Remote Sensing of Environment 301 (2024) 113932

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

An AI approach to operationalise global daily PlanetScope satellite imagery for river water masking

Samuel J. Valman [a,b,*], Doreen S. Boyd [b], Patrice E. Carbonneau [c], Matthew F. Johnson [b], Stephen J. Dugdale [b]

[a] Nottingham Geospatial Institute, University of Nottingham, Nottingham NG7 2TU, UK
[b] School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK
[c] Department of Geography, Durham University, South Road, Durham DH1 3LE, UK

ARTICLE INFO

Editor: Menghua Wang

Keywords:
Earth observation
Artificial intelligence
Rivers
CubeSats
Hydrology
Neural networks

ABSTRACT

Monitoring rivers is vital to manage the invaluable ecosystem services they provide, and also to mitigate the risks they pose to property and life through flooding and drought. Due to the vast extent and dynamic nature of river systems, Earth Observation (EO) is one of the best ways to measure river characteristics. As a first step, EO-based river monitoring often requires extraction of accurate pixel-level water masks, but satellite images traditionally used for this purpose suffer from limited spatial and/or temporal resolution. We address this problem by applying a novel Convolutional Neural Network (CNN)-based model to automate water mask extraction from daily 3 m resolution PlanetScope satellite imagery. Notably, this approach overcomes radiometric issues that frequently present limitations when working with CubeSat data. We test our classification model on 36 rivers across 12 global terrestrial biomes (as proxies for the environmental and physical characteristics that lead to the variability in catchments around the globe). Using a relatively shallow CNN classification model, our approach produced a median F1 accuracy score of 0.93, suggesting that a compact and efficient CNN-based model can work as well as, if not better than, the very deep neural networks conventionally used in similar studies, whilst requiring less training data and computational power. We further show that our model, specialised to the task at hand, performs better than a state-of-the-art Fully Convolutional Neural Network (FCN) that struggles with the highly variable image quality from PlanetScope. Although classifying rivers that were narrower than 60 m, anastomosed or highly urbanised was slightly less successful than our other test images, we showed that fine tuning could circumvent these limitations to some degree. Indeed, fine tuning carried out on the Ottawa River, Canada, by including just 5 additional site-specific training images significantly improved classification accuracy (F1 increased from 0.81 to 0.90, $p < 0.01$). Overall, our results show that CNN-based classification applied to PlanetScope imagery is a viable tool for producing accurate, temporally dynamic river water masks, opening up possibilities for river monitoring investigations where high temporal variability data is essential.

1. Introduction

Rivers provide a multitude of ecosystem services, but also pose risks to property and life through flooding and drought events (Rinke et al., 2019; Arnell and Gosling, 2016). Despite their importance, data about key characteristics of rivers (Hannah et al., 2011; Gardner et al., 2021) at spatial and temporal scales amenable to their management are not readily available. The dominance of field measurements in the river sciences often means cost and time constraints prohibit the kind of extensive records that would be required to understand such large and dynamic systems (Piégay et al., 2020). Consequently, river managers have increasingly turned to Earth Observation (EO) science, using remote sensing (Piégay et al., 2020; Pavelsky and Smith, 2008), to provide the necessary data.

The crucial first step in calculating any river characteristic from EO data is extracting an accurate water mask. This foundational component governs all further analysis by defining which pixels contain exclusively water. Therefore, a typical data pipeline for the extraction of river habitat data from satellite imagery involves taking a satellite image, defining a water mask, and subsequently using the combination of this

* Corresponding author at: Nottingham Geospatial Institute, University of Nottingham, Nottingham NG7 2TU, UK.
E-mail address: samuel.valman@nottingham.ac.uk (S.J. Valman).

https://doi.org/10.1016/j.rse.2023.113932
Received 29 June 2023; Received in revised form 22 November 2023; Accepted 27 November 2023
Available online 7 December 2023

water mask and the original image to calculatecharacteristics such as discharge, width, or temperature. (Pavelsky and Smith, 2008). Currently, many studies rely on historical water masks from pre-processed EO datasets (e.g., the Global River Width from Landsat dataset (GRWL; Allen and Pavelsky, 2018) or the Surface Water and Ocean Topography mission River Database (SWORD; Altenau et al., 2021) that are static in time and are assumed to be centred on average bankfull river width (Allen and Pavelsky, 2015). However, in reality, water masks are dynamic, varying as a function of hydromorphic changes that occur in all river systems. Because manual delineation of water masks is laborious and time consuming (Mahoney et al., 2020), this leads to a need for automated procedures for generating water masks.

Traditional methods of water mask creation include use of the Normalised Difference Water Index (NDWI; Sekertekin et al., 2018). Since the NDWI uses the green and near-infrared wavelengths (Mcfeeters, 1996), it is suitable to use with data from the majority of satellite EO missions. Extracting water masks based on the NDWI involves using histogram thresholding methods that generate very little computational demand; however, this technique relies on the choice of specific threshold values that vary with region (Frazier and Hemingway, 2021). Even when automating the choice of threshold for individual images (Otsu, 1979), segmentation errors persist. This unwanted spectral response occurs most often in heavily vegetated or urbanised regions where shadows, and reflection patterns similar to water, disrupt the expected response (Zheng et al., 2021) which led to the development of the Modified NDWI (Xu, 2006). However, the MNDWI requires a Short-Wave infrared band, precluding its use with higher-resolution VIS-NIR commercial satellites (Planet Labs, 2018; Gleyzes et al., 2012). Algorithms such as RivWidth (Pavelsky and Smith, 2008) use computer vision-based techniques to smooth the results of these threshold methods (Ziou and Tabbone, 1998; Dougherty, 2020). By combining these methods with cloud infrastructure, the process can be extended globally in the form of RivWidthCloud and operationalised to be reapplied to new imagery (Allen and Pavelsky, 2018; Yang et al., 2019). However, the RivWidth method is still built on the MNDWI water mask foundation (Zou et al., 2018) and thus inherits its limitations. Alternatively, some authors use fractional pixel analysis to help separate some of the complexity in the output from OTSU NDWI thresholding (Cooley et al., 2017) but these do not address the root problems with NDWI.

As an alternative to thresholding methods, a variety of machine learning methods (Abburu and Golla, 2015) such as Support Vector Machines (Foody and Mathur, 2006) or cluster analysis (Genitha and Vani, 2013) have been considered for water mask generation. However, while these approaches generate acceptable results, especially at coarse resolutions, they have not as yet produced masks with the precision required for segmentation of higher resolution imagery. Furthermore, these algorithms are rarely sufficiently generalisable to be extended to 'big data' (Ling et al., 2019) derived from satellite constellations. Yuan et al. (2021) hypothesised that Artificial Neural Networks (ANN) perform better than these other machine learning approaches by being able to "learn" sensor discontinuity. However, on their own, ANNs will suffer from the same issues with generalisability due to their tendency to specialise to their training location (Foody et al., 2003).

While there are some drawbacks associated with these methods, these approaches have also demonstrated the power of remote sensing for water resource science and management. These studies have tended to use open-source data from Sentinel and Landsat sensors, which have provided spatially extensive research avenues for provision of river data. However, these EO data are intrinsically restrained by their coarseness in spatial and temporal resolution (Gleason and Durand, 2020; Junqueira et al., 2021), which consequently limits their usefulness in characterising dynamic freshwater environments that change at very fine timescales. However, the recent development of EO satellite constellations, such as PlanetScope, offer potential for mapping rivers at much higher temporal resolution via daily data collection (Frazier and Hemingway, 2021). This, coupled with the relatively fine spatial resolution of PlanetScope imagery, means that the required characteristics of river systems can potentially be measured and monitored across more river systems globally (Junqueira et al., 2021; Feng et al., 2019). PlanetScope satellites have already been used for tasks such as measuring water quality parameters (Niroumand-Jadidi et al., 2020), such as suspended sediment concentration (Wirabumi et al., 2021), in various waterbodies. In the context of rivers, such studies could be enhanced and extended with dynamic water masks. Unsurprisingly, the effectiveness of these satellites does not match Landsat or Sentinel in understanding water quality (Mansaray et al., 2021) but by combing these more 'traditional' satellites with high resolution water masks extracted from PlanetScope, the best of both products could potentially be achieved (Gabr et al., 2020).

Working with satellite constellations raises a range of considerations for the creation of accurate water masks. An important one, is that the ability of satellite constellations to observe rivers globally at high spatio-temporal resolution increases the variety of conditions over which a model is required to generalise. This need for generality demands robust water mask extraction methods capable of consistently extracting water masks from across the continuum of rivers that exist globally. Another important consideration is that often, satellite constellations such as PlanetScope constitute smaller, low-cost sensors, whose radiometric quality and lower signal to noise ratio (in comparison to 'conventional' satellites) means that water mask extraction methods must be robust to variance in image quality (Haq, 2022).

Convolutional Neural Networks (CNN) show promise for providing the robust approach required by taking a deep learning, neighbourhood based, approach (Moortgat et al., 2022; Marochov et al., 2021; Carbonneau et al., 2020; Yasir et al., 2023; Qayyum et al., 2020). CNNs tile a fixed number of pixels and then pass a smaller 3D kernel over these tiles; the height and width of the kernel is the number of image rows and columns, as specified by the user, and the depth of the kernel is the number of spectral bands in the image. The neural network uses this kernel to "learn" the space-intensity relationship between pixels within a class. The goal of this operation is to learn the high-level features which dictate which class (i.e., land or water) the tiles belong to. The output of a CNN is a prediction, returned in the same form as the input: a set of tiles which have been given a single class value (Reina et al., 2020). Due to the fixed number of pixels, the real-world spatial extent of this tile is directly related to the ground sampling distance of the pixels in the image. Therefore, unless the CNN is predicting from hyperspatial imagery (<10 cm) the output can appear pixelated. Two predominant methods are often employed to overcome the pixelation issue. Fully Convolutional Neural Networks (FCNs) are often currently considered state of the art (e.g., Tiramisu; Jégou et al., 2017), using traditional CNN architecture to learn 'deep' characteristics but combine this with up-sampling to provide a pixel level result (Long et al., 2015). These have been applied successfully in water classifications (Carbonneau and Bizzi, 2023; Isikdogan et al., 2017) and can be relatively efficient for CNNs (Moortgat et al., 2022). A variety of structures of FCN have been developed for water masking (e.g., Li et al., 2021), however, all these methods still require significant computational resources and deep learning expertise for effective implementation. Moreover, most FCNs also require more detailed training data (e.g., Moortgat et al., 2022); which adds an additional challenge that can create barriers for practitioners aiming to build water masks. Some authors have overcome this by using Open Street Map annotations to confirm water presence for model training (Mazhar et al., 2022) but along with using water presence maps (Pekel et al., 2016) these methods cannot be used with PlanetScope due to PlanetScope's georeferencing accuracy and edge effect errors caused by scale/resolution mismatches. Another concern stems from the high variability of radiometric quality in PlanetScope data, meaning that an FCN model will encounter challenges in effectively addressing radiometric differences between images without particularly large training sets.

A second strand of research, termed 'CNN supervised classification' (CSC) has been proposed (Carbonneau et al., 2020). This uses a pre-trained CNN to provide localised (within test image) training data for an ANN Multi-Layer Perceptron (MLP) model which in turn provides pixel-level semantic segmentation (Yuan et al., 2021; Carbonneau et al., 2020), thus potentially overcoming significant radiometric differences between images. This method has been highly successful for mapping river habitat types in hyperspatial RGB imagery (Carbonneau et al., 2020), but other applications (including glacier mapping using medium resolution Landsat imagery; (Marochov et al., 2021) demonstrate CSCs viability for segmenting satellite data where the target size (be it river or glacier width) is commensurate with these larger resolution satellite inputs. The CSC method only requires relatively imprecise training data, drawing patches over land or water on an image, which can be collected very quickly with limited a priori skills, as opposed to multi-step GIS workflows (Moortgat et al., 2022) or working with semi-automatic classifiers. We therefore hypothesise that such an approach is well-suited to producing water masks from PlanetScope imagery, and thus test the use of a relatively shallow model architecture (i.e., fewer layers than in other recent publications on FCNs and CSCs), with a view to demonstrating the viability of our proposed methodology for non-specialists.

In this paper, we employ daily PlanetScope individual images and develop a novel Artificial Intelligence (AI) algorithm capable of generating water masks for these high temporal resolution data at a spatial resolution of 3 m. This builds on previous CSC applications (Marochov et al., 2021; Carbonneau et al., 2020) but the method has not been previously tested with the large ratio between target size and pixel resolution we present here. Moreover, this study constitutes the first application of this method used for the extraction of daily river water masks, requiring vastly more imagery to be classified by practitioners than would be available with traditional satellites (e.g., Landsat's 16-day return period). In turn this produces an added computational requirement that we hypothesise would be better met by the shallow CSC built here as opposed to more traditional architectures used in earlier CSC applications such as the very deep VGG16 model. This represents yet another reduction in processing requirements when compared to the computationally intensive needs of very deep FCNs favoured in alternative large water masking studies (Isikdogan et al., 2017; Isikdogan et al., 2020; Carbonneau and Bizzi, 2023). Our proposed AI algorithm enables the automation of water mask development, so future EO-based river studies can be attempted without the requirement for manually extracting water masks from each image in a temporal stack. As noted, the fine spatio-temporal resolution of PlanetScope, and difficulties with using CubeSats more generally (e.g., limited spectral resolution, issues calibrating within a constellation (NASA, 2020)), requires the model to be able to generalise to a much greater degree than other EO-based water mask algorithms (Feng et al., 2019; Junqueira et al., 2021; Moortgat et al., 2022; Isikdogan et al., 2017). We thus also hypothesise that the 'self-contained' nature of the CSC method would cope with these radiometric issues better than a FCN would.

