RiverScope: High-Resolution River Masking Dataset

Rangel Daroya[1], Taylor Rowley[1], Jonathan Acero Flores[1], Elisa Friedmann[1], Fiona B Bennitt[1]
Heejin An[1], Travis Thomas Simmons[1], Marissa Hughes[2], Camryn L Kluetmeier[2]
Solomon Kica[2], J. Daniel Vélez[2], Sarah E. Esenther[3], Thomas E. Howard[3]
Yanqi Ye[3], Audrey J. Turcotte[4], Colin Gleason[1], Subhransu Maji[1]
[1]University of Massachusetts Amherst
[2]University of North Carolina at Chapel Hill
[3]Brown University
[4]University of Colorado Boulder

Abstract

Surface water dynamics play a critical role in Earth's climate system, influencing ecosystems, agriculture, disaster resilience, and sustainable development. Yet monitoring rivers and surface water at fine spatial and temporal scales remains challenging—especially for narrow or sediment-rich rivers that are poorly captured by low-resolution satellite data. To address this, we introduce RiverScope, a high-resolution dataset developed through collaboration between computer science and hydrology experts. RiverScope comprises 1,145 high-resolution images (covering 2,577 square kilometers) with expert-labeled river and surface water masks, requiring over 100 hours of manual annotation. Each image is co-registered with Sentinel-2, SWOT, and the SWOT River Database (SWORD), enabling the evaluation of cost-accuracy trade-offs across sensors—a key consideration for operational water monitoring. We also establish the first global, high-resolution benchmark for river width estimation, achieving a median error of 7.2 meters—significantly outperforming existing satellite-derived methods. We extensively evaluate deep networks across multiple architectures (e.g., CNNs and transformers), pretraining strategies (e.g., supervised and self-supervised), and training datasets (e.g., ImageNet and satellite imagery). Our best-performing models combine the benefits of transfer learning with the use of all the multispectral PlanetScope channels via learned adaptors. RiverScope provides a valuable resource for fine-scale and multi-sensor hydrological modeling, supporting climate adaptation and sustainable water management.

Code — https://github.com/cvl-umass/riverscope-models
Dataset — https://github.com/cvl-umass/riverscope
Extended version — https://arxiv.org/abs/2509.02451

1 Introduction

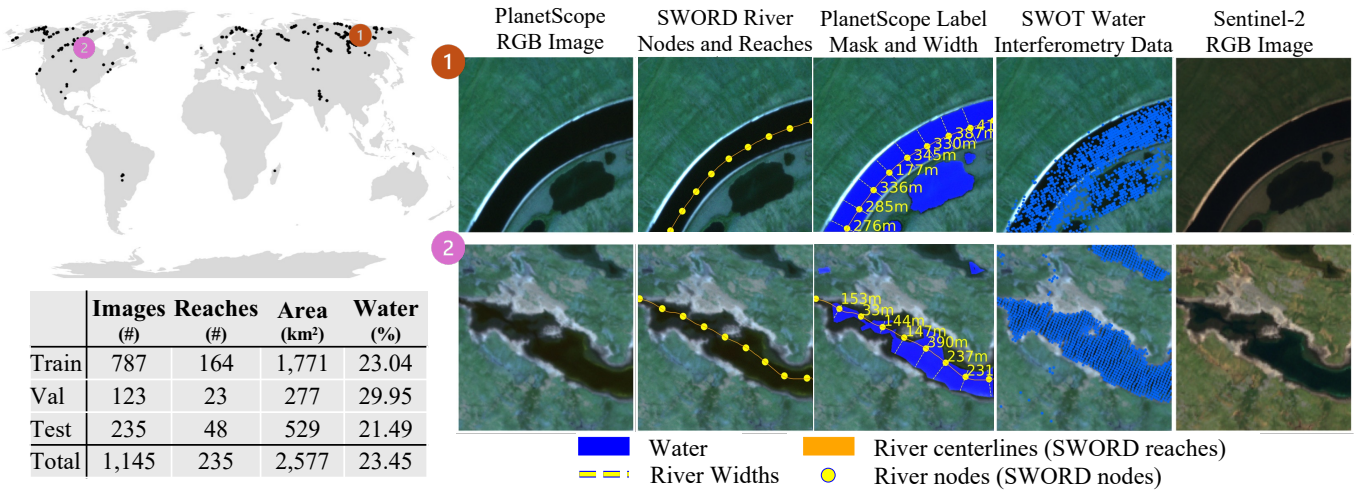
Global surface waters form the circulatory system of the climate, playing a vital role in transporting water and the materials it carries across the planet (Alsdorf, Rodríguez, and Lettenmaier 2007). These dynamics are shaped not only by natural phenomena like floods and erosion but also by human activities like dam construction and irrigation (Yang et al. 2022). Understanding and mapping surface water dynamics is increasingly important for climate resilience and environmental sustainability due to their role in agriculture (Tian et al. 2015), hydropower (Wasti et al. 2022), ecological services (Zhao et al. 2003), urban planning (Ellis 2013), and transportation (Opher and Friedler 2010).

While the hydrology community has made progress in locating and mapping surface water (i.e., hydrography) (Valman et al. 2024), characterizing water dynamics like extent and flow remains challenging. Satellite imagery is effective at capturing water extent, but properties such as flow are not directly observable (Feng et al. 2019). Stream gauges, which directly measure flow, are sparse and insufficient to provide a global view (Gleason and Hamdan 2017).

These limitations have driven the launch of the Surface Water and Ocean Topography (SWOT) mission (Biancamaria et al. 2016), which uses radar interferometry to measure surface water elevation with unprecedented precision (Vinogradova et al. 2025). However, river discharge (flow) is still inferred, a calculation that critically depends on accurate river width measurements (Bjerklie et al. 2018). To support this, datasets like the SWOT River Database (SWORD) (Altenau et al. 2021) were developed, but they rely on width estimates from coarser Landsat imagery (Allen and Pavelsky 2018).

