A globally sampled high-resolution hand-labeled validation dataset for evaluating surface water extent maps

Rohit Mukherjee[1, *], Frederick Policelli[2, *], Ruixue Wang[3], Elise Arellano-Thompson[3], Beth Tellman[3], Prashanti Sharma[3], Zhijie Zhang[3], and Jonathan Giezendanner[4]

[1]Pacific Northwest National Laboratory, Richland, WA, United States of America
[2]University of Arizona, Tucson, Arizona, United States of America
[3]NASA Goddard Space Flight Center, Maryland, United States of America
[3]Massachusetts Institute of Technology, Boston, Massachusetts, United States of America
[*]These authors contributed equally to this work.

Correspondence: Rohit Mukherjee (rohitmukherjee@live.com)

Abstract. Effective monitoring of global water resources is increasingly critical due to climate change and population growth. Advancements in remote sensing technology, ~~especially~~ specifically in spatial, spectral, and temporal resolutions, ~~have revolutionized~~ are revolutionizing water resource monitoring, leading to more frequent and high-quality surface water extent maps using various techniques such as traditional image processing and machine learning algorithms. However, satellite imagery datasets contain trade-offs that result in inconsistencies in performance~~. For example, the disparity~~ , such as disparities in measurement principles between optical (e.g. Sentinel-2) and radar (e.g. Sentinel-1) sensors, and differences in spatial and spectral resolutions among optical sensors. Therefore, developing accurate and robust surface water mapping solutions requires independent validations from multiple datasets ~~in order~~ to identify potential biases within the imagery and algorithms. However, high-quality validation datasets are expensive to build, and few contain information on water resources. For this purpose, we introduce a globally sampled, high spatial resolution dataset labeled using ~~3m~~ 3-meter PlanetScope imagery. Our surface water extent dataset comprises ~~of 90~~ 100 images, each with a size of 1024x1024 pixels, which were sampled using a stratified random sampling strategy ~~. We covered~~ covering all 14 biomes~~and also~~ . We highlighted urban and rural regions, lakes, and rivers, including braided rivers and ~~shorelines. To demonstrate the usability of our dataset, we evaluated our novel Sentinel-1 algorithm called the Equal Percent Solution (EPS) for~~ coastal regions. We evaluated two surface water extent ~~delineation. Our method produced an overall accuracy of 88% , with low commission error. However, EPS also had a high omission error. While investigating the source behind this issue using our hand labels, we found evidence that water signals in Sentinel-1 are affected by turbulence and muddiness. Further, mountainous regions distorted the signals from the water in river valleys leading to inaccuracies. Similar to our evaluation, we expect our dataset to be used for analyzing~~ mapping methods using our dataset - Dynamic World (Brown et al., 2022) based on Sentinel-2, and the NASA IMPACT model (Paul and Ganju, 2021) based on Sentinel-1. Dynamic World achieved a mean IoU of 72.16% and F1 score of 79.70%, while the NASA IMPACT model had a mean Intersection over Union (IoU) of 57.61% and F1 score of 65.79%. Performance varied substantially across biomes, highlighting the importance of evaluating models on diverse landscapes to assess their generalizability and robustness. Our dataset can be used to analyze satellite products and methods~~to gain~~ , providing insights into their advantages and drawbacks.

~~We expect our high-quality dataset to improve our understanding of the accuracy, spatial generalizability, and robustness of existing surface water products and methods to promote efficient monitoring of our natural resources~~Our dataset offers a unique tool for analyzing satellite products, aiding in the development of more accurate and robust surface water monitoring solutions.

1 Introduction

Mapping surface water is becoming increasingly ~~significant~~ important due to the ~~impact~~ impacts of climate change, as ~~several~~ many regions face the prospect of droughts (Dai, 2013) and floods (Tellman et al., 2021). Timely, accurate, and reliable monitoring of surface water extent is critical for better management, conservation, and risk reduction practices~~has been~~, but remains a growing challenge for researchers. Remotely sensed satellite data ~~products~~ have provided a unique vantage point for measuring surface water extent (Bijeesh and Narasimhamurthy, 2020; Mueller et al., 2016) using different measurement principles such as optical and radar sensors (Markert et al., 2018). Recent advances in satellite sensors have increased spatial, spectral, and temporal resolutions, ~~which led to a significant rise in interest in~~ leading to significant growth in methods for monitoring surface water using multiple satellite products ~~(Markert et al., 2020; Pekel et al., 2016)~~ (Pekel et al., 2016; Martinis et al., 2022; Giezendanner et al., 2023). Among these methods, machine learning and deep learning algorithms ~~have become increasingly popular in recent years as they are able to take advantage of~~ gained popularity due to their ability to leverage large volumes of satellite ~~products (~~data (both public and commercial) to accurately map the ~~earth~~Earth's surface ~~(Isikdogan et al., 2017; Martinis et al., 2022; Wieland et al., 2023)~~(Isikdogan et al., 2017; Wieland et al., 2023).

