Convolutional Neural Networks for Water segmentation using Sentinel-2 Red, Green, Blue (RGB) composites and derived Spectral Indices

Thomas James [1], Calogero Schillaci [2*], Aldo Lipani [1]

*corresponding author

[1] Department of Civil, Environmental and Geomatic Engineering, Gower Street, London WC1E 6BT, University College London

[2] Department of Agricultural and Environmental Sciences, Milan Via Celoria 2, University of Milan

Abstract

Near-real time water segmentation with medium resolution satellite imagery plays a critical role in water management. Automated water segmentation of satellite imagery has traditionally been achieved using spectral indices. Spectral water segmentation is limited by environmental factors and requires human expertise to be applied effectively. In recent years the use of convolutional neural networks (CNN's) for water segmentation has been successful when used on high-resolution satellite imagery, but to a lesser extent for medium resolution imagery. Existing studies have been limited to geographically localised datasets and reported metrics have been benchmarked against a limited range of spectral indices. This study seeks to determine if a single CNN based on Red, Green, Blue (RGB) image classification can effectively segment water on a global scale and outperform traditional spectral methods. Additionally, this study evaluates the extent to which smaller datasets (of very complex pattern e.g harbor megacities) can be used to improve a globally applicable CNN's within a specific region. Multispectral imagery from the European Space Agency, Sentinel-2 satellite (10 m spatial resolution) was sourced. Test sites were selected in Florida, New York, and Shanghai to represent a globally diverse range of water body typologies. Region-specific spectral water segmentation algorithms were developed on each test site, to represent benchmarks of spectral index performance. DeepLabV3-ResNet101 was trained on 33,311 semantically labelled true-colour samples. The resulting model was retrained on three smaller subsets of the data, specific to New York, Shanghai and Florida. CNN predictions reached a maximum mean intersection over union result of 0.986 and F1-Score of 0.983. At the Shanghai test site, the CNN's predictions outperformed the spectral benchmark, primarily due to the CNN's ability to process contextual features at multiple scales. In all test cases, retraining the networks to localised subsets of the dataset improved the localised region's segmentation predictions. The CNN's presented are suitable for cloud-based deployment and could contribute to the wider use of satellite imagery for water management.

1. Introduction

Near-real time mapping of water bodies from satellite imagery plays a critical role in water management. The continuous monitoring of environmental change over time, such as estimation of water availability, prediction of floods, and droughts, is essential to human activities such as agriculture, hydrology, and management (Molden, 2007; Schanze, et al., 2006, Ferral, et al., 2019). As a result, there has been significant interest in identifying methods of accurately automating the water segmentation of satellite imagery.

A large body of research has been devoted to the development of Spectral Indices (SIs) to automate water mapping tasks (McFeeters, 1996; Feyisa, et al., 2014; Xu, 2006; Jain, et al., 2020, Zhou et al 2017). SIs are the most prominent tool for automated water mapping and are widely integrated with geospatial software platforms and application programming interfaces (API's). SIs classify each image pixel independently without processing the contextual features of an image. Subsequently, the performance of SIs is hindered by features such as shadow or bright objects such as roofs and solar panels. Additionally, the process of selecting and optimally thresholding a SI is a complex arduous task that must be performed by an experienced professional.

In recent years, the expansive growth in the availability and capabilities of graphics processing units (GPU's) has driven the development of sophisticated deep learning (DL) architectures, and more specifically, convolutional neural networks (CNN's). Innovations in CNN architecture has enabled multiscale contextual detection of features within a scene (Chen, et al., 2017). This has led to a surge of interest in state-of-the-art CNN applications to classify land with semantic segmentation (Hoeser and Kuenzer, 2020; Tsagkatakis, et al., 2019). CNN's been hugely successful when used on very high-resolution imagery (< 1 m × pixel), with reported overall accuracy scores that exceed 99% (Talal, et al., 2018; Chen, et al., 2018). CNN's have been less successful on medium resolution imagery, achieving segmentation results ranging from 84% to 97% overall accuracy (Isikdogan, et al., 2017; Wang, et al., 2020; Wieland and Martinis, 2020). Medium resolution imagery contributes to the majority of land mapping activities due to their typically higher spatial and temporal resolution, highlighting a need for further development within this field (Belward, A. and Skoien, J., 2015).

Studies tend to be localized to specific geographic regions and have benchmarked CNN predictions against a small group of spectral water segmentation indices, most often Normalized difference water index (NDWI) and Modified Normalized Difference Water Index mNDWI (Isikdogan, et al., 2017; Wang, et al., 2020; Guo, et al., 2020). This study seeks to determine how effective CNN's are in on a global scale, and if CNN's are able to outperform a wide range of spectral methods. Regarding the use of machine learning (ML) for water mapping, Land Remote-Sensing Satellite (Landsat) imagery was used in a ML framework in Nepal (Acharya et al., 2018, 2019) and China (Jiang et al., 2018), where the latter assessed also the performance of the surface water extraction for the entire scene. While the subpixel surface water coverage in urban environments was object of investigation in Sun et al. (2017), and a focus on detection of subpixel-scale inundation was proposed by Jones (2019).

Regarding CNN for Remote sensing classification, a recent increase in the output of literature could be seen by a systematic search carried out in Scopus; the query included title abstract and keywords, ("water AND segmentation AND with AND convolutional AND neural AND networks") and it was limited for document type (articles and reviews) and subject area earth and environmental sciences, resulted in 66 research paper between the 2017 and 2020. Out of this papers (Wieland and Martinis, 2020) used CNN and Sentinel-2 multispectral imagery to describe a methodology to map large-scale surface water change after drought in Germany. Hughes et al., (2020) used the CNN for classify Synthetic Aperture Radar (SAR) imagery.

The first water separation index developed for a multispectral sensor was the (NDWI (McFeeters, 1996). The index was built initially for a Landsat Thematic Mapper (TM) and uses the Near Infrared (NIR) band and Green band to delineate open water features, excluding soil and terrestrial vegetation. There are significant challenges associated with mapping shallow water due to shadow from large physical structures from built-up areas. Xu, (2006) modified the NDWI with mNDWI, replacing the NIR band with short-wavelength infrared (SWIR) band to better partition built-up areas. The resolution performance of mNDWI is limited by the typically lower resolution of the SWIR band. The mNDWI also produces a higher occurrence of false positives in shadow areas, such as cloud shadow, or on dark surfaces such as roads.

Feyisa et al. (2014) addressed this shadowing problem with two automated water extraction indices: $AWEI_{nsh}$ and $AWEI_{sh}$, optimized for environments with no shadow and shadow. The $AWEI_{sh}$ removes shadow pixels, while $AWEI_{nsh}$ has been designed specifically for urban areas. An alternative method was proposed by Mishra and Prasad (2015) to improve detail the detection of shallow water. This was achieved simply through the addition of an index using blue and NIR band. Jain, et al. (2020) built upon I, with PI, demonstrating a reduction in noise with the SWIR band instead of the NIR band. Errors often occur from spectral diversity within the water. Turbid water has higher reflectance in the NIR and above bands due to high concentrations of suspended sediment. This can be corrected by integrating the normalized difference built-up (NDBI) index (Zha, et al., 2004). False negatives can occur from water bodies that contain high concentrations of phytoplankton (Chen, et al., 2015). This can be corrected using the normalized difference vegetation index (NDVI) (Tarpley, et al., 2015). Table 1 summarizes all SIs described in this literature review.

Table 1: A summary table of all Spectral Indices related to water segmentation.

Indices	Equation	Merit	Limitation	Reference
NDWI	$(\rho_{\text{Green}}$ - (NIR)) / $(\rho_{\text{Green}}$ + (NIR))	NIR channel has higher resolution capabilities that other sensors.	Less capable of delineating between built-up areas and water.	(McFeeters, 1996)
mNDWI	$(\rho_{\text{Green}}$ – $(SWIR_1)$) / $(\rho_{\text{Green}}$ + $(SWIR_1)$)	Use of the SWIR band offers greater contrast between built-up areas and water bodies.	The SWIR bands are less capable at higher resolutions. Typically produces false positives on roads, shadows and dark surfaces.	(Xu, 2006)
$AWEI_{nsh}$	4 $(\rho_{\text{Green}}$ – $(SWIR_1)$) – 0.5 (NIR) + 2.75 $(SWIR_2)$	Capable of delineating water and dark surfaces that occur from shadow in built up urban areas	Typically produces false positives on roads, shadows and dark surfaces.	(Feyisa, et al., 2014)
$AWEI_{sh}$	$(\rho_{\text{Blue}}$ + 2.3 ρ_{Green} - 1.5 ((NIR) + $(SWIR_2)$)) / ((ρ_{Green} + (NIR) + $(SWIR_1)$ + $(SWIR_2)$)	Removed shadow pixels.	High albedo surfaces such as snow, white roofs and crop-coverings can produce false positives.	(Feyisa, et al., 2014)
I	$(\rho_{\text{Green}}$ – (NIR)) / $(\rho_{\text{Green}}$ + (NIR)) + $(\rho_{\text{Blue}}$ – (NIR)) / $(\rho_{\text{Blue}}$ + (NIR))	Improves the detail of shallow water detection.	Excess spectral noise.	(Mishra and Prasad, 2015)
PI	$(\rho_{\text{Green}}$ – $(SWIR_1)$) / $(\rho_{\text{Green}}$ + (NIR)) + $(\rho_{\text{Blue}}$ – $(SWIR_1)$) / $(\rho_{\text{Blue}}$ + (NIR))	Noise reduction resulting from SWIR use.	SWIR bands are less capable at higher resolutions.	(Jain, et al., 2020)
NDBI	(($SWIR_1$) – (NIR)) / (($SWIR_1$) + (NIR))	Identifying built up areas. Capable of isolating narrow water bodies.	Only applicable in areas of dense Noise occurs from any vegetation.	(Zha, et al., 2004)
NDVI	((NIR) - ρ_{Red}) / ((NIR) + ρ_{Red})	Can be used for calibrating against high water phytoplankton content.	Water bodies with low reflectance in both red and NIR can produce false positives.	(Tarpley, et al., 2015)

The rationale of this work is to develop a widely usable application Copernicus Sentinel-2 multispectral and true-color imagery, Red, Green and Blue (RGB) composite, for selected sites will be accessed and labelled manually, existing spectral algorithms will be fine-tuned to generate benchmarks that represent the optimal capabilities of spectral water segmentation methods, sites containing complex and diverse waterbodies were selected. This study also investigates the potential of geographically localizing CNN's with the use of smaller subsets of data through transfer learning. The results of this study hope to contribute to the development of automated water segmentation tools to streamline access to earth observation analytics.