In order to create a model that can generalise globally, we also examine the potential to include a 'human-in-the-AI-loop' in this context. Here, we test the potential to improve classification performance in a specific use case by fine tuning it with limited additional training data. Our overarching aim was therefore to develop and evaluate an AI algorithm that automates the extraction of river water masks from PlanetScope imagery. We aimed to overcome computational limitations associated with previous methods while ensuring practitioners could easily create and enhance training data, by pioneering a robust shallow CSC algorithm. In order to achieve this, we developed the following three objectives which provided not only insights into the feasibility of our proposed method but also highlights its viability as a tool for water mask generation on which to base further EO river research and analysis:

1. Develop a CSC-based classifier, with minimal processing demands, capable of extracting river water masks from PlanetScope imagery and compare its accuracy to 'conventional' image segmentation algorithms and a state-of-the-art Fully Convolutional Neural Network.
2. Apply the CSC approach to a range of global rivers to understand when, where and how effective our approach might be across 12 global biomes.
3. Test the extent to which the inclusion of limited additional training data improves classification accuracy at a specific river, shedding light on the viability of a 'human-in-the-AI-loop' classification strategy.

2. Methods

2.1. PlanetScope imagery

Planet Labs PBC operates the PlanetScope constellation of 200 CubeSat (i.e., small, low-cost) satellites (Planet Labs, 2022), imaging a large proportion of the globe every day (Planet Labs, 2018). There have been 3 generations of these PlanetScope "Dove" satellites, improving the radiometric quality of images as well as adding 4 additional bands in the most recent 2021 "SuperDoves" (coastal blue, green 1, yellow, and red edge; Le Roux et al., 2021). Here, we focus solely on four-band imagery enabling interoperability across the whole 5-year range of historical and current PlanetScope imagery. Given the new SuperDoves do not include a shortwave infrared band (shown to improve water delineation; Xu, 2006) we do not expect that the increased complexity of 8 bands would provide significant classification improvement.

PlanetScope satellites capture imagery at 3 to 5 metre resolution depending on the individual satellite's altitude (Planet Labs, 2018). This imagery is resampled to 3 m making PlanetScope the highest resolution continuously recording satellite system (Frazier and Hemingway, 2021). Higher resolution imagery is obtainable but hindered by requiring the user to task it to a specific limited location and time at additional cost (Cornebise et al., 2022). EO data captured by these satellites is pre-processed before provision to the end-user. In this pre-processing the imagery is calibrated against ground stations and MODIS data to remove the effects of the atmosphere on reflectance received at the sensor (Planet Labs, 2018). As such, all imagery utilised in this study used the PlanetScope's pre-processed surface reflectance product. This enables a reduction in the client-side model processing pipeline and also reflects the trend in satellite remote sensing towards the use of analysis-ready products (Gorelick et al., 2017). To help correct for different relative responses from individual satellites in the constellation and other adverse climatic effects, the PlanetScope product can be downloaded normalised against Sentinel 2 (Kington and Collison, 2022). However, interoperability between the constellation is difficult because there is a high turnover of satellites due to their short lifespans. In addition, their limited payload capacity requires smaller technology placed closer together than traditional satellites, which leads to relatively high potential error between PlanetScope sensors (Frazier and Hemingway, 2021; NASA, 2020).

2.2. Training and testing imagery

To develop a water mask algorithm able to generalise across a diverse range of rivers, training data (Lew and Schumacher Jr, 2020) covers a variety of global rivers. Traditional river classification systems (Kasprak et al., 2016; Rosgen, 1994; Brierley and Fryirs, 2013) could provide the framework for this training set. However, the algorithm was also required to distinguish rivers that exist in the context of different land uses, meaning that the training selection framework must be globally holistic. We therefore used World Wildlife Fund (WWF) global biomes to dictate globally representative landscapes from which to select test and training data for our algorithm (Olson et al., 2001). Of the 14 WWF biomes, Tropical Coniferous Forest and Mangrove Forest were

removed because their relative size and positioning limited the number of rivers that could confidently be placed within their boundaries. The Natural Earth 10 m wide river centre line vector file (Kelso and Patterson, 2010) was used to select rivers in the remaining 12 biomes. This threshold ensured that rivers were resolvable within the 3 m resolution of PlanetScope, allowing for the inclusion of narrower rivers than other coarser datasets (e.g., GRWL) whilst still ensuring the presence of large perennial streams. A range of river types and sizes were included to incorporate the full range of potential study areas (Fig. 1). The final selection of rivers comprised water courses between 25 and 7400 m wide, selected to provide global coverage, with 3 rivers from each of the 12 biomes resulting in a total of 36 individual rivers.

For each river, a minimum of 3 scenes were delineated and downloaded. These were from different dates and different positions on the river. These dates were distributed throughout different seasons at all sites to maximise potential applications of the water masks generated, with the requirement that the river channel contained water (if an intermittent river) and this water was not entirely frozen. Two scenes were used as training data and the third used as hold-out test images, conferring a nominal test/train split of 66/33 (Yin et al., 2021; Carbonneau et al., 2020). This is weighted slightly more towards testing than the 80/20 split used in some studies (Moortgat et al., 2022). However, the aim of CSC is to be able to predict many different scenarios from the training set and therefore there should be a larger weighting for test data (Carbonneau et al., 2020). As such, 5 additional hold-out images were added to further test the algorithm. These contained anomalies such as snow, excessive shadow, or cloud cover that had evaded the Planet Explorer cloud mask.

2.3. Data labelling

Training data were labelled manually using QGIS 3.16.5 with the GRASS 8.7.5 plug in (Baghdadi et al., 2018). This workflow involved delineating land and water in a shapefile by drawing polygons over the clear land and water sections of an image. There was no precise requirement for the extent of polygons but a reasonable coverage of the unambiguous land classes across the image, taking 3–5 min, was expected (see supplementary materials for examples). This was then saved as a raster of the same dimensions as the original image. Labels consisting of large clear polygons were drawn to simulate the requirement for training data to be quick to assemble. Test images were labelled with the "Semi-Automatic classification QGIS plug-in" version 7.10.10 (Congedo, 2021). A minimum of 8 sample polygons were delineated per image to train this semi-automatic classifier, and more were included where necessary when a visual inspection of the accuracy of the result was not considered sufficient. This method was used because every pixel in an image was classified. Alternative attempts at manual delineation often excluded the hardest to classify, channel edge pixels which were most important for understanding model success. The resulting training data consisted of 72 scenes which were then divided into 20 by 20-pixel training tiles starting in the top left corner of each image for CSC model calibration. Of these, only training tiles which were purely water or purely land were included in the training data (Marochov et al., 2021). These were balanced against the smallest class to prevent overfitting (Gavrilov et al., 2018), resulting in 393,000 training tiles. Additional training sets consisting of 10 by 10 and 32 by 32 tiles were also produced to enable comparison between models trained using different tile sizes.

2.4. CSC model architecture

Our water mask classification model was run on a PC with a 12-core Intel i7-12700K with 32 Gb RAM and a NVIDIA GeForce RTX 3070 GPU with 8 Gb RAM and 5888 CUDA cores. TensorFlow and Keras (Abadi et al., 2016) were chosen to develop the model. The classification models produced here were based on similar approaches developed and used in a variety of remote sensing fields known as CNN-supervised classification (CSC) models (Carbonneau et al., 2020; Marochov et al., 2021). CSC models require training a fully connected sequential CNN model which predicts a tiled test image, analogous to a rough sketch of the river channel due to these larger CNN tiles. The results from this CNN are thereafter used to train a localised MLP model which enables semantic (i.e., pixel level) segmentation specific to the spectral signatures of the scene in question.

Three relatively simplistic CNN models were developed with tile sizes of 32 × 32, 20 × 20 and 10 × 10 pixels respectively, hereafter referred to as M32, M20, M10. Each of these three sequential models is made up of 9 fully connected layers including two convolutional layers and two dense layers (Fig. 2). All layers used the 'RELU' activation function with the exception of the final 'SoftMax' output layer (Géron, 2022). The convolutional layers and first dense layer have 32 neurons each. The convolutional layers have a kernel size of 3 by 3 which dictates the dimensions of the convolutional window which moves across the tile. The models are compiled with the 'sparse categorical cross entropy' loss function, and an 'Adams' optimiser, both of which are broadly adopted in image classification (Goodfellow et al., 2016). These values were selected based on an assessment of the loss and accuracy outputs during grid based hyperparameter tuning training runs which used the TensorFlow TensorBoard call back to assess which combinations worked best. All models were run for 10 epochs after this was found to be enough to result in convergence (see supplementary material). The

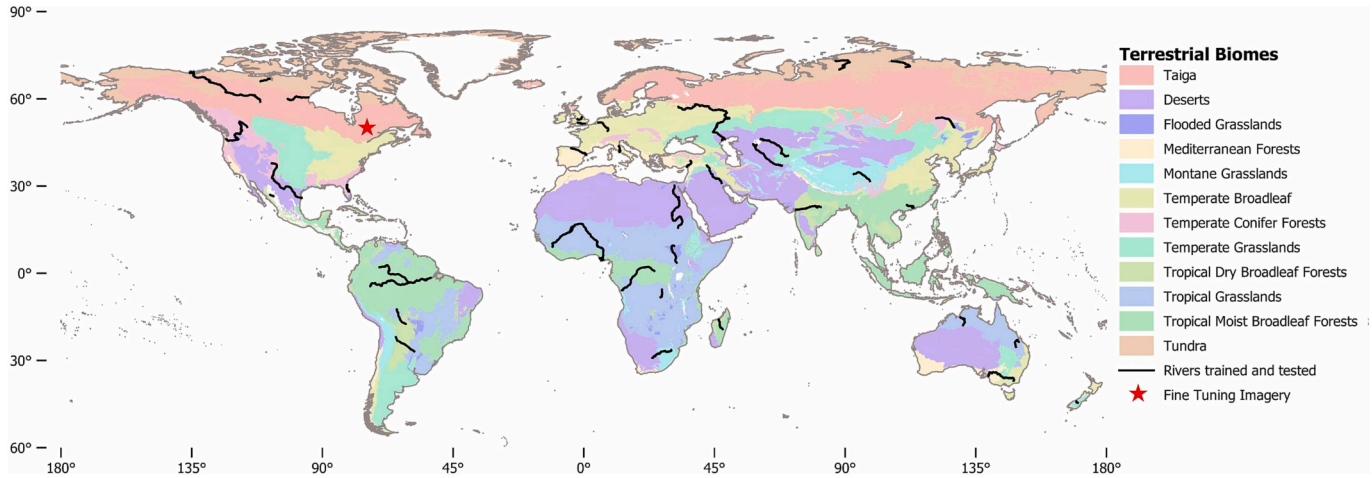

Fig. 1. The Global terrestrial biomes presented by (Olson et al., 2001) overlayed with the rivers used in this study for testing and training. 12 biomes were selected for this study with 3 rivers selected from each biome. Additional sites used for testing the ability of the algorithm to be fine-tuned were in the area here.

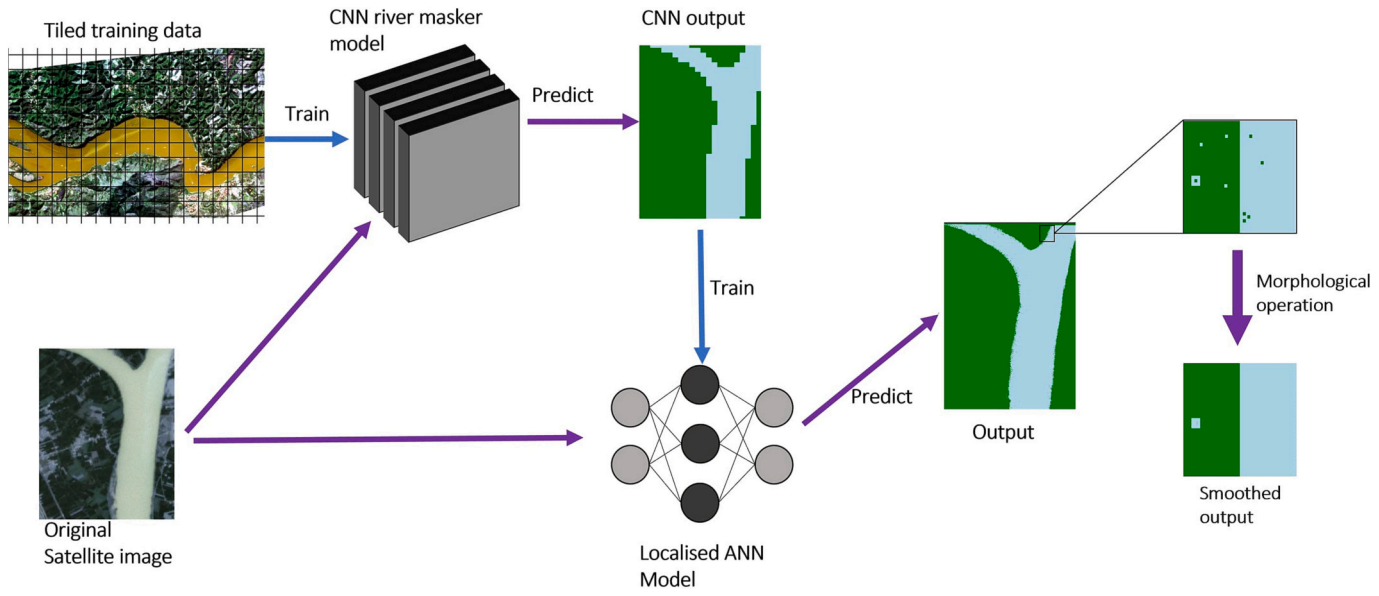

Fig. 2. Conceptual diagram of the CSC structure used in this study. A pre-trained 9-layer CNN which uses its predictions to train a 5-layer MLP. The final results are processed by a morphological operator to remove any remaining speckle. The CNN architecture consisted of a convolutional layer, a batch normalisation layer, a max pool layer, a second convolutional layer, a second max pool layer, a flattening layer, a dense layer, a drop out layer and the final output dense layer. The MLP architecture consisted of a normalisation layer, 3 dense layers and then a final output dense layer.

output of these CNNs was in the form of the likelihood of each tile being land or water.