Spatial resolution remains a key bottleneck, especially for small rivers and fine-scale changes (Filippucci et al. 2022). Sentinel-2 (ESA 2022) and Landsat (Observation and Center 2020) imagery (10-30m/pixel) often miss these features (Flores et al. 2024), limiting the accuracy of river models. This motivates the need for higher-resolution, expertly labeled datasets.

We introduce RiverScope, a densely annotated, global-scale dataset of high-resolution (3 m/pixel) PlanetScope (PBC 2024) imagery of rivers and adjacent water bodies. The dataset covers 2,577 km^2, comprising 500x500-pixel images sampled across diverse geographic and hydrological contexts. Each image is co-located with Sentinel-2, SWOT, and SWORD data within a $\pm$12-hour window (see Figure 1). This unique alignment enables the comparison of sensor performance on hydrologically relevant tasks, quantifying trade-offs between accuracy and cost—a critical consideration for scaling monitoring systems at agencies like NASA or ESA. Surface water masks were created by hydrology and machine learning experts through



	Images (#)	Reaches (#)	Area (km²)	Water (%)
Train	787	164	1,771	23.04
Val	123	23	277	29.95
Test	235	48	529	21.49
Total	1,145	235	2,577	23.45

Figure 1: RiverScope presents a global, high-resolution satellite image dataset focused on rivers using PlanetScope (PBC 2024) and co-registered with SWOT (Vinogradova et al. 2025), SWORD (Altenau et al. 2021), and Sentinel-2 (ESA 2022). To the left we show the distribution and splits of our expert-labeled dataset, covering various geographic and hydrological contexts.

100 hours of manual annotation, ensuring high-quality ground truth data. We leverage this dataset to establish benchmarks using hydrologically relevant evaluation metrics.

Rivers are an important, underexplored domain for machine learning, with complex natural features distinct from structured objects in typical urban datasets. RiverScope captures complex hydrological features such as riverbanks, sandbars, and varying morphologies—elements that are often lost in coarser datasets. By leveraging high-resolution multispectral imagery, models can better delineate river boundaries, improving width estimates essential for discharge modeling.

Our benchmark serves two key purposes:

1. **Water segmentation.** We evaluate diverse deep learning models on PlanetScope imagery, including ImageNet-pretrained models and remote sensing-specific methods. We show that lightweight input adapters using all the multispectral PlanetScope bands outperform RGB-only models. Our models outperform established baselines including NDWI (McFeeters 1996) and custom CNN architectures (Valman et al. 2024). PlanetScope's fine spatial resolution enables detection of detailed river structures often missed by lower-resolution sensors like Sentinel (see Figure 2). Models trained on coarse imagery often struggle with precise boundary delineation, highlighting the potential of fusing complementary modalities in our dataset to improve performance of Sentinel estimates.
2. **River width estimation.** We introduce a benchmark for river width—a hydrologically meaningful metric poorly captured by pixel-level segmentation metrics like IoU. Our best model, trained on PlanetScope imagery, achieves a median width error of 7.2 meters, substantially outperforming the prior state-of-the-art of 30 meters (Valman et al. 2024). Landsat- and Sentinel-derived widths show median errors of 45.0 and 39.0 meters, respectively. Notably, we also provide the first evaluation of SWOT-derived widths, which exhibit a median error of 41.9 meters at "nodes" spaced every 200 meters along the river.

By publicly releasing RiverScope, we aim to foster machine learning research on domain-specific challenges in river monitoring—problems with direct implications for climate adaptation and sustainable water management.

2 Related Work

Satellite image datasets. Previous works have introduced large-scale datasets based on satellites with global coverage, such as Sentinel and Landsat (Bastani et al. 2023; Claverie et al. 2018). For example, SatlasPretrain (Bastani et al. 2023) presents a diverse dataset of Sentinel and NAIP (USDA Farm Service Agency n.d.) imagery labeled for object detection, scene classification, and segmentation. EuroSAT (Helber et al. 2019) and BigEarthNet (Sumbul et al. 2019) provide Sentinel-based datasets for image classification, while NASA periodically releases segmentation labels through the Harmonized Landsat-Sentinel program (Jones 2019), including water body mapping products (Daroya et al. 2025).

While Sentinel and Landsat offer broad spatial and temporal coverage, their resolutions (10 m for Sentinel, 30 m for Landsat) limit their utility for fine-grained hydrological analysis, such as tracking narrow river dynamics. High-resolution datasets remain limited and are often focused on urban or industrial settings, relying on aerial imagery with sparse global coverage and infrequent revisit rates (Christie et al. 2018).

The PlanetScope mission (PBC 2024) addresses these limitations by offering globally available, high-resolution (3 m per pixel), near-daily imagery. This has enabled recent work on surface water monitoring using PlanetScope data (Valman et al. 2024; Flores et al. 2024). However, these efforts have largely focused on specific regions and *do not* publicly release annotated datasets.

To address this gap, we introduce RiverScope, a densely labeled, global-scale dataset of high-resolution PlanetScope imagery for water segmentation focused on rivers. While

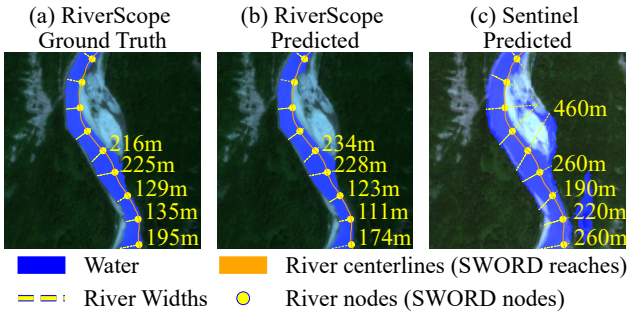


Figure 2: RiverScope can be used to precisely segment rivers and water bodies (a-b). Existing low-resolution images like Sentinel (c) tend to over segment narrow rivers, inflating river width estimates due to less detail in the images. Yellow dots mark SWORD nodes; the orange line represents a section of the SWORD reach used as the river centerline.

