Key Points:

- Satellite video derived flood extents and velocities successfully validate 2D hydraulic model predictions
- Test-time augmentation during deep learning inference improved flood extent delineation and enhanced 2D model validation metrics
- Incorporating characterization of discharge uncertainty into hydraulic model predictions resulted in more accurate model validation

Author Contributions:

Conceptualization: Christopher Masafu, Richard Williams
Data curation: Christopher Masafu, Richard Williams
Formal analysis: Christopher Masafu
Investigation: Christopher Masafu
Methodology: Christopher Masafu, Richard Williams
Supervision: Richard Williams
Visualization: Christopher Masafu
Writing – original draft: Christopher Masafu
Writing – review & editing: Richard Williams

Satellite Video Remote Sensing for Flood Model Validation

Christopher Masafu[1] and **Richard Williams[1]**

[1]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK

Abstract Satellite-based optical video sensors are poised as the next frontier in remote sensing. Satellite video offers the unique advantage of capturing the transient dynamics of floods with the potential to supply hitherto unavailable data for the assessment of hydraulic models. A prerequisite for the successful application of hydraulic models is their proper calibration and validation. In this investigation, we validate 2D flood model predictions using satellite video-derived flood extents and velocities. Hydraulic simulations of a flood event with a 5-year return period (discharge of 722 m^3 s^{-1}) were conducted using Hydrologic Engineering Center—River Analysis System 2D in the Darling River at Tilpa, Australia. To extract flood extents from satellite video of the studied flood event, we use a hybrid transformer-encoder, convolutional neural network (CNN)-decoder deep neural network. We evaluate the influence of test-time augmentation (TTA)—the application of transformations on test satellite video image ensembles, during deep neural network inference. We employ Large Scale Particle Image Velocimetry (LSPIV) for non-contact-based river surface velocity estimation from sequential satellite video frames. When validating hydraulic model simulations using deep neural network segmented flood extents, critical success index peaked at 94% with an average relative improvement of 9.5% when TTA was implemented. We show that TTA offers significant value in deep neural network-based image segmentation, compensating for aleatoric uncertainties. The correlations between model predictions and LSPIV velocities were reasonable and averaged 0.78. Overall, our investigation demonstrates the potential of optical space-based video sensors for validating flood models and studying flood dynamics.

Plain Language Summary Videos of the Earth surface recorded by satellites can enable us to observe and characterize dynamic moving features, such as floods, that would otherwise be very difficult or dangerous to investigate from the ground. Hydrologists often rely on using physics-based computer models to simulate flood events, but require observational data to make sure these reflect reality accurately. We use artificial intelligence techniques to automatically detect flood extents from satellite video, and track surface features from frame to frame in order to measure how fast the water surface is flowing. Satellite video was collected during opportunistically clear skies in January 2022, along a 6.5 km length of the River Darling in Australia. The flood extent and flow velocities were used to improve numerical model predictions of the flood event. Our findings demonstrate the considerable promise of satellite video to complement existing flood mapping and modeling approaches, and to provide insight into the earth's hydrosphere, particularly in remote locations and during extreme conditions.

1. Introduction

Satellite video presents a significant advance in remote sensing. This novel technology enables the acquisition of high-definition video from space (Xiao et al., 2018), offering the potential for novel hydrological applications and providing unique insights into flood dynamics, flow pattern, and disaster management and recovery. As more of the world's population faces exposure to floods (Rentschler & Salhab, 2020; Tellman et al., 2021) there is an ongoing need to assess how fine-scale earth observation data, such as satellite video, can be used to improve the calibration and validation of two-dimensional flood inundation models. There are, however, inherent tradeoffs between the challenge of acquiring stable video from a non-geostationary platform in sun-synchronous orbit, the duration and pixel size of satellite videos, and the integration and ingestion of satellite video derived data products into existing hydrological modeling frameworks (Ghamisi et al., 2018; McCabe et al., 2017; Shu et al., 2021). To harness the value of these big data, deep learning, a facet of artificial intelligence, offers potential to handle the volumes of data associated with satellite video and to automatically learn and extract flood extent patterns in this data, opening up a plethora of opportunities to progress research in water sciences (Shen, 2018; Sit et al., 2020; Z. Sun et al., 2022). Although traditional machine learning techniques have been widely used to delineate flood extents in traditional satellite imagery (Qi et al., 2020; Su et al., 2020), satellite video scenes

present unique challenges in the form of limited annotated data and complexity of both spatial and temporal information, which calls for advanced architectures suited for understanding the complex spatial and temporal relationships present in dynamic video scenes. Further developments in optical flow monitoring techniques, a facet of computer vision, are enabling the analysis of the motion of objects within a sequence of images or video frames. Specifically, non-intrusive optical flow-based techniques for estimation of velocities in rivers have been demonstrated as viable tool for acquiring spatially distributed flow velocity information in natural environments (Pearce et al., 2020; Perks et al., 2020). Estimating surface river velocities from satellite-based video is a promising domain for flood science, although issues around the low temporal resolution of satellite video, image pre-processing and satellite platform drift demand further investigation.

Two dimensional hydraulic models are an integral tool for understanding flood dynamics. Outputs from these models are actively used for flood risk management (Tsakiris, 2014), infrastructure design (Shrestha et al., 2022), disaster preparedness (Nkwunonwo et al., 2020) as well as in modeling the future of flooding under different climate change scenarios (Mishra et al., 2018), thus aiding in the design of long-term adaptation strategies. Inundation modeling using hydraulic models serves as the principal tool for understanding the intensity of riverine flood hazards. Despite their extensive applicability, two dimensional hydraulic models have rarely undergone rigorous validation against observed data to assess the skill of their predictions (Pasternack, 2011; Wing et al., 2021). In general, 2D modeling studies often use minimal data to assess model accuracy, meaning that models typically target the most basic performance benchmarks, and in many cases, models have simply not been validated (Molinari et al., 2019). Indeed, many regions around the world lack comprehensive and high-quality data of commensurate coverage that can be used to validate these models (Rollason et al., 2018). A handful of studies that exemplify the validation of 2D models include those by Bernhofen et al. (2018), Eilander et al. (2023), and Wing et al. (2017). These studies utilized satellite-derived flood extents to validate global flood models, which offer a macro-level understanding of flood risk and do not capture the fine-scale intricacies of local flood dynamics. Even fewer studies report the validation of two-dimensional hydraulic model simulations using velocities, such as Barker et al. (2018), Fischer et al. (2015), and Williams et al. (2013) who relied upon traditional point-velocity measurements for assessing model skill. In fact, whilst some two-dimensional models might accurately replicate flood extents, substantial deviations of simulated velocity might be observed when compared to field observations (G. Li et al., 2022). Further, there is wide recognition that both two-dimensional models and observations come with their own set of uncertainties such as those linked to extreme discharge measurements, terrain data accuracy and observation field data errors which can introduce discrepancies when validating models (Grimaldi et al., 2016; Schumann, 2017). A systematic investigation of the role of alternative spatially distributed data sets for validating 2D flood models is pertinent.

Remote sensing for flood inundation studies relies on two categories of sensors for monitoring surface water dynamics—microwave and optical sensors (Dasgupta et al., 2018; Grimaldi et al., 2020). Flood water pixel identification from optical satellite imagery has conventionally largely relied on spectral water indices (e.g., the Normalized Difference Water Index (NDWI) (McFeeters, 1996) and the modified NDWI (H. Xu, 2006)) as well as supervised and unsupervised classification. These techniques have been known to misclassify (overestimate) water bodies (Khalid et al., 2021). Machine learning methods, such as Support Vector Machine classifiers and Random Forest algorithms have also been adopted in several floodplain mapping studies (e.g., Mobley et al., 2021; Nandi et al., 2017) with significant contributions to near real-time flood hazard mapping (Ho et al., 2021). Big data analytics and computer vision techniques (specifically, deep learning) are now paving the way for automated delineation of flood extents with high accuracy (J. Wang et al., 2022). Deep convolutional neural networks (CNNs) have revolutionized binary and multi-class image classification and are especially relevant in time-sensitive applications such as flood inundation mapping (Shastry et al., 2023). CNNs overcome several key limitations of traditional machine learning in image classification tasks; they are highly scalable and can process large amounts of complex data with little human intervention. CNNs can also leverage on transfer learning; the use of networks pre-trained on extremely large data sets, then fine-tuned for new tasks, enhancing model generalization, and avoiding overfitting (Tan et al., 2018). Convolutional neural networks (CNNs) have benefited greatly from the rapid development of large labeled data sets, such as ImageNet, which offer high-quality training images at an unprecedented scale (1.3 million training images, 50,000 validation images and 100,000 test images spanning 1,000 classes), allowing generic features learned to be used in complex classifications of presumably disparate data sets (Huh et al., 2016; Ridnik et al., 2021; Yamashita et al., 2018).

Semantic image segmentation using CNN-based networks entails pixel-level identification, classification, and labeling. Applications of deep learning networks for semantic segmentation of floods in remote sensing images

have mostly been demonstrated on fully convolutional networks (FCNs) and to a lesser extent, encoder-decoder architectures. Hashemi-Beni and Gebrehiwot (2021) utilized FCN-8s to generate binary classification maps of flood inundated areas. Gebrehiwot et al. (2019) applied a CNN-based network (FCN-16s) to extract flooded regions from Unmanned Aerial Vehicle (UAV) imagery. Basnyat et al. (2021) utilized a modified version of the U-Net architecture for binary segmentation tasks on their flood detection system while Girisha et al. (2019) successfully utilized both FCN-32s and U-Net for semantic segmentation of UAV videos within an urban zone.

Transformers, a class of neural network architecture originally built to solve sequence to sequence problems in natural language processing (e.g., the transformer based-chatbot ChatGPT [Generative Pre-trained Transformer] (Y. Liu et al., 2023)), have now been adapted as a complement to CNNs for semantic segmentation of remotely sensed imagery attaining state-of-the-art performance (see, e.g., Gu et al., 2022; Z. Xu et al., 2021; Zhang et al., 2022). Transformers, unlike CNNs, rely on "self-attention" mechanisms which allows them to extract and use information from arbitrarily large contexts of the input data (e.g., pixels in an image) simultaneously. It enables the network to capture long-range dependencies and consider global context, making it well-suited for understanding complex patterns and relationships within the data. CNNs on the other hand can only exploit local information correctly due to their small convolutional kernel sizes. However, it's worth noting that due to the quadratic complexity of self-attention, Transformers can be computationally more expensive, especially for large images, and do not generalize well when trained on insufficient amounts of data, as compared to CNNs. Consequently, hybrid architectures that combine the strengths of transformers and CNNs to achieve a balance between global context modeling and computational efficiency for semantic segmentation tasks have been proposed (e.g., A. He et al., 2023; Q. He et al., 2023; Z. Zhou et al., 2023). Although the adoption of these architectures is still evolving, their robustness in understanding complex scenes which exhibit large variations within the same class and subtle differences between different classes, such as in video remote sensing, is yet to be explored.

Traditional methods for measurement of instantaneous water flow velocity, including impellor-type current meters, electromagnetic flow meters, acoustic Doppler velocimeters and acoustic Doppler current profilers (aDcp) remain limited during flood conditions due to logistical challenges such as inaccessibility of flooded areas, flow turbulence as well as limitations in instrument measurement ranges. Image velocimetry, a non-contact method used to measure fluid flow velocities by analyzing images of flow patterns, such as those from video, has gained traction as a method of collecting river velocity data (e.g., Dal Sasso et al., 2021; Pearce et al., 2020). Large Scale Particle Image Velocimetry (LSPIV) is a frequently adopted technique for water-surface velocity analysis and relies on tracking the motion of appropriate artificial or natural "seeding" particles, such as bubbles or debris, between successive images in a time series. LSPIV based on UAV video has been demonstrated in several studies (e.g., C. Chen et al., 2021; Lewis et al., 2018; W.-C. Liu et al., 2021). Legleiter and Kinzel (2021) successfully derived surface river flow velocities from satellite video to within 8.65% of independent radar gage based measurements, building upon the work of Kääb et al. (2019) who utilized near-simultaneous satellite still images (acquired with a nominal time lapse of ~90 s between each other) to estimate river surface velocities in the Yukon River, Alaska. Depending on the video acquisition frame rate, video sequences of between 8 and 30 s are sufficient for image velocimetry over large spatial extents in rapid fashion (Legleiter & Kinzel, 2020; Pearce et al., 2020; Strelnikova et al., 2023). The next step in the domain of satellite video-based image velocimetry is the deployment of these velocity estimates in an operationally useful manner—the validation of hydraulic model predictions.

The aim of this investigation is to present the first attempt at validating flood model simulations using satellite video observations. The principal contributions of this work are threefold. First, we leverage on the temporal richness of satellite video to fine-tune a hybrid CNN-Transformer network for semantic segmentation of flood extents. Specifically, we evaluate the accuracy of six variants of a transformer encoder—CNN decoder architecture in segmenting our test video image. Second, we utilize optical flow techniques to analyze the motion of naturally occurring features between consecutive video frames to estimate the velocity of flow in our study reach. Finally, we utilize these flood extents and velocities to validate 2D flood model outputs, explicitly accounting for uncertainty in our 2D modeling. As non-intrusive optical flow measurement and deep-learning based techniques for delineation of flood pixels in near-real time evolve, and further constellations are established, satellite video presents a potential opportunity to exploit a temporally rich data source, capable of providing data to comprehensively validate flood model predictions. Though limited to cloud-free acquisitions, the unmatched temporal resolution of satellite video can compensate for the limitation of optical atmospheric imaging windows.
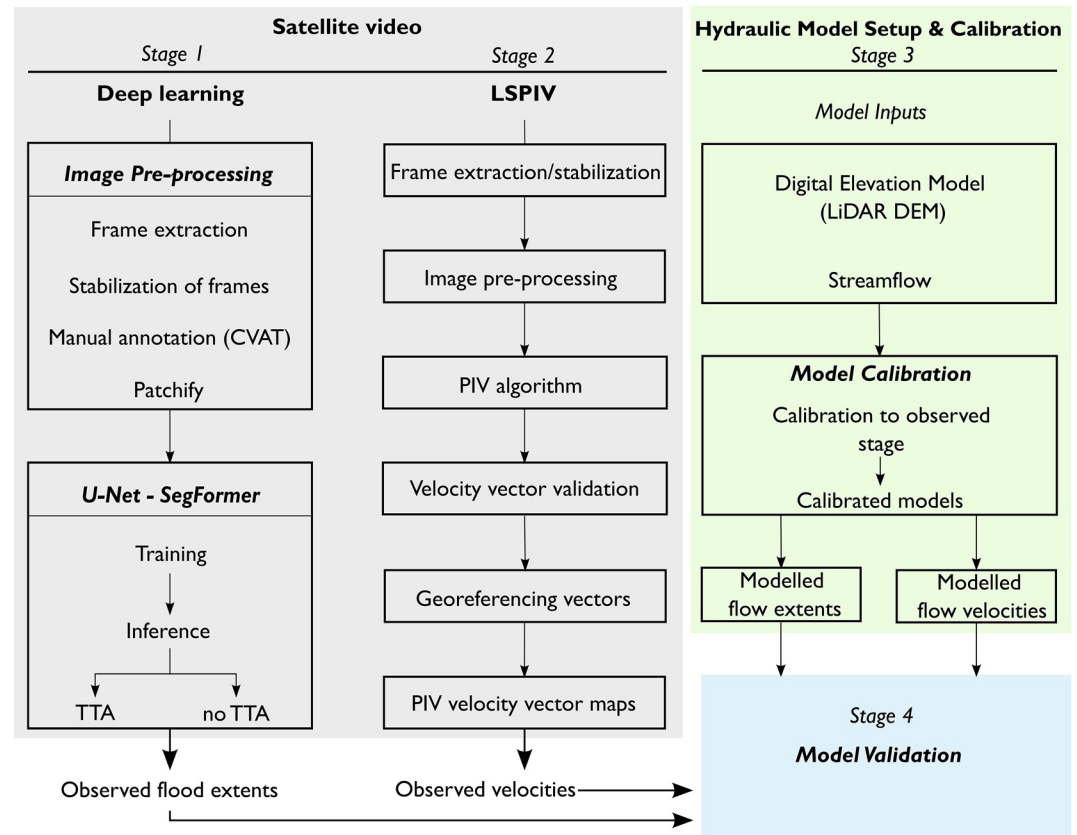
Figure 1. Experimental framework. CVAT, Computer Vision Annotation Tool; LiDAR, Light Detection and Ranging.