The aim of this work is twofold: i) Determine if water segmentation using CNN's on Sentinel-2 row imagery can outperform multispectral water segmentation indices and, ii) determine if transfer learning with small geographically localized datasets can improve CNN water segmentation's performance in specific regions.

2. Material and methods

2.1 Data Preparation

Suitable areas of interest were selected using google earth imagery. Satellite data from the sites was downloaded from the Sentinel-hub API and labelled. Test sites from Florida, New York and Shanghai were set aside for testing.

As part of the Copernicus programme of the European Commission (EC), the European Space Agency (ESA) has launched the Sentinel-2 constellation (Drusch et al.,2012). The Copernicus programme aims to enable atmospheric, land and marine environment monitoring, climate change research, emergency management, and support security. The constellation consists of two satellites, 2A and 2B (Drusch, et al., 2012). The purpose of the Sentinel-2 mission is to monitor global land surfaces and coastal waters continuously. The Sentinel-2 constellation systematically acquires imagery between −56° to 84° latitude. Sentinel-2 is sun-synchronous at 786 km altitude with 14 + 3/10 revolutions per day. The Sentinel-2 satellites are equipped with filter-based push-broom imager multispectral (MSI) sensors. The bands at *10* m resolution are the blue (458 to 523 nm), green (543 to 578 nm), red (650 to 680 nm) and near-infrared (NIR) (785 to 900 nm). There are 6 bands of 20 m spatial resolution, four of which are narrow bands (689 to 713 nm, 733 to 748 nm, 773 to 793 nm, 855 to 875 nm), primarily used for vegetation characterisations, and two SWIR-1 (1565 to 1655 nm), SWIR-2 (2100 to 2280 nm) used for detecting clouds, snow and ice and vegetation moisture measurements. There are 3 bands of 60 m spatial resolutions: aerosols (433 to 453 nm), important for analysing the oceanic ecosystem and water vapour (935 to 955 nm) Shortwave infrared for Cirrus detection (1360 to 1390 nm); these bands are used for atmospheric corrections (Gascon, et al., 2017). The Sentinel-2 scenes were accessed using the Sentinel-hub Web Coverage Service (WCS).

The Sentinel hub application programming interface (API) was used to source all data used within this study. The API enabled programmatic processing and integration of satellite data into a Python environment. The data is made available through two different levels: Level-1C (L1C) and Level-2A (L2A). L1C corresponds to top-of-atmosphere (TOA) observations, while the L2A is an atmospherically corrected bottom-of-atmosphere (BOA) product. Specific layer configurations were set up to generate the data. Bands 1 to 12 were sourced from both L1C and L2A products and corresponding true-colour composite were accessed for each selected scene.

2.1.1 Site Selection and Data Acquisition

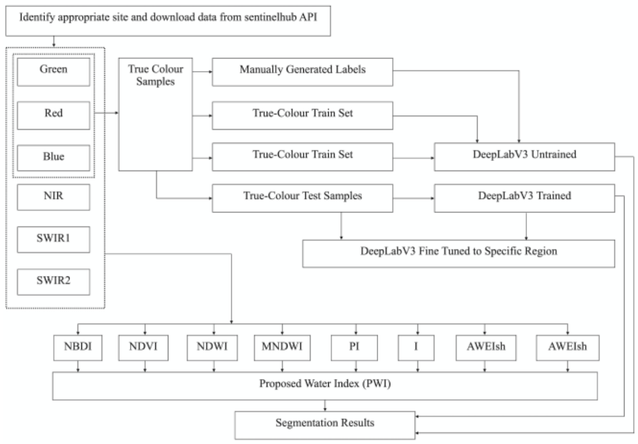

Figure 1. Flowchart to show the stages of model training and proposed water segmentation index development.

The workflow is summarised in a flowchart (Figure 1). Sites were identified with the aid of the google earth imagery platform. Areas of heavily built-up and complex sea to land interfaces or locations with densely packed diverse inland water bodies were selected. The training data was selected from the Netherlands, Osaka, New York, and Florida (Figure 2). The Sentinel-hub EO Browser was used to quick search for suitable acquisition dates. From selected sample sites and acquisition dates, the 10 m resolution, 12 multi-spectral layers data and a corresponding true colour reconstruction was requested from the Sentinel-hub API. Samples downloaded for the labelling purposes were both L1C and L2A products and were filtered at 0% cloud coverage to prevent any incorrect labelling. The samples intended for labelling were also selected from summer months, to prevent mislabelling due to periodic snow or ice within the scene. Once labelled, additional scenes at the same location were downloaded at both the L1C and the L2A processing levels, with a maximum cloud cover filter of 20%.

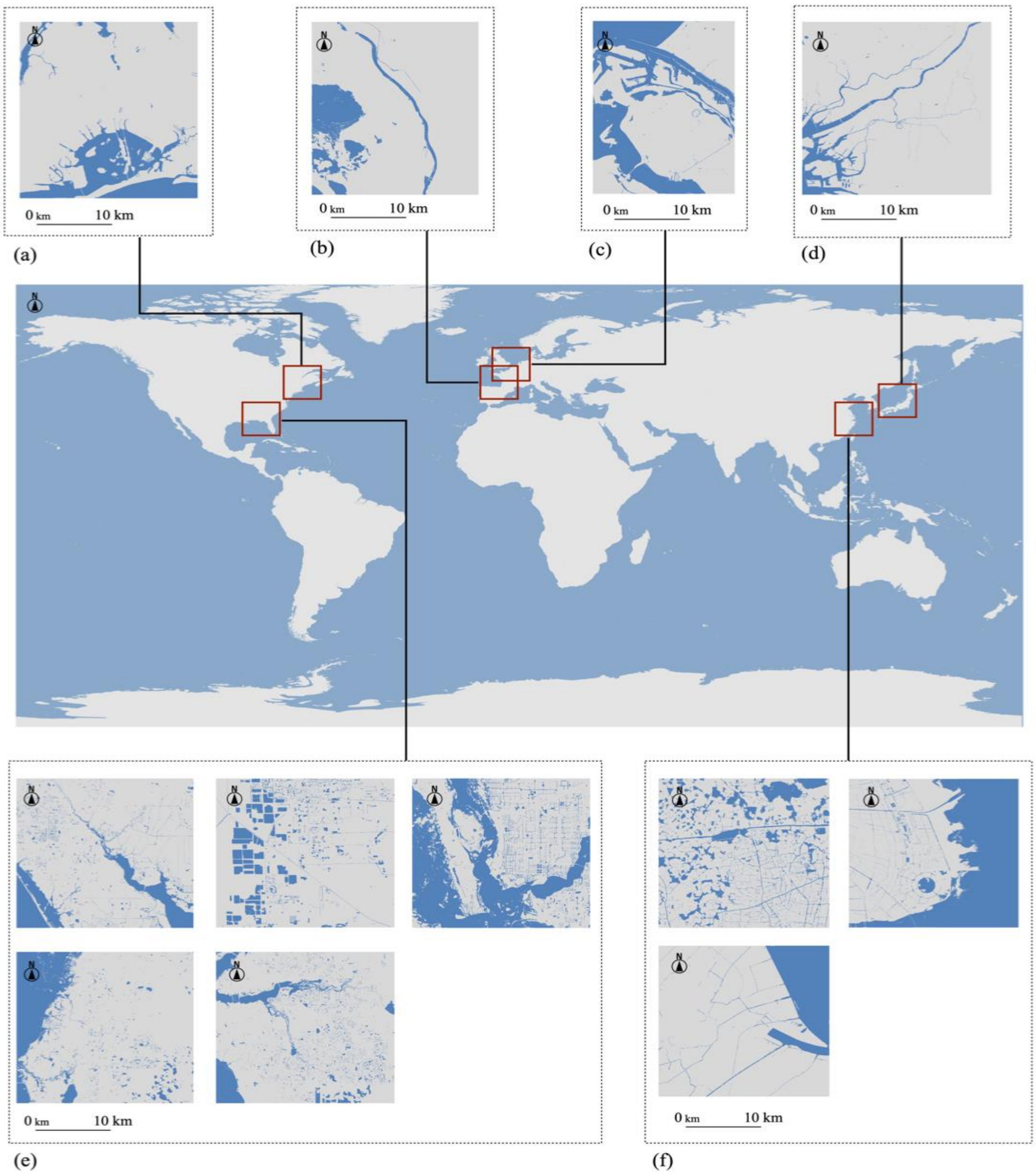

Figure 2. Labels for all samples used in the training process. a) New York, b) Marseille, c) Rotterdam, d) Osaka, e) Florida region, d) Shanghai and surrounding region.