Valman et al. (2024) presents a similar large-scale water segmentation dataset, RiverScope extends this line of work by incorporating co-registered measurements from complementary sensors—including Sentinel (ESA 2022), Surface Water and Ocean Topography (SWOT) (Biancamaria et al. 2016), and SWOT River Database (SWORD) (Altenau et al. 2021)—and by *openly* releasing the annotated data.

The recently launched SWOT satellite provides interferometric measurements of surface water elevation and extent, while SWORD offers reach-level hydrologic variables such as river width and slope (See § 3 for details). By aligning these with PlanetScope and Sentinel imagery, RiverScope enables benchmarking across multiple tasks and facilitates evaluating cost-accuracy trade-offs of different sensors—a key consideration for scaling monitoring systems.

Satellite image models. While early studies in satellite image modeling focused primarily on low-resolution data (Bastani et al. 2023; Manas et al. 2021; Daroya et al. 2025), a growing body of work now targets applications requiring high-resolution imagery (Flores et al. 2024). For instance, Flores et al. (2024) used PlanetScope imagery to segment headwater streams—small, often unnamed tributaries that require fine spatial details. Beyond stream segmentation (Valman et al. 2024), PlanetScope has been applied to river discharge estimation, reservoir monitoring, and water quality assessment (Wang and Vivoni 2022; Flores et al. 2024).

Building on prior work (Valman et al. 2024), we focus on water segmentation using PlanetScope, and further extend to estimating river widths and validating outputs of complementary sources such as Sentinel, SWOT, and SWORD. Despite increased interest in high-resolution applications, there has been limited exploration of pretraining strategies (e.g., ImageNet (Deng et al. 2009), CLIP (Radford et al. 2021), MoCov3 (Chen, Xie, and He 2021)) and segmentation architectures (Ronneberger, Fischer, and Brox 2015; Ranftl, Bochkovskiy, and Koltun 2021) specifically tailored for this domain. Many prior approaches rely on randomly initialized convolutional models (Valman et al. 2024) or simple thresholding methods (McFeeters 1996).

In this work, we benchmark a range of segmentation architectures and pretraining methods for high-resolution river segmentation and width estimation. Across all models and training strategies, higher-resolution imagery delivers higher accuracy, highlighting the clear advantage of fine spatial detail for both river segmentation and width estimation.

3 RiverScope Construction

3.1 Data Sources

We collect satellite image data from the PlanetScope and the Sentinel-2 constellation of satellites. PlanetScope data were selected to cover areas that spatially and temporally intersect with SWOT, SWORD, and Sentinel-2. This is done to give access to multiple observation modalities in the same location and time, while enabling cross-validation across different sources. PlanetScope images were used for annotating water masks to capture fine details that high-resolution imagery better reveals. We discuss the collection process for each of the data sources below (more details in § A.4 (Appendix)).

Surface Water and Ocean Topography (SWOT) is a satellite mission launched by NASA and CNES in December 2022, and orbits with a Ka-band Radar Interferometer (KaRIn) that revisits almost the entire globe every 21 days. SWOT produces a three-dimensional point cloud of the Earth's water surfaces by timing microwave pulses and comparing the phase returned to its twin antennae. Ground processing first classifies each radar sample as water or non-water, then aggregates these classifications along rivers at fixed nodes (i.e., points every 200 m) and reaches (i.e., line segments around 10 km long). Hydrological variables are derived from these aggregates such as effective width (water area divided by effective length), surface elevation, and slope.

Since Ka-band radar is less sensitive to clouds, SWOT provides large-scale, repeatable monitoring of rivers, lakes, and reservoirs wider than about 50 m worldwide, offering an alternative to traditional gauge-based systems. Previous studies (Yao et al. 2025) show SWOT can adequately resolve rivers, but mask quality is affected by factors such as vegetation, wetlands, and nearby urban areas. Our dataset provides the first systematic evaluation of river width estimates derived from SWOT on a global scale.

SWOT River Database (SWORD) is a global database of rivers designed to support SWOT river products. It integrates multiple data sources including river widths estimated from Landsat imagery (30 m/pixel resolution), river flow characteristics extracted from digital elevation maps (Yamazaki et al. 2019), and identified human-made obstructions along river paths. SWORD structures rivers using points (**nodes**) spaced approximately 200 m apart, connected by line segments (**reaches**) around 10 km long (see Figure 2). These nodes and reaches serve as reference points from which SWOT data are obtained. SWORD also provides river width estimates at the node level, which are part of our dataset.

Sentinel-2 data from the European Space Agency (ESA) provides 10 m/pixel to 60 m/pixel multispectral imagery, depending on the spectral band (the red, blue, green, and near infrared (NIR) bands have 10 m/pixel resolution). We

download tiles that correspond to the same location and approximate time as PlanetScope imagery, while selecting tiles with the least cloud cover. Although water segmentation labels are available from WorldCover (Van De Kerchove et al. 2021), these only provide data for 2020 and 2021 with a 10 m/pixel resolution. There are 13 spectral bands available for Sentinel imagery. We include all available spectral bands as part of our dataset.