2. Data and Methodology

2.1. Overview

The investigation consisted of four stages (Figure 1). In stage 1, Jilin-1 satellite video-based flood extents are derived using a state-of-the-art deep-learning network. We adopt a hybrid architecture consisting of a transformer-based encoder and a CNN-based decoder, this allows us to integrate the strengths of both transformers and CNNs. Whilst CNNs excel at local feature extraction, transformers are stronger at capturing global context and long-range dependencies. This hybrid approach for semantic segmentation allows us to benefit from the strengths of each architecture. Specifically, we adopt SegFormer (Xie et al., 2021) as the encoder, a lightweight yet efficient and powerful model that has attained state-of-the-art semantic segmentation performance on popular benchmarks like the Cityscapes data set and ADE20K (Cordts et al., 2016; B. Zhou et al., 2017). We utilize a U-Net decoder which allows the model to leverage multi-level information, reconstruct spatial details, and efficiently integrate low-level and high-level features to generate accurate and fine-grained segmentation masks. We evaluate the segmentation capabilities of six different variants of the SegFormer encoder, from SegFormer B0–B5 and narrow down on one model series. We train the selected deep learning model, relying on transfer learning and Test-Time Augmentation (TTA) for improved flooded class prediction accuracy. TTA involves the creation of augmented copies of a test data set, following which the deep-learning network returns ensemble predictions of class labels which are averaged to produce segmentation maps.

In Stage 2, we utilize LSPIV to compute river surface velocity vectors over two cloud-free subsets of our video (Figure 2). Stage 3 involves simulation of a recent flood event using a hydraulic model (Hydrologic Engineering Center—River Analysis System [HEC-RAS] version 6.0) driven by uncertain discharge estimates derived from Bayesian analysis of stage-discharge relations (see Section 2.3.2). This HEC-RAS 2D model is calibrated using stage height from the gauge at Tilpa (Figure 2). Stage 4 validates the HEC-RAS 2D model predictions using both the deep-learning segmented images from Stage 1, and the velocity vectors estimated in Stage 2.
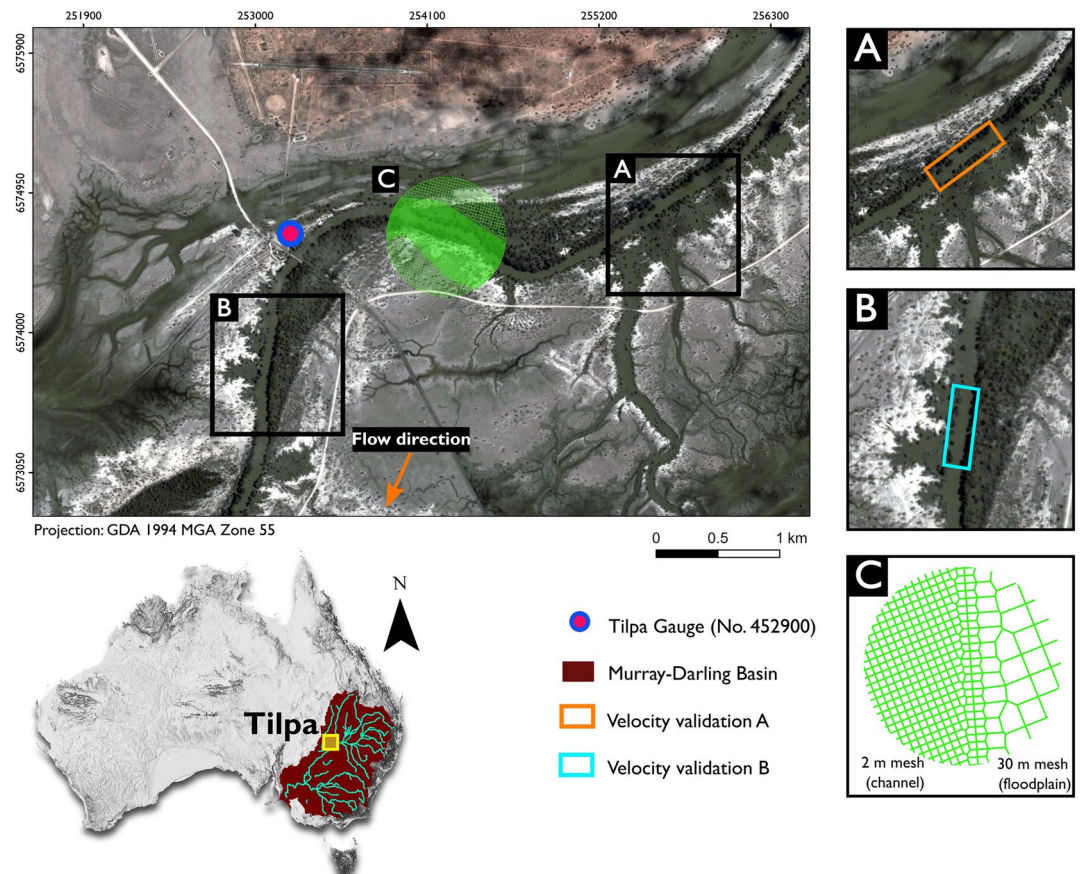
Figure 2. The Darling River at Tilpa study area, located in the Murray-Darling basin (shaded inset), New South Wales, Australia, with a basemap of Jilin-1 satellite video acquired at 23:12 UTC on 5 February 2022. Panels A and B indicate flood model validation locations. Panel C presents the hybrid mesh used in hydraulic modeling.

2.2. Study Area

The study area covers a 6.5-km-long section of the middle Darling River at Tilpa floodplain, located in north-western New South Wales (NSW), which is part of the Murray-Darling River basin in south-eastern Australia (Figure 2). A gauging station at Tilpa (Station number 425900) records water course discharge every 15 min from 1995 to present, for an upstream catchment area of 502,500 km^2. Extreme multi-day rainfall, caused by a series of deep low-pressure systems, resulted in intense storms and major flooding in eastern NSW from 22 February to 9 March 2022. Between 19 January and 11 February 2022, the Darling River at Tilpa was above major flood level (11.5 m), with the flood wave peaking at 12.3 m on 31 January 2022 (Bureau of Meteorology, 2022). The town of Tilpa experienced extensive floodplain inundation.

The Darling River at Tilpa has a river style (Brierley et al., 2002) that is meandering, is planform controlled and features discontinuous floodplain. Creeks are connected to the Tilpa's main channel and fill with water when the Tilpa's stage is high. This complex floodplain configuration demands high-resolution 2D flood inundation modeling. We calibrated our HEC-RAS 2D model using stage height observations from the gauging station at Tilpa (Figure 2) and validated our model simulations at two locations (A and B, Figure 2), geographical extents that adequately represent the complexity of the floodplain and enabled us to capture fine scale flood hydraulics at a resolution sufficient to comprehensively assess inundation dynamics. Additionally, using these two locations, rather than the whole study area, overcomes the potential problem of jagged prediction patches, which are artifacts associated with the reconstruction of large sized deep learning-based prediction mosaics. This is a limitation when making flood pixel predictions using neural networks which generally results in data loss around the border of large image patches (Heller et al., 2018; Yuan et al., 2021).

Table 1
Jilin-1 GF03C02 Satellite Sensor Specifications and Video Product Information

| Attribute	Details
Product components and format	Video file—AVI
	Individual video frames—TIFF Images
Image mode	Gaze imaging
	Conventional and multi-target pushbroom imaging/space target imaging
Spectral bands	RGB (Bayer pattern)
	Red B1: 580–730 nm
	Green B2: 490–580 nm
	Blue B3: 430–520 nm
Scene size	14.4×6 km
Orbit	Type: Sun-synchronous
	Altitude: 535 km
Spatial resolution (ground sampling distance)	1.2 m
Video duration	28 s

2.3. Data

2.3.1. Satellite Video

At present, two commercial companies operate constellations with satellite-based optical video sensors. Planet's SkySat offer satellite video products in panchromatic mode at a Ground Sampling Distance of 1.1 m, a frame rate of 30 frames per second (fps) with a duration between 30 and 120 s (Bhushan et al., 2021). Jilin-1 GF-03, launched by the Chang Guang Satellite Technology Company, is another high-resolution commercial satellite video constellation. Jilin-1 collects 0.5–1.2 m 4K high definition color video at frame rates of up to 10 fps and revisit capability of up to six times a day globally (Y. Chen et al., 2022; European Space Agency, 2022). Currently, both platforms are limited to 30–120 s duration video captures. For our study, Satellite video (Table 1) was acquired on 5 February 2022 at 23:12 UTC by Jilin-1 GF-03. The video has a spatial resolution of 1.22 m and was acquired at five frames per second for a duration of 28 s, yielding 140 frames, with each frame measuring $12,000 \times 5,000$ pixels.

2.3.2. Rated Discharge

2.3.2.1. Rating Curve Uncertainty Estimation

The influence of rating curve uncertainties on streamflow time series estimates is particularly pronounced in natural river systems, especially during floods, when a rating curve is typically extrapolated beyond the maximum gauging in the rating, resulting in significant systematic errors (Horner et al., 2018). The propagation of errors from flood model forcing data (i.e., streamflow) to eventual model outputs warrants explicit consideration of uncertainties associated with the rating curve. Kiang et al. (2018) investigated different techniques for estimation of stage-discharge rating curve uncertainty concluding that the choice of methods is fully dependent upon the constraints of the specific application. Bayesian inference has however been suggested as a robust technique to handle independent gauging errors and provide precise discharge series uncertainty envelopes (Ocio et al., 2017) and was adopted here.

Rating curve uncertainty in this investigation is assessed using the BaRatin (Bayesian Rating Curve) method (Le Coz et al., 2014) which combines uncertain gaugings and prior hydraulic knowledge to derive uncertain stage-discharge relations. The BaRatin framework defines stage and discharge measurement uncertainties as Gaussian distributions with a mean of zero and is composed of three main components; (a) a measurement error model, consisting of prior estimates of parameters based on preliminary hydraulic analysis of a gauging station, (b) Posterior rating curves, which are a derived from a simulation consisting of gauging data, and (c) The application of Markov Chain Monte Carlo and Bayesian inference to sample the posterior distribution of the rating curve
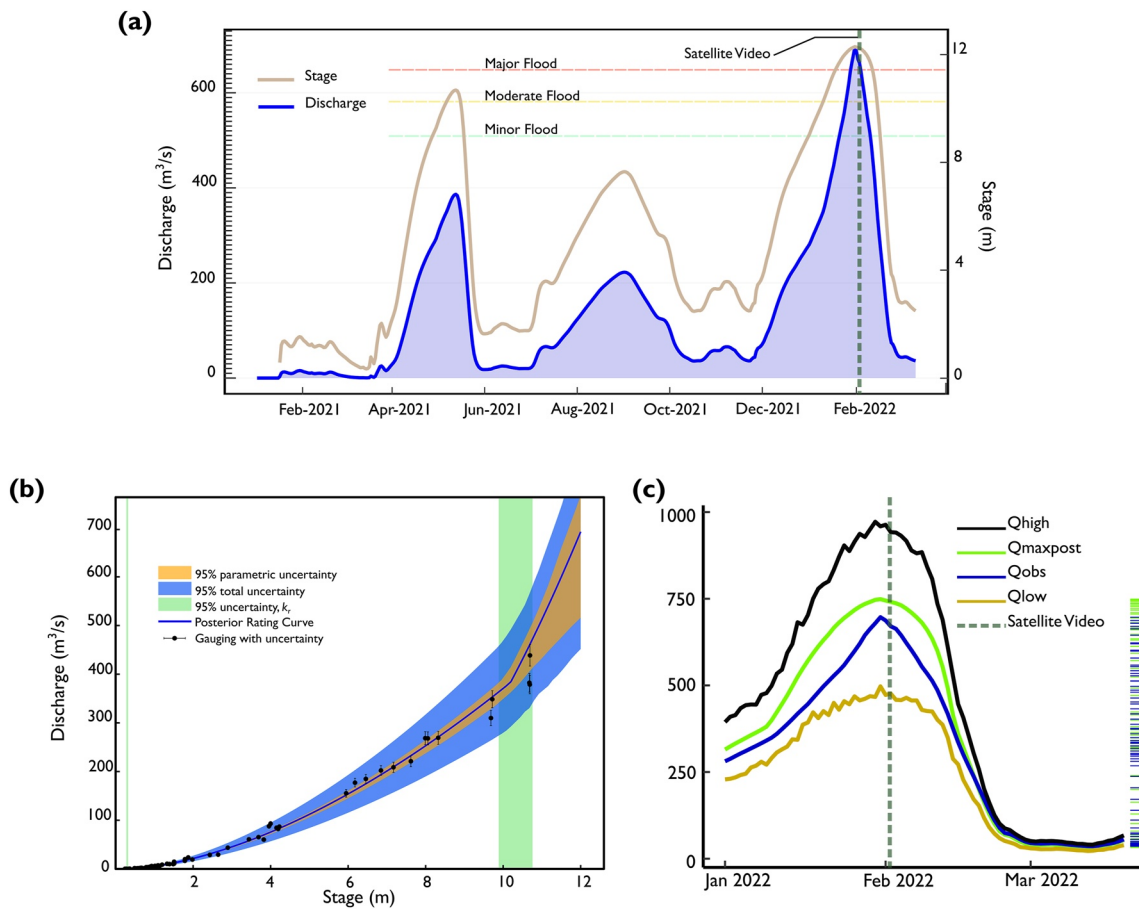
Figure 3. (a) Stage-discharge relationship for the Darling River at Tilpa gauging station from February 2021–2022. (b) Posterior rating curves and associated uncertainties derived using the BaRatin method. (c) Discharge time series. Q_{obs} is the observed discharge. Q_{high} and Q_{low} are discharges with associated stage (non-systematic and systematic stage measurement errors) and rating curve (parametric errors and structural/remnant errors) uncertainty. $Q_{maxpost}$ is the upper 95% confidence band of a streamflow timeseries based on an uncertainty analysis of the rating curve. The rug plot to the right depicts the distribution of Q_{obs} and $Q_{maxpost}$ discharges.

parameters relying on information contained in observed gaugings. The eventual rating curve equation is based on a matrix of hydraulic controls that relates discharge Q to stage h using power functions:

$$Q(h) = \sum_{r=1}^{N_{segment}} \left(1_{[k_{r-1}; k_r](h) \times} \sum_{j=1}^{N_{control}} M(r,j) \times a_j (h - b_j)^{c_j} \right) \qquad (1)$$

In the above equation, $M(r,j)$ is the matrix of controls, and the notation $1_I(h)$ denotes a function equal to 1 if h is included in the interval I, and 0 otherwise. Segments in the rating curve ($N_{segment}$) are user defined while segment limits k_r, coefficients a_j and exponents c_j are inferred.

For our flood model simulations, we forced our hydraulic model using streamflow time series (Figure 3) based on two scenarios: (a) Q_{obs}, observed discharged drawn directly from the Tilpa gauging station; and (b) $Q_{maxpost}$, discharge computed from measured stage and the MaxPost rating curve (a rating curve that maximizes the posterior distribution of a set of parameters inferred from the Bayes theorem (see Kiang et al., 2018 for detail)). The $Q_{maxpost}$ streamflow timeseries is based on uncertain rating curves (with 95% confidence bounds), computed by Bayesian analysis of prior hydraulic controls and gagings with individual uncertainties.