Transfer learning is a field of AI that takes advantage of the weights in pre-trained classification models developed from millions of training images (Zhuang et al., 2020). The desire to use 4-band imagery here prevents the use of transfer learning on models trained with 3-band RGB imagery and solutions that are currently available in the literature. Models that enable transfer learning using variable band numbers are not yet commonly accepted (de la Comble and Prepin, 2021). Hence, many methods take the architecture of these very deep CNN models developed in "Kaggle" programming competitions but do not keep the pre-trained weights (Marochov et al., 2021). However, models such as the commonly adapted VGG16 model, were developed with a thousand output classes (Simonyan and Zisserman, 2014). We hypothesised that a more simplistic model would be more effective when attempting to satisfy the performance requirements desired from a binary output (Nativi et al., 2021). By using a much shallower architecture we believe that our models will converge quicker because they are not using the large processing power of traditional methods to learn intricacies in 'high level' features, such as straight lines (Zoph et al., 2018). These high level features are largely superfluous to this study's requirements because the relatively low resolution of PlanetScope means that these higher features are often not visible in the imagery. For example, learning aspects of spectral combinations in riverbanks would be of use with higher resolution imagery but here the 3 m resolution edge pixels are often mixed, negating these lessons. Therefore, developing our own shallower models allows for reduced tile size, potentially improving results and reducing processing time. These factors make the CSC models developed here both more effective and applicable to a larger user base who lack high-performance machines (Boothroyd et al., 2021). Nonetheless, a 'control' CSC model was developed to test this theory, using this VGG16 architecture to generate a very-deep CNN with a 2-neuron 'SoftMax' output layer added to provide the desired output. This used a tile size of 32 by 32 pixels, the minimum size enabled by VGG16. This very deep network was only run for 50 epochs before convergence, where upon visual inspection the validation loss value ceased to be improved by additional training (see supplementary materials for loss curves).

The subsequent MLP stage was kept simple to account for the very different inputs that might be used to train the model, due to the global diversity of river environments. Therefore, this neural network consisted of 4 dense fully connected layers using the same 'ReLU' and 'SoftMax' activation functions as the CNN phase. The first 3 layers had 64 neurons and the last again had 2, one for each potential output. The same compilers were used in this stage as in the CNN stage. The MLP was only run for a single epoch as it became apparent that with many of the images with smaller rivers more epochs resulted in overfitting. The result from this section also used NumPy's argmax function to convert the softmax results into a binary classification. The final stage of our modelincludes an additional morphological operator to eliminate speckle from the water masks output by the MLP stage and constrain results to the main channel (Riggs et al., 2021; Pavelsky and Smith, 2008). For this, the OpenCV 'dilate' function was used with a kernel size of 3 (Bradski and Kaehler, 2000). Thus, any pixels not immediately adjacent to 2 or more other pixels were converted to the alternative classification in the same neighbourhood. This procedure removed the majority of speckle without removing smaller watercourses in the images.

To compare the models to a more basic classification algorithms, the Otsu segmentation method was also applied on all test images (Otsu, 1979). This uses histograms to statistically split the imagery into two classes without the need for training data or a priori knowledge of the imagery. To compare against state-of-the-art AI models the Tiramisu FCN (Jégou et al., 2017) was trained using the same images as the other models. However, this requires pixel level classification training data that is much more time consuming to develop. To assist with this process, the DoodleVerse package (Buscombe and Goldstein, 2022) for semi-automatic image classification was used. Any training images that were not satisfactorily classified or were too large for DoodleVerse were classified using the same QGIS Semi-automatic classifier previously used for holdout image classification. A tile size of 224*224 was used due to the greater information requirements of a FCN. Because this model requires considerably more training data than CSC, an augmentation script was applied that added additional altered (e.g., flipped/rotated) tiles containing at least 5% water in the image. To further reduce the impact of the limited water coverage in the training images the focal loss function was used to put greater importance on the weights for water tiles (Lin et al., 2017). The Tiramisu model was run for 20 epochs before

convergence.

2.5. 'Human-in-the-loop' model enhancement

Our CSC approach was expected to be able to use minimal processing power to provide accurate predictions despite the variability of PlanetScope imagery. However, the model is unlikely to be capable of accurately predicting the water mask for every river globally, at all times of year, regardless of the quality of the satellite imagery and any corrections applied. This is because the infinite variability of global river systems (Thoms and Sheldon, 2019; Frazier and Hemingway, 2021) means that it was not feasible to train our model on every eventuality or user requirement. However, the ability to train with minimal processing power enabled us to test an additional 'human-in-the-loop' computation stage to understand whether retraining (i.e., fine tuning) the model through the inclusion of a few additional images, quickly labelled using a low effort approach such as the QGIS 'magic wand' (Baghdadi et al., 2018), can effectively improve classification results in a specific river system not previously seen by the classifier.

To test the ability of this model to be fine-tuned in this manner, an independent study site in Southern Canada was chosen which the model had not already seen. 20 images from the Ottawa River were collected. 15 were used as test images and processed the original M20 CSC model developed here. The remaining 5 images were quickly labelled using the QGIS 'magic wand' tool, producing additional training data in less than an hour. The original CSC was then retrained with this additional training data included, for a total of 408,000 tiles, effectively biasing the training data to this new river system (which now has more training data than other rivers) with the same architecture, number of epochs, and learning rate. The 15 test images were predicted again to assess if this 'human-centric' CSC was more effective at correctly predicting these images.

2.6. Statistical analysis

Three validation metrics were employed to different degrees. *Loss* (see section 2.3) and *Accuracy* (eq. 1) were used during training of the CNNs to assess the ability of each model to predict randomly assigned test data (Goodfellow et al., 2016; Yuan et al., 2021). Although this was important for improving CNNs during the building phase, it only showed their ability to predict known tiles as water and land and was not necessarily linked to the final results. The main measure of classification accuracy was therefore the F1 score (Eq. 4), which measured the ability of a full CSC model to predict a holdout image at a pixel scale. This F1 score is sometimes considered the harmonic mean of *Precision* (eq. 2) and *Recall* (eq. 3; Carbonneau et al., 2020). It is a well-accepted method of balancing where a model correctly predicts a pixel and where it provides false positives or negatives (James et al., 2021; Goodfellow et al., 2016). Therefore, if a model predicts a single class for a whole image, it will not score highly despite, by process of elimination, having correctly predicted all instances of one of those two classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2\frac{P.R}{P + R} \quad (4)$$

Accuracy (Eq. 1) considers the True Positives (TP) and True Negatives (TN) which are respectively water and land pixels which have been identified correctly. It then divides these by the total predictions made, including False Positives (FP) and False Negatives (FN) that represent water and land pixels predicted incorrectly, thus giving the ratio of correctly identified pixels. For assessing predictions of holdout images, F1 score was used because it takes account of potential class imbalances. It does this through combining *Precision*, how many water pixels identified were correct, and recall which is the ratio of correctly predicted water pixels to those that should have been predicted.

Statistical analysis was carried out using the SciPy package (Virtanen et al., 2020), which was also used to compare F1 scores against river characteristics (e.g., river width and planform, anthropogenic influence) and imagery metadata, to understand whether differences between F1 scores for biomes were caused by intrinsic landscape characteristics or were artifacts of the satellite imagery itself. River width was expected to be a limiting factor on predictive ability, due to the requirement of tile-sized sections of water being needed to provide pixels to train the MLP. Width was manually extracted from the widest river reach in the image, provided that the width was maintained in this reach for at least 20 pixels (60 m) downstream to represent the width of a single tile at the CNN stage of classification. Width measurements were log transformed based on the hypothesis that the differences in the narrower widths would have a greater impact than differences between larger width values. The NIR band is integral to NDWI measurements because NIR wavelengths are strongly absorbed by water (Mcfeeters, 1996), and this was also tested in a similar manner to understand its influence on the results. The 5th to 95th percentile range of NIR values in an image was used as this represented the difference between land or water excluding the most over- or under-exposed areas of the image. We used linear regression to test the potential relationship between these environmental characteristics/imagery metadata and classification accuracy (F1). For the 'human-in-the-AI-loop' tests, we used a paired samples *t*-test to differentiate between F1 scores before and after the inclusion of the fine-tuned test data. A Box-Cox transformation was used before testing to achieve a normalised distribution, removing the effects of skew towards the better-predicted imagery.

3. Results

3.1. Model selection

The various CSC models took between 36 min and 2 h 11 min to train, excluding the time taken to tile imagery and save it as TensorFlow Record file types. In contrast, the Tiramisu FCN model required 40 h at ~2 h an epoch. The simpler models (M10, M20 and M32 tiles) all converged within 10 epochs. The very deep layered network based on the VGG16 architecture (also 32 × 32 tiles) converged over 50 epochs, while the comparator Tiramisu model converged after 20 epochs. All models produced some highly accurate water masks but also struggled with some imagery and biomes (Fig. 3). Training metrics describe the ability of the CNN stage of the model to predict a random selection of the training tiles providing an indication of the internal ability to learn during training. Here these training metrics corresponded with tile size and depth. The very deep model (VGG16) yielded the highest accuracy (0.94) compared to M32 (0.87), M20 (0.85) and M10 (0.84), and the lowest loss (0.13) compared to 0.31, 0.34 and 0.35 for M32, M20 and M10 respectively. By way of comparison, the Tiramisu FCN achieved a training accuracy of 0.92 and a loss of 0.04.

The final learning curves for each of these models are included in the supplementary material. However, training metrics only measure the ability of the CNN to predict tiles taken from the training set, which could be related to those used to train the model. These training metrics do not provide an assessment of the CNN's ability to predict hold-out image tiles or the full CSC model's ability to predict the whole of a new hold-out image, and surprisingly, these training metrics were found to not correspond to the best final predictions. In predicting hold out images, the M20 model produced the highest median F1 score (0.93), making it the most successful model, while the M10, M32, and very deep CSC models were all similarly effective but slightly lower scoring (0.90,

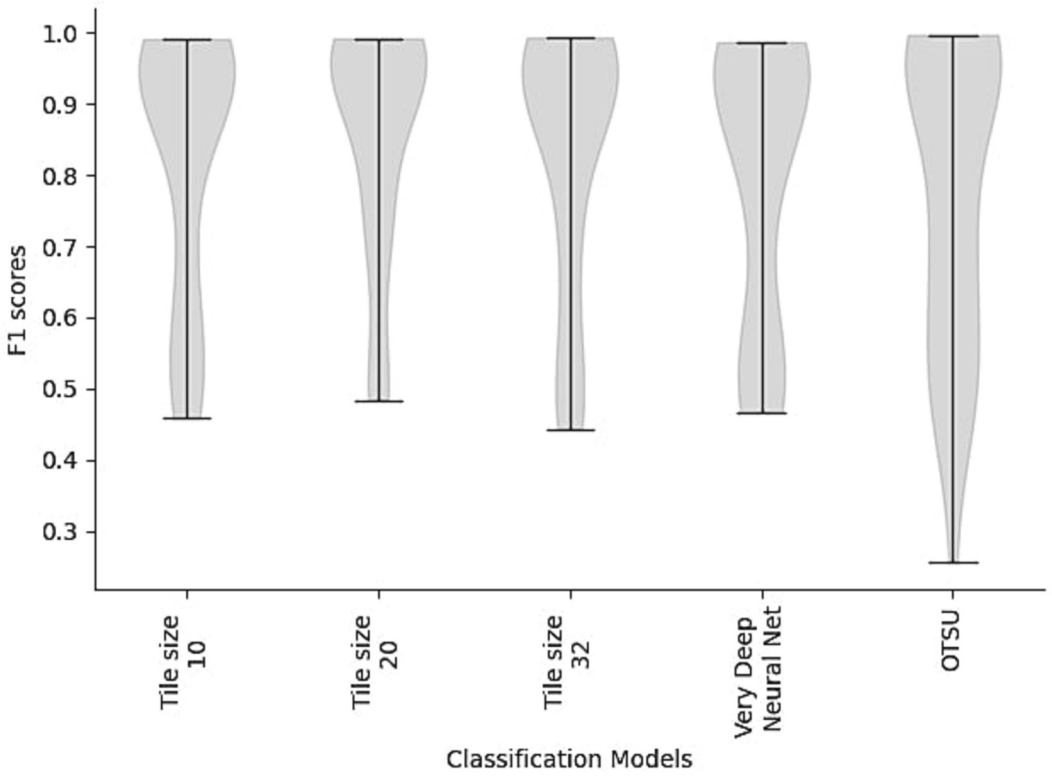

Fig. 3. Comparison of the ability of tested models to predict hold-out images. All models shown here are skewed towards high F1 scores. There are different distributions of the scores for the weakest predictions made by each model. M20 in particular never predicts as poorly as the other models and has a thinner tail suggesting no clustering of poorly predicted images.