~~Even though there are several~~ However, the effectiveness of satellite water products based on ~~multiple sensors , their effectiveness~~ different sensors is not consistent across all conditions~~. Every satellite data product has several~~, as each product involves trade-offs between spatial, spectral, and temporal resolutions (Wulder et al., 2015). Higher spatial resolution ~~satellite products , e.g.~~ products like PlanetScope (PS) ~~,~~ often produce more accurate maps than lower resolution Sentinel-2 (10 ~~meter~~m) or Landsat 8 (30 ~~meter) (Acharki, 2022). Similarly~~m) (Acharki, 2022). Moreover, radar and optical sensors measure surface water properties ~~using different measurement principles~~ differently, leading to variations in accuracy and suitability (Martinis et al., 2022) even at similar spatial resolutions. The study by Ghayour et al. (2021) compared Landsat 8 and Sentinel-2 and found ~~performances to vary across multiple methods. Alternatively, Wolpert (2002) asserts that there is~~ performance varied across methods. As Wolpert (2002) asserted, no single algorithm ~~that is guaranteed to perform well in all situations. Li et al. (2022) compared widely used methods for surface water detection~~ is expected to perform optimally in every situation. The study by Li et al. (2022) summarizes the current common methods of water extraction based on optical and radar images.

~~Evaluating these~~ Independently evaluating satellite products and ~~surface water~~ methods using independent validation datasets is crucial ~~to increase trust~~ for increasing trust in the results (Bamber and Bindschadler, 1997). However, ~~these datasets are expensive to build and hence valuable, while existing datasets might~~ such datasets are resource-intensive to create and existing ones may not be suitable for all needs. For example, BigEarthNet ~~(Sumbul et al., 2019) contains close to~~ (Sumbul et al., 2019) contains around 600,000 multi-labeled Sentinel-2 image patches~~where 83, 000 of which~~, of which 83,000 contain water bodies. This dataset ~~will confirm~~ confirms the presence of water within a patch ~~, but will not identify each pixel . For a more high-resolution (1m) large-scale solution, the~~ but does not delineate it at the pixel level. The Chesapeake Conservancy Land Cover dataset ~~(Robinson et al., 2019) contains a water class labeled per pixel. LandCoverNet~~ (Chesapeake Bay Program, 2023) provides high-resolution (1 m) per-pixel water labels for the Chesapeake Bay watershed regional area. LandCoverNet (Alemohammad and Booth, 2020) contains global ~~data based on 10m~~ 10-meter resolution data from Sentinel-2 ~~data also containing~~ with a water class. ~~Apart from surface water, mapping floods have~~ Flood mapping has also been a strong research ~~direction with Sentinel-1 (S1) based NASA Flood Extent Detection dataset~~ focus, with datasets like the Sentinel-1-based NASA Flood Detection (Gahlot et al., 2021), Sen1Floods11 ~~-~~(Bonafilia et al., 2020), Sen12-Flood ~~-~~(Rambour et al., 2020), and C2S-MS Floods ~~datasets (Cloud to Street et al., 2022) with~~ (Cloud to Street et al., 2022) that use both optical (Sentinel-2) and radar (~~S1) . Although these datasets are~~ Sentinel-1) imagery. While suitable for validating surface water maps, ~~in some cases they are based on publicly available satellite products that are 10m in resolution , or in case of high spatial resolutionoptions, they are not globally distributed. Further, the~~ some of these datasets rely on 10-meter resolution public satellite imagery or lack global coverage at high resolution. The ephemeral nature of floods ~~requires a specific model for accurate detection even though flood water~~ also requires specialized detection models even though floodwater is technically surface water ~~(Bonafilia et al., 2020) .~~ (Bonafilia et al., 2020). Wieland et al. (2023) developed a semi-automated global binary surface water reference dataset ~~that contains~~ with 15,000 tiles (256 × 256 pixels) ~~. Their dataset is based on high spatial resolution data (around 1m) and uses a stratified random sampling technique providing a crucial dataset for benchmarking purposes~~sampled from high-resolution (≤1 m) imagery. However, this dataset uses weak labels generated by a model rather than manual labeling, making it less suitable for validation.