2.4. Deep Learning Model

2.4.1. Transformer-Based Encoder

Transformer based models were initially designed for Natural Language Processing tasks and excelled over CNNs and Recurrent Neural Network models (e.g., Long-Short Term Memory that process sequence elements recursively and can only attend to short-term context) thanks to their "self-attention mechanism." In the context of semantic segmentation, the self-attention mechanism helps the model understand the relationships between different spatial locations (pixels) in an image. By using self-attention, Transformers can capture long-range dependencies and understand global context, which is essential for accurate semantic segmentation, where the label of a pixel may depend on distant regions in the image. Although the field of deep learning is evolving rapidly, the most successful transformer architectures adapted for semantic segmentation tasks include Vision Transformer (ViT) (Dosovitskiy et al., 2021), Swin Transformer (Z. Liu et al., 2021) and SegFormer (Xie et al., 2021). Here we leverage on SegFormer, which uses a hierarchical Transformer architecture (called "Mix Transformer") as its encoder and a lightweight decoder for segmentation. SegFormer's encoder leverages tokenization, self-attention, and hierarchical aggregation to efficiently capture important visual information from input images, making it well-suited for semantic segmentation tasks. Here, we fine-tune six variants of SegFormer (B0–B5) with increasing model sizes offering improved performance at the cost of increased computational requirements.

2.4.2. Convolutional Neural Network (CNN)-Based Decoder

Convolutional neural networks (CNNs) are a widely used architecture in deep learning, initially proposed by Fukushima (1980) and refined by Lecun and Bengio (1995). CNNs, a class of artificial neural networks, are mostly defined as a series of layers, with the initial layers performing feature learning and the final layers performing classification. CNNs consist of three types of layers: (a) convolutional layers, which are the first layers to extract features from an input image whose outputs (feature maps) are then passed on to sequential layers; (b) pooling layers, which take feature maps as inputs and progressively reduce the spatial size of the feature maps, controlling model overfit; and (c) an activation function applied to the outputs of the CNN that enable the model to capture non-linear behavior in the input data (Hosseiny, 2021).

Whilst a plethora of deep-learning CNNs exist, the most prominent architectures that have attained state-of-the art performance in semantic segmentation of remote sensing images are U-Net (Ronneberger et al., 2015), Google's DeepLab (L.-C. Chen et al., 2016) and PSPNet (Zhao et al., 2017). In a comparative study of semantic segmentation of remote sensing images using the three aforementioned models, J. Hu et al. (2019) reported U-Net achieving the best accuracy. A more recent investigation by Y. Sun et al. (2022), who similarly carried out an intercomparison of the three CNN networks, found U-Net outperforming the other three models in the segmentation of remotely sensed images.

Originally developed for biomedical image segmentation, U-Net (and its variants) has gained prominence in diverse fields for its ability to leverage on data augmentation to efficiently learn from a small number of annotated images. U-Net's fairly simple architecture, consisting of a downsampling (encoder) and upsampling (decoder) path allows for precise localization of segmented pixels. The U-Net decoder is a crucial component that contributes to the strength of the U-Net architecture in semantic segmentation tasks. The decoder is responsible for upsampling the low-resolution feature maps from the encoder to the original input image resolution while also fusing multi-level features to create final segmentation maps. Thanks to the U-Net decoder's low memory requirements as well as its ability to be trained end-to-end (meaning both the encoder and decoder are learned jointly during the training process) we deployed a U-Net decoder for our segmentation tasks.

2.4.3. Neural Network Training

Training deep neural networks typically requires a vast amount of data, which may not always be available. Here, rather than train our model from scratch, we fine-tuned our encoder to improve its segmentation capabilities on our data set. Fine tuning entails taking a pre-trained model's encoder (in our case, SegFormer models B0–B6) and further training it on a specific data set, the objective being to adapt the learned features to perform well on a new task. For instance several studies have utilized pre-trained models, such as VGG16 or ResNets, which have learned to recognize general image features from ImageNet, then fine-tuned the models on much smaller data sets for specific image segmentation tasks (see, e.g., Hashemi-Beni & Gebrehiwot, 2020; Tong et al., 2020).

We leverage on the temporal richness of video to extract images used for fine-tuning our model. Although our video was acquired at a native frame rate of 5 Hz (resulting in 140 images in sequence), we subsample our video at a much lower frame rate of 1 Hz by retaining only every tenth frame from the original 5 Hz series resulting in a sequence consisting of 28 images. Skipping frames enabled us to decouple temporal information in our images. When frames are skipped, the information captured in one frame is temporally further apart from the information in the skipped frames. As a result, the sequential order of frames is disrupted, and the temporal information becomes less tightly coupled, reducing the strength of autocorrelation. Also, because there was minimal movement or low temporal variation in consecutive satellite video scenes, by skipping frames, we omit redundant or near-duplicate information, which leads to a reduction in autocorrelation since repetitive patterns are less prominent. To ensure that all our extracted frames were aligned and free from any motion artifacts as a result of the satellite platform's vibrations, we aligned our images in TrakeEM2 (Cardona et al., 2012) using the Scale-Invariant Feature Transform algorithm (see detail in Text S2 in Supporting Information S1). Stabilized frames were exported to Intel's open-source image annotation tool Computer Vision Annotation Tool and were manually labeled. Labeled masks were then converted to binary format in Python (with background = 0, flood = 1). These annotated masks served as the ground truth for training and validation of our deep neural network.

At the training stage, the full-sized satellite images could not be loaded on to the network due to memory limitations, a common challenge faced when training deep learning models which require extensive memory to store input images, weight parameters and activations as images are channeled through the network. Therefore, the original images were split into patches of $256 \times 256 \times 3$. To artificially diversify and increase the size of our training data set, we deployed data augmentation. Data augmentation is a technique that reduces generalization errors (model overfitting) by adding a range of deformations and noise to the training data. To implement data augmentation, we used the Albumentations library (Buslaev et al., 2020) in Python, where we applied vertical and horizontal flips, transposition, grid distortions, elastic transforms and random gamma to both the images and corresponding binary masks (Figure 4). A total of 5,660, 1,698, and 566 image patches ($256 \times 256 \times 3$) were used for training, validation and testing the network.

To initiate training, pre-trained SegFormer weights were imported and hyper-parameters set as follows: we used a batch size of eight and a learning rate of 1×10^{-4} with a call back to reduce the learning rate by a factor of 2×10^{-4} when the validation metric (Intersection-Over-Union/Jaccard Index) stopped improving. The deep learning model was implemented within the PyTorch deep-learning framework and trained on an NVIDIA GeForce RTX 2070 Super Graphical Processing Unit for 150 epochs. The trained network learned to associate images and masks and make an independent prediction on a test image which was independent of the 28 frames used in the original training/validation/testing pipeline.

2.4.4. Neural Network Evaluation Metrics

Model segmentation performance was assessed using two metrics, Binary Cross Entropy loss (L_{BCE}) and the Intersection Over Union (IoU; also known as the Jaccard Index).

L_{BCE}, compares model predicted probabilities to the ground truth labels (see Section 2.3.1), which can either be 0 or 1. It then computes a score that penalizes the probabilities based on the distance from the expected values. L_{BCE} calculates the difference between the actual and precited probability distributions for predicting class 1 (Jadon, 2020). The score is minimized, and a perfect value is zero:

$$L_{BCE} = -\frac{1}{n} \sum_{i=1}^{n} \left(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log\left(1 - \hat{Y}_i\right) \right) \tag{2}$$

where Y denotes the ground truth label while $\hat{Y}$ is the precited probability of the classifier.

IoU is a simple ratio that compares how close a pixel is between the training sample and predicted regions (Bebis et al., 2016):

$$IoU = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

where A and B denote the ground truth and the prediction segmentation maps respectively. IoU ranges between 0 and 1. If the model prediction is perfect, IoU = 1. The lower the IoU, the worse the predicted result.
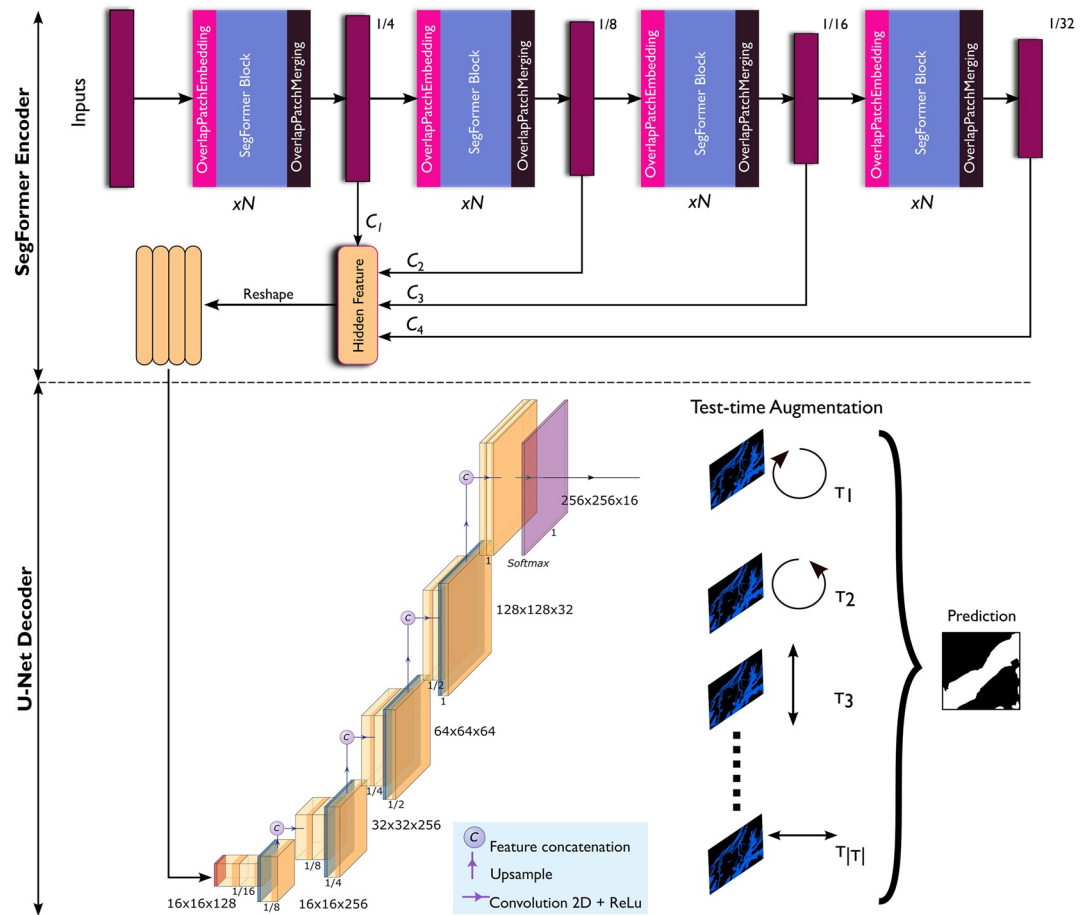
Figure 4. End-to-end deep learning network training, validation and testing pipeline.

Since our binary segmentation task involved a data set with an inherent class imbalance, we utilize L_{BCE} and IOU as our loss functions during model training. Although we report other metrics including recall, precision and the F1 score (see Text S1 in Supporting Information S1, for a description of these metrics) we do not use them as loss functions during training as these metrics would strongly bias our results to the class that occupies a large portion of our images (flood imagery has disproportionately few pixels per image being identified as flooded).

2.4.5. Test-time Augmentation (TTA)

Although data augmentation has been widely used when training deep learning networks, a less common way of improving semantic segmentation prediction accuracy is TTA (see, e.g., Gonzalo-Martín et al., 2021; J. Liu et al., 2022; Moshkov et al., 2020; Wieland et al., 2023). TTA involves applying augmentation transformations on the test image to create an ensemble of predictions, which are then averaged, to improve prediction results. Assuming x to be our input image and τ a transformation operation, choosing $T = \{\tau_1, \tau_2, \ldots, \tau_{|T|}\}$ as a set of augmentations at model inference time, we can formulate TTA as:

$$y_{\text{tta}} = \frac{1}{|T|} \sum_{i=1}^{|T|} \Theta_{\text{target}}(\tau_1(x)) \tag{4}$$

where Θ_{target} is the neural network trained on our target data set (satellite video frames).

In our application, we supplied the network with a full-sized satellite video scene, which is patched into $256 \times 256 \times 3$ sized patches without overlap. The network then made predictions on multiple transformations of the image patches, creating an ensemble of predictions. The transformations were then reverted, a process known as dis-augmentation, following which predictions were averaged, thresholded then stitched back to the original full-size image using the Patchify library in Python.

Image parsing which assigns semantic class labels, in this case flooded versus non-flooded pixels, for our trained deep learning network returned probability maps, which were thresholded to binarize final pixel classes. Assuming a base (true) segmentation probability map is returned by the network during testing, binary segmentation is done by thresholding as follows (Hong et al., 2021):

$$P = \delta(p(y)), \tau \tag{5}$$

where $P \in \{0,1\}^{WxH}$ is a predicted segmentation label and $p(y)$ is the true probability map of the potential segmentation labels $y \in \{0,1\}^{WxH}$ for an image with height H, and width W. δ is the thresholding function with a threshold τ. We set a fixed τ value of 0.5 for our model training and testing.

2.4.6. LPSIV Velocity Estimation

Whilst a variety of image velocimetry algorithms have been deployed for river surface velocity estimation (see Perks et al., 2020 for a summary), we rely on the frequently used PIVLab algorithm (Thielicke & Sonntag, 2021; Thielicke & Stamhuis, 2014) for our LSPIV analysis. Our general LSPIV workflow is depicted in Figure 1. For ease of processing, we crop our original video, focusing on the reaches A and B (Figure 2). We sub-sample our videos to lower frame rates of 1, 0.5 and 0.25 Hz by retaining only the fifth, tenth, and twentieth frames respectively from the original 5 Hz series. By lowering the frame rate, the time interval between consecutive images is increased. Resultantly, tracked features move farther between frames, leading to larger displacements which translate to larger feature sizes in the PIV analysis, making it easier for the PIV algorithm to accurately track the features. We then stabilize all our images in TrakEM2 to counter any residual motion effects from the satellite platform.

We import and process our images in PIVLab. To optimize the quality of our velocimetry results, we pre-process the images by applying contrast limited adaptive histogram equalization (Pizer, 1990; Yadav et al., 2014), which enhances the contrast and improves the visibility of details in an image. We also applied an intensity high-pass filter which removed low-frequency background noise and enhanced the visibility of flow features in the images, making it easier for the PIV algorithm to accurately detect and track them (Thielicke & Stamhuis, 2014). A crucial parameter input to the PIVLab algorithm is the interrogation area (IA) usually measured in pixels. Whilst a large IA may improve the accuracy of velocity measurements, it can also lead to a loss of spatial resolution. Conversely, a small IA may provide higher spatial resolution but could also be sensitive to noise or errors in particle tracking. A step-size, which is usually 50% of the IA, determines the spacing of PIV output vectors. Here, we utilized an IA of 128 pixels with reducing sizes over 4 passes (i.e., 128–64–32–8 pixels). For our PIV analysis, we utilized the FFT window deformation (direct Fourier transform correlation with multiple passes and deforming windows) cross-correlation algorithm (for details, see Astarita, 2008; Thielicke & Sonntag, 2021; Thielicke & Stamhuis, 2014). We postprocessed our computed velocities to remove spurious velocity vectors due to poor particle tracking, image artifacts, or other issues. Our post-processing specifically entailed setting a standard deviation filter which was used to filter out noisy or erroneous velocity vectors from the calculated velocity field based on their standard deviation values (set to PIVLab's default value of 8 in our study). Velocity vectors were georeferenced within PIVlab from an image coordinate system back into a projected coordinate reference system (GDA 1994 MGA Zone 55) and exported to ArcGISPro for analysis.