Table 1
Comparison of water mask model averages and their performance in different biomes. For each variable, the best performing model has been highlighted. Mean F1 score was used for comparing biomes because each of these had a limited numbers of images, limiting the usefulness of median F1 score.

Biome (mean average F1 score)	Tile Size 10 CSC	Tile Size 20 CSC	Tile Size 32 CSC	Very Deep (VGG16) CSC	Otsu NDWI thresholding
Overall Median	0.92	0.93	0.92	0.92	0.90
Overall Mean	0.83	0.86	0.83	0.81	0.78
Overall Range	0.53	0.51	0.55	0.52	0.74
Tundra	0.90	0.89	0.77	0.76	0.82
Taiga	0.98	0.98	0.96	0.97	0.99
Montane Grassland	0.64	0.74	0.75	0.64	0.63
Temperate Coniferous Forest	0.77	0.65	0.81	0.65	0.79
Temperate Broadleaf Forest	0.68	0.77	0.70	0.69	0.58
Temperate Grassland	0.97	0.97	0.97	0.97	0.97
Tropical Moist Broadleaf	0.96	0.98	0.98	0.98	0.98
Tropical Dry Broadleaf	0.66	0.69	0.67	0.64	0.80
Tropical Grassland	0.87	0.99	0.81	0.96	0.98
Mediterranean Forest	0.82	0.87	0.87	0.86	0.50
Deserts	0.79	0.82	0.77	0.78	0.67
Flooded Grasslands	0.90	0.91	0.92	0.84	0.70

0.92, 0.92). While the Otsu segmentation approach scored a similar median classification result (0.90), its much lower mean value (Table 1) and longer-tailed error distribution (Fig. 3) highlights the considerably poorer performance of this 'conventional' classification approach in comparison to the CSC-based models. Using the comparator Tiramisu FCN, only 4 of 36 images were predicted with sufficient accuracy to justify recording F1 scores (F1 = 0.52 to 0.87; see supplementary material), highlighting how good training/validation performance does not guarantee accurate predictions of hold-out imagery. The FCN was therefore not included in further model comparisons.

Overall, M20 produced water masks with F1 $\geq$ 0.9 for over 60% of the hold-out images making it the best model in this study (Fig. 4). While it produced the best F1 scores in 6 of the 12 biomes (Table 1), other models performed better in some other biomes with varying degrees of departure from M20. In fact, the Otsu thresholding method (the simplest used here, which produced some highly inaccurate results [e.g., average F1 of 0.5 for Mediterranean Forest]), also produced near perfect F1 scores in polar-type biomes such as the Taiga (average F1 of 0.99). The clear histogram response facilitated separation of the water from non-water pixels which also allowed M20 to perform similarly well (F1 = 0.98). The Otsu method also exceeded the accuracy of all other methods in the Tropical Dry Broadleaf (0.80 compared to 0.66–0.69) biome despite still producing less than desirable results. Nevertheless, in all other cases it was not better than M20.

3.2. Variability within biomes

Water mask generation was consistently more accurate for rivers in some biomes (such as Tropical Moist Broadleaf) than others (e.g., Tropical Dry Broadleaf biome; Fig. 5, Table 1). Understanding these inconsistencies requires analysis of within-biome causes of poor model performance. Ecoregions were plotted separately for M20 to reveal those that produced similar results, possibly due to internal factors (Fig. 5). The 12 biomes can be visually categorised into 3 groups in terms of their performance: consistently good, varied, and poor (Fig. 5). In the 'poor' group Montane Grassland, Temperate Broadleaf and Temperate Coniferous biomes yield some water masks that are highly accurate (Fig. 6a), but along with Tropical Dry Broadleaf, also generate water masks of an unacceptable quality (Fig. 6b).

Descriptive statistics of underlying imagery components were tested against F1 scores to understand the causes of the variable results (Fig. 7). While no significant relationship was found between F1 score and the range of any of the spectral bands, any of the metadata characteristics provided by PlanetScope, or any of the physical river characteristics visible from the imagery ($p>0.05$ in all cases), we did observe a moderately significant trend between log-transformed river width and F1 score ($r^2 = 0.34$, $p<0.01$) which is consistent regardless of biome, indicating that this control (i.e., width) is a local, rather than regional, driver of differences in F1 score.

Nine poorly predicted images were isolated, with F1 scores ranging

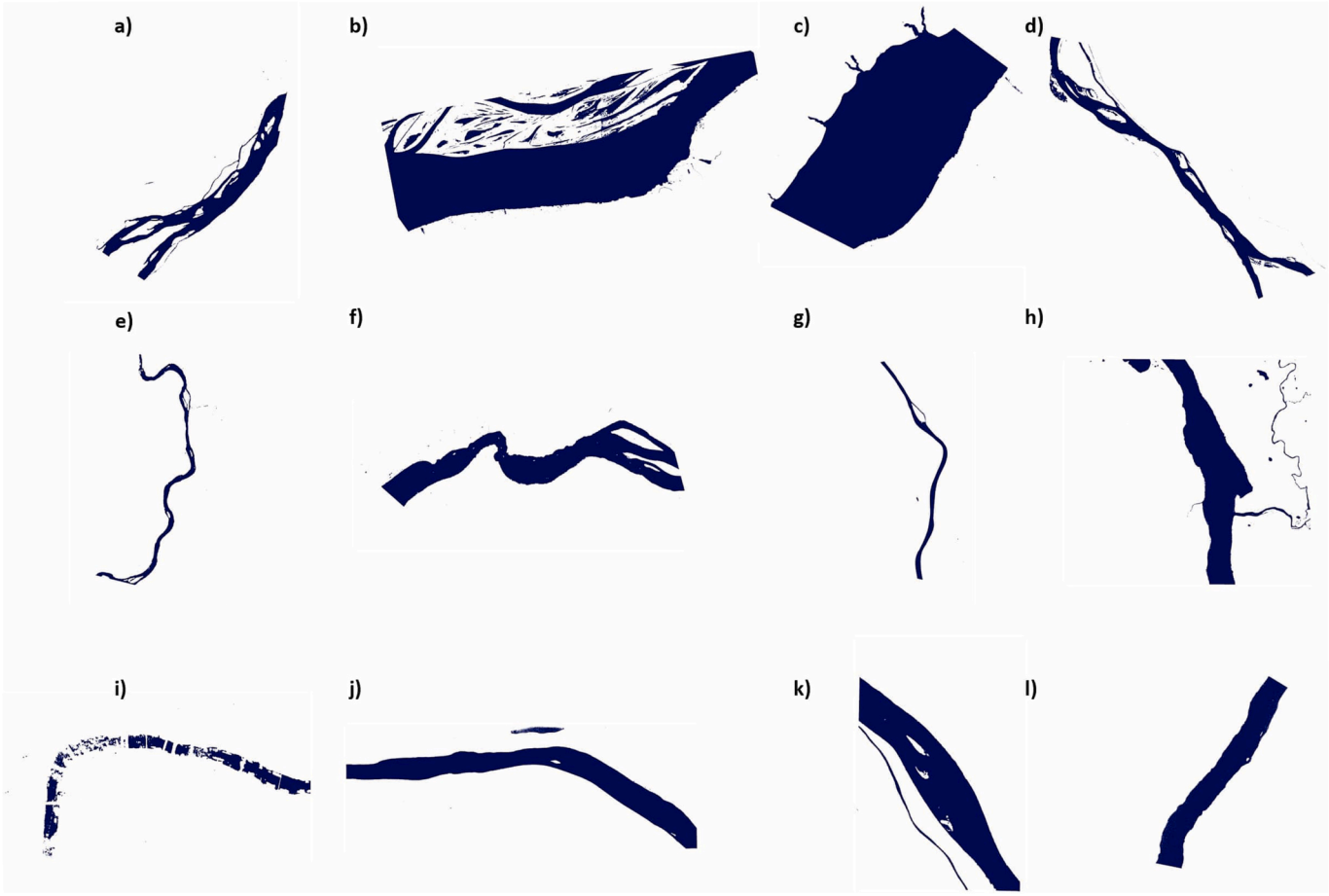

Fig. 4. Selection of water masks produced using M20. The majority of these rivers scored an F1 accuracy >0.97. However, the River Thames, through central London, is also included with a score of 0.85 to display model results in the most complex urban environments with multiple watercraft and bridges in a small spatial area (i). In this case, it has still produced a fairly accurate water mask in demanding circumstances. Rivers shown here a: Niger (Mali), Tropical Grasslands . b: Amazon (Brazil), Tropical Moist Broadleaf . c: Volga (Russia), Temperate Grasslands . d: Nile (Egypt), Flooded Grasslands . e: Betsiboka (Madagascar), Tropical Moist Broadleaf . f: Xi (China), Tropical Moist Broadleaf . g: Victoria (Australia), Tropical Grasslands . h: Pyasina (Russia), Tundra . i: Thames (UK), Temperate Broadleaf, j: Mackenzie (Canada), Taiga. K: Slave (Canada), Taiga . l: Congo (Democratic Republic of the Congo), Tropical Grasslands.

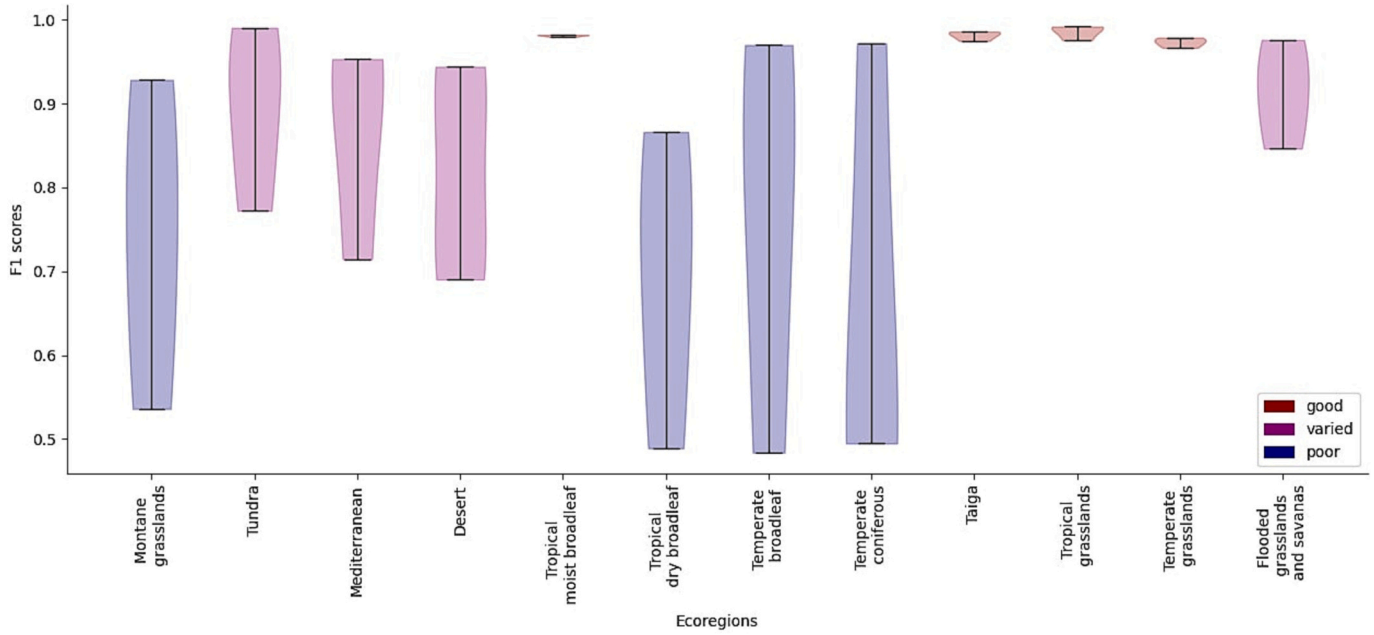

Fig. 5. Within region F1 scores for biomes predicted by M20. Based on these results, biome predictability was classified as good, varied, or poor.