To ~~establish the~~ thoroughly evaluate a product's effectiveness and robustness~~of a product, there needs to be multiple independent evaluations~~ , multiple independent assessments are needed since high accuracy ~~scores~~ on one dataset ~~are not indicative of similar performances on other datasets. One reason is that any single dataset cannot be representative of~~ does not guarantee similar performance on others. No single dataset can fully represent the real world ~~(Paullada et al., 2021) . Similarly, hand-labeled datasets include subjectivity from labelers (Misra et al., 2016)which means that there could be no single ground truth label. In addition to multiple validation datasets, independent evaluations are necessary due to~~ (Paullada et al., 2021) and manual labels inevitably contain some subjectivity (Misra et al., 2016). Independent evaluations also help mitigate the issue of data leakage ~~(Vandewiele et al., 2021). Data leakage is experienced when researchers involve their validation set as a part of their trainingprocess which leads to an overfit model. Multiple and~~ , where the validation set is improperly used during model training, leading to overfitting (Vandewiele et al., 2021). Multiple independent validation datasets are therefore ~~required for thorough evaluation and increased~~ essential for comprehensively evaluating and building trust in remote sensing-based surface water products and methods.

In this ~~research~~study, we present ~~our hand-labeled~~ a high-quality, globally sampled, ~~high-spatial-resolution dataset based on 90 samples of 3m PS images , each of size 1024x1024 pixels~~high-resolution surface water dataset consisting of 100 hand-labeled 1024×1024 pixel PlanetScope images at 3-meter resolution. Our work builds upon existing satellite-based ~~remotely sensed datasets for~~ datasets for validating surface water extent~~validation. Our motivation for this work~~. The motivation is to provide a higher resolution ~~independent dataset to evaluate~~ hand-labeled dataset for evaluating surface water products ~~based on publicly available~~ derived from medium-resolution ~~datasets, such as~~ public satellites like Landsat and Sentinel ~~system of satellites. Our objective is to understand the advantages and drawbacks of each of these products and methods through our validation dataset. We expect our dataset to contribute towards an improved understanding of the accuracy~~ and commercial higher resolution Planet imagery. Our dataset addresses some of the limitations of existing datasets by providing pixel-level water hand labels at a higher resolution (3 meters) compared to some other datasets and encompassing diverse biomes and contexts (urban/rural, mountains/plains, rivers/lakes) for comprehensive evaluations. We evaluate two state-of-the-art surface water extent mapping methods using our dataset: the Dynamic World land use and land cover product based on optical Sentinel-2 imagery and the NASA IMPACT inundation mapping model based on radar Sentinel-1 data, which was the winning solution in a recent flood detection challenge. By applying our validation dataset to these products and methods, we aim to better understand their advantages and limitations. We anticipate our dataset will contribute to improved accuracy assessment, spatial generalizability analysis, and robustness evaluation of existing surface water products and methods~~to promote effective monitoring of our natural resources. In the following sections, we describe our sampling strategy in selecting our labels, explain the processing of PS imagery required for labeling, analyze our hand-labeled images, and demonstrate the usefulness of our dataset by evaluating our novel surface water mapping method using Synthetic Aperture Radar (SAR) imagery from the ESA S1 satellite constellation~~. These advancements can ultimately benefit by promoting more effective monitoring and management of water resources, especially in the face of climate change and population growth.