2.5. Flood Model

Two dimensional (2D) flood flow modeling was accomplished using HEC-RAS version 6.0, building on previous flood modeling studies using this code (e.g., Mokhtar et al., 2018; Navarro-Hernández et al., 2023; Pradhan et al., 2022). The model domain was 4.4×7.7 km. Topography was defined from a 1 m resolution LiDAR-based bare-earth digital surface model acquired when the river channel was dry (Geoscience Australia, 2022). A heterogenous 2D computational mesh was generated using a cell size of 30×30 m within the floodplain and 2×2 m between channel banks (Figure 2). Model upstream inflow boundary conditions were set using data from the gauge located at Tilpa, with 15-min interval gauged streamflow being used. We simulated two scenarios with different upstream boundary forcing: model M_{obs}, based on observed discharge data; and model $M_{qmaxpost}$ using uncertain streamflow estimates (see Section 2.3.2.1). An energy slope value was used as the outflow boundary condition; the gradient, equivalent to the normal depth, was estimated by computing the bed slope along the terrain profile. Unsteady 2D flow simulations were conducted from 00:00 4

Table 2
Confusion Matrix of Cell Descriptors in Binary Classification of Flood Masks

	Wet in observed data	Dry in observed data
Wet in model simulation	A_1	B_2
Wet in model simulation	C_3	D_4

February to 23:59 6 February 2022 using HEC-RAS's diffusive wave equations since the flood wave was not highly dynamic (Brunner et al., 2020; Yalcin, 2020).

Model calibration data were available in the form of stage height from the gauge at Tilpa (Figure 2). 2D model runs (see Section 2.3.2.1) were calibrated by adjusting a spatially discretized manning's roughness coefficient n over a parameter space between 0.025 and 0.033, with 0.002 increments until model simulations closely matched observed stage height data. Modeled stage height accuracy was assessed using two commonly used metrics (e.g., Alghafli et al., 2023; Moriasi et al., 2007), the percent bias (PBIAS) and root-mean-square error (RMSE).

2.5.1. Analysis: Flood Model Validation Using Observed Extents

We validated models M_{obs} and $M_{qmaxpost}$ against flood masks derived from the observed satellite video frames. HEC-RAS modeled, and satellite video-derived flood extents were converted into binary masks (wet/dry) representing only flood extent, then overlapped, using ArcGISPro software. To preserve the detail of the model output's higher resolution, specifically close to the floodplain where model performance matters most due to the occurrence of complex inundation dynamics, and because comparison of validation performance needed to be done at the same spatial resolution, binary masks derived from the satellite video frames were resampled using the nearest neighbor method to 1 m. During resampling, binary pixel interpolation did not yield new additional values, avoiding false accuracy errors. Since we were resampling a binary raster with only one of two values (0 or 1), each new pixel in the target 1 m resolution raster was assigned the value of the nearest corresponding pixel, meaning extents of the binary raster remained unchanged because no new information was introduced. The overlap between the modeled inundation boundaries and the observed satellite video, were calculated according to the number of pixels that showed model agreement, overprediction and underprediction, as per the states in Table 2. Data from these computations were then used to calculate performance scores.

To analyze model skill in reproducing the observed flood inundation extent, we rely on well-established spatial performance measures which account for the most critical attributes of model simulation precision: model bias (E), hit rate (HR)—proportion of the observed flood event that was successfully predicted by the model, the false alarm ratio (F) and a rigorous composite measure (Critical Success Index [CSI]) which penalized model overpredictions (Bates et al., 2021; Bernhofen et al., 2018; Wing et al., 2017).

Error bias (E) which indicates whether the model is biased toward overprediction or underprediction:

$$\text{Bias} = \frac{F_m \cap F_o + F_m}{F_m \cap F_o + F_o} - 1 \tag{6}$$

where F_m is the total modeled flood extent. A Bias score of 0 indicates an unbiased model while positive and negative scores indicate a tendency toward overprediction and underprediction respectively.

The HR which measures proportion of observed flood that was simulated by the model, ignoring whether the observed flood extents were exceeded. HR can range from 1 (entire flood captured) to 0:

$$\text{HR} = \frac{F_m \cap F_o}{F_o} \tag{7}$$

where F_o is the total observed flood extent. The HR ranges from 1 (entire flood captured) to 0.

The false alarm ratio (F), a measure of whether the model has a tendency to overpredict flood extent and can range from 0 (no false alarms) to 1 (all false alarms):

$$\text{FAR} = \frac{\text{False alarm area}_{area}}{\text{False alarm}_{area} + \text{hit}_{area}} \tag{8}$$

The CSI, a measure of the accuracy of a flood model, accounts for both overprediction and underprediction and can range from 0 (no match between modeled and observed data) to 1 (perfect match between modeled and

Table 3
Segmentation Accuracy for Increasing Sizes of SegFormer Encoders (B0–B5) Coupled With a U-Net Decoder

	L_{BCE}	IoU	F1 score	Recall	Precision
SegFormer-BO/U-Net	0.2522	0.8178	0.8789	0.8826	0.9170
SegFormer-B1/U-Net	0.2597	0.8192	0.8848	0.8959	0.9138
SegFormer-B2/U-Net	0.3515	0.7383	0.8171	0.8396	0.8253
SegFormer-B3/U-Net	0.3218	0.8186	0.8791	0.8801	**0.9212**
SegFormer-B4/U-Net	0.3456	0.8196	0.8819	0.8889	0.9161
SegFormer-B5/U-Net	**0.2518**	**0.8276**	**0.8878**	**0.8972**	0.9181

Note. Bold font highlights optimal values.

observed data). CSI omits regions that are dry in both predictions and observations since the flood model can accurately forecast these:

$$\text{CSI} = \frac{F_m \cap F_o}{F_m \cup F_o} \quad (9)$$

where $F_m \cap F_o$ is the intersection of the modeled and observed flood extent, or number of correct predictions, and $F_m \cup F_o$ is the union of modeled and observed extent. The CSI ranges from 1 (best) to 0 (worst).

2.5.2. Analysis: Flood Model Validation Using Observed Velocities

Whereas the HEC-RAS 2D model is developed to estimate depth-averaged water velocities, satellite video-based LSPIV computations yielded surface river velocities. For a like to like comparison, we depth-average our LSPIV velocities. A well-established technique for converting surface velocities into depth averaged velocity is by utilizing published values of depth-averaging constants, also referred to as alpha coefficients, α (Biggs et al., 2021; Creutin et al., 2003). Rantz (1982) proposed α values of between ~0.85 and 0.86 for natural channels. Hauet et al. (2018) recommend α values of 0.8 for water depths of less than 2 m and 0.9 for greater water depths with an uncertainty of ±15% at 90% confidence level within natural rivers. Vigoureux et al. (2022) suggested α values of between 0.85 and 1.2 for their LSPIV analysis and experimented with values between 0.8 and 1.0 for their depth averaging constants. To quantify the impact of choice of α coefficient on our estimated velocities, we utilize α coefficients of between 0.7 and 1.

We compare HEC-RAS model predictions with LSPIV velocities via linear regression where we fit a linear trend line and use R^2 and slope as indicators of model performance. The coefficient of determination, R^2, value between HEC-RAS 2D model velocities and LSPIV remains unaffected when the LSPIV velocities are adjusted by varied α values. However, other statistical indicators are sensitive to the choice of α coefficient, including the mean velocity difference between model and LSPIV-based velocity (with an optimum difference of 0 m s^{-1}), which we use to evaluate the accuracy of our validation.
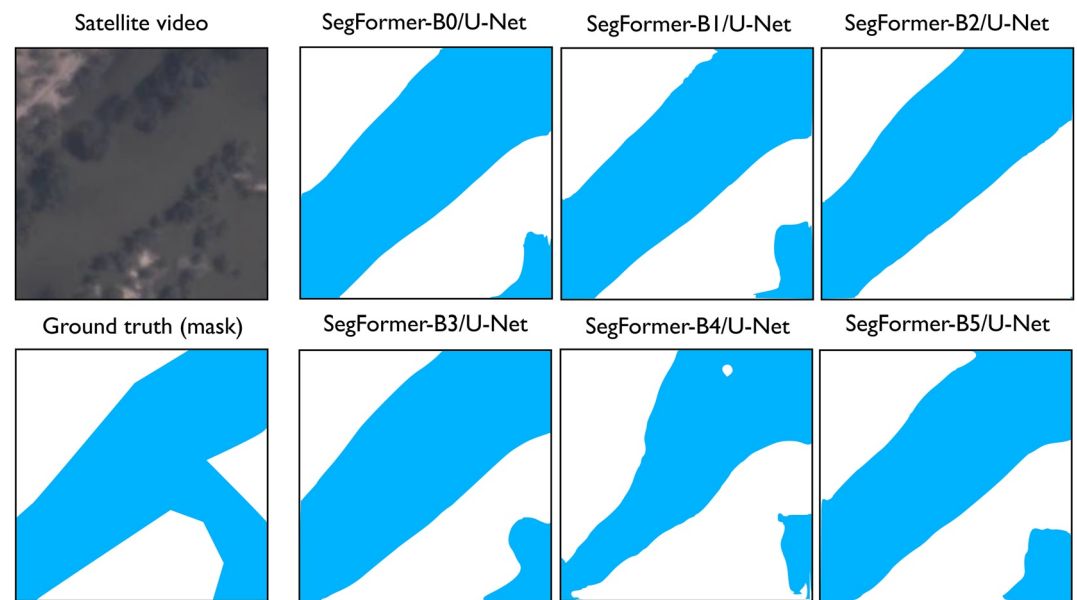
3. Results

3.1. Segmentation Accuracy

We evaluated the segmentation performance of six variants of our model based on SegFormer encoders (B0–B5) of increasing sizes, all pre-trained on ImageNet, to select an architecture that robustly segments flood pixels from a satellite video test image set. We retained the same hyperparameters of learning rate, batch size and number of epochs for all models during the testing phase to attain a fair comparison between the different backbones.

Quantitative results of model performance are detailed in Table 3. Overall, we find segmentation performance of all models to be satisfactory (IoU > 0.7). In general the performance of the deep-learning networks was on par with values reported in literature where hybrid CNN/transformer networks have been deployed for semantic segmentation (e.g., T. Wang et al., 2022). Unsurprisingly, we found the SegFormer-B5/U-Net network attained the most accurate scores across all metrics, except for the precision metric where the SegFormer-B3/U-Net network returned a better score of 0.9212. This can be explained by the higher number of parameters which leads to better segmentation accuracy, but at the cost of additional computational demand. Visual test results are presented in Figure 5 and further support the quantitative scores from model training and testing.

3.2. Flood Model Calibration

The 2D hydraulic model was calibrated to observed stage at the Tilpa gauging station by varying Manning's n roughness. In our investigation, both models were considered to be adequately calibrated with RMSE (0.46, 0.45 m) and PBIAS (0.77, 0.16) for models M_{obs} and $M_{qmaxpost}$ respectively (Table 4), comparable to ranges reported in other similar studies (Timbadiya et al., 2011; Zeiger & Hubbart, 2021). Median error (Figure 6c) increased marginally for model M_{obs} with both models showing a strong correlation between modeled and observed stages (R^2 of 0.72 and 0.71) for models M_{obs} and $M_{qmaxpost}$ respectively (Figure 6).

Figure 5. Results of the semantic segmentation of a sample satellite video patch using varied combinations of SegFormer encoders coupled with a U-Net decoder. The annotated mask serves as the ground truth with water pixels labeled blue, background pixels labeled white.

3.3. Flood Extent Validation Using Deep Learning-Based Flood Extents

Results from validation of both models at locations A and B (Figure 2) using satellite-video derived flood extents are detailed in Table 5. The predictive accuracy of both models was tested by choice of uncertainty associated with model forcing data (streamflow) and the influence of TTA. We report both results in detail in the following sections.

3.3.1. Model Comparison Scores

We assess the performance of both models at validation locations A then B, followed by a summary of the key findings from both reaches.

At validation location A, the predictive skill of both model M_{obs} and $M_{qmaxpost}$ was nearly similar, with CSIs ranging between 0.873 and 0.91. For both models, CSI declined when the models were validated against lower-quality observations, without implementation of TTA. When the model was forced using uncertain streamflow data, CSI averaged 0.918 as compared to 0.891 when using observed streamflow, a 3% improvement in performance. Average model bias declined marginally (by 3.3%) when forcing our model with uncertain streamflow. At validation location B, we observed lower values of CSI, ranging between 0.700 and 0.862, representing, on average, a 16% decline in model accuracy. Similar to validation location A, model performance scores improved across the board when TTA was applied. We however observe a consistent drop in model HR when TTA is applied at both validation location A and B.

For both models at both validation locations, underestimation of the flood extent is concentrated in the downstream end of the study reach (validation location B) where, on average, the HR was 85.5% as compared to 96.7% in validation location A. To contextualize these results, the skill of both models is in line with previous studies (e.g., Ongdas et al., 2020 using Sentinel-1B) who reported HR scores of between 59% and 77%. The models forced with uncertain streamflow data yielded superior results for all assessment metrics except for HR which declined whenever we applied TTA. Validation bias overall ranged between 0.808 and 1.152, with our scores being in line with those from other comparable validation studies in literature ranging from −1 to 5 (e.g., Wing et al., 2017, 2021 using a combination of HEC-RAS model outputs and Special Flood Hazard Maps from the United States Geological Service).

Table 4
2D Model Calibration Metrics for Models M_{obs} and $M_{qmaxpost}$

Manning's n	Model M_{obs}		Model $M_{qmaxpost}$	
	RMSE, m	PBIAS	RMSE, m	PBIAS
0.025	1.30	−8.08	0.81	−4.65
0.027	1.11	−6.64	0.53	−0.32
0.029	0.72	−1.94	0.56	2.29
0.031	0.46	0.77	0.45	0.16
0.033	0.51	2.48	0.61	2.20

Note. Model predictions are calibrated against observed stages over a time step of 15 min on 5 February 2022 (04:00–12:30).
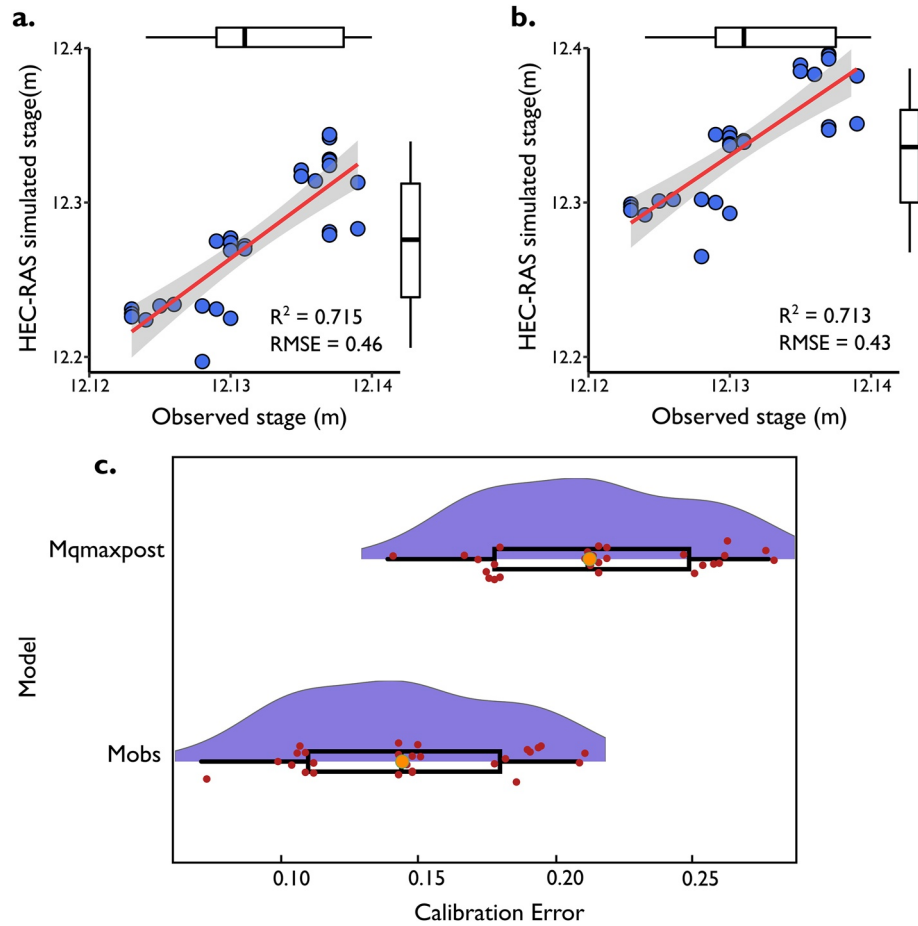
Figure 6. Hydrologic Engineering Center—River Analysis System (HEC-RAS) calibration results. Observed water surface elevations are compared against calibrated model M_{obs} (a) and $M_{qmaxpost}$ (b) predictions over a time step of 15 min on 5 February 2022 (04:00–12:30). Coefficient of determination (R^2) and root-mean-square error represent the results of regression analysis of the data. (c) Raincloud plot, a boxplot with a half-side violin plot, showing the error distributions for both models. In the boxplots, the orange dot represents the median score, the box encompasses the second and third quartiles, and the top and bottom whiskers respectively represent the largest and smallest values within 1.5 times the interquartile range.