PlanetScope has a constellation of satellites collecting 3 m/pixel resolution image data with a revisit frequency of 1 day. Each image has 4 bands corresponding to the red, green, blue, and NIR bands. Images of size 500×500 pixels were collected from 2023 to 2024 to be within the same time frame (± 12 hours) as SWOT. River sites spanning a broad range of widths were selected using width estimates from other sensors (e.g., SWORD) and were limited to SWOT's fast sampling orbit (i.e., the path where SWOT makes frequent measurements) to maximize the number of spatially intersecting samples. Figure 1 shows the distribution of the obtained 1,145 images from different geographic locations. Images with no cloud cover were selected to ensure optimal visibility of rivers. We manually label water pixels in all these images to cover an area of 2,577 km^2 (see § 3.2). We publicly release normalized versions (i.e., min-max normalized pixel values per image) of the multispectral data. Links to the original PlanetScope products with raw values are also available, and can be purchased directly from (PBC 2024). Models presented here are trained on the normalized data.

3.2 PlanetScope Data Labeling

Unlike Sentinel images, PlanetScope images provide higher temporal and spatial resolution, enabling more detailed fine-grained image analyses. Consequently, these higher-resolution images were selected for manual annotation. Each PlanetScope image was manually annotated by one of 15 hydrology and river experts using a multi-scale pixel-wise annotator (Tangseng, Wu, and Yamaguchi 2017). Annotators were provided with the RGB image of the river to label, corresponding SWORD reaches and nodes overlaid on the images, a zoomed-out contextual view surrounding the target area, and a Google Maps link of the location. These supplementary information were included to enhance context and accuracy in labeling. Annotators also had interactive abilities such as zooming in and out of the target area, and the option of using a variety of labeling tools. Labels can be provided using a paintbrush, a polygon selector, and a superpixel selection tool that clusters similar pixels together. These interactive features streamlined the annotation process, ensuring comprehensive and efficient label coverage. A detailed description and visualization of the tool is provided in § A.3 (Appendix).

Each annotator provided labels for all water pixels in the image, differentiating between river and non-river water. However, since the primary objective was detailed river analysis, most collected satellite imagery predominantly featured river water. Thus, our focus was on accurate segmentation of all water pixels to facilitate further river-specific analyses. Annotating each image required at least 5 minutes, with more complicated river configurations taking more time, totaling more than 100 hours of expert annotation effort for all collected PlanetScope images.

3.3 Data splits

We split the data geographically so the training data cover different areas than the validation and test. River reaches (i.e., 10 km river sections) were separated so that each river appears exclusively in one data split. Specifically, the dataset contains 164 unique river reaches in the training set, 23 in the validation set, and 48 in the test set. The splits were constructed to have a similar average number of water pixels per image, ensuring balanced representation (see Figure 1).

3.4 Tasks

Using the available data in RiverScope, we evaluate performance of existing architectures and pretraining methods on water and river-specific tasks. In particular, we investigate (1) water segmentation performance and (2) river width estimation performance of existing models.

Water segmentation. We investigate water segmentation performance on high-resolution satellite imagery using the labeled PlanetScope images. RiverScope models are trained on the PlanetScope train set, with hyperparameter selection done using F1 score on the corresponding validation set. All models are then evaluated on the held out PlanetScope test set. To evaluate Sentinel models, we predict on the corresponding Sentinel images taken in the same location and time as the PlanetScope test images. The F1 score metric is used to evaluate model accuracy, since it effectively accounts for class imbalance caused by the relatively small proportion of water pixels (around 20%, see Figure 1).

River width estimation. For each node defined in SWORD, the ground truth river width is derived from PlanetScope water segmentation labels (see Figure 2). Given a satellite image, the corresponding river reach that spatially intersects the image is retrieved from SWORD. For each node along the river reach, the local slope is estimated, and a perpendicular (orthogonal) line is computed. River widths are then calculated by counting the number of water pixels along this orthogonal line and multiplying by the image resolution (3 m/pixel for PlanetScope) (Yang et al. 2019). These measurements serve as the ground truth river widths for SWORD nodes and provide the reference for evaluating other methods.

To evaluate performance, we use bias, % bias, mean absolute error, and median absolute error. The bias is computed by subtracting the ground truth widths y_i from the predicted widths $\hat{y}_i$ and getting the average ($\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$). The % bias is computed by dividing the ground truth width for each prediction: $\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{y}_i - y_i}{y_i}$. The mean and median absolute errors are computed as the mean and median of $|y_i - \hat{y}_i|\ \forall i$, respectively. Model predictions are compared against width estimates from Landsat (via SWORD), Sentinel, and SWOT across all river nodes. The evaluation is limited to rivers with widths of 500 m or less since the labeled images are 500x500 pixels in size and do not reliably capture wider rivers. A total of 445 nodes were used exclusively for evaluation.

4 RiverScope Models

4.1 Training Details

We evaluate 27 models on RiverScope tasks based on varying segmentation models, backbones, and pretraining methods (Iakubovskii 2019) to see the effect of training with our dataset across different settings. We experiment on four semantic segmentation models covering different ways of handling features and multi-scale information: FPN (Long, Shelhamer, and Darrell 2015), DeepLabv3 (Chen et al. 2017), UNet (Ronneberger, Fischer, and Brox 2015), DPT (Ranftl, Bochkovskiy, and Koltun 2021). For each segmentation model, we then experiment with different backbones and available pretrained weights which are the basis of extracting features for each segmentation model. FPN, DeepLabv3, and UNet are applied to CNN and Swin-based backbones, while DPT is applied to ViT-based backbones. ResNet50 (RN50) (He et al. 2016), MobileNetv2 (MV2) (Sandler et al. 2018), Swin-T (Liu et al. 2021), Swin-B (Liu et al. 2021), ViT-B/16 (Dosovitskiy et al. 2021), and ViT-L/16 (Dosovitskiy et al. 2021) backbones are used. Pretraining methods also vary from supervised methods using SatlasNet (Bastani et al. 2023) and ImagetNet1k (Deng et al. 2009), and self-supervised methods using SeCo (Manas et al. 2021), MoCov3 (Chen, Xie, and He 2021), CLIP (Radford et al. 2021), Prithvi (Jakubik et al. 2023), and DINO (Caron et al. 2021). Among these, SeCo, SatlasNet, and Prithvi use satellite images for pretraining. Each of these configurations serves as the starting point before further fine-tuning with RiverScope.