2 Data Preparation

2.1 Sampling

Our objective was to build a dataset that ~~is close to a true representation of~~ closely represents the true distribution of surface water features using only 100 samples. A representative dataset ~~will enable testing~~ enables testing the spatial generalizability and accuracy of surface water extent products~~for their spatial generalizability in addition to their accuracy. Since~~ . However, achieving a true representation is ~~a nearly impossibletask (Paullada et al., 2021), we approached this problem~~ nearly impossible (Paullada et al., 20 . We approached this challenge by sampling from different biomes, as defined by Olson et al. (2001), ~~as they have multiple climate~~ which encompass various climates and land conditions, giving a better chance of providing high variance within samples. ~~In addition to biomes, we balanced our samples in both urban and rural regions as urban regions have a higher density of built-up environments. Similarly, we highlighted lakes and rivers. We added some examples of braided rivers and shorelines to improve representation.~~

~~In addition to our stratified sampling approach, we incorporated randomness in our sampling~~We employed a stratified random sampling strategy to ensure the representativeness of our dataset. First, we created a ~~buffer of~~ 2 km ~~around shapefiles by World Wildlife Fund (2005) defining~~ buffer around global rivers and lakes ~~. Next, we~~ shapefiles provided by World Wildlife Fund (2005) using Quantum GIS (QGIS). We then clipped these buffers with the shapefiles of each of the 14 biomes. ~~Then~~Within each biome, we randomly placed 50 points ~~on each of the biomes and randomly~~ using QGIS's random point generator and selected at least 5 of them as samples. ~~Finally, to ensure that we address several~~

To address the various contexts in which surface water exists, we randomly selected ~~some samples within~~ additional samples from urbanized regions (Patterson and Kelso, 2012), braided rivers, and ~~shorelines. Table ??,~~ coastal regions. Urban areas are spatially heterogeneous, often resulting in increased complexity for water detection. We also separately sampled from lakes and rivers to ensure a balanced representation of both water body types. Braided rivers and coastal areas were included.

Figure 1 shows the number of samples for each biome~~. Figure?? shows the~~ , while Figure 2 illustrates the global spatial distribution of the samples~~globally.~~ . The number of samples from Tropical & Subtropical Dry Broadleaf Forests and Tropical & Subtropical Coniferous Forests ~~cover less area, the number of samples was therefore limited to ensure fair representation. Two-thirds~~ was limited due to their smaller area coverage. Approximately two-thirds of our labels are from rivers, ~~while the rest are lakes.~~ and the remaining one-third are from lakes. We sampled a larger portion from Deserts and Xeric Shrublands (16 samples) because water extraction methods generally perform worse in these regions, especially when using radar imagery (Martinis, 2017).

~~**Biome Number of Samples** Tropical & Subtropical Dry Broadleaf Forests 2 Tropical & Subtropical Moist Broadleaf Forests 12 Tropical & Subtropical Coniferous Forests 3 Temperate Broadleaf & Mixed Forests 8 Temperate Conifer Forests 7 Boreal Forests/Taiga 8 Tropical & Subtropical Grasslands, Savannas & Shrublands 7 Temperate Grasslands, Savannas & Shrublands 6 Flooded Grasslands & Savannas 5 Montane Grasslands & Shrublands 5 Tundra 8 Mediterranean Forests, Woodlands & Scrub 5 Deserts & Xeric Shrublands 9 Mangroves 5 Number of samples per biome selected from our stratified random sampling for a total of 90 samples.~~ The temporal distribution of our samples spans from 2021 to 2023, covering different seasons to capture seasonal variations in surface water extent. While our sampling strategy aimed to maximize representativeness within the constraints of labeling resources, we acknowledge that the limited number of samples (100) may not fully capture all global surface water variations.

During the sampling process, we implemented quality control measures to ensure that the selected locations were suitable for labeling and analysis. We downloaded the Planet scene for each location, divided the scene into 1024×1024 sized images, and then selected the image that contained sufficient water and no cloud cover.