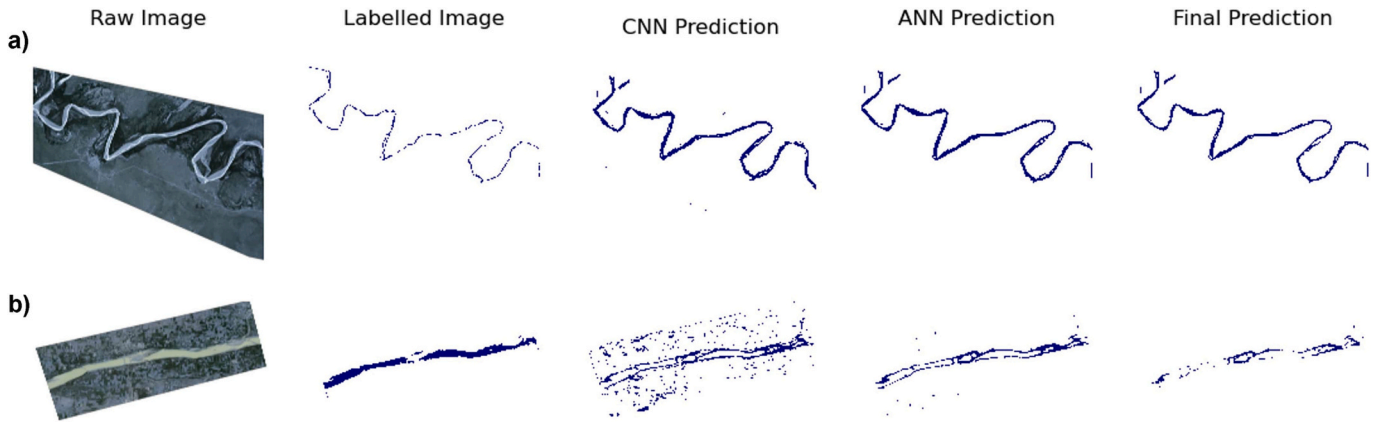

Fig. 6. Comparison of CSC predictive workflows for well and poorly predicted images in the poor performing biomes. Despite narrow sinuous channels, the Vaal River in South Africa, from the Tropical Dry Broadleaf biome, scored an F1 score of 0.86 (a). Other images in the same biome struggled with issues such as sand bars and urban areas, including the Narmada River in India (b) with an F1 score of 0.72. It is clear that if the CNN prediction is particularly poor, then despite the predictive improvements made by the ANN section, the final prediction will suffer.

from 0.48 to 0.85 (see supplementary materials). Visual inspection of these images identified three issues common to these locations that could have contributed to the low F1 scores yielded by these data: 1 . braided river types without a dominant main channel, 2. land-use types that are unclear even to a human operator (i.e., where colour/textural differences between water and non-water pixels are difficult to discriminate), and c) the presence of bridges or instream features (e.g., boats). These issues all appeared to cause issues for model predictions at the first CNN phase, which were then propagated through the MLP to the final result. Additional potential causes of error result from the presence of cloud or ice in imagery. For example, despite choosing cloud-free images from the PlanetScope database, some images still included small amounts of cloud (Fig. 8.a). Where there was a limited amount of ice and snow in an image the model still performed well (Fig. 8.b), but this was not the case with deep shadow covering parts of the image (Fig. 8.c).

3.3. Fine tuning

M20 produced acceptable F1 scores on the Ottawa River test images. As with results in the original holdout images, there was a tendency towards successful predictions with 9 images predicted with a F1 score over 0.9, 1 had a score of 0.85, and 5 had scores below 0.55 producing an overall median of 0.97. After the inclusion of the additional 'human-in-the-loop' fine tuning labels there was only a marginal increase in median to 0.98. However, improvements were found to be focused on the images which performed worst without fine tuning. This caused the mean F1 score to increase from 0.81 to 0.90 after this fine tuning. The increase was found to be significant after Box-Cox transformation to remove skew ($n=15$, $t=8.4$, $p<0.01$). There was some variation in F1 scores with a number of images producing marginally lower score in the fine-tuned run, in comparison to the base M20 model. However, three of the most poorly performing images saw greater improvement from <0.7 to >0.95 (Fig. 9). Two images still performed poorly, one likely due to the presence of a dam and the other due to an unexplained, red-coloured artefact potentially caused by shallow water (both factors potentially

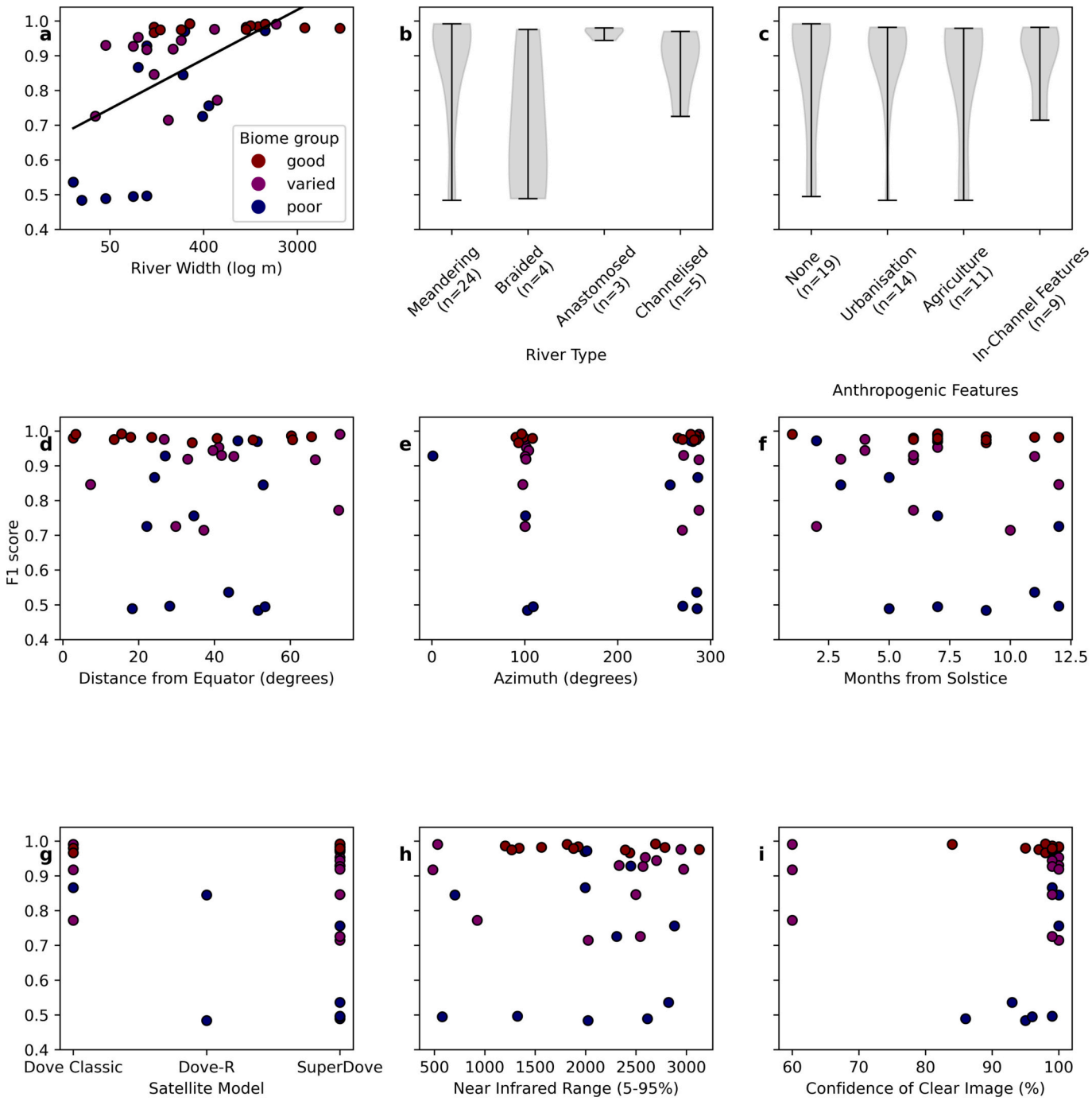

Fig. 7. F1 score as function of variability in external (e.g., physical river characteristics) and internal (image metadata) properties. Log transformed width correlates significantly to F1 score ($r^2 = 0.34, p < 0.01$), which is to be expected due to the greater importance of differences in smaller river widths, as opposed to differences in the widths of larger rivers. No other variables were found to be significantly correlated.

lacking in the training data; see supplementary material). The original test dataset (all images, all biomes) was then re-predicted using this fine-tuned model to test if the increase in the Ottawa River-specific training data had reduced its generalizability. Results show no impact on median F1 score but a minor (but significant) decrease in overall mean F1 score (0.83 to 0.81), indicating a slight reduction in the classification model's generalisability, at the expense of improved classification results for the Ottawa River.

4. Discussion

4.1. Applicability of the CSC model

The CSC models developed here successfully produced water masks from highly diverse satellite images, from rivers across a gradient of biomes and seasons. The novel use of the CSC architecture with a CubeSat constellation highlights the potential for the incorporation of daily PlanetScope satellite imagery in studies relating to dynamic river systems. The inclusion of a fine-tuning step (section 4.5) adds further capacity to refine the model on the sites of most importance and interest

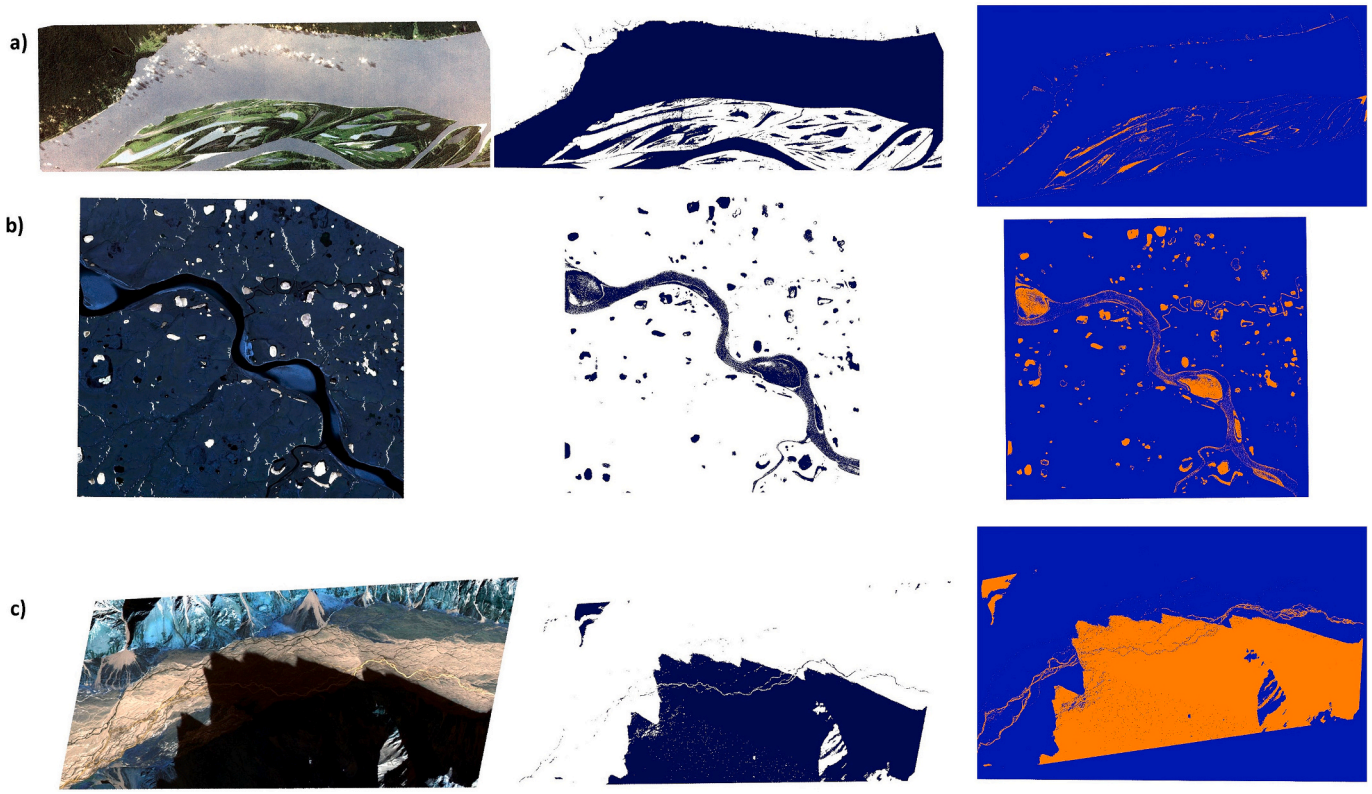

Fig. 8. Examples highlighting where the M20 CSC model coped or struggled with more difficult images. The first column displays the original satellite image, the second displays the prediction, and the third shows correct predictions in blue and incorrect predictions in red. a: Amazon, Brazil with unprocessed cloud. b: Popigay, Russia with snow covered frozen pools . c: Ahuri, New Zealand, deep shade which resulted in a poor prediction. However, further research is needed because the model still separated the river in both shaded and unshaded sections of the image. Example c was not included in the analysis when selecting hold out images but was added as an additional test image to see how the model coped with more extreme examples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for a user, allowing the generation of highly accurate water masks with minimal labour or expertise required. The median F1 score of M20 (0.93) is a particularly notable given that our approach is effectively an automated system being tested across multiple global domains, and in some biomes our approach performs extremely well scoring over >0.95 F1 on all test images. Although we acknowledge that the lack of field data prevented formal ground validation, the global nature of this study precludes such an approach. Instead, more manually intensive labelling methods were used for the assembly of validation data thatalso come with their own associated error which inherently prevents perfect F1 scores (see Moortgat et al., 2022 for further discussion). Therefore, the high F1 scores are encouraging given the limitations of the labelled data these test images were compared against.