Table 5
Validation Metrics for Models M_{obs} and $M_{qmaxpost}$ Against Observed Satellite Video Data at Validation Locations A and B (See Figure 2)

Validation area	Model	Error bias (E)	Critical Success Index (CSI)	False alarm ratio (F)	Hit rate (HR)
A	M_{obs}	1.085	0.910	0.085	0.993
A	M_{obs} TTA	1.050	0.941	0.053	0.994
A	$M_{qmaxpost}$	1.020	0.873	0.077	0.941
A	$M_{qmaxpost}$ TTA	0.986	0.896	0.049	0.938
B	M_{obs}	1.072	0.700	0.204	0.853
B	M_{obs} TTA	0.808	0.803	0.003	0.805
B	$M_{qmaxpost}$	1.152	0.716	0.221	0.898
B	$M_{qmaxpost}$ TTA	0.868	0.862	0.004	0.865

Note. M_{obs}, Model forced using observed streamflow; $M_{qmaxpost}$, Model forced using uncertain streamflow; TTA, Test-time Augmentation applied to validation images.

3.3.2. Test-Time Augmentation (TTA) and Model Performance Scores

Figure 7 shows an overview of the influence of TTA in the binarization of flood pixels in the observed data. Results differed substantially when TTA was applied during validation of both models M_{obs} and $M_{qmaxpost}$ at both locations. At both reaches, CSI scores ranged from 0.7 to 0.91, and improved to between 0.803 and 0.941 when TTA was implemented. Overall, these CSI results are in line with other flood model validation studies which range from 0.1 to 0.9 (e.g., Mester et al., 2021; Wing et al., 2017).

HR scores generally followed a negative trend when TTA was implemented as compared to the other performance measures. Since this metric only considers the proportion of wet observed pixels, ignoring whether observed flood extents were exceeded, increasing the flood pixel class had an inverse impact on the flood capture. For both models at both validation locations, average HR scores decreased by 0.1% (at location A) and 4.8% (at location B). With TTA implemented, both models Models M_{obs} and $M_{qmaxpost}$ reached a
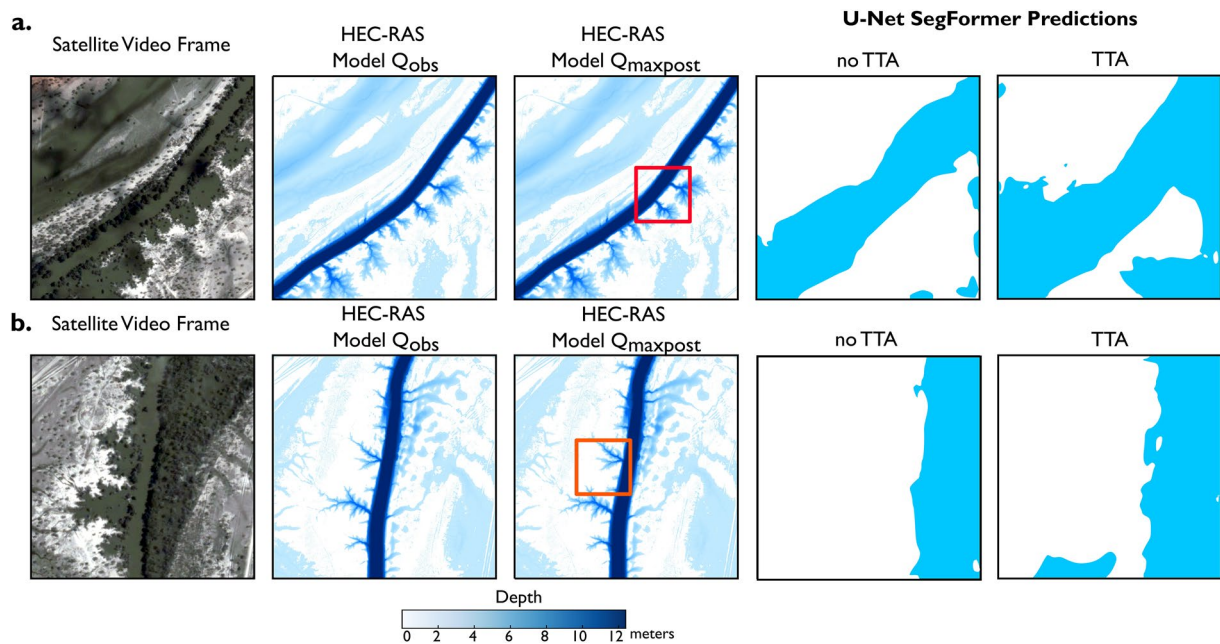
Figure 7. Semantic segmentation results at reach A (a) and B (b): The first panel to the top left shows the observed satellite video flood image followed by Hydrologic Engineering Center—River Analysis System (HEC-RAS) model flood outputs from Models M_{obs}, $M_{qmaxpost}$ and corresponding binary segmentation maps from zoomed-in insets with no Test-Time Augmentation (TTA) then TTA applied.

nearly similar skill level across all metrics, indicating the importance of TTA in our segmentation strategy, which resulted in better agreement between model predicted and observed flood extents.

3.4. Flood Velocity Validation Using LSPIV

Figure 8 shows the comparison of satellite-derived velocity estimates against 2D model predictions, depth averaged using varying α coefficients (see Section 2.5.2). Gard (2008) provide criterion for assessing whether 2D model predictions are validated or not based on the correlation coefficient between measured and simulated velocities, with a correlation of 0.6–0.8 being moderately strong and 0.8 to 1 being very strong. Further, Ballard et al. (2010) proposed that an R-value of 0.6 ($R^2 = 0.36$) constitutes a validated 2D model, with Pasternack (2011) recommending R^2 of between 0.4 and 0.8 between observed versus 2D model predicted velocities. Correlations between both model predictions and LSPIV velocities were reasonable and averaged 0.78. Water velocities predicted by both model variants at both reaches were generally within the observed variability of the LSPIV data.

We find that accounting for discharge uncertainty had minimal influence on velocity validation metrics. A key influence on the accuracy of 2D model velocity predictions, however, was choice of α coefficient. We report consistently low values of mean absolute error as α approaches 1, with an average value of 0.12 m s^{-1}. We partly attribute these results to the use of high-quality topography for our 2D modeling. These results affirm the findings of previous work, such as that by Lane et al. (1999) who similarly found that topographic specification has a larger role to play in constraining model velocity predictive ability as compared to inflow data.

4. Discussion

4.1. Flood Extent Segmentation Using a Hybrid Transformer/CNN Network

Satellite video, unlike traditional still imagery, is composed of both static and motional context (temporal information), with static context being the scenes that remain relatively still or unchanged, with minimal or no movement over a period of time and motional context encompassing dynamic elements. Both static and motion-based contexts share a significant correlation, with image semantic segmentation particularly benefiting from the static context (e.g., Hashemi-Beni & Gebrehiwot, 2020; Leach et al., 2022) and the latter in video semantic segmentation (P. Hu et al., 2020; Y. Li et al., 2018). Although research into means for simultaneously
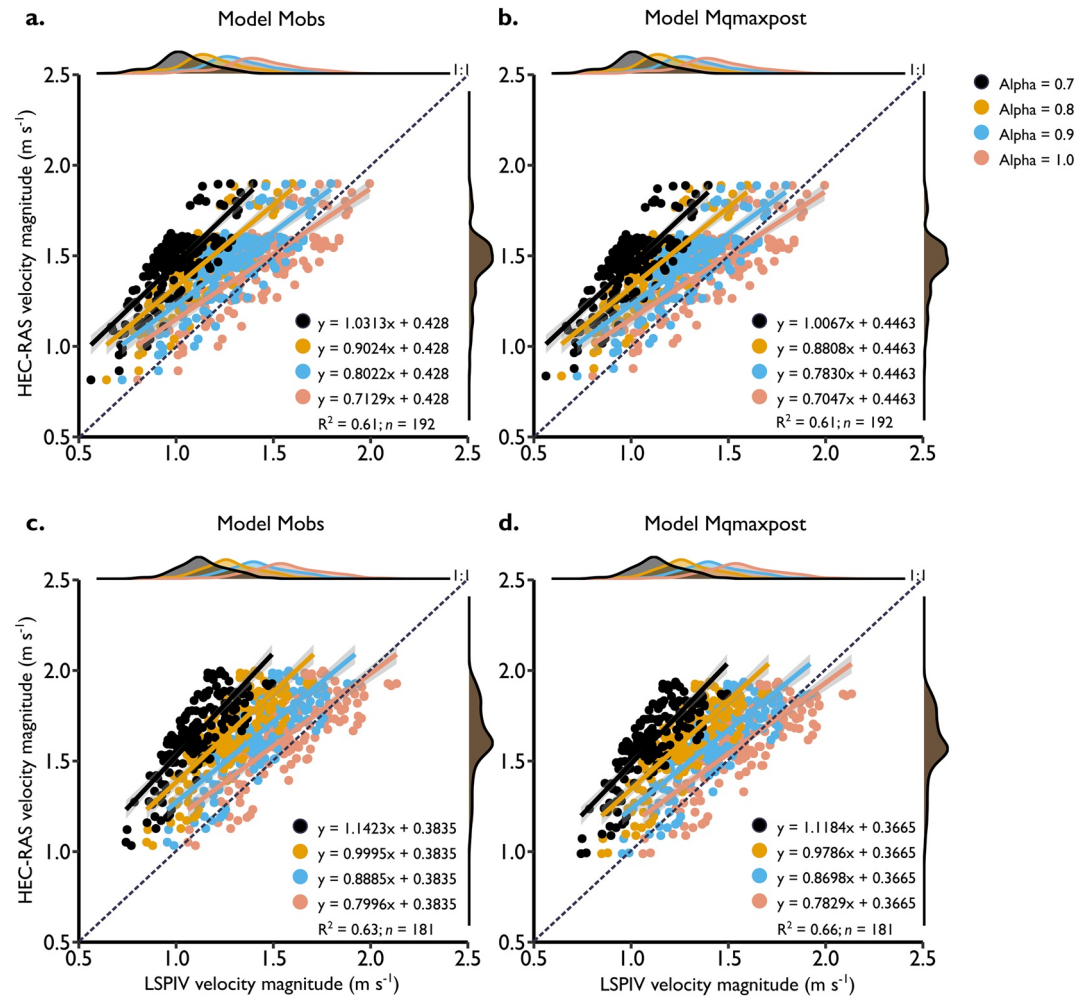
Figure 8. Satellite video-based Large Scale Particle Image Velocimetry (LSPIV) velocities versus Hydrologic Engineering Center—River Analysis System (HEC-RAS) 2D model simulations for Models M_{obs} and $M_{qmaxpost}$ at reach A (a, b) and reach B (c, d) respectively.

exploiting both the static and motional contexts has been attempted (G. Sun et al., 2022), we did not pursue this line of inquiry for the reasons we outline forthwith.

While video segmentation can be valuable for analyzing dynamic changes during a flood event and for various research purposes, image segmentation provides a focused and practical approach for flood model validation tasks, where the primary goal is to assess model accuracy based on specific flood extents at critical time points. Hydraulic flood model predictions are typically produced as snapshots in time, representing flood extents at specific time points. By using a single image for validation, our validation process aligns naturally with the instantaneous nature of the 2D hydraulic model outputs. Additionally, state-of-the-art deep learning models, such as 3D convolutional networks or spatio-temporal transformers, can be computationally intensive for video semantic segmentation and may require specialized hardware to achieve real-time performance. Moreover, many current deep learning models process video frames independently, leading to limited temporal consistency in segmentation results. We therefore opted to exploit image semantic segmentation.

In evaluating the performance of our hybrid transformer-encoder, CNN-decoder structure, we evaluated segmentation performance with increasing SegFormer encoder depths (i.e., SegFormer B0–B5). The training and inference performance of our six models were largely comparable (Table 3), with the SegFormerB5 encoder yielding the best performance, thanks to it having the greatest number of trainable parameters (81M as compared to 60, 44, 24, 13 and 3M for the models B0 to B4). It is essential to clarify that our intention was

not to directly compare a standalone U-Net with a transformer-based model. Instead, we aimed to emphasize the performance of our hybrid architecture, which leverages the transformer's capabilities as an encoder and the U-Net as a decoder. Although it has been shown that transformers make strong encoders for semantic segmentation tasks (L. Wang et al., 2022), they come with the cost of increased computational requirements. For our decoder, we find the original (vanilla) U-Net architecture still allowed us to obtain acceptable performance. In fact, while several variants of the U-Net architecture have been proposed (e.g., Attention U-Net, Inception U-Net, U-Net++) to improve segmentation performance, studies have found that the marginal segmentation performance gains of these new architectures may not be worthwhile as they come at significantly increased complexity and computational demand (e.g., Kugelman et al., 2022). Our results demonstrate that a vanilla U-Net decoder, combined with a transformer backbone can still provide comparable and competitive segmentation results. This finding has significance as the operationalization of these segmentation techniques will have greater impact when the complexity of model structures is low and accuracy high, which would encourage adoption in flood mapping studies. We note that although other hybrid structures composed of transformer-based encoder and CNN-based decoder architectures exist (e.g., TransUNet based on the ViT (J. Chen et al., 2021) and DC-Swin (L. Wang et al., 2022) based on the Swin transformer (Z. Liu et al., 2021)), it would be unmanageable to test them in this investigation. We speculate that any performance gains would be marginal at best.

Image segmentation performed by deep neural networks exploits two powerful features; (a) the capacity of deep learning architectures to overcome data scarcity by leveraging on previously trained networks and, (b) data augmentation techniques, which are used to increase the diversity and size of the training data set by applying various transformations to the original images which improves the model's generalization performance by reducing overfitting and capturing more representative features of the underlying data distribution. A recent study by Wieland et al. (2023) attributed improvements to their segmentation performance as a result of data augmentation, which helped introduce variations, as would be expected in real-world scenarios, that were otherwise not present in their training data set. Gonçalves et al. (2023) and Wieland et al. (2023) also found that adapting pre-trained models (trained on non-remote sensing specific databases such as ImageNet) which are fine-tuned on a custom data set, like in our case, led to superior model performance as opposed to initializing a model with random weights. Although domain-specific data sets of flood images with labeled masks exist, such as Sen1Floods11 based on Sentinel 1 and 2 (Bonafilia et al., 2020), SEN12-FLOOD based on a set of multimodal Synthetic Aperture Radar (SAR) and multispectral satellite imagery (Rambour et al., 2020), WorldFloods based on Senitnel-2 images (Mateo-Garcia et al., 2021) and FloodNet (Rahnemoonfar et al., 2021) training a good deep-learning model on data from disparate sensors remains a complex task due to issues including the fact that data from different sensors may have significant domain shifts, meaning that statistical characteristics (e.g., color distribution, resolution, lighting conditions) of the data can vary widely. Deep learning models are sensitive to such variations, and when faced with data from disparate sensors, they may struggle to generalize well to unseen sensor data. Further, data from disparate sensors may have inconsistencies in labeling or annotation standards with ground truth labels not being directly comparable between sensors, leading to challenges in creating a consistent and accurate training data set.