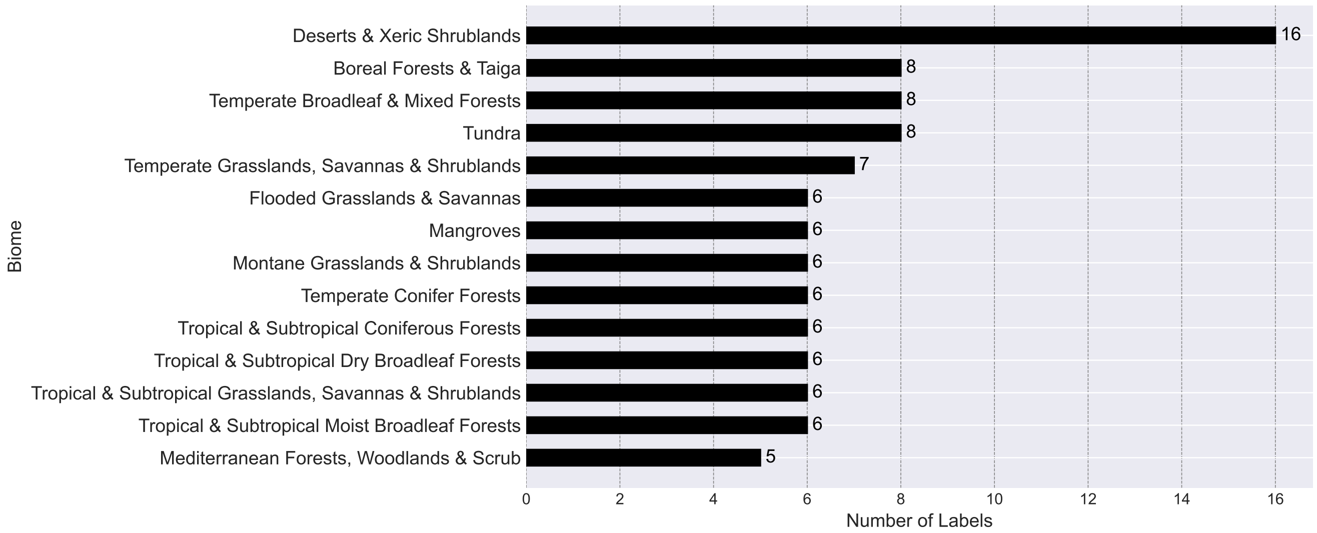

Figure 1. ~~Location~~ Distribution of ~~surface water labels~~ sampled ~~globally~~labels across different biomes. The bar chart illustrates the number of surface water labels ~~have been sampled to be representative~~ collected from each of ~~1)~~ the ~~diverse global~~ 14 biomes ~~(c~~defined by Olson et al. (2001). ~~f. table ??) and 2)~~ The sampling strategy aimed to ensure a balanced representation of surface water features across diverse ecological regions while accounting for the ~~global spatial distribution~~areal coverage of each biome.

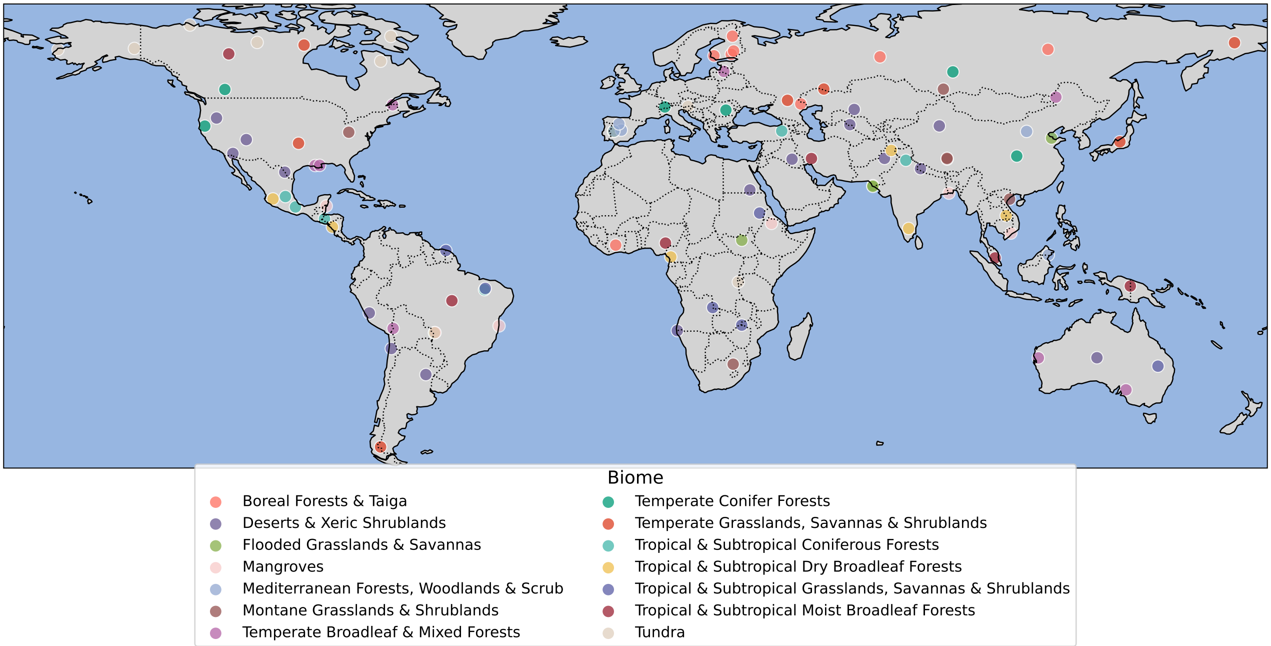
Biome
● Boreal Forests & Taiga
● Deserts & Xeric Shrublands
● Flooded Grasslands & Savannas
● Mangroves
● Mediterranean Forests, Woodlands & Scrub
● Montane Grasslands & Shrublands
● Temperate Broadleaf & Mixed Forests
● Temperate Conifer Forests
● Temperate Grasslands, Savannas & Shrublands
● Tropical & Subtropical Coniferous Forests
● Tropical & Subtropical Dry Broadleaf Forests
● Tropical & Subtropical Grasslands, Savannas & Shrublands
● Tropical & Subtropical Moist Broadleaf Forests
● Tundra