Every sample used in the training process and for evaluation purposes was labelled by photointerpretation and validated using at least 3 images throughout the year. The labelling process was aided by photoshop tools, primarily the ‘magic wand’ tool. The magic wand tool accepts a colour value of the selected pixels and expands the selection area to all neighbouring pixels of a similar colour value to build the “Region of

Interest" ROI. Large water bodies with homogenous colour could be quickly labelled, however scenes containing variable water texture required a more fragmented and attentive approach. Small localised water bodies required meticulous examination to ensure they were not missed. To reduce the number of falsely identified pixels, the scene was checked against 0.3 m resolution imagery obtained by miscellaneous sources through google earth imagery at a scale of at least 1:1000.

Once the sites boundaries were defined and were fully labelled, imagery of the same region were downloaded and matched with the original scene labels. The additional scenes were within the closest possible time periods to the original image to reduce any changes that may have occurred over time. These duplicates were chosen to represent the variability in both atmospheric and surface properties that can be expected within a scene.

The CNN used required input channels with dimensions of **3 × 244 × 244**, and labels with dimensions **1 × 244 × 244**. To preserve the 10 m resolution of the original samples, the images and the labels were spliced into sub-samples with dimensions of **224 × 244** (Figure 3). The number of spliced subsets equates to the number of samples stated for model training.

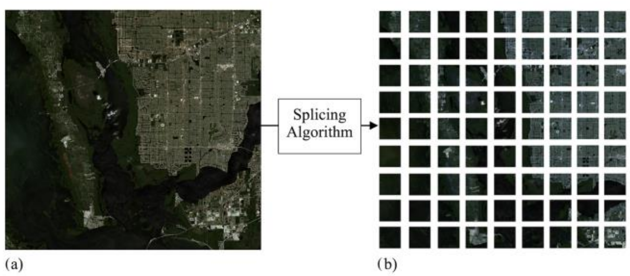

Figure 3. Input and output of splicing algorithm used to generate 244 x 244 pixel samples compatible with the DeepLabV3 model. a) Original true-colour image of Fort Myers, Florida. b) Corresponding true colour images spliced into 244 x 244 samples.

2.1.2 Test Sites

Three areas were delegated and preserved specifically as a benchmark for evaluating water mask predictions. The areas were not exposed to the CNN at any stage of the training process. These three sites have been displayed in Figure 4 and Table 2. All three test sites cover a mixture of heavily urbanised and rural land use. The first evaluation area was a 21.96 km × 19.52 km region covering Jacksonville Florida. The area was chosen due to the extraordinary density of small lakes within the land, and the complex meandering inland river network. The second area was a 19.52 km × 19.52 km region covering New York. The area was chosen primarily due to the densely packed tall buildings with extensive shadowed regions. The third area chosen was a 21.96 km × 19.52 km region of the northern section of Shanghai City, this area enabled the model to be evaluated on a transient intertidal zone with high levels of suspended sediment. Additionally, the area has a very high density of both large and small boats.

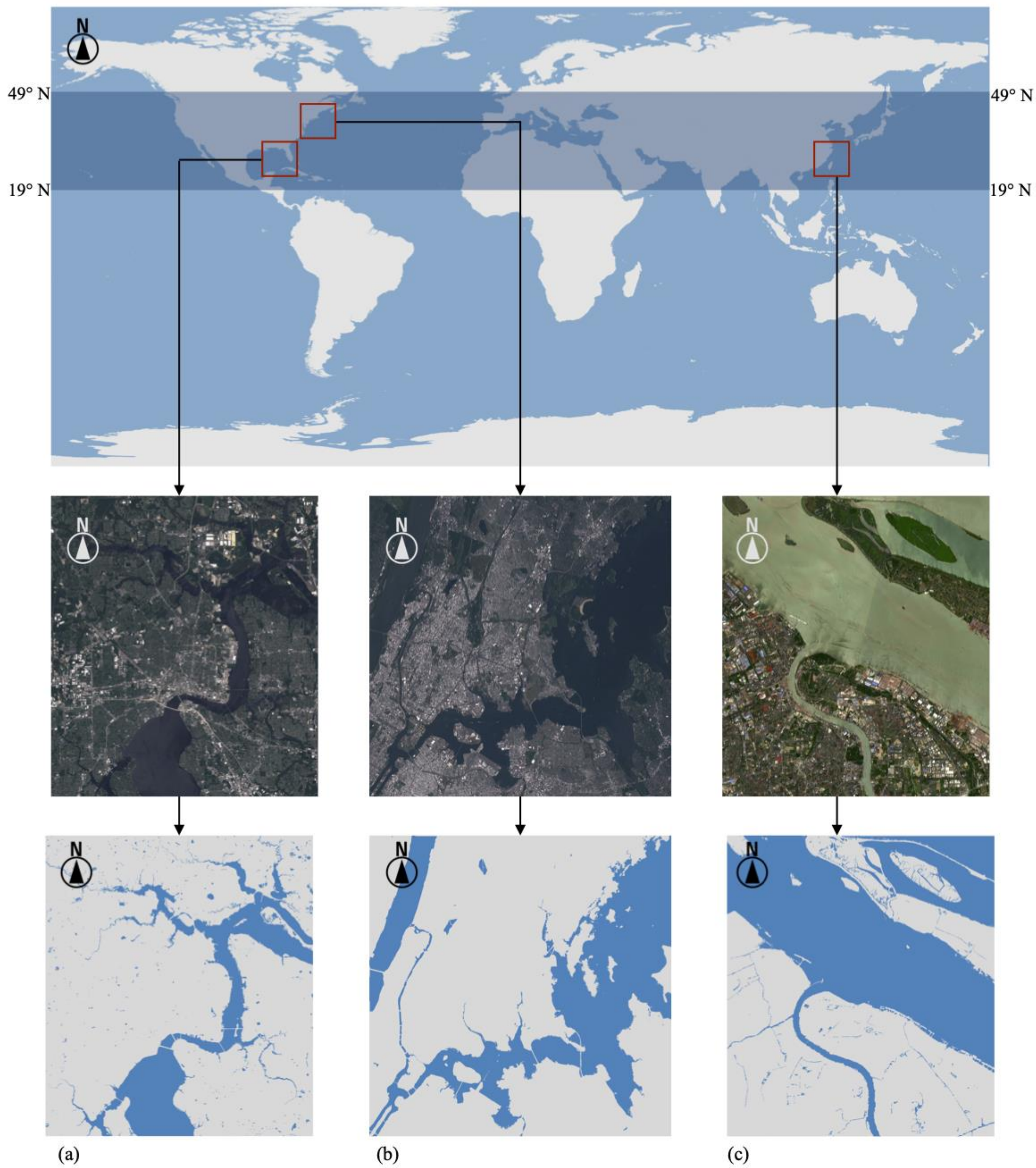

Figure 4. Test sites chosen to test the quality of the model for a) Jacksonville, Florida, b) New York, c) Shanghai. Each test site depicted in true colour form and corresponding binary classification label.

2.2. Benchmarks

The mIoU values for a range of SIs were calculated. SIs were developed for each test site through parameterisation with NBDI and NVDI indices.

2.2.1 Development of Spectral Benchmark

To test the performance of the model against benchmark SI, an optimised SI was generated for each evaluation sample to represent a theoretical best-case performance for what could be achieved through spectral methods. The process has been summarised in figure 5.

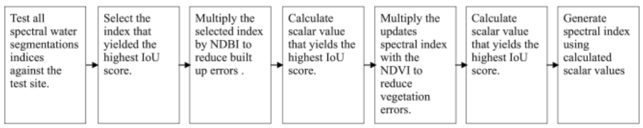

Figure 5. Flowchart to describe the stages of development for a spectral benchmark at each test site.

2.2.2 Algorithm tuning

A novel method was developed using regression curves was used to ‘fine-tune’ the algorithms. A preliminary assessment of the available SI showed that the index I, AWEIsh, and NDWI, demonstrated the highest-ranking performance for Florida, Shanghai and New York respectively. To ‘fine-tune’ this algorithm, it was multiplied by the NDBI and NDVI at different scalars. The scalars were plotted against mIoU results in a regression curve. The optimised scalar values were derived from the regression curve to produce an optimised SI for each test site. The finalised optimised spectral algorithms have been denoted in equations 1, 2 and 3.

$$\mathrm{PWI}_{\mathrm{Florida}} = -0.4\,\frac{(\mathrm{SWIR}_2)-(\mathrm{NIR})}{(\mathrm{SWIR}_2)+(\mathrm{NIR})} + \frac{\rho_{\mathrm{Green}}-(\mathrm{NIR})}{\rho_{\mathrm{Green}}+(\mathrm{NIR})} + \frac{\rho_{\mathrm{Blue}}-(\mathrm{NIR})}{\rho_{\mathrm{Blue}}+(\mathrm{NIR})} + 0.2\,\frac{(\mathrm{NIR})-\rho_{\mathrm{Red}}}{(\mathrm{NIR})+\rho_{\mathrm{Red}}} \quad (1)$$