The uncertainty of inundation data used to validate our models has the potential to introduce observational bias. That said, our bias scores were all positive and small, indicating that satellite video derived flood extents either largely matched or were slightly overestimated by model predictions. Satellite-based observations, both optical and radar, face well documented limitations including cloud cover as well as uncertainties associated with the timing of sensor overpass relative to the advancement of the flood wave. Incorporation of some of these observational uncertainties (such as timing of the sensor overpass) would clearly offset some of the error related to our validation. Previous studies have relied on machine learning models (e.g., Tanim et al., 2022), thresholding techniques (e.g., Tiampo et al., 2022) and histogram-based models (e.g., Singh & Kansal, 2022) to derive flood extents for model validation. Departing from these well-established methods, our trained neural network extracted flood extents from red, green, and blue (RGB) images in a topographically complex natural river floodplain. Our results demonstrate that satellite video RGB imagery, which is essentially a stack of images rich in temporal dynamics rather than a snapshot in time, can attain acceptable accuracies for flood model validation and can be used in place of multispectral/hyperspectral imagery. Further, as satellite video remote sensing is emerging as a key technology in earth observation, there is great scope for advancing the

synergy between deep learning algorithms and high spatial/temporal resolution RGB data in operational flood hazard mapping studies.

4.2. Effects of TTA on Segmentation Performance

Although implementing TTA introduced additional computational overhead during inference, since we had to process multiple augmented versions of each test image, the performance gains offered outweighed the computational cost. Similar to the findings of J. Liu et al. (2022) who utilized TTA during semantic segmentation of their remote sensing images, we found that the skill of both our models improved. In general, both models M_{obs} and $M_{qmaxpost}$ were highly sensitive to TTA with significant disparity between model skill at both validation locations reported (Table 5). CSI, model bias and F scores all improved (10.3%, 67.7%, and 2.4%), while the flood capture/HR decreased, indicating an increase in model overprediction as the segmented flood pixel class reduced. The sensitivity of model CSI performance to TTA was heighted more in validation location B where flood extent was less constrained by topography leading to more favorable F scores (mean of 0.004) as compared to location A (mean of 0.051). It is further manifest from our results that where model performance matters most, in zones of complex channels with localized spill, our HR, CSI and F scores remained high, demonstrating satisfactory flood extent prediction accuracy by our hydraulic model. Overall, we find that TTA helped our model to generalize better (with IoU >0.8) and reduced overfitting to our test image set.

4.3. 2D Model Validation Using Satellite Video Derived Flood Extents

The performance of our 2D hydrodynamic models in reproducing observed flood extent varied significantly, highlighting a key result—that discharge data uncertainty is a fundamental driver of out-of bank flood processes and plays a crucial role in model evaluation. It is well established in literature that an assessment of the uncertainty of input boundary data is cardinal to any model validation exercise (Grimaldi et al., 2016; Hoch & Trigg, 2019). Streamflow data from stage-discharge relations has been reported to have errors as much as 20% or more during extreme floods (McMillan et al., 2012), further underlining the importance of quantifying discharge data uncertainty when deriving inundation maps from hydraulic models (Bales & Wagner, 2009). We observed that the input boundary condition (inflow hydrograph) had a strong effect on the modeled flood extent at both the lower and upper reaches of our study area where flood capture varied by an average of 5% when using uncertain discharge estimates to drive our model, which also explained the variance in other validation scores. Although inundation extent by itself may not be sufficient for the assessment of model skill, and is primarily effective in flat extensive floodplains (Hesselink et al., 2003), our results account for uncertainty typically ignored in most inundation modeling studies. At a minimum, we propose that an appropriate application of 2D model simulations for generation of flood inundation maps must account for discharge uncertainty, especially in highly variable terrain, if reliable extent maps are to be expected.

Although most of the differences between the skill of our models could be attributed to streamflow uncertainty, there remains some uncertainty linked to the satellite observation data, likely linked to the timing of image acquisition. Previous research (Horritt, 2006) found that satellite inundation data may have added uncertainty if the overpass does not coincide with the advancement of the flood wave, which was the case in our investigation as the satellite video was acquired during the receding phase of the flood event when the floodplain was gradually dewatering. Nevertheless, an improvement in identifying flooded pixels correctly when using uncertain streamflow estimates shows that our analysis adopted a reasonable compromise with regards to total uncertainties and that the presence of uncertainty in the observed satellite video did not significantly impact our findings. These conclusions further advance the evidence rendered by Grimaldi et al. (2016) who similarly singled out the uncertainties of upstream boundary conditions as the prime propagator of error when calibrating and validating hydraulic models. Therefore, evaluating the skill of models remains a challenging task due to the residual uncertainty in validation data coupled with the inconsistencies inherent in evaluation metrics, whose reliability varies depending on flood magnitude.

4.4. 2D Model Validation Using Satellite Video Derived Velocities

When validating 2D model predictions using satellite video LSPIV velocities, the moderate R^2 values (ranging between 0.61 and 0.63) indicated that a substantial portion of the variance in the observed LSPIV velocities was

explained by the model predictions. However, the fact that there was little improvement in model performance even when accounting for discharge uncertainty underscores the complexities involved in accurately capturing every aspect of fluid dynamics during flood events. Although our 2D model predictions performed on par with results from scientific literature (Barker et al., 2018; Pasternack, 2011), the uncertainty associated with the choice of a depth-averaging constant remains. Complexities surrounding spatial variations of flow velocities, channel geometry, bed roughness and the transient nature of river flows means that published depth-averaging constants must be used judiciously. Similar to other studies where LSPIV velocities were depth-averaged using α coefficients (e.g., Le Coz et al., 2010; Masafu et al., 2022; Vigoureux et al., 2022), we find that α value of between 0.9 and 1 to be more realistic. We acknowledge that the gold standard for defining an α coefficient is to use aDcp observations but field-based river velocity measurements are logistically challenging and often impossible during high flow. Here we aim to build upon previous studies that have estimated river surface velocities (e.g., Legleiter & Kinzel, 2021) and further show that an independent satellite-video data set can be used to validate 2D model simulations.

4.5. Limitations in Applying Satellite Video for Flood Model Validation

Our study evidenced the utility of satellite video data sets in assessing reach scale 2D hydraulic model simulations. The likely future proliferation of nanosatellite constellations will lead to more opportunities to acquire satellite video. However, satellite video data sets come with limitations that can affect their applicability and accuracy in analyzing river dynamics. While video data sets capture dynamic changes over the time of recording, their spatial resolution is still limited compared to dedicated high-resolution still imagery. Moreover, limited spectral bands in satellite video means that some information relevant to river dynamics is not available, especially when compared to multispectral still imagery and SAR data. Near-infrared (NIR) bands are particularly beneficial in providing crucial information for delineating flood extents. SAR imagery, which operates in the microwave region of the electromagnetic spectrum and can capture data under various atmospheric conditions and at different times (night and day) make it highly reliable during stormy or cloudy conditions when optical sensors are ineffective. Similar to other commercial satellite data, satellite video is not yet open access and requires tasking for image acquisition. Further, the current satellite video imaging catalog is very limited compared to other publicly available still image satellite data.

Our use of temporally and spatially autocorrelated data from the same event for both training and testing can lead to overestimation of the deep neural network's performance. This is because the network was not challenged with data that significantly differed from what it had already encountered. This narrow data set scope limited our neural network's exposure to diverse flood characteristics like varying water levels, different terrain types, or urban versus rural flood dynamics. In urban settings, complexity increases due to diverse features like buildings, roads, and varied land use. Our hybrid deep neural network fine-tuned with rural catchment data, might not perform as well in urban landscapes without additional training on urban-specific features. To address the potential overestimation of our network's performance due to the use of autocorrelated data, we fine-tuned a deep neural network that had been pre-trained on ImageNet, a large and diverse data set encompassing a wide range of scenes and objects. This preliminary training provided our model with a foundational understanding of varied features and textures, which is beneficial for initial feature detection in flood scenarios.

Despite this initial step, the use of more diverse and distinct flood scenes for training and testing remains crucial. While the pre-training on ImageNet partially compensated for the lack of diversity in our flood-specific data set, it cannot entirely substitute for the direct input of varied flood scenes. Future research could focus on incorporating data sets that include multiple flood events from different geographical locations and times. This will significantly improve the ability of deep neural networks to generalize and perform accurately across various real-world flooding scenarios. Additionally, the challenge of determining optimum threshold values in semantic segmentation was amplified by our data set's limited diversity. Addressing this limitation by testing and optimizing deep neural networks across a variety of flood events would aid in identifying more universally applicable threshold values. The development of more adaptable neural network models, capable of adjusting threshold parameters based on the characteristics of each unique flood event, remains a key area for future improvement. Nonetheless, our approach is pioneering in its use of satellite video for flood analysis, providing a valuable proof of concept and a baseline for future models that could be trained and validated on more varied data sets.

4.6. Satellite Video for Flood Risk Science: Current Status and Future Perspectives

High resolution full-color earth observation video presents a fundamental paradigm shift away from the conventional periodic snapshots offered by most current satellite platforms with imagery sensors. Although full-motion satellite video offers previously unavailable temporal insights for monitoring flood dynamics, this data will not necessarily replace conventional earth observation missions but rather they will augment current sensors to foster further advances in hydrological process understanding. Moving from single image analysis to processing video streams presents a unique challenge due to the sheer size of satellite video. However, parallel developments in data analytics such as machine and deep-learning techniques are making processing of such big data sets less computationally intensive. Satellite video sensors suffer from the key constraint of cloud cover which impairs their ability to provide quality imagery during potentially critical storms which are typically accompanied by cloudiness. However, as demonstrated here, for large catchments where a period of clear skies follows a storm and the passage of a flood wave has relatively long duration, daylight acquisition satellite video is feasible. Although spatial and temporal-based cloud removal methods in optical satellite imagery have been proposed (e.g., Huang et al., 2015; Z. Li et al., 2019), their application in satellite video scenes is yet to be explored and changes in cloud obscuration are likely to be minimal during a single video sequence.

Nanosatellite constellations, such as CubeSats, which are highly modular and inexpensive imagers, are rapidly allowing real-time delivery of RGB video with very high-frequency revisits, up to 15 min for some constellations (Ivliev et al., 2022; Liddle et al., 2020; Lomaka et al., 2022; Marinan et al., 2013). As satellite video becomes increasingly available to commission, new opportunities for applications such as direct estimation of discharge from space are now in the horizon. Legleiter and Kinzel (2021) used satellite video to estimate flow velocities of the Tanana River directly from space. The production of digital surface elevation models using satellite video and Structure-from-Motion (SfM) photogrammetry type techniques was proposed by d'Angelo et al. (2014, 2016). Combining space-derived velocities with quality digital elevation and bathymetry data presents the potential for estimating discharge from flash floods and ungauged catchments where in-situ gauges might be too expensive to install and operate. The UK's Center for Ecology and Hydrology FluViSat study ("FluViSat—Hydrological Flow Measurements from Satellite Video, n.d.," https://www.ceh.ac.uk/our-science/projects/Fluvisat, last access: 16 August 2023) currently underway, is another application that will aim to demonstrate computation of surface velocities and river discharges using satellite video. Nanosatellite constellations, which are any satellite with a mass between 1 and 10 kg including Cubesats and SunSats, are widely believed to the future of low Earth orbit observations. These satellites are cheap to launch and provide high revisit capabilities as compared to other larger satellite missions. Such disruptive earth observation technologies, combined with rapidly advancing big data analytics will transform satellite video applications in flood hydrology.

5. Conclusion

Using satellite video-derived flood extents and velocities, we were able to assess the skill of a two-dimensional hydraulic model to prediction a flood event. Two sets of simulations were undertaken, with one set focused on simulating the flood event whilst accounting for discharge uncertainty while the other set did not account for discharge uncertainty. Flood extents were derived from satellite video scenes using a hybrid transformer-encoder CNN-decoder deep neural network. Leveraging on the transformer's self-attention mechanism and CNN's effectiveness at local spatial feature detection, we attained robust predictions of flood pixels. Implementation of TTA while delineating flood extents resulted in further improvements in segmentation performance. Validation of both models using flood extents showed good model performance. Models were also validated using satellite based LSPIV; to the best of our knowledge, this study presents the first attempt of this nature to use satellite derived LSPIV velocities for validating 2D model predictions. Although there remains uncertainty with regards to choice of a depth-averaging constant (α), our linear regression models still showed that model-predicted and LSPIV velocities had a significant statistical relationship. This demonstrates the notable benefits of non-contact-based velocity estimation from space, especially during high flow conditions when the use of traditional river velocity measurement techniques is precluded. Although discrepancies between model and satellite-derived flood extents can be linked to other uncertainties, such as those associated with the satellite data itself, topography, and the model, this study emphasized the importance of accounting for uncertainty in discharge; a dominant yet often neglected source of uncertainty which is especially heightened during flood events. We find that accounting for discharge uncertainty had less impact on metrics associated

with velocity-based validation than those associated with inundation extent. This is likely due to the fact that model velocity estimates are subject to greater spatial variability from factors such as hydraulic parameters, roughness coefficients and channel geometry. The wider implication of this study is the demonstration that high resolution satellite video has significant potential as a source of temporally rich data for the validation of velocity and extent predictions. With rapid advances in remote sensing sensor technologies and constellations, satellite video is likely to become more straightforward and cheaper to commission. In many regions of the world, where ground-based hydrological observations are not routine, this will open the door for wide-scale observations of flow velocities and extents in real-time, enabling further progress in the science of flood modeling.

Data Availability Statement

Data used in this study are available at Masafu and Williams (2023).

References

Alghafli, K., Ali, A. M., Shi, X., Sloan, W., Obeid, A. A. A., & Shamsudduha, M. (2023). Evaluation of runoff estimation from GRACE coupled with different meteorological gridded products over the Upper Blue Nile Basin. *Journal of Hydrology: Regional Studies*, *50*, 101545. https://doi.org/10.1016/j.ejrh.2023.101545

Astarita, T. (2008). Analysis of velocity interpolation schemes for image deformation methods in PIV. *Experiments in Fluids*, *45*(2), 257–266. https://doi.org/10.1007/s00348-008-0475-7

Bales, J. D., & Wagner, C. R. (2009). Sources of uncertainty in flood inundation maps. *Journal of Flood Risk Management*, *2*, 139–147. https://doi.org/10.1111/j.1753-318X.2009.01029.x

Ballard, E., Gard, M., & Pelle, B. (2010). *Flow–habitat relationships for spring and fall-run Chinook Salmon and steelhead/Rainbow Trout spawning in the Yuba River*. US Fish and Wildlife Service, Sacramento Fish and Wildlife Office.