Figure 2. Global distribution of the 100 surface water labels sampled for the dataset. The map depicts the geographical locations of the sampled labels, which were sampled to represent diverse global biomes (refer to Table 1 for the number of labels per biome) and ensure a representative dataset of water features. The sampling approach also aimed to capture the variability of surface water features across urban areas, braided rivers, and coastal regions.

2.2 Data Processing

After ~~sampling 90~~ selecting 100 locations based on ~~various criteria~~our sampling strategy, we downloaded 8-band~~SuperDove PS imagery from the years~~ , 3-meter resolution SuperDove PlanetScope (PS) imagery from 2021 ~~and 2022 using our~~ to 2023 using our access to the NASA Commercial Smallsat Data (CSDA) Program~~access~~. As our objective ~~is~~ was to evaluate most ~~medium resolution~~ medium-resolution satellite sensors, including Sentinel-1 (S1), we ensured that the ~~loss of~~ failure of the Sentinel-1B satellite on December 23, 2021, did not create a large temporal gap between the label and the last available scene from the satellite. ~~Therefore, for locations found to be~~ For locations only covered by Sentinel-1B ~~satellite and not by~~ and not Sentinel-1A, we acquired PS scenes before the ~~date of failure , which is on Dec 23, 2021. The rest of the samples are all from 2022. In selecting these 90 scenes, we discarded~~ Sentinel-1B failure date.

During the scene selection process, we excluded areas with perennially frozen water. If a location contained seasonal ice, we ~~substituted~~ replaced that PS image with ~~an image from summer when~~ a summer image when the water was not frozen. This approach ensured that our dataset focused on liquid water surfaces, which are more relevant for surface water extent mapping.

From each larger PS scene, we ~~selected an image of~~ extracted a 1024x1024 ~~pixels~~ pixel image, covering an area ~~equal to~~ of approximately 9.4 square ~~km. We decided on keeping the labels large enough to ensure that the corresponding spatially coincident~~ kilometers. We chose 1024x1024 pixel images to ensure sufficient pixels and spatial context for comparison with medium-resolution ~~imagery from Landsat or Sentinel had sufficient pixels in the image for comparison . For example, 30m~~ imagery (e.g., Landsat, Sentinel). For instance, a 30-meter Landsat image corresponding to our labels ~~will be~~ would have around 100x100 pixels, while a ~~Sentinel at 10m will be~~ 10-meter Sentinel image would have approximately 376x376 ~~, where Landsat and Sentinel have enough spatial context~~pixels. Figure 3 showcases two examples of the PS images selected for labeling, displayed in False Color Composite (near-infrared, red, and green bands).