Equation 1: Proposed spectral index for the Florida test scene: $\mathrm{PWI}_{\mathrm{Florida}}$

$$\mathrm{PWI}_{\mathrm{Shanghai}} = 0.5\,\frac{(\mathrm{SWIR}_2)-(\mathrm{NIR})}{(\mathrm{SWIR}_2)+(\mathrm{NIR})} + 4\left(\rho_{\mathrm{Green}} - (\mathrm{SWIR}_1)\right) - \frac{0.25\,(\mathrm{NIR})+2.75\,(\mathrm{SWIR}_2)}{\rho_{\mathrm{Green}}+(\mathrm{SWIR}_1)+(\mathrm{SWIR}_2)+(\mathrm{NIR})} - 3.4\,\frac{(\mathrm{NIR})-\rho_{\mathrm{Red}}}{(\mathrm{NIR})+\rho_{\mathrm{Red}}} \quad (2)$$

Equation 2: Proposed spectral index for the Shanghai test scene: $\mathrm{PWI}_{\mathrm{Shanghai}}$

$$\mathrm{PWI}_{\mathrm{New\ York}} = -0.1\,\frac{(\mathrm{SWIR}_2)-(\mathrm{NIR})}{(\mathrm{SWIR}_2)+(\mathrm{NIR})} + \frac{\rho_{\mathrm{Green}}-(\mathrm{NIR})}{\rho_{\mathrm{Green}}+(\mathrm{NIR})} + \frac{\rho_{\mathrm{Blue}}-(\mathrm{NIR})}{\rho_{\mathrm{Blue}}+(\mathrm{NIR})} + 0.3\,\frac{(\mathrm{NIR})-\rho_{\mathrm{Red}}}{(\mathrm{NIR})+\rho_{\mathrm{Red}}} \quad (3)$$

Equation 3: Proposed spectral index for the New York test scene: $\mathrm{PWI}_{\mathrm{NewYork}}$

2.2.3 Benchmark Algorithm Threshold Optimisation

The optimum threshold value for the benchmark water indices was determined by plotting the result of the probability density function of pixel intensity values of the proposed water index. This enabled visual inspection of the distribution of pixel intensity. The segmentation threshold was chosen by visually identifying the lowest point between the two intensity peaks. This was performed iteratively in unison with the algorithm tuning step (Figure 6).

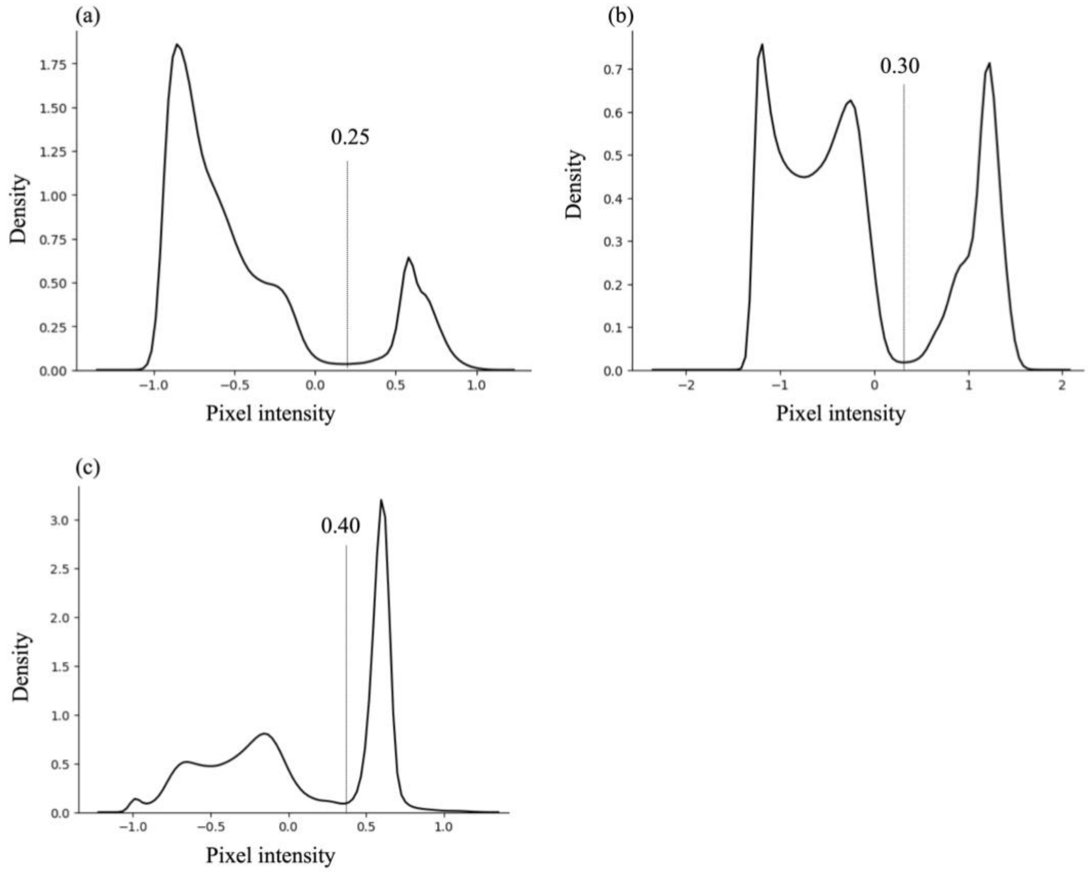

Figure 6. Distribution plots to show the results of the probability density function of pixel intensity values of the proposed water index for a) Florida, b) New York, c) Shanghai. Includes indication of the values of optimal thresholds.

Table 2: Summary table for test site data

	Jacksonville Florida	**New York**	**Shanghai**
Coordinates (WGS84)	-81.761727, 30.241694, -81.507889, 30.454348	-73.957111, 40.717802, -73.703274, 40.930456	120.69626, 30.873884, 120.950097, 31.086538
Sample features	Small densely distributed lakes and urban water bodies. Narrow, convoluted, intertidal rivers. Variable water texture and reflectivity.	Densely packed tall buildings and with extensive shadowing.	Large intertidal zone with complex tributaries and patched of sediment rich water. Variable water texture and reflectivity.
Dimensions	2196 pixels × 1952 pixels 21.96 (km) × 19.52 (km)	1952 pixels × 1952 pixels 19.52 km × 19.5 2km	2196 pixels × 1952 pixels 21.96 km × 19.52 km
Product	L1C (BOA)	L1C (BOA)	L1C (BOA)

2.3. Image Segmentation and CNN

2.3.1 Semantic segmentation

The semantic segmentation refers to the process of making pixel-wise predictions for a given image (Long, et al., 2015). The potential methods of scene classification have been depicted in figure 7. Semantic segmentation differs from image recognition, object detection and instance segmentation in that every pixel in the image is given a classification, in this case, black pixels represent planet earth, while pink pixels in. For EO classification tasks, semantic segmentation is the classification method of choice, due to their applicability to land surface classification and change detection tasks (Jain, et al., 2020; Hoeser and Kuenzer, 2020).

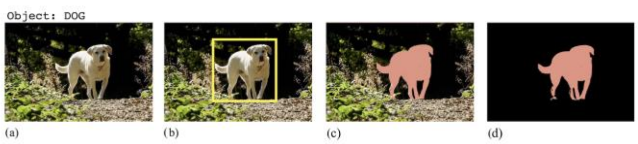

Figure 7. Depiction of various computer vision classification tasks. a) Object Detection, b) Object Localisation, c) Instance Segmentation, d) Semantic Segmentation.

2.3.2 Evaluating Segmentation

The similarity of the segmentation prediction and 'ground truth' indicates the quality of the prediction. Many evaluation criteria have been proposed to evaluate the quality of the performance of a given segmentation method. Intersection over Union (IoU) and F1-scores are the two most frequently used metrics of evaluation for computer vision semantic segmentation tasks. All the metrics used within this study have been outlined in table 3.

IoU computes a ratio between the intersection and the union of the prediction and the ground truth. This returns a value between 0 and 1. A value of 1 indicates a segmentation result that perfectly matches the ground truth (Figure 8).

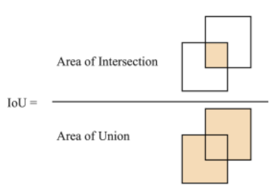

Figure 8. Visualised formula for the computation of the intersection over union (IoU) metric.

Where more than one class exists, the mean Intersection over Union (mIoU) can be calculated by taking the mean of the IoU values across all the classes (Garcia-Garcia, et al., 2018).

The F1 score is a statistical metric for evaluating classification that represents the harmonic mean between the precision and recall. The metric returns a result between 0 and 1, where 1 indicates both the precision and the recall was perfect (Sasaki, 2007).

It is important to note that there is no metric that perfectly represents the quality of a semantic prediction. The IoU penalizes instances of incorrect classification more than the F1-score. Therefore, a lower IoU score can be expected.

Table 1: Table to describe metrics used within this thesis.