All models are trained on the PlanetScope labeled images from RiverScope. Binary cross entropy loss (Eq. 1) is used to train all models. A predicted segmentation mask $\hat{\mathbf{Y}}_i \in \mathbb{R}^{W \times H}$ with pixels $\hat{y}_{i,(j,k)}$ is compared with the ground truth mask $\mathbf{Y}_i \in \mathbb{R}^{W \times H}$ using Eq. 1. The Adam optimizer is used for training with the learning rate chosen from 10^{-1} to 10^{-6} based on the validation set performance.

$$\mathcal{L}_{bce}\left(\mathbf{Y}_i, \hat{\mathbf{Y}}_i\right) = \frac{1}{WH}\sum_{j=1}^{W}\sum_{k=1}^{H} -\Big(y_{i,(j,k)} \log \hat{y}_{i,(j,k)} + \left(1 - y_{i,(j,k)}\right) \log\left(1 - \hat{y}_{i,(j,k)}\right)\Big) \quad (1)$$

4.2 Baseline Methods

To evaluate (1) the effect of training on high-resolution satellite images and (2) the advantage of using pretrained models on the tasks, we compare results on several existing baselines. For (1), we compare performance against the same models trained on Sentinel data which have lower resolution. For (2), we compare on two models: a recent river width estimation model trained on PlanetScope data by Valman et al. (2024), and a widely used water segmentation algorithm NDWI (McFeeters 1996).

Sentinel models are trained on sampled Sentinel data from 2020-2021 with WorldCover (Van De Kerchove et al. 2021) labels for water segmentation. The same segmentation models, backbones, and pretraining methods from § 4.1 are also used to evaluate the effect of using low-resolution data on the downstream tasks. Multispectral data are used as input for all Sentinel models. We include more details in the Appendix.

PlanetScope CNN (Valman et al. 2024) is a recent work that similarly labeled rivers from PlanetScope images. They use a randomly initialized custom CNN that was trained to segment water from satellite images, and was designed to take four bands as input.

PlanetScope NDWI (McFeeters 1996) uses the Normalized Difference Water Index (NDWI) to detect water from satellite images, which is widely used in remote sensing. NDWI is a value between -1 and 1 that uses the difference in the green and NIR reflectance values to detect water (Eq. 2), since water reflects almost no NIR. To find the threshold t for water (i.e., NDWI $> t$), Otsu thresholding (Otsu et al. 1975) is applied.

$$\text{NDWI} = \frac{\text{green} - \text{NIR}}{\text{green} + \text{NIR}} \quad (2)$$

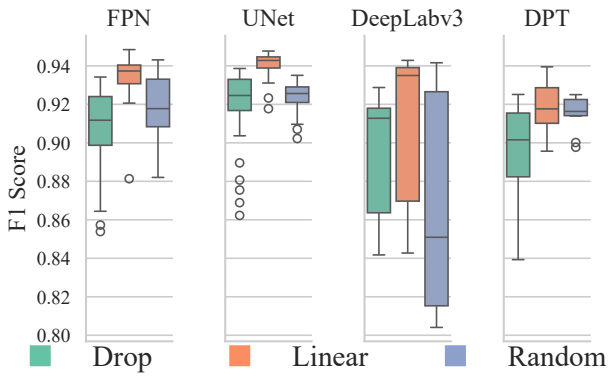


Figure 3: Segmentation performance of different methods for adapting a 4-channel satellite image to RGB to utilize existing RGB pretrained models. 'Drop' refers to dropping the NIR channel, 'Linear' refers to applying a linear layer to convert 4 channels to 3 channels, and 'Random' refers to training 4-channel models without any pretraining applied. We find that linearly projecting the input from 4-channels to 3-channels worked best (raw numbers in Table A1 (Appendix)).

5 Experiments

5.1 Water Segmentation

Linear adaptation of multispectral data to RGB pretrained models yields optimal performance. Since RiverScope labeled image data is composed of four channels, we evaluate the most effective way to use pretrained models that take three-channel RGB images as input. We look into (1) dropping the additional NIR channel to effectively keep only RGB (**Drop**), (2) linearly projecting the 4-channel input to 3-channel (**Linear**), and (3) training 4-channel models from scratch without using pretrained models (**Random**). Figure 3 shows the performance across these different settings. Our results indicate that the linear adaptor yields the best performance in terms of average F1 score across different architectures, outperforming other methods. There is also generally less variance in the performance of 'Linear'. This suggests that leveraging RGB pretrained models leads

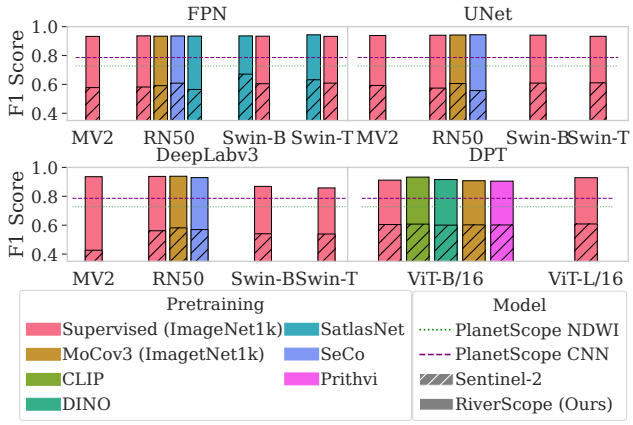


Figure 4: RiverScope trained models more accurately segment river pixels compared to Sentinel trained models. Each subplot shows the average F1 score improvement of a given segmentation model across multiple runs. For each architecture and pretraining combination, hatched bars represent the performance of Sentinel trained models, while solid bars represent the performance of RiverScope trained models. We show raw numbers in Table A4, A3 (Appendix).

to better performance than training from scratch, even when using additional channels. Based on these findings, we adopt the linear adaptor for all subsequent experiments.