The method proposed here is not intended to compete with the GRWL water masks (Allen and Pavelsky, 2018) or similar products, which are employed regularly in satellite studies of river systems (Lin et al., 2019; Rosentreter et al., 2021; Yang et al., 2020). Rather, our approach is intended as a method for developing water masks in dynamic environments at (up to) daily frequency, as opposed to using a static product. We propose that a continuation of the method developed here could refine existing GRWL RivWidth predictions, which require a binary water mask as input (Pavelsky and Smith, 2008). Placing the model developed here in a similar cloud computing pipeline (Yang et al., 2019) may encourage users to adopt concurrent development of water masks rather than use static, assumed bankfull datasets such as the GRWL (Allen and Pavelsky, 2018). Adaptive water masks could push remote sensing breakthroughs in temporally sensitive aspects of river systems such as flow dynamics and large-scale variations in water quality across a range of discharges. Our method also provides water masks and therefore river widths at a higher resolution than the GRWL. The GRWL application is limited to rivers ≥90 m wide (Allen and Pavelsky, 2018) which could be overcome with the 20*20 3 m pixel cells used here, enabling 60 m rivers and sometimes smaller to be fed into RivWidth. This ability to monitor smaller rivers, especially at a higher temporal resolution, would vastly increase the applicability of these methods for studying flooding scenarios where water progressively accumulates downstream into larger rivers.

4.2. Constraints on model success

Despite the success of the M20 model, Median F1 scores in our investigation are not as high as those in similar AI-driven water mask studies (e.g., Moortgat et al., 2022; Carbonneau et al., 2020; Qayyum et al., 2020). These studies differ from ours in a variety of key ways because of constraints we had to place on this model to work with PlanetScope imagery. Therefore, the very fact that the accuracy of our M20 model approaches these other studies, in spite of these limitations, makes us consider it effective.

All satellites represent a trade-off between spatial resolution and frequency which directly relates to the quality of water mask predictions which can be achieved with their imagery. CubeSats are released at a relatively low altitude which decays further over time (Planet Labs, 2018), enabling them to acquire medium-high resolution imagery. Having a constellation of CubeSats facilitates daily return period imagery at this 3 m resolution, which is vital for the task of monitoring river systems. This means the spatial resolution we are able to achieve for our resulting water masks is superior to models focusing solely on ‘conventional’ EO satellites (e.g., Landsat, Sentinel). Conversely, proprietary satellites such as Worldview or China’s GF-2 produce imagery with a resolution many times finer than PlanetScope is capable of

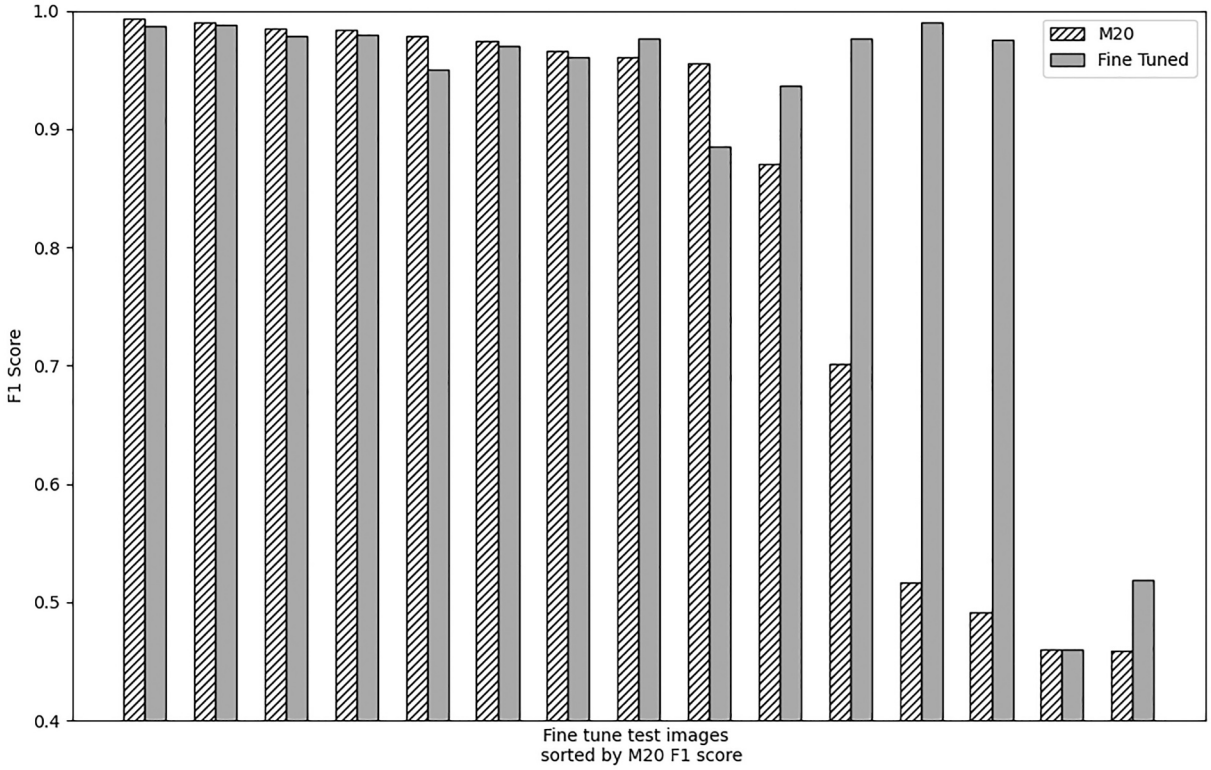

Fig. 9. Difference in F1 scores between images predicted by the M20 model and after 5 images from the same river were included in the training data to fine tune it. Improvements are not universal but are dramatically better for the worst performing images.

(Moortgat et al., 2022; Li et al., 2021). The same can be found of aerial imagery where the proximity to the target results in very high-resolution data. Water mask classification models produced using these data understandably perform very well because each CNN tile is able to include more pixels for the same ground area enabling better characterisation of water or land structures (Carbonneau et al., 2020; Moortgat et al., 2022; Li et al., 2021). While these high- and hyperspatial resolution data capture methods have potential to produce better results than those presented here, it is not practically or financially feasible to obtain the temporal frequency of imagery required for generation of dynamic water masks. PlanetScope is thus the only system we are aware of that provides both relatively high temporal *and* spatial resolution. While we acknowledge that PlanetScope is proprietary, it is less expensive than other high-resolution satellite or airborne imagery (Cornebise et al., 2022). Furthermore, PlanetScope offers 5000 km^2 a month of free imagery for individuals/groups affiliated with an educational institution under their Education and Research Programme, meaning that the techniques developed here will be open and useable by a large academic community. Additionally, it is currently available for free for some NGOs and Government employees through ESAs third party mission scheme (ESA, 2023) and NASAs Commercial Smallsat Data Acquisition program (Maskey et al., 2021).

A trade-off of the high return period provided by CubeSats is their much lower radiometric quality, an issue that our classification model would not have had to contend with if developed with data from other satellite sources. Larger, high quality sensor systems used in these other platforms create radiometrically stable data with a much higher signal-to-noise ratio and therefore less erroneous values for the model to cope with (Ling et al., 2019; Yin et al., 2021; Isikdogan et al., 2017). Similarly, 'conventional' aerial imagery is largely immune from the disruptive effects of the atmosphere (Smith et al., 2021), with the resulting image quality being several orders of magnitude higher. Conversely, with PlanetScope, each platform, each sensor on the platform, and the calibration of these sensors differs as a result of the sheer quantity of satellites and their limited payload capacity (Frazier and Hemingway, 2021). Although the metadata provided by Planet does show no correlation with F1 score (Fig. 7), unreported differences in illumination conditions and weather conditions add additional complexities. The intricacy of hosting this assemblage of satellites means that, despite proprietary corrections made by PlanetScope (Kington and Collison, 2022) to normalise for radiometric variability between satellites, differences in absolute radiometric values still abound (Wilson et al., 2017). This in part may be the cause of the lack of correlation between F1 score and NIR band metrics. This aligns with findings that CNN models can delineate water masks relatively well using just panchromatic imagery (Moortgat et al., 2022) when NIR values would otherwise normally be expected to be the chief determinate of ability to depict water from 4-band imagery (Mcfeeters, 1996). Based on these factors, a relatively simple classification model might ostensibly be considered unviable, but our M20 CSC method nevertheless appears to be able to adequately handle these complexities. These findings are particularly interesting because CubeSats are increasingly common in the EO sector (De et al., 2022). If CNNs (and by extension, CSC) can produce accurate classifications despite this variation in CubeSat image quality, then they could be applicable to a wide variety of cases beyond water masking.

By comparison, the much poorer performance of the comparator state-of-the art Tiramisu FCN may come as a surprise. However, work using FCNs generally involves much higher quality (i.e., radiometrically stable) imagery (Moortgat et al., 2022; Carbonneau and Bizzi, 2023) than PlanetScope, meaning that strong radiometric differences (which our CSC approach is robust to) are not accounted for by the FCN model architecture. The FCN requires considerably larger tile sizes than CSC, and while these tiles can be mixed (water and land), the requirement for larger tiles makes it more difficult to balance water and land pixels, even with a focal loss function "punishing" the model more for errors in water predictions. Trying to predict directly to the pixel level requires a large training set (Jégou et al., 2017), which is also inherently difficult with larger tile sizes (i.e., fewer training tiles per image). This necessitated

the use of augmentation to gain more training tiles, potentially biasing the model to expect images of similar radiometric variability to this relatively narrow training base of 72 images. We hypothesise that this meant that when predicting hold-out images, the model was subsequently unable to cope with the extreme radiometric differences between these 'unseen' PlanetScope images. The additional complexity of the semantic segmentation thus led the FCN to perform poorer than our CSC which, through the inclusion of the secondary ANN stage, was able to achieve high quality semantic segmentation for the image quality directly related to the input image.

4.3. Model structural and testing choices

Our classification model is reasonably robust to levels of internal sensor noise and poor radiometric quality, but testing the model globally introduced further external variability. Many methodological studies into the use of EO for mapping river characteristics constrain their datasets to single regions and limited numbers of images (Feng et al., 2019; Junqueira et al., 2021; Moortgat et al., 2022; Qayyum et al., 2020). This limits the image variability resulting from different climates, vegetation, and river types (Thoms and Sheldon, 2019; Malhi et al., 2022). As a result, when site-specific classification methods are expanded globally, they are generally ineffective (Foody et al., 2003). A large part of the lower overall accuracy found here, in comparison to similar studies (e.g., Moortgat et al., 2022; Marochov et al., 2021), is explained by our inclusion of rivers from across different biomes. It is important to reiterate that this model was not intended nor expected to work universally and without adaption.

To maximise potential operational use of the model, structural decisions were made which could have impacted prediction scores. Spatial resolution and the associated tile size used in different models is key to understanding how structural differences have impacted the relative performance of models. Increasing tile size facilitates higher per-tile information content, allowing the CNN phase of the CSC to better learn contextual features such as texture and geometry (Carbonneau et al., 2020); the improved training metrics on a random allocation of test tiles in each epoch (0.87 accuracy and 0.31 loss for M32 compared to 0.84 and 0.36 for M10) thus fit with what we know about increasing tile size (Reina et al., 2020). This enables a better prediction for images with a higher range of different pixel brightness values. However, these improved accuracy and loss metrics in training did not correspond to increased F1 scores on hold out images (Table 1). Due to the resolution of our imagery, the larger the tile size, the more difficult it is to pick out enough pure-water tiles when predicting an image. This in turn leads to a poorly trained MLP and a poorer final CSC outcome. Therefore, there is clearly a balance between information and resolution for choosing the best tile size. Higher resolution imagery can increase pixels in a tile whilst maintaining the tile's footprint and therefore, ability to predict narrow streams whilst also increasing F1 scores (Carbonneau et al., 2020; Moortgat et al., 2022).

It might be surprising that a model structure often considered universally useful such as VGG16 (Theckedath and Sedamkar, 2020), did not perform well here. VGG16 requires a minimum tile size of 32 due to the number of max pooling layers in the architecture (Simonyan and Zisserman, 2014). In part, the lack of success here might be attributable to this balance between tile size, resolution, and river size. However, the depth and complexity of the VGG16 could also be too large for the quantity of training data available and the binary classification task. VGG16 was created with 1000 classes and 138 million parameters which can lead it to very easily overfit with practical (in this case binary) tasks and relatively small training data sets (Wu et al., 2017), which explains why it was limited in value here.

4.4. Ecoregion differences

Median F1 scores are shown to obscure considerable variation within the results for different biomes (Fig. 5). This is a common trend in image classification, with potential for individual results to be considerably worse than their aggregated scores (Carbonneau et al., 2020; Buscombe and Ritchie, 2018). A selection of smaller rivers simply did not have enough pixels for the CNN step to be effective, which results in unsuccessful predictions. However, there is a large degree of variation in some results from rivers with widths of ~60 m. We hypothesise that these rivers are large enough to have 'disruptive' features (e.g., large islands and bridges), which impact model accuracy, yet they lack sufficient pure water pixels required for the model to overcome these errors, weakening the correlation between river width and F1 score.