Barker, J. R., Pasternack, G. B., Bratovich, P. M., Massa, D. A., Wyrick, J. R., & Johnson, T. R. (2018). Kayak drifter surface velocity observation for 2D hydraulic model validation. *River Research and Applications*, *34*(2), 124–134. https://doi.org/10.1002/rra.3238

Basnyat, B., Roy, N., & Gangopadhyay, A. (2021). Flood detection using semantic segmentation and multimodal data fusion. In *Presented at the 2021 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom workshops)* (pp. 135–140). https://doi.org/10.1109/PerComWorkshops51409.2021.9430985

Bates, P. D., Quinn, N., Sampson, C., Smith, A., Wing, O., Sosa, J., et al. (2021). Combined modeling of US fluvial, pluvial, and coastal flood hazard under current and future climates. *Water Research*, *57*(2), e2020WR028673. https://doi.org/10.1029/2020WR028673

Bebis, G., Boyle, R., Parvin, B., Koracin, D., Porikli, F., Skaff, S., et al. (Eds.) (2016). *Advances in visual computing: 12th international symposium, ISVC 2016, Las Vegas, NV, USA, December 12–14, 2016, proceedings, Part I, lecture notes in computer science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-50835-1

Bernhofen, M. V., Whyman, C., Trigg, M. A., Sleigh, P. A., Smith, A. M., Sampson, C. C., et al. (2018). A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique. *Environmental Research Letters*, *13*(10), 104007. https://doi.org/10.1088/1748-9326/aae014

Bhushan, S., Shean, D., Alexandrov, O., & Henderson, S. (2021). Automated digital elevation model (DEM) generation from very-high-resolution Planet SkySat triplet stereo and video imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *173*, 151–165. https://doi.org/10.1016/j.isprsjprs.2020.12.012

Biggs, H., Smart, G., Doyle, M., Holwerda, H., McDonald, M., & Ede, M. (2021). *River discharge from surface velocity measurements–A field guide for selecting alpha*. Envirolink Advice Report.

Bonafilia, D., Tellman, B., Anderson, T., & Issenberg, E. (2020). Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 210–211).

Brierley, G., Fryirs, K., Outhet, D., & Massey, C. (2002). Application of the river styles framework as a basis for river management in New South Wales, Australia. *Applied Geography*, *22*(1), 91–122. https://doi.org/10.1016/S0143-6228(01)00016-9

Brunner, G. W., United States Army Corps of Engineers, Institute for Water Resources (U.S.), & Hydrologic Engineering Center (U.S.). (2020). *HEC-RAS river analysis system: Hydraulic reference manual*. US Army Corps of Engineers, Institute for Water Resources, Hydrologic Engineering Center.

Bureau of Meteorology. (2022). Special climate statement 76 – Extreme rainfall and flooding in south-eastern Queensland and eastern New South Wales.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Cardona, A., Saalfeld, S., Schindelin, J., Arganda-Carreras, I., Preibisch, S., Longair, M., et al. (2012). TrakEM2 software for neural circuit reconstruction. *PLoS One*, *7*(6), e38011. https://doi.org/10.1371/journal.pone.0038011

Chen, C., Liang, D., Wang, X., & Li, Z. (2021). Application of UAV-based PIV to rivers. In *Presented at the 2021 7th international conference on hydraulic and civil engineering & smart water conservancy and intelligent disaster reduction forum (ICHCE & SWIDR)* (pp. 1707–1712). https://doi.org/10.1109/ICHCESWIDR54323.2021.9656476

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1606.00915

Chen, Y., Tang, Y., Yin, Z., Han, T., Zou, B., & Feng, H. (2022). Single object tracking in satellite videos: A correlation filter-based dual-flow tracker. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 1–13. https://doi.org/10.1109/JSTARS.2022.3185328

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1604.01685

Creutin, J. D., Muste, M., Bradley, A. A., Kim, S. C., & Kruger, A. (2003). River gauging using PIV techniques: A proof of concept experiment on the Iowa River. *Journal of Hydrology*, *277*(3–4), 182–194. https://doi.org/10.1016/S0022-1694(03)00081-7

Dal Sasso, S. F., Pizarro, A., & Manfreda, S. (2021). Recent advancements and perspectives in UAS-based image velocimetry. *Drones*, *5*(3), 81. https://doi.org/10.3390/drones5030081

d'Angelo, P., Kuschk, G., & Reinartz, P. (2014). Evaluation of Skybox video and still image products. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XL–1*, 95–99. https://doi.org/10.5194/isprsarchives-XL-1-95-2014

d'Angelo, P., Máttyus, G., & Reinartz, P. (2016). Skybox image and video product evaluation. *International Journal of Image and Data Fusion*, *7*(1), 3–18. https://doi.org/10.1080/19479832.2015.1109565

Dasgupta, A., Grimaldi, S., Ramsankaran, R. A. A. J., Pauwels, V. R. N., & Walker, J. P. (2018). Towards operational SAR-based flood mapping using neuro-fuzzy texture-based approaches. *Remote Sensing of Environment*, *215*, 313–329. https://doi.org/10.1016/j.rse.2018.06.019

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929

Eilander, D., Couasnon, A., Leijnse, T., Ikeuchi, H., Yamazaki, D., Muis, S., et al. (2023). A globally applicable framework for compound flood hazard modeling. *Natural Hazards and Earth System Sciences*, *23*(2), 823–846. https://doi.org/10.5194/nhess-23-823-2023

European Space Agency. (2022). Technical note on quality assessment for Jilin-1 SP and GF03C video missions.

Fischer, J. L., Bennion, D., Roseman, E. F., & Manny, B. A. (2015). Validation of a spatial model used to locate fish spawning reef construction sites in the St. Clair–Detroit River system. *Journal of Great Lakes Research*, *41*(4), 1178–1184. https://doi.org/10.1016/j.jglr.2015.09.019

FluViSat—Hydrological Flow Measurements from Satellite Video. (n.d.). Retrieved from https://www.ceh.ac.uk/our-science/projects/Fluvisat

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gard, M. (2008). Flow-habitat relationships for juvenile spring/fall-run Chinook salmon and steelhead/rainbow trout rearing in the Yuba River. Draft report prepared by the Energy Planning and Instream Flow Branch of the USFWS, Sacramento, California.

Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P., & Langan, T. E. (2019). Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. *Sensors*, *19*(7), 1486. https://doi.org/10.3390/s19071486

Geoscience Australia. (2022). Elvis - Elevation and depth - Foundation spatial data.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., et al. (2018). Multisource and multitemporal data fusion in remote sensing (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1812.08287

Girisha, S., MM, M. P., Verma, U., & Pai, R. M. (2019). Semantic segmentation of UAV aerial videos using convolutional neural networks. In *Presented at the 2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE)* (pp. 21–27). IEEE. https://doi.org/10.1109/AIKE.2019.00012

Gonçalves, D. N., Junior, J. M., Carrilho, A. C., Acosta, P. R., Ramos, A. P. M., Gomes, F. D. G., et al. (2023). Transformers for mapping burned areas in Brazilian Pantanal and Amazon with PlanetScope imagery. *International Journal of Applied Earth Observation and Geoinformation*, *116*, 103151. https://doi.org/10.1016/j.jag.2022.103151

Gonzalo-Martín, C., García-Pedrero, A., & Lillo-Saavedra, M. (2021). Improving deep learning Sorghum head detection through test time augmentation. *Computers and Electronics in Agriculture*, *186*, 106179. https://doi.org/10.1016/j.compag.2021.106179

Grimaldi, S., Li, Y., Pauwels, V. R. N., & Walker, J. P. (2016). Remote sensing-derived water extent and level to constrain hydraulic flood forecasting models: Opportunities and challenges. *Surveys in Geophysics*, *37*(5), 977–1034. https://doi.org/10.1007/s10712-016-9378-y

Grimaldi, S., Xu, J., Li, Y., Pauwels, V. R. N., & Walker, J. P. (2020). Flood mapping under vegetation using single SAR acquisitions. *Remote Sensing of Environment*, *237*, 111582. https://doi.org/10.1016/j.rse.2019.111582

Gu, X., Li, S., Ren, S., Zheng, H., Fan, C., & Xu, H. (2022). Adaptive enhanced Swin transformer with U-net for remote sensing image segmentation. *Computers & Electrical Engineering*, *102*, 108223. https://doi.org/10.1016/j.compeleceng.2022.108223

Hashemi-Beni, L., & Gebrehiwot, A. (2020). Deep learning for remote sensing image classification for agriculture applications. In *The international archives of photogrammetry, remote sensing and spatial information sciences* (pp. 51–54).

Hashemi-Beni, L., & Gebrehiwot, A. A. (2021). Flood extent mapping: An integrated method using deep learning and region growing using UAV optical data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 2127–2135. https://doi.org/10.1109/JSTARS.2021.3051873

Hauet, A., Morlot, T., & Daubagnan, L. (2018). Velocity profile and depth-averaged to surface velocity in natural streams: A review over alarge sample of rivers. *E3S Web of Conferences*, *40*, 06015. https://doi.org/10.1051/e3sconf/20184006015

He, A., Wang, K., Li, T., Du, C., Xia, S., & Fu, H. (2023). H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, *42*(9), 1. https://doi.org/10.1109/TMI.2023.3264513

He, Q., Yang, Q., & Xie, M. (2023). HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation. *Computers in Biology and Medicine*, *155*, 106629. https://doi.org/10.1016/j.compbiomed.2023.106629

Heller, N., Dean, J., & Papanikolopoulos, N. (2018). Imperfect segmentation labels: How much do they matter? In D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, et al. (Eds.), *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis* (pp. 112–120). Springer International Publishing. https://doi.org/10.1007/978-3-030-01364-6_13

Hesselink, A. W., Stelling, G. S., Kwadijk, J. C. J., & Middelkoop, H. (2003). Inundation of a Dutch river polder, sensitivity analysis of a physically based inundation model using historic data: Sensitivity analysis of an inundation model. *Water Resources Research*, *39*(9), 1234. https://doi.org/10.1029/2002WR001334

Ho, J. C., Vu, W., Tellman, B., Dinga, J. B., N'diaye, P. I., Weber, S., et al. (2021). From cloud to refugee camp: A satellite-based flood analytics case-study in Congo-Brazzaville. In *Earth observation for flood applications* (pp. 131–146). Elsevier.

Hoch, J. M., & Trigg, M. A. (2019). Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models. *Environmental Research Letters*, *14*(3), 034001. https://doi.org/10.1088/1748-9326/aaf3d3

Hong, S., Bonkhoff, A. K., Hoopes, A., Bretzner, M., Schirmer, M. D., Giese, A.-K., et al. (2021). Hypernet-ensemble learning of segmentation probability for medical image segmentation with ambiguous labels.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., & Pierrefeu, G. (2018). Impact of stage measurement errors on streamflow uncertainty. *Water Resources Research*, *54*(3), 1952–1976. https://doi.org/10.1002/2017WR022039

Horritt, M. S. (2006). A methodology for the validation of uncertain flood inundation models. *Journal of Hydrology*, *326*(1–4), 153–165. https://doi.org/10.1016/j.jhydrol.2005.10.027

Hosseiny, H. (2021). A deep learning model for predicting river flood depth and extent. *Environmental Modelling & Software*, *145*, 105186. https://doi.org/10.1016/j.envsoft.2021.105186

Hu, J., Li, L., Lin, Y., Wu, F., & Zhao, J. (2019). A comparison and strategy of semantic segmentation on remote sensing images. https://doi.org/10.48550/ARXIV.1905.10231

Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., & Perazzi, F. (2020). Temporally distributed networks for fast video semantic segmentation. In *Presented at the proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8818–8827).

Huang, B., Li, Y., Han, X., Cui, Y., Li, W., & Li, R. (2015). Cloud removal from optical satellite imagery with SAR imagery using sparse representation. *IEEE Geoscience and Remote Sensing Letters*, *12*(5), 1046–1050. https://doi.org/10.1109/LGRS.2014.2377476

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning?

Ivliev, N., Evdokimova, V., Podlipnov, V., Petrov, M., Ganchevskaya, S., Tkachenko, I., et al. (2022). First earth-imaging CubeSat with harmonic diffractive lens. *Remote Sensing*, *14*(9), 2230. https://doi.org/10.3390/rs14092230

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *Presented at the 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)* (pp. 1–7). IEEE. https://doi.org/10.1109/CIBCB48159.2020.9277638

Kääb, A., Altena, B., & Mascaro, J. (2019). River-ice and water velocities using the Planet optical CubeSat constellation. *Hydrology and Earth System Sciences*, *23*(10), 4233–4247. https://doi.org/10.5194/hess-23-4233-2019

Khalid, H. W., Khalil, R. M. Z., & Qureshi, M. A. (2021). Evaluating spectral indices for water bodies extraction in western Tibetan Plateau. *The Egyptian Journal of Remote Sensing and Space Science*, *24*(3), 619–634. https://doi.org/10.1016/j.ejrs.2021.09.003

Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., et al. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, *54*(10), 7149–7176. https://doi.org/10.1029/2018WR022708

Kugelman, J., Allman, J., Read, S. A., Vincent, S. J., Tong, J., Kalloniatis, M., et al. (2022). A comparison of deep learning U-Net architectures for posterior segment OCT retinal layer segmentation. *Scientific Reports*, *12*(1), 14888. https://doi.org/10.1038/s41598-022-18646-2

Lane, S. N., Bradbrook, K. F., Richards, K. S., Biron, P. A., & Roy, A. G. (1999). The application of computational fluid dynamics to natural river channels: Three-dimensional versus two-dimensional approaches. *Geomorphology*, *29*(1–2), 1–20. https://doi.org/10.1016/S0169-555X(99)00003-3

Leach, N. R., Popien, P., Goodman, M. C., & Tellman, B. (2022). Leveraging convolutional neural networks for semantic segmentation of global floods with PlanetScope imagery. In *Presented at the IGARSS 2022-2022 IEEE international geoscience and remote sensing symposium* (pp. 314–317). IEEE.

Le Coz, J., Hauet, A., Pierrefeu, G., Dramais, G., & Camenen, B. (2010). Performance of image-based velocimetry (LSPIV) applied to flash-flood discharge measurements in Mediterranean rivers. *Journal of Hydrology*, *394*(1–2), 42–52. https://doi.org/10.1016/j.jhydrol.2010.05.049

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., & Le Boursicaud, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology*, *509*, 573–587. https://doi.org/10.1016/j.jhydrol.2013.11.016

Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. MIT Press.

Legleiter, C. J., & Kinzel, P. J. (2020). Inferring surface flow velocities in sediment-laden Alaskan rivers from optical image sequences acquired from a helicopter. *Remote Sensing*, *12*(8), 1282. https://doi.org/10.3390/rs12081282

Legleiter, C. J., & Kinzel, P. J. (2021). Surface flow velocities from space: Particle image velocimetry of satellite video of a large, sediment-laden river. *Frontiers in Water*, *3*, 652213. https://doi.org/10.3389/frwa.2021.652213

Lewis, Q. W., Lindroth, E. M., & Rhoads, B. L. (2018). Integrating unmanned aerial systems and LSPIV for rapid, cost-effective stream gauging. *Journal of Hydrology*, *560*, 230–246. https://doi.org/10.1016/j.jhydrol.2018.03.008

Li, G., Liu, J., & Shao, W. (2022). Flood risk assessment using TELEMAC-2D models integrated with multi-index analysis in Shenzhen River Basin, China. *Water*, *14*(16), 2513. https://doi.org/10.3390/w14162513

Li, Y., Shi, J., & Lin, D. (2018). Low-latency video semantic segmentation. In *Presented at the proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5997–6005).

Li, Z., Shen, H., Cheng, Q., Li, W., & Zhang, L. (2019). Thick cloud removal in high-resolution satellite images using stepwise radiometric adjustment and residual correction. *Remote Sensing*, *11*(16), 1925. https://doi.org/10.3390/rs11161925

Liddle, J. D., Holt, A. P., Jason, S. J., O'Donnell, K. A., & Stevens, E. J. (2020). Space science with CubeSats and nanosatellites. *Nature Astronomy*, *4*(11), 1026–1030. https://doi.org/10.1038/s41550-020-01247-2

Liu, J., Liu, G., & Zhao, Y. (2022). Improve semantic segmentation of high-resolution remote sensing imagery with RS-TTA. In *Presented at the 2022 5th international conference on pattern recognition and artificial intelligence (PRAI)* (pp. 817–825). IEEE.