Water segmentation improves by training on high-resolution data. Figure 4 displays the difference in performance when models are trained on high-resolution RiverScope data compared to models trained on low-resolution Sentinel data. RiverScope's best model is a SeCo pretrained RN50 UNet, whereas Sentinel's is a Satlas pretrained Swin-B FPN. Performance is likely positively affected by the presence of more details in high-resolution images (Figure 2) that enable models to better distinguish between water and non-water. For example, sand bars around rivers could be better seen if more details were present in the image. In addition, since high-resolution images can more precisely define the boundaries of water bodies, models trained on high-resolution data can learn these boundaries with better accuracy. This is supported by additional results in Table A4 (Appendix) showing the very low precision of Sentinel models (see Figure A5 (Appendix) for visualizations).

Using pretrained models improve segmentation performance. While PlanetScope CNN and PlanetScope NDWI also use high-resolution multispectral images, their performance is still lower compared to any of the RiverScope models that utilize transfer learning (Figure 4). Due to the rule-based nature of NDWI, it is sensitive to man-made land features and generally overestimates water boundaries which leads to falsely identifying pixels as water (Xu 2006) (see Table A3, Figure A4 (Appendix)). PlanetScope CNN tends to miss a significant number of water pixels, resulting in more false negatives and lower recall (see Figures A4, A5 (Appendix)). Pretraining on large datasets such as ImageNet and SatlasPretrain can speed up the process of learning image features and result in better performance.

PS†		Mean	Median	Bias	% Bias
	SWOT	94.4	41.9	28.2	38.0
	GRWL (SWORD/Landsat)	77.0	45.0	-6.9	25.0
	Sentinel	152.8	39.0	119.8	202.1
✓	NDWI (McFeeters 1996)	194.6	51.0	160.2	176.3
✓	CNN (Valman et al. 2024)	87.2	30.0	39.3	14.0
✓	RiverScope (Ours)	**15.3**	**7.2**	**5.7**	**11.4**

Table 1: River width estimation errors (m) for the best performing Sentinel (Satlas pretrained Swin-B FPN) and RiverScope (ImageNet pretrained RN50 FPN) models compared to other baselines. The RiverScope trained model has the lowest errors overall. Raw numbers are in Table A6 (Appendix). †PS: PlanetScope is used as input

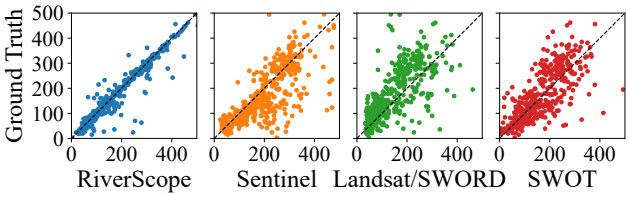


Figure 5: Distribution of width estimates. The RiverScope model predicted widths that cluster closely to the $y = x$ line.

RiverScope and Sentinel trained models can be improved by reducing false positives. Table A4 (Appendix) shows that for both types of trained models, precision trails behind recall, confirming that false positives—not missed detections—are the performance bottlenecks. Although a Sentinel model generally performs poorly compared to a RiverScope model, the former's recall is competitive, indicating that it can locate water pixels properly; it mainly struggles with precision. As detailed in Table A5 (Appendix), most false positives come from 'Tree Cover', 'Grassland', and 'Herbaceous Wetland' which typically occur along the boundary of rivers, and can look similar to water. Augmenting the dataset with additional examples of these three land cover types could potentially improve both Sentinel and RiverScope trained models. Alternatively, aggregating predictions across multiple dates for the same location might resolve false positives, since it is likely that a falsely identified water-like pixel would only appear as water in one time snapshot.

5.2 River Width Estimation

RiverScope trained models have lower river width estimation error than their low-resolution counterparts. Table 1 shows the RiverScope trained model has the lowest error, while Figure 5 shows it overpredicts less than other models. This improvement can be linked to the better water segmentation performance of the RiverScope model (Figure 4). With more accurate identification of river waters (as a result of the finer details and the more precise water boundaries in high-resolution satellite images), the estimated widths are also likely to be more precise. Figure 6 shows comparisons

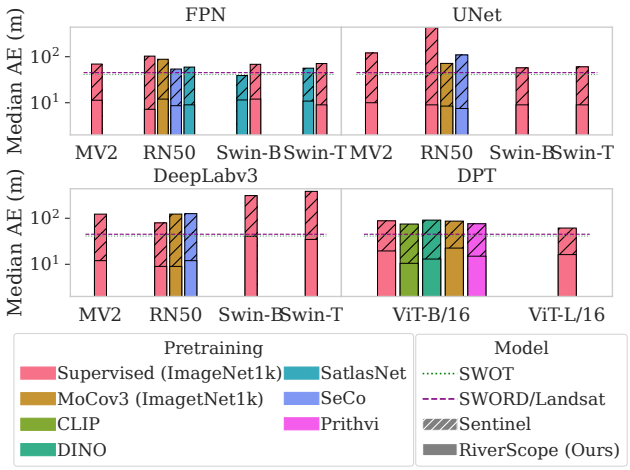


Figure 6: RiverScope trained models improve performance on river width estimation. Each subplot shows the median absolute error (m) for a given segmentation model. For each architecture and pretraining combination, the hatched bar is the performance of a Sentinel trained model, while the solid bar is of a RiverScope trained model. We additionally show estimates from sensors SWOT and Landsat. Across the board, we see significantly lower error when using a model trained with high-resolution RiverScope images. We show the raw numbers in Table A6 (Appendix).

across different architectures and settings, showing that in all configurations, training with RiverScope leads to superior river width estimation performance.