Metric	**Description**	**Formula**
Total number of classes	Number of classes defined, (2 for a binary task).	K
True Positive	Sum of correctly identified water pixels.	TP
True Negative	Sun of correctly identified non-water pixels.	TN
False Positive	Sum of pixels incorrectly identified as water	FP
False Negative	Sum of pixels incorrectly identified was non-water	FN
Precision (P)	The proportion of water detected water pixels	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall (R)	The proportion of ground truth water pixels detected	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
F_1 Score	The harmonic mean between the precision and recall	$2 \times \frac{R \times P}{R + P}$
Intersection over union (IoU)	The ratio between the intersection and the union of the predicted segmentation and the ground truth.	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$
Mean Intersection over Union (mIoU)	The mean of the IoU values across all the classes	$\frac{1}{K}\sum_{k=1}^{K} \mathrm{IoU}_k$

2.3.3 Choice of Model: DeepLabV3

A CNN that is tasked with the segmentation of these water bodies must have the capability to learn features that are spatially invariant and complex in nature. Based on a review of existing architectures for semantic segmentation, a naïve decoder architecture was chosen over an encoder-decoder. The results reported by Guo, et al. (2020) in particular showed that the models that used bilinear upsampling were better suited to water segmentation tasks. The model chosen for this task was the DeepLabV3-ResNet101 model (Figure 9). This is a state-of-the-art CNN with that is currently ranked the third highest performing network on the PASCAL-VOC 2012 t*est dataset* for semantic segmentation and the second highest performing naïve decoder (Hoeser and Kuenzer, 2020). The segmentation of water bodies is possible primarily due to the atrous spatial pyramid pooling section of the network combined with 6.091×10^7 trainable parameters, enabling the CNN to understand features at depth, across multiple scales.

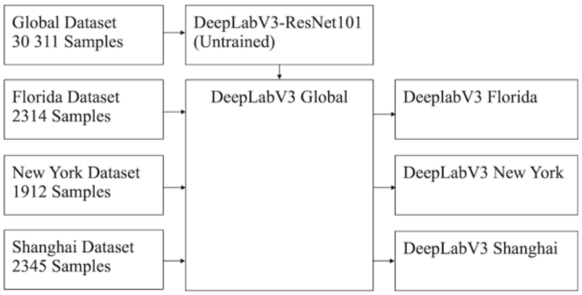

Figure 9. Model training flow chart.

2.3.4 Models

This study presents four models that were trained and evaluated within this study (figure 7), train: DeepLabV3 was retrained with the 33,331 samples collected in step one.

1. DeepLabV3_Global:
   The DeepLabV3 model was loaded in a 'untrained' form. The hyperparameters were adjusted and the model was retrained with all 33,311 training samples. Intended for water segmentation tasks independent of location.
2. DeepLabV3_Florida:
   DeepLabV3_global was retrained with 2314 samples from the state of Florida. intended to complete water segmentation tasks in Florida.
3. DeepLabV3_New_York:
   DeepLabV3_global was retrained with 1912 training samples within New York and Pennsylvania. Intended to perform water segmentation tasks in the New York area.
4. DeepLabV3_Shanghai:
   DeepLabV3_global was retrained with 2345 training samples within Shanghai and small surrounding cities. Intended to perform water segmentation tasks in the New York area.

2.3.5 Dataset Manipulation

The non-test dataset was randomly split so 80% of the training samples were training data, and the remaining 20% of the data was validation data. A train-loss and a validation-loss was computed for each batch. Where the train-loss exceeded the validation loss, the model was considered to be underfitting, and if they validation loss exceeded to the train loss then the model was considered to be overfitting.

The dataset was augmented to enhance the size, quality, and diversity of the training data set. This acts as a regularizer to reduce overfitting. Synthetically created duplicates of the training set were created by combinations of horizontally and vertically flipped images before splicing. Further augmentation was done with 'salt and pepper' noise and by blurring the samples. By training the network on deliberately noisy data, it was hoped that the model would better generalise when tested on noisy data.

The model hyperparameters were 'fine-tuned' to find an optimal trade-off between bias and variance. The model training was performed on a smaller subset of the dataset containing 1000 samples. The training loss and validation loss were recorded at each batch and plotted in training logs. The relationship between train-loss and validation loss was used to fine tune the hyperparameters. The final training hyperparameters have been summarised in table 4.

Trial runs of model training with learning rates of × 10^{-2}, × 10^{-3}, × 10^{-4}, and × 10^{-5} were tested. A learning rate of × 10^{-3} was chosen for all the training. This learning rate was found to be an optimal trade-off between large gradient descent step sizes that fail to identify global minima by overshooting, and gradient descent step sizes cause convergence on local minima and require impractical training time periods.

The number of epochs refers to the number of times an algorithm will train through the entire dataset. Gradient descent is an iterative optimisation algorithm, therefore, requires more than one epoch. Each epoch is comprised of at least on batch. The size of a batch is determined by how many training samples are present within the batch and the number of iterations is defined at the number of batches required to complete one epoch (Smith, et al., 2017; Masters and Luschi, 2018). Figure 10 demonstrates the trajectory of the train-loss and validation-loss outputs when a model is trained for too many epochs. Each model was trained for between 60 and 90 epochs. At a specific point during the training process, the validation loss would start to increase. This was the indicator that the model was overfitting to the training data. This point was identified, and the model was configured to stop training at the identified epoch (Figure 10). The loss function used for the model was mean squared error (MSE).

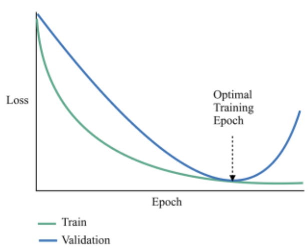

Figure 10. Visual representation of the how the optimal training epoch is chosen.

2.4. Transfer Learning

The resulting model from step three was retrained three times with smaller subsets of data from Florida, New York and Shanghai. Once the DeepLabV3 Global was successfully trained and evaluated, the model was loaded with the weights and re-trained with the smaller region-specific datasets. The Florida samples made up of sites limited to the state of Florida, the New York samples were limited to the sites within the State of New York and Pennsylvania. Shanghai training samples were limited to Shanghai and neighbouring cities Suzhou and Nantong.

Table 4: Summary of Model Hyperparameters, all parameters were set using the validation score.

Model name	**Optimiser**	**Learning rate**	**Loss function**	**Number of training sample**	**Epochs stop**	**Max epoch**	**Training-validation split**	**Batch size**
DeepLabV3 Global	Adam	1×10^{-3}	MSE	33331	10	60	80/20	20
DeepLabV3 Florida	Adam	1×10^{-3}	MSE	2314	25	90	80/20	20
DeepLabV3 New York	Adam	1×10^{-3}	MSE	1912	25	90	80/20	20
DeepLabV3 Shanghai	Adam	1×10^{-3}	MSE	2345	25	90	80/20	20

2.5. Model Evaluations

The trained models were used to generate water mask predictions for each test site. The predictions were compared to the ground truths and the water mask predictions made using the SIs developed in step two. The predictions were quantitatively and qualitatively analysed.

The evaluation of a semantic segmentation output is conventionally done using metrics. However, the evaluation can benefit from a parallel qualitative analysis to visually identify the relationships and patterns that may exist.

To evaluate quantitatively the CNN's performance against spectral water segmentation methods, the F1-Scores and mIoU results were computed for all the CNN and SI predictions on all three test sites. Each metric was calculated by comparing the prediction to the manually generated ground truths. Comparisons are made between the CNN's and the SI benchmark.
To determine whether transfer learning improved the results with respect to a specific region, the performance of DeepLabV3_Florida, DeepLabV3_New_York and DeepLabV3_Shanghai was compared to DeepLabV3_Global for each test site.

Qualitative observations were made about the overall prediction quality and how the CNN's responded to contextually dependent features. Special attention was focused on transient features such as boats and intertidal zones or wetlands. This was achieved by visually identifying specific sources of false positives and false negatives for all the segmentation methods. The characteristics that were exclusive to each method or common to both methods were noted.

It is important to note the following assumptions made within this study:

1. The evaluations made within this study are made on the assumption that the test sites were labelled within an error margin of 10 m (1 pixel). The use of high-resolution third-party validation data combined with a manual pixel-wise classification enabling accuracy to be maximised, however there is no existing benchmark to validate the accuracy of the test site labels.
2. The additional scenes used to expand the data, was assumed to have equal water body limits as the original labelled sample.
3. The spectral benchmark algorithms presented within this study are assumed to be the best possible representation of what can be achieved using SI.

3. Results

3.1 Test Site: Jacksonville, Florida

The results of the segmentation predictions for DeepLabV3_Global, DeepLabV3_Florida and, $PWI_{Florida}$ have been displayed in Figure 11 and table 5. All SIs performed better than the CNN predictions (Figure 12). The mIoU results for DeepLabV3_Global, DeepLabV3_Florida and the $PWI_{Florida}$ were 0.913, 0.918, 0.93, respectively with F1-Scores of 0.923, 0.927, 0.943, respectively. Retraining DeepLabV3_Global on the 2314 sample Florida dataset increased the mIoU result and F1-Score by 0.004 respectively.

Predictions made on the Jacksonville Florida test site yielded the lowest segmentation performance of all the test sites for all segmentation methods.

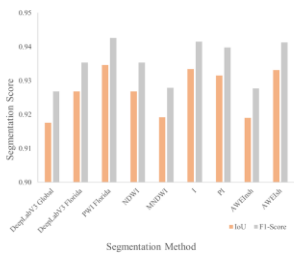

Figure 11. Bar chart to compare the F1-Score and the IoU scores for all water segmentation methods for test site Jacksonville Florida.

Table 5: Florida Segmentation Results, the best performing

	mIoU	F_1 Score
DeepLabV3_Global	0.91759	0.9268
DeepLabV3_Florida	0.9268	0.9353
$PWI_{Florida}$	**0.9346**	**0.9426**

Both DeepLabV3_Global and DeepLabV3_Florida identified all the boats (supplemental materials figure 1) within the scene as water, however both CNN's were unable to segment the protruding structures such as pontoons, jetties and harbours. The boundaries were spatially consistent but appeared to be generalised.