Braided streams were found to produce lower F1 scores, which was especially the case within the worst performing biome (Tropical Dry Broadleaf). Despite the overall quantity of pure water pixels across the theoretical bankfull area, individual stretches of water were often too narrow for the CNN phase to pick out. Seasonality therefore effects these narrow channels, which would only be measurable when flows are high enough to increase the water surface area to the minimum requirements of the model (Ashmore et al., 2011). Therefore, although the model produced here may not be viable for average discharge measurements in these braided streams, the return period of PlanetScope makes it a viable to still be an important flood monitoring tool (Feng et al., 2019).

Some conditions were not accounted for in the training data, but we included separate test images for these (see supplementary material). Snow, ice, and frozen river systems proved problematic for the model, although further research is needed to determine if and how our model should delineate frozen watercourses. Encouragingly, however, the model does not appear to be disrupted by winter conditions on the banks such as snow-covered land in the Taiga biome (Fig. 8.b). Similarly, the model appears able to cope with artifacts relating to minor cloud or shadow, which could negatively impact F1 scores but not to a great extent (Fig. 8.a). The model was not trained for these specifically, yet when applied to extreme cases of each it was still able to function to some degree. In one heavily shaded image, the model delineates dark mountain shadow and land as different classes, yet in each the river is considered different from the surrounding land type (Fig. 8.c). This shadow issue is present in other studies of PlanetScope classification in mountains (Qayyum et al., 2020) and should be investigated further but nonetheless clearly shows that the CNN phase is based on more information content than simply pixel values.

Other conditions where the model was expected to struggle did not lead to catastrophic failure. For example, results from the Temperate Broadleaf biome were lower than several other biomes but still around F1 $\approx$ 0.8. Upon inspecting images associated with these F1 scores, it became clear that they comprised urban areas, such as central London. We might expect the model to struggle in these areas due to the conflicting NIR reflectance values from other urban land use types (Xu, 2006). However, the model coped fairly well barring errors around bridges and watercraft. This is promising because it suggests that limited additional specific test data could improve urban results dramatically.

The categorisation of the different biomes into good, variable, and poor is clearly simplistic for the task at hand but could be useful. These categories highlight characteristics that the model struggles with. For example, Montane Grasslands often have braided streams due to the high gradient environment (Montgomery and Buffington, 1998), more ice and snow, and smaller first order streams (Strahler, 1957). These lead to fewer seasons in which the model can be applied because it is expected to produce poorer results when low angle sun hits steep valley sides, low flow confines braided streams to separate small channels, or snow covers banks. This provides useful information for river managers on our model's applicability. However, when the 'flashy' nature of upland hydrological cycles limits other forms of flood monitoring, these braided streams would combine to become wide enough for our method to identify and produce robust river masks. This means that while the model might be sub-optimal for braided streams during particularly low flows, it is likely to work effectively during large flood events when

other remote sensing methods cannot be deployed as rapidly (Ballesteros-Cánovas et al., 2015).

4.5. Fine tuning

We have shown where and when M20 is likely to work best but intend it to be considered as a 'base' model, to be further fine-tuned to the region or task required. This will enable users to save vast amounts of time when creating water masks for rivers, by allowing them to achieve highly accurate classification results with the labelling of only a handful of images. The main strength of the CSC method is that the base CNN has learnt to smooth the difficulties and differences associated with CubeSat data. Therefore, with minimum effort individuals can operationalise the high spatial and temporal resolution of PlanetScope to process water masks.

'Interventional model training' can be used to bias the model towards the biome in question whilst adding to the overall learning and generalizability of the model (Wu et al., 2022). Here we use this to integrate human knowledge and experience to keep humans 'in the loop' of model training and usage (Wu et al., 2022; Nunes et al., 2015). The training data pipeline used in this study only required the drawing of simple polygons enabling "no data" classifications (see https://github.com/SamValman/Public_RiverTwin for a step-by-step guide). Increasing the speed and ease of labelling thus encourages practitioners to invest the time to use and improve the classification model (through the creation of additional labels). This simplification step during training could have inhibited our model in comparison to similar studies with more complex methods of creating training data (Moortgat et al., 2022). This was especially the case with the FCN comparison attempted here, which in turn required considerably more training data and pixel level classifications, which required the use of semi-automatic classifiers, sometimes with multiple iterations to obtain sufficient quality water masks. The simple polygon-based labelling method of CSC enables model training to be targeted to capture the right information at the right place and time, which in turn necessitates non-experts to develop additional training data (Rabaey, 2020).

Testing of the theory that these models could be improved with limited additional training data was carried out on the Ottawa River, situated in the Taiga biome. Although the Taiga was the best performing biome, the Ottawa River was not one of the initial 36 rivers in the study. Therefore, some of the characteristics that make up the Ottawa River could have been learnt by the model from other similar rivers, but it will not have explicitly seen this river. Despite this, the M20 model first predicted the majority of the 15 Ottawa River test images well (Fig. 9). However, fine-tuning M20 with 5 additional training images from the Ottawa River significantly improved predictions on these same 15 hold-out images. This is an important demonstration of just how little user input is required to create useful water mask predictions once the initial global model has been improved.

In this study, we chose to retrain the model from scratch whilst including the fine-tuned images. This was done to prevent 'catastrophic forgetting' whereby transfer learning causes a model to be unable to accurately make predictions of the original training task (Kirkpatrick et al., 2017). In an ideal world the methods chosen would enable 'lifelong learning' whereby with each user-fed fine-tuning run, the model would increase its training set (Parisi et al., 2019). This would reduce the time taken for users to update the model and allow the model to continuously improve by a concerted community effort. Methods of overcoming this 'catastrophic forgetting' problem are beyond the scope of this paper, but bringing these methods from the computer science domains to EO and river science should be a research priority. On top of this call for an increase in community efforts to train truly global models, there need to be easier ways to access image classification models such as those used here. Cloud computing would enable access to multiple individuals with diverse site knowledge and training data (Rabaey, 2020). This is partly why the model was built using the TensorFlow API, which has good compatibility with Google Earth Engine (Gorelick et al., 2017; Abadi et al., 2016). Future work should focus on bringing this lifelong learning into the cloud to collectivise the best models for maximum benefit.

5. Conclusion

Rivers are large, dynamic systems which cannot be monitored by field measurements at the spatial and temporal scale required for management. EO has the spatial coverage to provide a potential solution to monitoring issues but traditionally used open-source satellites, such as Landsat, do not always have the spatial or temporal resolution to be useful for practical applicability. The PlanetScope constellation operates a daily return rate with 3 m spatial resolution, but it has previously been limited to localised studies due to its variable radiometric quality. Our novel application of CNN-Supervised Classification (CSC) was shown to produce accurate water masks from PlanetScope imagery, overcoming the inherent radiometric issues associated with imagery from CubeSat constellations to provide automated water masking at a finer temporal resolution than any other current EO method. We found that CNN-based architecture is a promising option to analyse the increasing quantity of images coming from CubeSat satellites with increasing numbers of bands. By being the first model to explicitly develop accurate water masks from PlanetScope data, it has created the potential for river management applications to be operationalized at a global scale with daily medium-high resolution imagery.

This model developed here was demonstrated to generalise reasonably well across different global biomes, with a median F1 score of 0.93 and max F1 scores of 0.99. Narrow anastomosed streams, deep shadow, and some urban influences were shown to limit the accuracy of model predictions. We showed that these limitations were concentrated in some biomes, but also demonstrated that the model can be fine-tuned to improve its applicability in specific locations. Indeed, our results indicate that very marginal increases in the training data can significantly improve the results of the base model in case study regions. Where it works particularly well, we also believe that our model could be extended to differentiate between different hydromorphic units such as vegetation, visible sediment, or hydraulic features. This extension would be a directly applicable management tool for environmental assessments which often request mapping of these features which might change with different discharge levels (Fryirs and Brierley, 2022). We therefore call for more research into making web-hosted 'lifelong learning' models that utilise the increasing availability of satellite data and the ability to continuously retrain these models, improving both their accuracy and allowing them to classify other riverine features. This would help the entire field move forwards towards the practical operationalization of this imagery.

Funding

The project was supported by the Geospatial Systems Centre for Doctoral Training [EP/S023577/1], funded by the UK's Engineering and Physical Sciences Research Council (EPSRC). The authors would like to acknowledge Planet Education and Development programme which provided data for this study.

CRediT authorship contribution statement

Samuel J. Valman: Conceptualization, Formal analysis, Investigation, Software, Writing – original draft, Writing – review & editing. **Doreen S. Boyd:** Conceptualization, Supervision, Writing – review & editing. **Patrice E. Carbonneau:** Methodology, Software, Writing – review & editing. **Matthew F. Johnson:** Writing – review & editing. **Stephen J. Dugdale:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code is provided in a GitHub repository linked in the supplementary material. This includes a list of all satellite imagery which is proprietary but can be accessed through Planet Research program

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2023.113932.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., ISARD, M., 2016. {TensorFlow}: A System for {Large-Scale} Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283.

Abburu, S., Golla, S.B., 2015. Satellite image classification methods and techniques: A review. Int. J. Comput. Appl. 119.

Allen, G.H., Pavelsky, T.M., 2015. Patterns of river width and surface area revealed by the satellite-derived North American River Width data set. Geophys. Res. Lett. 42, 395–402.

Allen, G.H., Pavelsky, T.M., 2018. Global extent of rivers and streams. Science 361, 585–588.

Altenau, E.H., Pavelsky, T.M., Durand, M.T., Yang, X., Frasson, R.P.D.M., Bendezu, L., 2021. The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A global river network for satellite data products. Water Resour. Res. 57 e2021WR030054.

Arnell, N.W., Gosling, S.N., 2016. The impacts of climate change on river flood risk at the global scale. Clim. Chang. 134, 387–401.

Ashmore, P., Bertoldi, W., Tobias Gardner, J., 2011. Active width of gravel-bed braided rivers. Earth Surf. Process. Landf. 36, 1510–1521.

Baghdadi, N., Mallet, C., Zribi, M., 2018. QGIS and Applications in Water and Risks. John Wiley & Sons.

Ballesteros-Cánovas, J.A., Czajka, B., Janecka, K., Lempa, M., Kaczka, R., Stoffel, M., 2015. Flash floods in the Tatra Mountain streams: Frequency and triggers. Sci. Total Environ. 511, 639–648.

Boothroyd, R.J., Williams, R.D., Hoey, T.B., Barrett, B., Prasojo, O.A., 2021. Applications of Google Earth Engine in fluvial geomorphology for detecting river channel change. Wiley Interdiscip. Rev. Water 8, e21496.

Bradski, G., Kaehler, A., 2000. OpenCV. Dr. Dobb's J. Softw. Tools 3, 120.

Brierley, G.J., Fryirs, K.A., 2013. Geomorphology and River Management: Applications of the River Styles Framework. John Wiley & Sons.

Buscombe, D., Goldstein, E., 2022. A reproducible and reusable pipeline for segmentation of geoscientific imagery. Earth Space Sci. 9 e2022EA002332.

Buscombe, D., Ritchie, A.C., 2018. Landscape classification with deep neural networks. Geosciences 8, 244.

Carbonneau, P.E., Bizzi, S., 2023. Global mapping of river sediment bars. Earth Surf. Process. Landf. https://doi.org/10.1002/esp.5739.

Carbonneau, P.E., Dugdale, S.J., Breckon, T.P., Dietrich, J.T., Fonstad, M.A., Miyamoto, H., Woodget, A.S., 2020. Adopting deep learning methods for airborne RGB fluvial scene classification. Remote Sens. Environ. 251, 112107.

de la Comble, Prepin, K., 2021. Efficient transfer learning for multi-channel convolutional neural networks. In: 2021 17th International Conference on Machine Vision and Applications (MVA), 25-27 July 2021, pp. 1–6.

Congedo, L., 2021. Semi-Automatic Classification Plugin: A Python tool for the download and processing of remote sensing images in QGIS. J. Open Sourc. Softw. 6, 3172.

Cooley, S.W., Smith, L.C., Stepan, L., Mascaro, J., 2017. Tracking dynamic northern surface water changes with high frequency Planet CubeSat Imagery. Remote Sens. 9, 1306.

Cornebise, J., Oršolić, I., Kalaitzis, F., 2022. Open high-resolution satellite imagery: the worldstrat dataset–with application to super-resolution. Adv. Neural Inf. Proces. Syst. 35, 25979–25991.

De, R., Abegaonkar, M.P., Basu, A., 2022. Enabling science with CubeSats—Trends and prospects. IEEE J. Miniaturizat. Air Space Syst. 3, 221–231.

Dougherty, E.R., 2020. Digital image processing methods. CRC Press.

European Space Agency, 2023. Terms and Conditions for the Utilisation of Data under ESA's Third Party Missions scheme between the EUROPEAN SPACE AGENCY and the Principal Investigator. ESA-EOPG-PDGS-PR-2 [Online]. Available at: https://earth.esa.int/eogateway/documents/20142/1560778/ESA-Third-Party-Missions-Terms-and-Conditions.pdf [Accessed 26th September 2023].

Feng, D., Gleason, C.J., Yang, X., Pavelsky, T.M., 2019. Comparing discharge estimates made via the BAM algorithm in high-order Arctic rivers derived solely from optical CubeSat, Landsat, and Sentinel-2 data. Water Resour. Res. 55, 7753–7771.

Foody, G.M., Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103, 179–189.

Foody, G.M., Boyd, D.S., Cutler, M.E., 2003. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. Remote Sens. Environ. 85, 463–474.