Higher-resolution training yields more accurate predictions. The Riverscope model trained on 3 m PlanetScope imagery (highest resolution among models in Table 1), achieves the best width estimation performance. River width is estimated by counting water pixels along the orthogonal (Figure 2), so a one pixel misclassification at each riverbank side gets an error of $2\sqrt{2}\Delta$ if the orthogonal cuts the pixel diagonally (2Δ if edge-aligned), where Δ is a sensor's resolution. For Sentinel this is 28 m (20 m if edge-aligned), so the observed 39 m median absolute error (Table 1) implies roughly 2 pixels of error. PlanetScope's 3 m/px resolution lowers this to 8.5 m (6 m), and our trained model achieves 7.2 m error—essentially a 1 pixel error. RiverScope models are not only more precise due to the increased spatial resolution, but also better at resolving misclassifications of water.

5.3 Future Directions

Sentinel predictions could potentially be further improved by removing outliers. While Sentinel has higher resolution than Landsat (10 m/px vs 30 m/px), the mean absolute error of Landsat (via SWORD) estimated widths is lower (Table 1). This may be attributed to the additional post-processing steps—such as basin-specific distribution fitting—applied to Landsat products, which can suppress outliers in the width estimates (Allen and Pavelsky 2018). Nonetheless, looking at median absolute error, a metric less sensitive to outliers, Sentinel widths still result in lower error compared to Landsat.

Combining SWOT and Sentinel predictions might offer a more robust approach for width estimation. Table 1 shows that SWOT and Sentinel achieve comparable median absolute error, but Sentinel underperforms in mean absolute error and bias. This bias stems from the high false positive rate in Sentinel's water masks as detailed in § 5.1. Tying Sentinel widths to co-located SWOT measurements might help lessen this overestimation and reduce errors whenever optical observations are degraded by shadows, haze, clouds, or sunglint (Zhu and Helmer 2018). Incorporating SWOT's KaRIn point cloud interferometry data—which provides geophysical data and radar echo power of samples—with Sentinel's multispectral data may provide a physically grounded constraint on rivers that can reject optical/spectral outliers (Lin et al. 2020). Exploring this direction in future work may yield more reliable river width estimates when the optical data are limited or compromised.

6 Limitations

While our dataset presents rivers of varying widths, most rivers come from the Northern Hemisphere, providing an opportunity to expand to rivers in the Southern Hemisphere. The PlanetScope imagery used in our dataset has high resolution, but it is a commercial product—purchasing requirements limit access to additional data. Future large-scale applications can use our work to evaluate the trade-off between cost and accuracy. Additionally, all optical sensors remain vulnerable to haze, sunglint, and cloud cover. § A.1 (Appendix) discusses these limitations in detail.

7 Conclusion

We introduced RiverScope, a high-resolution dataset bridging machine learning and critical challenges in hydrology. It provides expertly annotated 3 m/pixel PlanetScope imagery co-registered with public data from SWOT, SWORD, and Sentinel to benchmark different sensor capabilities for monitoring Earth's river systems.

Our benchmark establishes a new state-of-the-art for river width estimation, achieving a median error of 7.2 meters. This result also highlights a critical trade-off: while our RiverScope dataset is released publicly for research, scaling this high-accuracy approach requires purchasing additional PlanetScope imagery. This cost must be weighed against the lower performance of freely available Sentinel and Landsat data, which yield larger errors. We additionally show that using linear adaptors to adapt pretrained models for multispectral data is key to accurate segmentation.

By providing the first high-resolution benchmark for river width—an essential variable for estimating river discharge—RiverScope enables the development of more reliable hydrological models. This is a significant step towards robust monitoring of global river dynamics. We invite the machine learning community to use this resource to advance the state-of-the-art in remote sensing hydrology and create new multi-sensor approaches to characterize our planet's rivers.

Acknowledgements

We thank William Tooley for participating in the labeling of images. RD, TS, CG, and SM were supported in part by NASA grant 80NSSC22K1487. Additionally, RD and SM are supported by NSF grant 2329927, CG by NASA grant 80NSSC24K1646, and TR by a contract from the Jet Propulsion Laboratory.

References

Allen, G. H.; and Pavelsky, T. M. 2018. Global extent of rivers and streams. *Science*, 361(6402): 585–588.

Alsdorf, D. E.; Rodríguez, E.; and Lettenmaier, D. P. 2007. Measuring surface water from space. *Reviews of geophysics*, 45(2).

Altenau, E. H.; Pavelsky, T. M.; Durand, M. T.; Yang, X.; Frasson, R. P. d. M.; and Bendezu, L. 2021. The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A global river network for satellite data products. *Water Resources Research*, 57(7): e2021WR030054.

Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; and Kembhavi, A. 2023. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16772–16782.

Biancamaria, S.; Lettenmaier, D.; Pavelsky, T.; Cazenave, A.; Champollion, N.; Benveniste, J.; and Chen, J. 2016. Remote Sensing and Water Resources.

Bjerklie, D. M.; Birkett, C. M.; Jones, J. W.; Carabajal, C.; Rover, J. A.; Fulton, J. W.; and Garambois, P.-A. 2018. Satellite remote sensing estimation of river discharge: Application to the Yukon River Alaska. *Journal of Hydrology*, 561: 1000–1018.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional map of the world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6172–6180.

Claverie, M.; Ju, J.; Masek, J. G.; Dungan, J. L.; Vermote, E. F.; Roger, J.-C.; Skakun, S. V.; and Justice, C. 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219: 145–161.

Daroya, R.; Lucchese, L. V.; Simmons, T.; Prum, P.; Pavelsky, T.; Gardner, J.; Gleason, C. J.; and Maji, S. 2025. Improving Satellite Imagery Masking using Multi-task and Transfer Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Ellis, J. 2013. Sustainable surface water management and green infrastructure in UK urban catchment planning. *Journal of Environmental Planning and Management*, 56(1): 24–41.