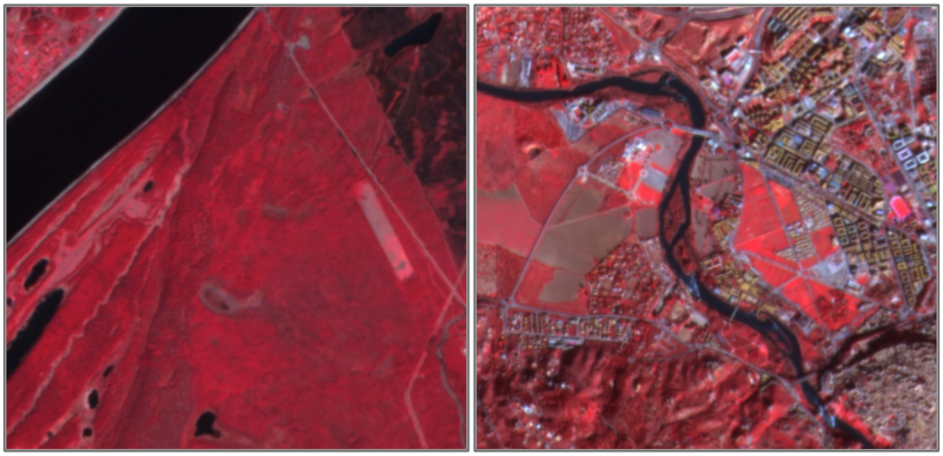

Figure 3. PlanetScope images selected for labeling are shown in False Color Composite (near infrared, red, and green). Left: Vilyuy River, Sakha Republic, Russia (SID09) and Right: Tagus River, Toledo, Spain (SID17).

2.3 Data Labeling

We ~~recognize that labeling data is a subjective process, and will thus contain certain biases from the labeler (Paullada et al., 2021) . Moreover, the labelers will inherit the biases from the data from which they are labeling. Our objective was to provide information as close to the ground truth as possible. One way to achieve this is to label using a higher spatial resolution dataset relative to the products that will potentially be evaluated. Therefore, we used PS data at 3m with a view to evaluating any satellite product that with lower spatial resolution, e.g., S1 at 10m~~used high-resolution 3-meter PlanetScope (PS) data for labeling, ideal for the evaluation of lower-resolution satellite products such as Sentinel-1 (S1), Sentinel-2 (S2) at 10 meters, or Landsat sensors at 30 meters.

~~We employed labelers with experience in analyzing and labeling surface waterin satellite imagery, and performed independent quality checks.~~ The labeling was performed by experienced analysts to distinguish between three classes: water, low-confidence water, and non-water. The water class represents areas with a clear presence of water, while the low-confidence water class marks pixels where the presence of water is uncertain but probable. The non-water class encompasses all other land cover types.

To assist the labelers, we ~~created a~~ provided true-color composite (TCC) and ~~a~~ false-color composite (FCC) ~~with~~ images using the near-infrared, red, and green bands, for each sample. ~~We explored several annotation tools for computer vision applications and decided to use Labelbox (Sharma et al., 2019) through an academic license. We found Labelbox to have highly efficient tools for creating quality labels.~~

~~Examples of PlanetScope imagery and corresponding labels (Top Row: Dong Tranh River, Ho Chi Min City, Vietnam, and Bottom Row: Siran River, Pakistan). The images are labeled with three categories: 1) non-water, 2) low-confidence water, and 3) water. The low-confidence water category marks pixels where delineating between water and no water is not straightforward, but the probability of water being present is high.~~

~~During labeling, we encountered several~~ In cases where the presence of water was ~~uncertain. However, whenever there was confusion~~ unclear in the PS imagery, we cross-referenced ~~with the~~ them with higher-resolution basemaps ~~provided by~~ from Bing and Google. ~~For features that were still not resolved, we introduced a ’~~Unresolved features were assigned to the low-confidence water ~~’ category. In total, we have three classes - ’water ’, ’low-confidence water’, and ’non-water’. However, during evaluation~~ category, ensuring that the water class only includes pixels with a high degree of certainty. During the evaluation process, the low-confidence water class can be ~~ignored since these corresponding features are not confidently identified, hence kept separate from the evaluation process . After~~ excluded or added to the water category as necessary.

To streamline the labeling process and ensure the creation of high-quality labels, we utilized the Labelbox platform (Sharma et al., 2019) , which provides efficient tools for data annotation. After the initial labeling, we performed ~~quality controls on each of the labels to ensure the accuracy of our labels.~~ several rounds of quality checks on each label to maintain accuracy and consistency across the dataset.