The $PWI_{Florida}$ predictions were sharper around the land-sea interface. However, large sediment-rich water in bays, inlets, and rivers as land was segmented as land.

In the case of DeepLabV3_Global small sections of false positives occurred in areas of forest vegetation. This error was reduced by retraining on the local dataset, and therefore not present in the DeepLabV3_Florida prediction, however sections of false negative predictions occurred larger water bodies where they had not occurred in the DeepLabV3_Global predictions. For example, a section of water at the mouth of the estuary, classified as land by DeepLabV3_Florida.

A common characteristic of all the predictions was the misidentification of riverine wetlands and failure to identify narrow (< 10 m) rivers and small (< 30 m) water bodies.

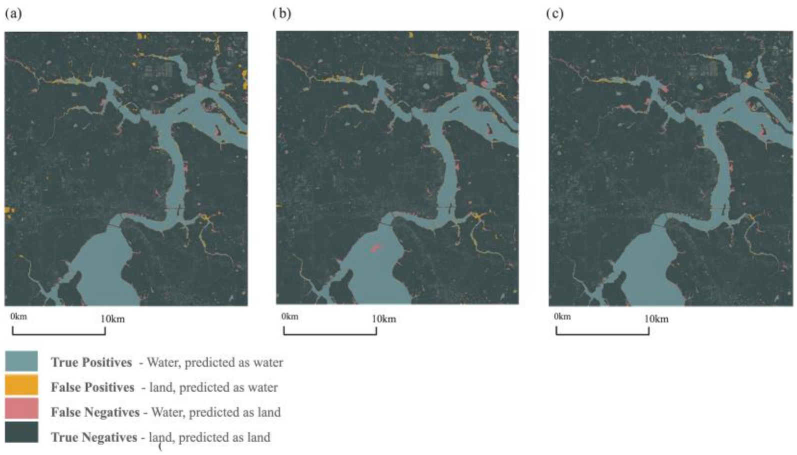

Figure 12. Plots to compare segmentation results of the DeepLabV3 Global Model trained on all 33,311 training samples, the DeepLabV3_Florida model retrained on imagery from the surrounding area and the PWIFlorida spectral index, fine-tuned specifically for the test site. The mask has been overlaid on the original true colour sample at 83% opacity.

3.2 Test Site: New York

The results of the segmentation predictions for DeepLabV3_Global, DeepLabV3_New_York and the $PWI_{Florida}$ have been displayed in Figure 13 and table 6. The CNN’s were outperformed by all SIs with the exception of the index I and AWEIsh (Figure 14). The mIoU results for DeepLabV3_Global, DeepLabV3_New_York and the $PWI_{NewYork}$ predictions were 0.969, 0.975, 0.976 respectively with F1-Scores of 0.978, 0.982, 0.983 respectively. The quality of the water segmentations in the New York test site were best of all the test sites, for all segmentation methods.

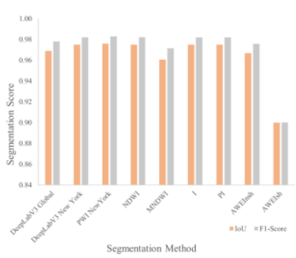

Figure 13. Bar chart to compare the F1-Score and the IoU scores for all water segmentation methods for test site New York.

Re-training of DeepLabV3_Global to samples local to New York increased the mIoU result and F1-Score by 0.006, 0.005 and respectively, this was the largest improvement of all the test sites.

Table 6: New York: Segmentation Results

	mIoU	F_1 Score
DeepLabV3_Global	0.969	0.978
DeepLabV3_New_York	0.975	0.982
$PWI_{NewYork}$	**0.976**	**0.983**

DeepLabV3_Global predictions in New York mirrored those of the Florida test site with very large false positive occurring over the forest areas, particularly in Van Cortlandt Park. This was reduced in the DeepLabV3_New_York predictions by retraining on localised imagery.

Both DeepLabV3_Global and DeepLabV3_New_York were unable to accurately label sections of river that were less than 30 m wide. The CNN's generalised across complex sections of the land-water interface, which mirrored the Florida test site results.

The DeepLabV3_New_York prediction demonstrated a reduced ability to detecting bridges and features compared to the DeepLabV3_Global prediction.

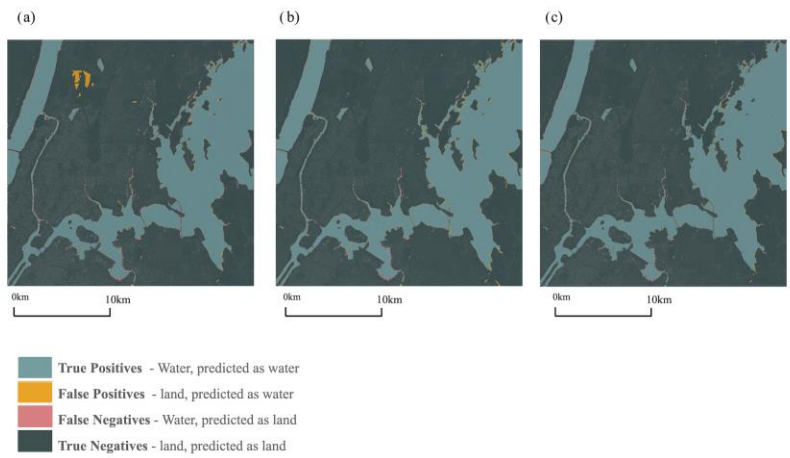

Figure 14. Plots to compare segmentation results of the DeepLabV3 Global Model trained on all 33,311 training samples, the DeepLabV3_New York model retrained on imagery from the surrounding area and the PWINewYork spectral index, fine-tuned specifically for the test site. The mask has been overlaid on the original true colour sample at 83% opacity.

3.3 Test Site: Shanghai

The results of the segmentation predictions for DeepLabV3_Global, DeepLabV3_Shanghai and the $PWI_{Shangahi}$ have been displayed in Figure 15 and table 7. Both CNN predictions outperformed all SIs (Figure 15). The mIoU results for DeepLabV3_Global, DeepLabV3_Shanghai and the $PWI_{Shanghai}$ predictions were 0.952, 0. 953, 0. 951 respectively with F1-Scores of 0.973, 0.974, 0.973 respectively. The transfer learning process increased the mIoU result and F1-Score by 0.001 respectively, indicating that the transfer learning process was the least effective in Shanghai than at any of the three test sites.

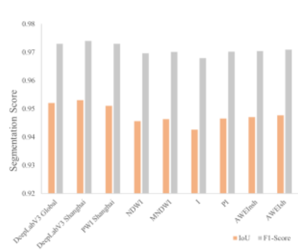

Figure 15. Bar chart to compare the F1-Score and the IoU scores for all water segmentation methods for test site Shanghai.

Table 7: Segmentation Results: Shanghai

	mIoU	F_1 Score
DeepLabV3_Global	0.952	0.973
DeepLabV3_Shanghai	**0.953**	**0.974**
$PWI_{Shanghai}$	0.951	0.973

$PWI_{Shanghai}$ predictions produced false positives at the locations of white roofs and solar panels. False negatives in areas of sediment rich water, particularly around features that are intertidally submersed. The $PWI_{Shanghai}$ algorithm was unable to predict at about 5.6×10^5 m^2 intertidal island in the middle of the coastal estuarine system, this island was however identified by DeepLabV3_Global. A zoomed in perspective of this has been displayed in Figure 16.

The contextually dependent features like the boats and the residual turbid water of the boats was mapped as land by the PWIShanghai index (supplemental materials figure 1). The CNN's demonstrated a capability to identify the boats, turbid water and map them as water. This has been demonstrated in closer detail in Figure 16.

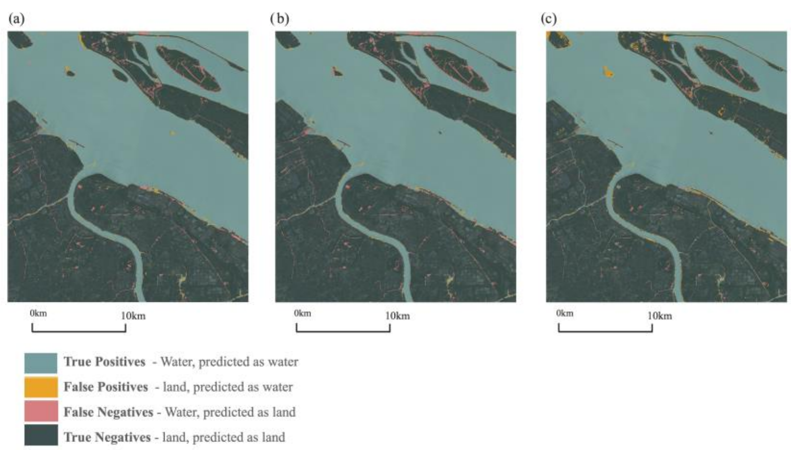

Figure 16. Plots to compare segmentation results of the DeepLabV3 Global Model trained on all 33,311 training samples, the DeepLabV3_Shanghai model retrained on imagery from the surrounding area and the PWIShanghai spectral index, fine-tuned specifically for the test site. The mask has been overlaid on the original true colour sample at 83% opacity.

A common characteristic of all segmentation predictions was the inability to identifying the narrow, sediment rich tributaries within the intertidal zone and the piers protruding into the estuary. This result was common to the CNN predictions of all test sites, Table 8.

Table 8: Results for all water segmentation predictions at all test sites. Numbers in bold are the highest scoring results.