Frazier, A.E., Hemingway, B.L., 2021. A technical review of planet smallsat data: practical considerations for processing and using planetscope imagery. Remote Sens. 13, 3930.

Fryirs, K., Brierley, G., 2022. Assemblages of geomorphic units: A building block approach to analysis and interpretation of river character, behaviour, condition and recovery. Earth Surf. Process. Landf. 47, 92–108.

Gabr, B., Ahmed, M., Marmoush, Y., 2020. PlanetScope and landsat 8 imageries for bathymetry mapping. J. Marine Sci. Eng. 8, 143.

Gardner, J.R., Yang, X., Topp, S.N., Ross, M.R.V., Altenau, E.H., Pavelsky, T.M., 2021. The Color of Rivers. Geophys. Res. Lett. 48 e2020GL088946.

Gavrilov, A.D., Jordache, A., Vasdani, M., Deng, J., 2018. Preventing model overfitting and underfitting in convolutional neural networks. Intern. J. Softw. Sci. Comput. Intell. (IJSSCI) 10, 19–28.

Genitha, C.H., Vani, K., 2013. Classification of satellite images using new fuzzy cluster centroid for unsupervised classification algorithm. In: 2013 IEEE Conference on Information & Communication Technologies. IEEE, pp. 203–207.

Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Sebastopol, USA.

Gleason, C.J., Durand, M.T., 2020. Remote sensing of river discharge: a review and a framing for the discipline. Remote Sens. 12, 1107.

Gleyzes, M.A., Perret, L., Kubik, P., 2012. Pleiades system architecture and main performances. Intern. Arch. Photogram. Remote Sens. Spatial Inform. Sci. 39, 537–542.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Hannah, D.M., Demuth, S., Van Lanen, H.A., Looser, U., Prudhomme, C., Rees, G., Stahl, K., Tallaksen, L.M., 2011. Large-scale river flow archives: importance, current status and future needs. Hydrol. Process. 25, 1191–1200.

Haq, M.A., 2022. Planetscope nanosatellites image classification using machine learning. Comput. Syst. Sci. Eng. 42.

Isikdogan, F., Bovik, A.C., Passalacqua, P., 2017. Surface water mapping by deep learning. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 10, 4909–4918.

Isikdogan, L.F., Bovik, A., Passalacqua, P., 2020. Seeing through the clouds with deep water map. IEEE Geosci. Remote Sens. Lett. 17, 1662–1666.

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. Int. J. Remote Sens. 42, 5338–5365.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 11–19.

Junqueira, A.M., Mao, F., Mendes, T.S.G., Simões, S.J.C., Balestieri, J.A.P., Hannah, D. M., 2021. Estimation of river flow using CubeSats remote sensing. Sci. Total Environ. 788, 147762.

Kasprak, A., Hough-Snee, N., Beechie, T., Bouwes, N., Brierley, G., Camp, R., Fryirs, K., Imaki, H., Jensen, M., O'brien, G., Rosgen, D., Wheaton, J., 2016. The blurred line between form and process: a comparison of stream channel classification frameworks. PLoS One 11, e0150293.

Kelso, N.V., Patterson, T., 2010. Introducing natural earth data-naturalearthdata.com. Geogr. Tech. 5, 25.

Kington, J., Collison, A., 2022. Scene level normalization and harmonization of Planet Dove Imagery. Planet Labs, San Francisco.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114, 3521–3526.

Le Roux, J., Christopher, S., Maskey, M., 2021. Exploring the use of PlanetScope data for particulate matter air quality research. Remote Sens. 13, 2981.

Lew, G., Schumacher Jr., R.M., 2020. Garbage in, Garbage out. AI and UX: Why Artificial Intelligence Needs User Experience. Springer.

Li, M., Wu, P., Wang, B., Park, H., Yang, H., Wu, Y., 2021. A deep learning method of water body extraction from high resolution remote sensing images with multisensors. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 14, 3120–3132.

Lin, P., Pan, M., Beck, H.E., Yang, Y., Yamazaki, D., Frasson, R., David, C.H., Durand, M., Pavelsky, T.M., Allen, G.H., 2019. Global reconstruction of naturalized river flows at 2.94 million reaches. Water Resour. Res. 55, 6499–6516.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. Proceed. IEEE Internat. Conf. Comput. Vision 2980–2988.

Ling, F., Boyd, D., Ge, Y., Foody, G.M., Li, X., Wang, L., Zhang, Y., Shi, L., Shang, C., Li, X., Du, Y., 2019. Measuring river wetted width from remotely sensed imagery at the subpixel scale with a deep convolutional neural network. Water Resour. Res. 55, 5631–5649.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and pattern Recognition., pp. 3431–3440.

Mahoney, C., Merchant, M., Boychuk, L., Hopkinson, C., Brisco, B., 2020. Automated SAR image thresholds for water mask production in Alberta's Boreal Region. Remote Sens. 12, 2223.

Malhi, R.K.M., Kiran, G.S., Srivastava, P.K., Bhattacharya, B.K., Mohanta, A., 2022. Spectral mixture analysis of AVIRIS-NG data for grouping plant functional types. Adv. Space Res. https://doi.org/10.1016/j.asr.2022.12.023.
Mansaray, A.S., Dzialowski, A.R., Martin, M.E., Wagner, K.L., Gholizadeh, H., Stoodley, S.H., 2021. Comparing PlanetScope to Landsat-8 and Sentinel-2 for sensing water quality in reservoirs in agricultural watersheds. Remote Sens. 13, 1847.
Marochov, M., Stokes, C.R., Carbonneau, P.E., 2021. Image classification of marine-terminating outlet glaciers in Greenland using deep learning methods. Cryosphere 15, 5041–5059.
Maskey, M., Hall, A., Murphy, K., Tucker, C., Mccarty, W., Kaulfus, A., 2021. Commercial Smallsat data acquisition: Program update. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 600–603.
Mazhar, S., Sun, G., Bilal, A., Hassan, B., Li, Y., Zhang, J., Lin, Y., Khan, A., Ahmed, R., Hassan, T., 2022. AUnet: A deep learning framework for surface water channel mapping using large-coverage remote sensing images and sparse scribble annotations from OSM data. Remote Sens. 14, 3283.
Mcfeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432.
Montgomery, D.R., Buffington, J.M., 1998. Channel processes, classification, and response. River Ecol. Manage. 112, 1250–1263.
Moortgat, J., Li, Z., Durand, M., Howat, I., Yadav, B., Dai, C., 2022. Deep learning models for river classification at sub-meter resolutions from multispectral and panchromatic commercial satellite imagery. Remote Sens. Environ. 282, 113279.
NASA, 2020. Earth Science Division Commercial SmallSat Data Acquisition Program Pilot Evaluation Report. NASA, Washington, DC, USA.
Nativi, S., Mazzetti, P., Craglia, M., 2021. Digital ecosystems for developing digital twins of the earth: the destination earth case. Remote Sens. 13, 2119.
Niroumand-Jadidi, M., Bovolo, F., Bruzzone, L., Gege, P., 2020. Physics-based bathymetry and water quality retrieval using planetscope imagery: Impacts of 2020 COVID-19 lockdown and 2019 extreme flood in the Venice Lagoon. Remote Sens. 12, 2381.
Nunes, D.S., Zhang, P., Silva, J.S., 2015. A survey on human-in-the-loop applications towards an internet of all. IEEE Commun. Surv. Tutor. 17, 944–965.
Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth. A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience 51, 933–938.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transact. Syst. Man Cybernet. 9, 62–66.
Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S., 2019. Continual lifelong learning with neural networks: a review. Neural Netw. 113, 54–71.
Pavelsky, T.M., Smith, L.C., 2008. RivWidth: a software tool for the calculation of river widths from remotely sensed imagery. IEEE Geosci. Remote Sens. Lett. 5, 70–73.
Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422.
Piégay, H., Arnaud, F., Belletti, B., Bertrand, M., Bizzi, S., Carbonneau, P., Dufour, S., Liébault, F., Ruiz-Villanueva, V., Slater, L., 2020. Remotely sensed rivers in the anthropocene: State of the art and prospects. Earth Surf. Process. Landf. 45, 157–188.
Planet Labs, 2018. Planet imagery product specifications. Planet Labs, San Francisco, CA, USA, p. 91.
Planet Labs, 2022. Planet To Launch 36 SuperDove Satellites With SpaceX. Available from: https://www.planet.com/pulse/planet-to-launch-36-superdove-satellites-with-spacex/ [Accessed 9th January 2023].
Qayyum, N., Ghuffar, S., Ahmad, H.M., Yousaf, A., Shahid, I., 2020. Glacial lakes mapping using multi satellite PlanetScope imagery and deep learning. ISPRS Int. J. Geo Inf. 9, 560.
Rabaey, J.M., 2020. Human-centric computing. IEEE Transact. Very Large Scale Integr. (VLSI) Syst. 28, 3–11.
Reina, G.A., Panchumarthy, R., Thakur, S.P., Bastidas, A., Bakas, S., 2020. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. Front. Neurosci. 14, 65.
Riggs, R.M., Allen, G.H., David, C.H., Lin, P., Pan, M., Yang, X., Gleason, C., 2021. RODEO: An algorithm and Google Earth Engine application for river discharge retrieval from Landsat. Environ. Model. Softw. 105254.
Rinke, K., Keller, P.S., Kong, X., Borchardt, D., Weitere, M., 2019. Ecosystem services from inland waters and their aquatic ecosystems. Atlas of Ecosystem Services. Springer.
Rosentreter, J.A., Borges, A.V., Deemer, B.R., Holgerson, M.A., Liu, S., Song, C., Melack, J., Raymond, P.A., Duarte, C.M., Allen, G.H., Olefeldt, D., Poulter, B., Battin, T.I., Eyre, B.D., 2021. Half of global methane emissions come from highly variable aquatic ecosystem sources. Nat. Geosci. 14, 225–230.
Rosgen, D.L., 1994. A classification of natural rivers. Catena 22, 169–199.
Sekertekin, A., Cicekli, S.Y., Arslan, N., 2018. Index-based identification of surface water resources using sentinel-2 satellite imagery. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 19-21 Oct. 2018, pp. 1–5.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition 1409 (1556), 1–14.
Smith, K.E., Terrano, J.F., Pitchford, J.L., Archer, M.J., 2021. Coastal wetland shoreline change monitoring: a comparison of shorelines from high-resolution WorldView satellite imagery, aerial imagery, and field surveys. Remote Sens. 13, 3030.
Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. EOS Trans. Am. Geophys. Union 38, 913–920.
Theckedath, D., Sedamkar, R., 2020. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks. SN Comput. Sci. 1, 1–7.
Thoms, M., Sheldon, F., 2019. Large rivers as complex adaptive ecosystems. River Res. Appl. 35, 451–458.
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272.
Wilson, N., Greenberg, J., Jumpasut, A., Collison, A., Weichelt, H., 2017. Absolute Radiometric Calibration of Planet Dove Satellites, Flocks 2p & 2e. Planet, San Francisco, CA, USA.
Wirabumi, P., Kamal, M., Wicaksono, P., 2021. Determining effective water depth for total suspended solids (TSS) mapping using PlanetScope imagery. Int. J. Remote Sens. 42, 5784–5810.
Wu, B., Liu, Z., Yuan, Z., Sun, G., Wu, C., 2017. Reducing overfitting in deep convolutional neural networks using redundancy regularizer. In: International Conference on Artificial Neural Networks. Springer, pp. 49–55.
Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L., 2022. A survey of human-in-the-loop for machine learning. Futur. Gener. Comput. Syst. 135, 364–381.
Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. 27, 3025–3033.
Yang, X., Pavelsky, T.M., Allen, G.H., Donchyts, G., 2019. RivWidthCloud: an automated Google Earth Engine algorithm for river width extraction from remotely sensed imagery. IEEE Geosci. Remote Sens. Lett. 17, 217–221.
Yang, X., Pavelsky, T.M., Allen, G.H., 2020. The past and future of global river ice. Nature 577, 69–73.
Yasir, M., Jianhua, W., Shanwei, L., Sheng, H., Mingming, X., Hossain, M., 2023. Coupling of deep learning and remote sensing: a comprehensive systematic literature review. Int. J. Remote Sens. 44, 157–193.
Yin, Z., Ling, F., Li, X., Cai, X., Chi, H., Li, X., Wang, L., Zhang, Y., Du, Y., 2021. A cascaded spectral-spatial CNN model for super-resolution river mapping with MODIS imagery. IEEE Trans. Geosci. Remote. 60, 1–13.
Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. 169, 114417.
Zheng, Y., Tang, L., Wang, H., 2021. An improved approach for monitoring urban built-up areas by combining NPP-VIIRS nighttime light, NDVI, NDWI, and NDBI. J. Clean. Prod. 129488.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proc. IEEE 109, 43–76.
Ziou, D., Tabbone, S., 1998. Edge detection techniques-an overview. Pattern Recogn. Image Analys. C/C Raspoznav. Obraz. Analiz Izobrazh. 8, 537–559.
Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable Image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710.
Zou, Z., Xiao, X., Dong, J., Qin, Y., Doughty, R.B., Menarguez, M.A., Zhang, G., Wang, J., 2018. Divergent trends of open-surface water body area in the contiguous United States from 1984 to 2016. Proc. Natl. Acad. Sci. 115, 3810–3815.