Liu, W.-C., Lu, C.-H., & Huang, W.-C. (2021). Large-scale particle image velocimetry to measure streamflow from videos recorded from unmanned aerial vehicle and fixed imaging system. *Remote Sensing*, *13*(14), 2661. https://doi.org/10.3390/rs13142661

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., et al. (2023). Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. https://doi.org/10.48550/arXiv.2103.14030

Lomaka, I. A., Elisov, N. A., Boltov, E. A., & Shafran, S. V. (2022). A novel design of CubeSat deployment system for transformable structures. *Acta Astronautica*, *197*, 179–190. https://doi.org/10.1016/j.actaastro.2022.05.027

Marinan, A., Nicholas, A., & Cahoy, K. (2013). Ad hoc CubeSat constellations: Secondary launch coverage and distribution. In *2013 IEEE aerospace conference. Presented at the 2013 IEEE aerospace conference* (pp. 1–15). IEEE. https://doi.org/10.1109/AERO.2013.6497174

Masafu, C., Williams, R., Shi, X., Yuan, Q., & Trigg, M. (2022). Unpiloted Aerial Vehicle (UAV) image velocimetry for validation of two-dimensional hydraulic model simulations. *Journal of Hydrology*, *612*, 128217. https://doi.org/10.1016/j.jhydrol.2022.128217

Masafu, C., & Williams, R. D. (2023). Satellite video remote sensing for flood model validation [Dataset]. https://doi.org/10.5525/gla.researchdata.1537

Mateo-Garcia, G., Veitch-Michaelis, J., Smith, L., Oprea, S. V., Schumann, G., Gal, Y., et al. (2021). Towards global flood mapping onboard low cost satellites with machine learning. *Scientific Reports*, *11*(1), 7249. https://doi.org/10.1038/s41598-021-86650-z

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., et al. (2017). The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017

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

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality: Benchmarking observational uncertainties for hydrology. *Hydrological Processes*, *26*, 4078–4111. https://doi.org/10.1002/hyp.9384

Mester, B., Willner, S. N., Frieler, K., & Schewe, J. (2021). Evaluation of river flood extent simulated with multiple global hydrological models and climate forcings. *Environmental Research Letters*, *16*(9), 094010. https://doi.org/10.1088/1748-9326/ac188d

Mishra, B. K., Rafiei Emam, A., Masago, Y., Kumar, P., Regmi, R. K., & Fukushi, K. (2018). Assessment of future flood inundations under climate and land use change scenarios in the Ciliwung River Basin, Jakarta. *Journal of Flood Risk Management*, *11*(S2), S1105–S1115. https://doi.org/10.1111/jfr3.12311

Mobley, W., Sebastian, A., Blessing, R., Highfield, W. E., Stearns, L., & Brody, S. D. (2021). Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: A pilot study in southeast Texas. *Natural Hazards and Earth System Sciences*, *21*(2), 807–822. https://doi.org/10.5194/nhess-21-807-2021

Mokhtar, E. S., Pradhan, B., Ghazali, A. H., & Shafri, H. Z. M. (2018). Assessing flood inundation mapping through estimated discharge using GIS and HEC-RAS model. *Arabian Journal of Geosciences*, *11*(21), 682. https://doi.org/10.1007/s12517-018-4040-2

Molinari, D., De Bruijn, K. M., Castillo-Rodríguez, J. T., Aronica, G. T., & Bouwer, L. M. (2019). Validation of flood risk models: Current practice and possible improvements. *International Journal of Disaster Risk Reduction*, *33*, 441–448. https://doi.org/10.1016/j.ijdrr.2018.10.022

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., & Horvath, P. (2020). Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports*, *10*(1), 5068. https://doi.org/10.1038/s41598-020-61808-3

Nandi, I, Srivastava, P. K., & Shah, K. (2017). Floodplain mapping through support vector machine and optical/infrared images from Landsat 8 OLI/TIRS sensors: Case study from Varanasi. *Water Resources Management*, *31*(4), 1157–1171. https://doi.org/10.1007/s11269-017-1568-y

Navarro-Hernández, M. I., Valdes-Abellan, J., Tomás, R., Tessitore, S., Ezquerro, P., & Herrera, G. (2023). Analysing the impact of land subsidence on the flooding risk: Evaluation through InSAR and modelling. *Water Resources Management*, *37*(11), 4363–4383. https://doi.org/10.1007/s11269-023-03561-6

Nkwunonwo, U., Whitworth, M., & Baily, B. (2020). A review of the current status of flood modelling for urban flood risk management in the developing countries. *Scientific African*, *7*, e00269. https://doi.org/10.1016/j.sciaf.2020.e00269

Ocio, D., Le Vine, N., Westerberg, I., Pappenberger, F., & Buytaert, W. (2017). The role of rating curve uncertainty in real-time flood forecasting: The role of RCU in flood forecasting. *Water Resources Research*, *53*(5), 4197–4213. https://doi.org/10.1002/2016WR020225

Ongdas, N., Akiyanova, F., Karakulov, Y., Muratbayeva, A., & Zinabdin, N. (2020). Application of HEC-RAS (2D) for flood hazard maps generation for Yesil (Ishim) river in Kazakhstan. *Water*, *12*(10), 2672. https://doi.org/10.3390/w12102672

Pasternack, G. B. (2011). *2D modeling and ecohydraulic analysis*. University of California at Davis.

Pearce, S., Ljubičić, R., Peña-Haro, S., Perks, M., Tauro, F., Pizarro, A., et al. (2020). An evaluation of image velocimetry techniques under low flow conditions and high seeding densities using unmanned aerial systems. *Remote Sensing*, *12*(2), 232. https://doi.org/10.3390/rs12020232

Perks, M. T., Dal Sasso, S. F., Hauet, A., Jamieson, E., Le Coz, J., Pearce, S., et al. (2020). Towards harmonisation of image velocimetry techniques for river surface velocity observations. *Earth System Science Data*, *12*(3), 1545–1559. https://doi.org/10.5194/essd-12-1545-2020

Pizer, S. M. (1990). Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Presented at the Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia* (p. 2).

Pradhan, D., Sahu, R. T., & Verma, M. K. (2022). Flood inundation mapping using GIS and hydraulic model (HEC-RAS): A case study of the Burhi Gandak River, Bihar, India. In R. Kumar, C. W. Ahn, T. K. Sharma, O. P. Verma, & A. Agarwal (Eds.), *Soft computing: Theories and applications, lecture notes in networks and systems* (pp. 135–145). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-0707-4_14

Qi, X., Li, K., Liu, P., Zhou, X., & Sun, M. (2020). Deep attention and multi-scale networks for accurate remote sensing image segmentation. *IEEE Access*, *8*, 146627–146639. https://doi.org/10.1109/access.2020.3015587

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., & Murphy, R. R. (2021). FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, *9*, 89644–89654. https://doi.org/10.1109/ACCESS.2021.3090981

Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., & Datcu, M. (2020). *Sen12-flood: A SAR and multispectral dataset for flood detection*. IEEE.

Rantz, S. E. (1982). *Measurement and computation of streamflow*. U.S. Department of the Interior, Geological Survey.

Rentschler, J., & Salhab, M. (2020). *People in Harm's way: Flood exposure and poverty in 189 countries*. Policy research working papers. The World Bank. https://doi.org/10.1596/1813-9450-9447

Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). ImageNet-21K pretraining for the masses.

Rollason, E., Bracken, L. J., Hardy, R. J., & Large, A. R. G. (2018). The importance of volunteered geographic information for the validation of flood inundation models. *Journal of Hydrology*, *562*, 267–280. https://doi.org/10.1016/j.jhydrol.2018.05.002

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, Part III 18*.

Schumann, G. J.-P. (2017). Remote sensing of floods. In *Oxford research encyclopedia of natural hazard science*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389407.013.265

Shastry, A., Carter, E., Coltin, B., Sleeter, R., McMichael, S., & Eggleston, J. (2023). Mapping floods from remote sensing data and quantifying the effects of surface obstruction by clouds and vegetation. *Remote Sensing of Environment*, *291*, 113556. https://doi.org/10.1016/j.rse.2023.113556

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Shrestha, A., Mascaro, G., & Garcia, M. (2022). Effects of stormwater infrastructure data completeness and model resolution on urban flood modeling. *Journal of Hydrology*, *607*, 127498. https://doi.org/10.1016/j.jhydrol.2022.127498

Shu, M., Zhong, Y., & Lv, D. (2021). Small moving vehicle detection via local enhancement fusion for satellite video. *International Journal of Remote Sensing*, *42*(19), 7189–7214. https://doi.org/10.1080/01431161.2021.1944694

Singh, S., & Kansal, M. L. (2022). Chamoli flash-flood mapping and evaluation with a supervised classifier and NDWI thresholding using Sentinel-2 optical data in Google earth engine. *Earth Science Informatics*, *15*(2), 1073–1086. https://doi.org/10.1007/s12145-022-00786-8

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. https://doi.org/10.48550/ARXIV.2007.12269

Strelnikova, D., Perks, M. T., Dal Sasso, S. F., & Pizarro, A. (2023). River flow monitoring with unmanned aerial system. In *Unmanned aerial systems for monitoring soil, vegetation, and riverine environments* (pp. 231–269). Elsevier.

Su, T., Liu, T., Zhang, S., Qu, Z., & Li, R. (2020). Machine learning-assisted region merging for remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *168*, 89–123. https://doi.org/10.1016/j.isprsjprs.2020.07.017

Sun, G., Liu, Y., Ding, H., Probst, T., & Van Gool, L. (2022). Coarse-to-fine feature mining for video semantic segmentation. In *Presented at the proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3126–3137).

Sun, Y., Bi, F., Gao, Y., Chen, L., & Feng, S. (2022). A multi-attention UNet for semantic segmentation in remote sensing images. *Symmetry*, *14*(5), 906. https://doi.org/10.3390/sym14050906

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., et al. (2022). A review of earth artificial intelligence. *Computers & Geosciences*, *159*, 105034. https://doi.org/10.1016/j.cageo.2022.105034

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Artificial neural networks and machine learning – ICANN 2018, lecture notes in computer science* (pp. 270–279). Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7_27

Tanim, A. H., McRae, C. B., Tavakol-Davani, H., & Goharian, E. (2022). Flood detection in urban areas using satellite imagery and machine learning. *Water*, *14*(7), 1140. https://doi.org/10.3390/w14071140

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., et al. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, *596*(7870), 80–86. https://doi.org/10.1038/s41586-021-03695-w

Thielicke, W., & Sonntag, R. (2021). Particle image velocimetry for MATLAB: Accuracy and enhanced algorithms in PIVlab. *Journal of Open Research Software*, *9*(1), 12. https://doi.org/10.5334/jors.334

Thielicke, W., & Stamhuis, E. (2014). PIVlab – Towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB. *Journal of Open Research Software*, *2*, e30. https://doi.org/10.5334/jors.bl

Tiampo, K. F., Huang, L., Simmons, C., Woods, C., & Glasscoe, M. T. (2022). Detection of flood extent using sentinel-1A/B synthetic aperture radar: An application for Hurricane Harvey, Houston, TX. *Remote Sensing*, *14*(9), 2261. https://doi.org/10.3390/rs14092261

Timbadiya, P. V., Patel, P. L., & Porey, P. D. (2011). Calibration of HEC-RAS model on prediction of flood for lower Tapi River, India. *Journal of Water Resource and Protection*, *3*(11), 805–811. https://doi.org/10.4236/jwarp.2011.311090

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, *237*, 111322. https://doi.org/10.1016/j.rse.2019.111322

Tsakiris, G. (2014). Flood risk assessment: Concepts, modelling, applications. *Natural Hazards and Earth System Sciences*, *14*(5), 1361–1369. https://doi.org/10.5194/nhess-14-1361-2014

Vigoureux, S., Liebard, L.-L., Chonoski, A., Robert, E., Torchet, L., Poveda, V., et al. (2022). Comparison of streamflow estimated by image analysis (LSPIV) and by hydrologic and hydraulic modelling on the French Riviera during November 2019 flood. In P. Gourbesville & G. Caignaert (Eds.), *Advances in hydroinformatics* (pp. 255–273). Springer Water. Springer Nature. https://doi.org/10.1007/978-981-19-1600-7_16

Wang, J., Wang, S., Wang, F., Zhou, Y., Wang, Z., Ji, J., et al. (2022). FWENet: A deep convolutional neural network for flood water body extraction based on SAR images. *International Journal of Digital Earth*, *15*(1), 345–361. https://doi.org/10.1080/17538947.2021.1995513

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008

Wang, T., Lan, J., Han, Z., Hu, Z., Huang, Y., Deng, Y., et al. (2022). O-Net: A novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification. *Frontiers in Neuroscience*, *16*, 876065. https://doi.org/10.3389/fnins.2022.876065

Wieland, M., Martinis, S., Kiefl, R., & Gstaiger, V. (2023). Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment*, *287*, 113452. https://doi.org/10.1016/j.rse.2023.113452

Williams, R. D., Brasington, J., Hicks, M., Measures, R., Rennie, C. D., & Vericat, D. (2013). Hydraulic validation of two-dimensional simulations of braided river flow with spatially continuous aDcp data. *Water Resources Research*, *49*(9), 5183–5205. https://doi.org/10.1002/wrcr.20391

Wing, O. E. J., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., & Erickson, T. A. (2017). Validation of a 30 m resolution flood hazard model of the conterminous United States: 30 m resolution flood model of conus. *Water Resources Research*, *53*(9), 7968–7986. https://doi.org/10.1002/2017WR020917

Wing, O. E. J., Smith, A. M., Marston, M. L., Porter, J. R., Amodeo, M. F., Sampson, C. C., & Bates, P. D. (2021). Simulating historical flood events at the continental scale: Observational validation of a large-scale hydrodynamic model. *Natural Hazards and Earth System Sciences*, *21*(2), 559–575. https://doi.org/10.5194/nhess-21-559-2021

Xiao, A., Wang, Z., Wang, L., & Ren, Y. (2018). Super-resolution for "Jilin-1" satellite video imagery via a convolutional network. *Sensors*, *18*(4), 1194. https://doi.org/10.3390/s18041194

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, *27*(14), 3025–3033. https://doi.org/10.1080/01431160600589179

Xu, Z., Zhang, W., Zhang, T., Yang, Z., & Li, J. (2021). Efficient transformer for remote sensing image segmentation. *Remote Sensing*, *13*(18), 3585. https://doi.org/10.3390/rs13183585

Yadav, G., Maheshwari, S., & Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. In *Presented at the 2014 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 2392–2397). IEEE.

Yalcin, E. (2020). Assessing the impact of topography and land cover data resolutions on two-dimensional HEC-RAS hydrodynamic model simulations for urban flood hazard analysis. *Natural Hazards*, *101*(3), 995–1017. https://doi.org/10.1007/s11069-020-03906-z

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yuan, F., Zhang, L., Xia, X., Huang, Q., & Li, X. (2021). A gated recurrent network with dual classification assistance for smoke semantic segmentation. *IEEE Transactions on Image Processing*, *30*, 4409–4422. https://doi.org/10.1109/TIP.2021.3069318

Zeiger, S. J., & Hubbart, J. A. (2021). Measuring and modeling event-based environmental flows: An assessment of HEC-RAS 2D rain-on-grid simulations. *Journal of Environmental Management*, *285*, 112125. https://doi.org/10.1016/j.jenvman.2021.112125

Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., & Wang, C. (2022). Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–20. https://doi.org/10.1109/tgrs.2022.3144894

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Presented at the Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641).

Zhou, Z., Zhang, J., & Gong, C. (2023). Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, *38*(17), 2491–2510. https://doi.org/10.1111/mice.13003

References From the Supporting Information

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., & Liu, L. (2016). Remote sensing image registration with modified SIFT and enhanced feature matching. *IEEE Geoscience and Remote Sensing Letters*, *14*(1), 3–7. https://doi.org/10.1109/lgrs.2016.2600858