	Florida		**New York**		**Shanghai**	
	mIoU	**F_1 Score**	**mIoU**	**F_1 Score**	**mIoU**	**F_1 Score**

DeepLabV3 Global	0.91328	0.9228	0.96106	0.9719	0.95165	0.9731
DeepLabV3 Fine Tuned	0.91759	0.9268	0.96873	0.9775	**0.95303**	**0.9738**
$PWI_{LOCATION}$	**0.9346**	**0.9426**	**0.9759**	**0.9827**	0.9512	0.9731
NWDI	0.9268	0.9353	0.975	0.9821	0.9456	0.9696
MNDWI	0.9192	0.9279	0.9605	0.9714	0.9463	0.9701
I	0.9334	0.9415	0.975	0.982	0.9426	0.9679
PI	0.9315	0.9398	0.975	0.982	0.9465	0.9702
AWEInsh	0.919	0.9277	0.9667	0.9758	0.947	0.9704
AWEIsh	0.9331	0.9413	0.8998	0.8998	0.9477	0.9709

Interestingly, despite an improved mIoU and F1-score, the DeepLabV3_Shanghai was less able to accurately identify islands in the middle of the estuary than DeepLabV3_Global (supplemental materials figure 2).

4. Discussions

From the obtained results, it was observed that CNN's are capable of outperforming SIs for water segmentation tasks. The CNN's demonstrated an ability to identify contextual features such as boats, turbid water and sediment rich intertidal water bodies. It was shown in all test cases that re-training the neural network to localised datasets improved prediction accuracy. This section explains these results and the associated successes and limitations. The results were placed within the context of existing literature with additional recommendations for further developments. The potential impact of these results on the field of earth observation will be discussed.

4.1 DeepLab_Global

The Shanghai test site results showed that the CNN's were capable of outperforming all available SIs. This was driven primarily by the intrinsic failures of spectral methods and CNN's ability to process context at multiple scales of an image.

The Shanghai test site was characterised by a large intertidal zone, sediment-rich water, and a high marine traffic volume. Suspended sediment within water bodies increases reflectance in NIR and SWIR radiation (Pham, et al., 2018). The gradient of reflectance between the VL and NIR and SWIR wavelengths was reduced causing large misclassification errors in the $PWI_{Shanghai}$ predictions in the East China Sea intertidal zone. The solar panels and white roofs were incorrectly mapped as water, this is likely attributed to the increased reflectance of VL, which in turn increases the gradient between VL and NIR/SWIR.

SI classify pixels on an individual basis without considering the context of the image. This accounted for the high quality prediction observed on the New York test site where the majority of water bodies are deep and there is a clear water-to-land interface due to the relatively small 0.5 m tidal range (Bowman, 1976). However, transient features such boats and turbid water are classified as non-water bodies. This an intrinsic error that could not be resolved through spectral methods.

The CNN approach is very different. The CNN's learned combinations of characteristics that make up a water body. These include edges, shapes, colour gradients and textural features (Zeiler and Fergus, 2014). The DeepLabV3 network used ASPP to examine convolutional feature layers with filters at multiple sample rates and fields of view (Chen, et al., 2017). This enabled the network to capture the various spatial contexts associated with water detection. The CNN was trained with 33,311 samples, this was a sufficient volume to develop a deep and rich contextual understanding of combinations of characteristics to make accurate prediction of sediment rich water bodies, boats and turbid water. The ability of the CNN to distinguish contextual features was the main driver of success when evaluated against a spectral methods of water segmentation at the Shanghai test site.

The CNN's were unable to accurately classify narrow meandering inland rivers, smaller water bodies (< 3 m) and complex structures protruding into the water. This was most observable in the Florida test case. The CNN's had a tendency to generalise across complex features, decreasing the prediction quality. This accounted for the poor overall segmentation predictions for the Florida test site where an extensive and complex land-sea interface exists. This was only partially reduced by retraining the model on region specific subsets of data, implying that the detection capabilities were partially limited by the CNN architecture. The 'black box' nature of deep neural networks makes drawing comparisons between results and network architecture difficult and speculative. However, it is clear that the 'smoothing' effect of the DeepLabV3 model precludes the model from achieving the same level of pixel-precision that is present in a SI segmentation. It is important to note that the DeepLabV3 model was built for computer vision tasks from terrestrial, close-range, side-view perspectives. The overhead perspective of EO imagery results in clustering and random distribution of features across which is very different from typical computer vision images.

4.2 Transfer Learning Performance

During the initial training stages, the DeepLabV3_Global model was shown a globally diverse range of water body typologies. The characteristics of these water bodies are heavily influenced by interdependent variables such as local geomorphology, weather patterns and human activities. These variables are often homogenous to a region. As an example, Florida is characterised by a porous plateau of karst limestone that allows water to move freely forming large wetlands and an extraordinary number of small lakes (Beck, 1986). The DeepLabV3_Global model learned the features to predict water bodies at all three sites based upon the generic characteristics of water bodies. Re-training the network with a small number of local samples reinforced the correct predications made by the DeepLabV3_Global model, while extracting characteristics that are specific to the local region and transferred the knowledge into the new network. This was particularly successful when applied to the New York test site; Large areas of forest were predicted as water by the DeepLabV3_Global model, but resolved in the DeepLabV3_New_York model prediction. It could be speculated that the DeepLabV3_Global model had fitted to the green texture rich water bodies in the Florida dataset and when knowledge was extracted and transferred to the DeepLabV3_New_York network, the error was eliminated.

In all test cases, the retraining of the networks resulted in some new errors that did not occur in the DeepLabV3_Global predictions. The most notable, unexpected error was the patch of water identified as land in the Florida test case. Deep learning models are known to be robust to label noise that is evenly distributed across a large dataset, yet highly sensitive to label noise that is concentrated within the dataset (Karimi, et al., 2020). Errors most likely arose from concentrated label noise within the smaller subsets of data. This noise would also be amplified in the augmentation process and transferred to the retrained network.

4.3 Comparisons with other CNN performances.

The Sentinel-2 mission has been in operation since 2015, which is still a relatively short time frame. As a result of this, the majority of studies covering water segmentation utilised different data sources. It is difficult to compare the performance of CNN's across different image resolutions. The most recent and closest matching study was the segmentation of water bodies within Sentinel-2 imagery exclusively in Germany by Wieland and Martinis, (2020). The results of the current study marginally improved upon this with more diverse and challenging urban test sites. The improved results could be attributed to the use of a more modern and sophisticated CNN that utilises ASPP. The use of complex and contextually rich training samples could also contribute to the small improvement in segmentation accuracy.

The findings of this study support the growing consensus that CNN's are becoming more capable than traditional SIs for land classification tasks. It is widely acknowledged that Deep Learning will be instrumental to sustainability and automation in the future (Gulati and Sharma, 2020). It can be expected that the development of network architectures will continue to improve and the subsequently, the quality segmentation tasks will follow.

4.4 Implications for the field

At a global scale Pekel et al. (2016) showed how long-term changes of water coverage are difficult to map and represent a societal challenge due to the documented reduction of inland water occurred in the Middle East, Central Asia, Australia and the USA. This is linked to drought and anthropic factors (Pekel et al., 2016) and reference therein. The different algorithms reported in the state-of-the-art were based on ML (Acharya et al., 2019), noise suppression methods used in order to mitigate the effect of landforms shadows and solid water forms (Jiang et al., 2018, 2020), and model fusion (Wagle et al., 2020).

For a long time, satellite imagery has been expensive and difficult to access for both individuals and organisations (Turner, et al., 2015). The barrier to entry has dropped significantly in recent years with the introduction high performance computing systems and large scale cloud- based computing frameworks, most notably 'Google Earth Engine' (GEE) (Gorelick, et al., 2017). However, to achieve reliable, high quality water mapping with SIs, expertise is required to select and optimise a SI. The CNN's developed within this study are easily deployable to cloud-based platforms. Very little skill is required to use a CNN within a platform like GEE. This could help broaden scope of the possibilities available to individuals and organisations who wish to use satellite imagery for water management.

SIs generally require EM radiation in the VL and the NIR and SWIR range. This adds a computational cost for image processing chains and a dependence on satellite sensors' multispectral capabilities. This study shows that state-of-the-art CNN's capabilities match and outperform SIs, potentially precluding the need for NIR and SWIR channel for water segmentation tasks. Alongside a large body of parallel research, this study could contribute to the development of streamlined satellite processing chains

5. Conclusion

Better results could be achieved through a redesign of the CNN architecture to better suit EO imagery. This could involve adjusting the dilation rates of the atrous convolution kernels to better suit the clustered nature of the water bodies. Alternatively, the use of an encoder-decoder network like DeepLabV3+ has the potential to improve segmentation performance. The incorporation of additional skip connections from the entry and middle blocks of the DeepLabV3+ encoder has been shown to sharpen segmentation outputs (Prabha, et al., 2020). Experimenting with this technique could make it possible to detect and localise very small, narrow and complex water bodies. Some recent studies have swapped RGB input channels for alternatives (Jain, et al., 2020). The performance of water segmentation with CNNs could be improved by replacing the RGB channels with band ratios or outputs of an existing spectral water index.

A further enhancement of the transfer learning aspect of this study could involve retraining DeepLabV3_Global to identify specific water typologies rather than geographic locations. For example, re-training DeepLabV3_Global on images collected in areas of karst limestone, instead of samples limited to Florida. CNN's trained to capture the characteristics of specific typologies would enable broader usage than a CNN retrained specifically to geographic location.

