang, Y., Colby, J. D., & Mulcahy, K. A. (2002). An efficient method for mapping flood extent in a coastal floodplain using Landsat TM and DEM data. International Journal of Remote Sensing, 23(18), 3681–3696. https://doi.org/10.1080/01431160110114484

Western, A. W., Grayson, R. B., Blöschl, G., Willgoose, G. R., & McMahon, T. A. (1999). Observed spatial organization of soil moisture and its relation to terrain indices. Water Resources Research, 35(3), 797–810. https://doi.org/https://doi.org/10.1029/1998WR900065

Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., & Zhang, B. (2018). Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. IEEE Transactions on Geoscience and Remote Sensing, 56(2), 937–949. https://doi.org/10.1109/TGRS.2017.2756851

Yakubovskiy, P., (2019)., Segmentation Models Pytorch, GitHub repository, Available from https://github.com/qubvel/segmentation_models.pytorch, [Accessed 10th August 2021]

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. Water Resources Research, 55(6), 5053–5073. https://doi.org/https://doi.org/10.1029/2019WR024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., & Bates, P. D. (2017). A high-accuracy map of global terrain elevations. Geophysical Research Letters, 44(11), 5844–5853. https://doi.org/10.1002/2017GL072874

Yang, X., Zhao, S., Qin, X., Zhao, N., & Liang, L. (2017). Mapping of Urban Surface Water Bodies from Sentinel-2 MSI Imagery at 10 m Resolution via NDWI-Based Image Sharpening. Remote Sensing, 9(6). https://doi.org/10.3390/rs9060596

Yu, X., Wu, X., Luo, C., & Ren, P. (2017). Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. GIScience & Remote Sensing, 54, 1–18. https://doi.org/10.1080/15481603.2017.1323377

Zhu, H., Ma, W., Li, L., Jiao, L., Yang, S., & Hou, B. (2020). A Dual–Branch Attention fusion deep network for multiresolution remote–Sensing image classification. Information Fusion, 58, 116–131. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.013

Zhu, X. X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geoscience and Remote Sensing Magazine, 5(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

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/https://doi.org/10.1016/j.rse.2014.12.014

Abbreviations

Adam - Adaptive Moment Estimation
ANN - Artificial Neural Network
ASPP - Atrous Spatial Pyramidal Pooling
BN - Batch Normalisation
CNN - Convolutional Neural Network
DEM - Digital Elevation Model
DSM - Digital Surface Model
DT - Decision Trees
DTM - Digital Terrain Model
EM - Electro Magnetic
EO - Earth Observation
FN - False Negative
FP - False Positive
FW - Flood Water
GEE - Google Earth Engine
HAND - Height Above Nearest Drainage
HS - Hyper-spectral
IoU - Intersection Over Union
LiDAR - Light Detection and Ranging
LULC - Land Use Land Cover
mIoU - Mean Intersection Over Union
ML - Machine Learning
MrVBF - Multi-Resolution Valley Bottom Flatness
MS - Multi-spectral
NDWI - Normalised Differential Water Index
nNIR - Narrow Near-Infrared
NRT - Near-Real-Time
OA - Overall Accuracy
ReLU - Rectified Linear Unit
RS - Remote Sensing
SAR - Synthetic Aperture Radar
SI - Spectral Indices
SVM - Support Vector Machine

SW - Surface Water

SWIR - Short-Wave Infrared

TI - Terrain Indices

TN - True Negative

TP - True Positive

TWI - Topographic Wetness Index