ESA. 2022. Sentinel-1-missions-Sentinel online-Sentinel online. *Eur. Sp. Agency*.

Feng, D.; Gleason, C. J.; Yang, X.; and Pavelsky, T. M. 2019. Comparing discharge estimates made via the BAM algorithm in high-order Arctic rivers derived solely from optical CubeSat, Landsat, and Sentinel-2 data. *Water Resources Research*, 55(9): 7753–7771.

Filippucci, P.; Brocca, L.; Bonafoni, S.; Saltalippi, C.; Wagner, W.; and Tarpanelli, A. 2022. Sentinel-2 high-resolution data for river discharge monitoring. *Remote Sensing of Environment*, 281: 113255.

Flores, J. A.; Gleason, C. J.; Brinkerhoff, C. B.; Harlan, M. E.; Lummus, M. M.; Stearns, L. A.; and Feng, D. 2024. Mapping proglacial headwater streams in High Mountain Asia using PlanetScope imagery. *Remote Sensing of Environment*, 306: 114124.

Gleason, C. J.; and Hamdan, A. N. 2017. Crossing the (watershed) divide: Satellite data and the changing politics of international river basins. *The Geographical Journal*, 183(1): 2–15.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Iakubovskii, P. 2019. Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models.pytorch.

Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. 2023. Foundation models for generalist geospatial artificial intelligence, 2023. *URL https://arxiv.org/abs/2310.18660*.

Jones, J. W. 2019. Improved automated detection of subpixel-scale inundation—Revised dynamic surface water extent (DSWE) partial surface water tests. *Remote Sensing*, 11(4): 374.

Lin, P.; Pan, M.; Allen, G. H.; de Frasson, R. P.; Zeng, Z.; Yamazaki, D.; and Wood, E. F. 2020. Global estimates of reach-level bankfull river width leveraging big data geospatial analysis. *Geophysical Research Letters*, 47(7): e2019GL086405.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; and Rodriguez, P. 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9414–9423.

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.

Observation, E. R.; and Center, S. E. 2020. Landsat 8–9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2. Dataset.

Opher, T.; and Friedler, E. 2010. Factors affecting highway runoff quality. *Urban Water Journal*, 7(3): 155–172.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

PBC, P. L. 2024. Planet Application Program Interface: In Space for Life on Earth.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ronneberger, O.; Fischer, P.; and 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*, 234–241. Springer.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Sumbul, G.; Charfuelan, M.; Demir, B.; and Markl, V. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904. IEEE.

Tangseng, P.; Wu, Z.; and Yamaguchi, K. 2017. Looking at outfit to parse clothing. *arXiv preprint arXiv:1703.01386*.

Tian, Y.; Zheng, Y.; Wu, B.; Wu, X.; Liu, J.; and Zheng, C. 2015. Modeling surface water-groundwater interaction in arid and semi-arid regions with intensive agriculture. *Environmental Modelling & Software*, 63: 170–184.

USDA Farm Service Agency. n.d. NAIP Imagery. https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/.

Valman, S. J.; Boyd, D. S.; Carbonneau, P. E.; Johnson, M. F.; and Dugdale, S. J. 2024. An AI approach to operationalise global daily PlanetScope satellite imagery for river water masking. *Remote Sensing of Environment*, 301: 113932.

Van De Kerchove, R.; Zanaga, D.; Keersmaecker, W.; Souverijns, N.; Wevers, J.; Brockmann, C.; Grosu, A.; Paccini, A.; Cartus, O.; Santoro, M.; et al. 2021. ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data. In *AGU Fall Meeting Abstracts*, volume 2021, GC45I–0915.

Vinogradova, N. T.; Pavelsky, T. M.; Farrar, J. T.; Hossain, F.; and Fu, L.-L. 2025. A new look at Earth's water and energy with SWOT. *Nature Water*, 1–11.

Wang, Z.; and Vivoni, E. R. 2022. Mapping flash flood hazards in arid regions using CubeSats. *Remote Sensing*, 14(17): 4218.

Wasti, A.; Ray, P.; Wi, S.; Folch, C.; Ubierna, M.; and Karki, P. 2022. Climate change and the hydropower sector: A global review. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2): e757.

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.

Yamazaki, D.; Ikeshima, D.; Sosa, J.; Bates, P. D.; Allen, G. H.; and Pavelsky, T. M. 2019. MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research*, 55(6): 5053–5073.

Yang, X.; Pavelsky, T. M.; Allen, G. H.; and Donchyts, G. 2019. RivWidthCloud: An automated Google Earth Engine algorithm for river width extraction from remotely sensed imagery. *IEEE Geoscience and Remote Sensing Letters*, 17(2): 217–221.

Yang, Y.; Zheng, J.; Zhang, H.; Chai, Y.; Zhu, Y.; and Wang, C. 2022. Impact of the Three Gorges Dam on riverbed scour and siltation of the middle reaches of the Yangtze River. *Earth Surface Processes and Landforms*, 47(6): 1514–1531.

Yao, J.; Xu, N.; Wang, M.; Liu, T.; Lu, H.; Cao, Y.; Tang, X.; Mo, F.; Chang, H.; Gong, H.; et al. 2025. SWOT satellite for global hydrological applications: accuracy assessment and insights into surface water dynamics. *International Journal of Digital Earth*, 18(1): 2472924.

Zhao, T.-q.; Ouyang, Z.-y.; Wang, X.-k.; Miao, H.; and Wei, Y.-c. 2003. Ecosystem services and their valuation of terrestrial surface water system in China. *Journal of Natural Resources*, 18(4): 443–452.

Zhu, X.; and Helmer, E. H. 2018. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. *Remote Sensing of Environment*, 214: 135–153.