This study has shown that CNN's are an effective tool for the segmentation of water bodies in medium resolution satellite imagery. This was done by training the DeepLabV3-ResNet101 network with manually labelled Sentinel-2 imagery.

Three main conclusions can be made based upon this research:

i) CNN's can be applied to medium resolution true-colour satellite imagery to effectively map water bodies on a global scale.
ii) Water segmentation using CNN's on medium resolution true colour satellite imagery can outperform multispectral water segmentation indices.
iii) Transfer learning with small geographically localised datasets can improve the performance of CNN water segmentation in specific geographic regions.

Further developments of the study could include adjusting the network to improve segmentation sharpness and feature localization in EO imagery. Results could be improved by replacing the RGB input channels with alternatives such as band ratios or SI outputs. Additionally, the model presented within this study could be 'fine-tuned' for specific water body typologies.

The results of this study could help broaden and streamline the use of EO imagery for water management by improving the efficiency of EO processing chains and lowering the skill barrier.

6. Acknowledgment

We thank the Academic Editor, Professor Jida Wang and the anonymous reviewers for the helpful comments on the drafts of the manuscript.

7. References

Acharya, T. D., Subedi, A. and Lee, D. H.: Evaluation of Machine Learning Algorithms for Surface Water Extraction in a Landsat 8 Scene of Nepal, Sensors 2019, Vol. 19, Page 2769, 19(12), 2769, doi:10.3390/S19122769, 2019.

Beck, B., 1986. A generalized genetic framework for the development of sinkholes and Karst in Florida, U.S.A.. *Environmental Geology and Water Sciences,* 8(1), pp. 5-18.

Belward, A. and Skoien, J., 2015. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites.. *ISPRS Journal of Photogrammetry and Remote Sensing,* Issue 103, pp. 115-128.

Bowman, M., 1976. The tides of the East River, New York. *Journal of Geophysical Research,* 81(9), pp. 1610-1616.

Chen, L. C. et al., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Computer Vision – ECCV 2018,* pp. 833-851.

Chen, L.-C.et al., 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 40(4), pp. 834-848.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.. *arxiv.org.*.

Chen, Y. et al., 2018. Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning. *Water,* 10(5), p. 585.

Drusch, Del Bello U., Carlier S., Colin O., Fernandez V., Gascon F., Hoersch B., Isola C., Laberinti P., Martimort P., Meygret A., Spoto F., Sy O., Marchese F., Bargellini P., 2012 Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment,. *Remote Sensing of Environment,* Volume 120, pp. 25-36.

Ferral, A., Luccini, E., Aleksinkó, A. and Scavuzzo, C. M.: Flooded-area satellite monitoring within a Ramsar wetland Nature Reserve in Argentina, *Remote Sens. Appl. Soc. Environ*., 15, 100230, doi:10.1016/j.rsase.2019.04.003, 2019.

Feyisa, G., Meilby, H., Fensholt, R. & Proud, S., 2014. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment,* Volume 140, pp. 23-35.

Gascon, F.; Bouzinac, C.; Thépaut, O.; Jung, M.; Francesconi, B.; Louis, J.; Lonjou, V.; Lafrance, B.; Massera, S.; Gaudel-Vacaresse, A.; Languille, F.; Alhammoud, B.; Viallefont, F.; Pflug, B.; Bieniarz, J.; Clerc, S.; Pessiot, L.; Trémas, T.; Cadau, E.; De Bonis, R.; Isola, C.; Martimort, P.; Fernandez, V.2017. Copernicus Sentinel-2A Calibration and Products Validation Status. *Remote Sensing,* 9(6), p. 584.

Gorelick, N. Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment,* Volume 202, pp. 18-27.

Gulati, S. and Sharma, S., 2020. Challenges and Responses Towards Sustainable Future Through Machine Learning and Deep Learning. *Data Visualization and Knowledge Engineering,* Volume 32, pp. 151-169.

Guo, H.; He, G.; Jiang, W.; Yin, R.; Yan, L.; Leng, W. 2020. A Multi-Scale Water Extraction Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images *ISPRS International Journal of Geo-Information,* 9(4).

Hoeser, T. and 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), p. 1667.

Hughes, L.H., Marcos, D., Lobry, S., Tuia, D., Schmitt, M., 2020. A deep learning framework for matching of SAR and optical imagery. *ISPRS J. Photogramm. Remote Sens*. 169, 166–179. https://doi.org/10.1016/j.isprsjprs.2020.09.012

Isikdogan, F., Bovik, A., Passalacqua, P., 2017. Surface Water Mapping by Deep Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 10(11).

Jain, P., Schoen-Pelan, B., Ross, R., 2020. Automatic flood detection in SentineI-2 images using deep convolutional neural networks. *SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing*, p. 617–623.

Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K. and Wang, G.: Multilayer Perceptron Neural Network for Surface Water Extraction in Landsat 8 OLI Satellite Images, *Remote Sens*., 10(5), 755, doi:10.3390/rs10050755, 2018.

Jiang, W., He, G., Pang, Z., Guo, H., Long, T. and Ni, Y.: Surface water map of China for 2015 (SWMC-2015) derived from Landsat 8 satellite imagery, Remote Sens. Lett., 11(3), 265–273, doi:10.1080/2150704X.2019.1708501, 2020.

Jones, J.: Improved Automated Detection of Subpixel-Scale Inundation—Revised Dynamic Surface Water Extent (DSWE) Partial Surface Water Tests, *Remote Sens*., 11(4), 374, doi:10.3390/rs11040374, 2019.

Karimi, D., Dou, H., Warfield, S., Gholipour, A., 2020. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis.*

Long, J., Shelhamer, E., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE transactions on pattern analysis and machine intelligence,* 39(4), p. 640–651..

Masters, D. and Luschi, C., 2018. Revisiting Small Batch Training for Deep Neural Networks. *arXiv:1804.07612 [cs, stat]*.

McFeeters, S., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features.. *International Journal of Remote Sensing*, 17(7), pp. 1-9.

Mishra, K., Prasad, P., 2015. Automatic Extraction of Water Bodies from Landsat Imagery Using Perceptron Model. *Journal of Computational Environmental Sciences,* p. 9.

Molden, D., 2007. *Water for Food Water for Life: A Comprehensive Assessment of Water Management in Agriculture.* London, Earth Scan.

Pekel, J. F., Cottam, A., Gorelick, N. and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, *Nature*, 540(7633), 418–422, doi:10.1038/nature20584, 2016.

Pham, Q.V., Ha, N.T.T., Pahlevan, N., Oanh, L.T., Nguyen, T.B., Nguyen, N.T. Using Landsat-8 Images for Quantifying Suspended Sediment Concentration in Red River (Northern Vietnam). *Remote Sensing*, 10(11), p. 1841.

Prabha, R., Tom, M., Rothermel, M., Baltsavias, E., Leal-Taixe, L., and Schindler, K. 2020. Lake Ice Monitoring With Webcams And Crowd-Sourced Images, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci*., V-, 549–556.

Sasaki, Y., 2007. *The truth of the F-measure,* Manchester: School of Computer Science, University of Manchester.

Schanze, J., Zeman, E., Marsalek, J., 2006. *Flood Risk Management: Earth and Environmental Sciences.* Dordrecht: Springer.

Smith, S., Kindermans, P.-J., Ying, C. & Le, Q. V., 2017. *Don't Decay the Learning Rate, Increase the Batch Size*. s.l., s.n.

Sun, W., Du, B. and Xiong, S.: Quantifying Sub-Pixel Surface Water Coverage in Urban Environments Using Low-Albedo Fraction from Landsat Imagery, *Remote Sens*. 2017, Vol. 9, Page 428, 9(5), 428, doi:10.3390/RS9050428, 2017.

Tarpley, J., Hirota, N., Kato, M. & Arakane, S., 2015. Combined Effect of an Atmospheric River and a Cut-off Low in Hiroshima Flooding Event on August 19, 2014. *AGUFM*, p. A53F–06.

Turner, W. Rondinini, C., Pettorelli, N., Mora, B., Leidner, A.K., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L.P., Leimgruber, P., Taubenboeck, H., Wegmann, M., Wikelski, M., Woodcock, C. 2017. Free and open-access satellite data are key to biodiversity conservation. *Biological Conservation*, Volume 182, pp. 173-176.

Wang, Y., Li, Z., Zeng, C., Xia G. -S., Shen H. 2020. "An Urban Water Extraction Method Combining Deep Learning and Google Earth Engine," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 769-782.

Wagle, N., Acharya, T. D., Kolluru, V., Huang, H. and Lee, D. H.: Multi-Temporal Land Cover Change Mapping Using Google Earth Engine and Ensemble Learning Methods, *Appl. Sci*., 10(22), 8083, doi:10.3390/app10228083, 2020.

Wieland, M. and Martinis, S., 2020. Large-scale surface water change observed by Sentinel-2 during the 2018 drought in Germany. *International Journal Of Remote Sensing*, Volume 41, pp. 4742-4756.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. . *International Journal of Remote Sensing*, 27(14), p. 3025–3033.

Zeiler, M. and Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014*, pp. 818-833.

Zha, Y., Gao, J. & Ni, S., 2004. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery.. *International Journal of Remote Sensing*,, 24(3), p. 583–594.

Zhou, Y., Dong, J., Xiao, X., Xiao, T., Yang, Z., Zhao, G., Zou, Z., Qin, Y. 2017. Open Surface Water Mapping Algorithms: A Comparison of Water-Related Spectral Indices and Sensors. *Water*, 9, 256.