In total, ~~combining time for labeling and~~ we labeled 100 images, each with a size of 1024×1024 pixels, covering a total surface area of 940 square kilometers. The labeling process, including quality control, ~~we spent~~ took approximately 2 hours ~~for each image, equalling 180~~ per image, resulting in a total of 200 hours of work~~and leading to 204 square km of~~. The labeled surface water ~~from a total surface areaof 850 square km~~accounts for nearly 250 square kilometers of the total area. Each label is ~~provided a sampled~~ assigned a unique sample ID (SID) ranging from 1 to ~~90 and contains~~ 100 and includes the date (YYYYMMDD) of the PS image ~~from which it was labeled~~used for labeling.

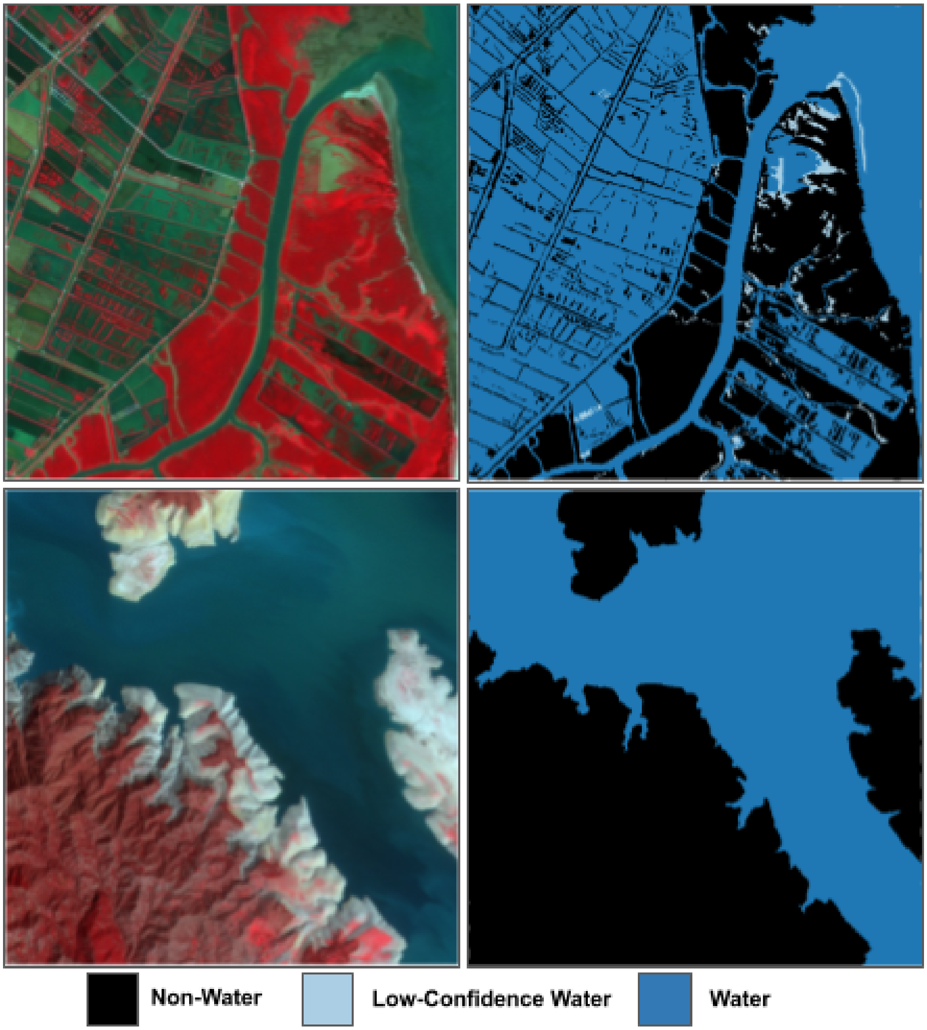

Figure 4. Examples of PlanetScope imagery and corresponding labels (Top Row: Dong Tranh River, Ho Chi Min City, Vietnam (SID46), and Bottom Row: Siran River, Pakistan (SID28)). The images are labeled with three categories: 1) non-water, 2) low-confidence water, and 3) water. The low-confidence water category marks pixels where delineating between water and no water is not apparent, but the probability of water being present is moderately high.

2.4 Dataset Analysis

We labeled a total of ~~90~~ 100 1024x1024 PS images at ~~3m, where water corresponds to 24% of all the~~ 3-meters, with the overall class distribution showing that covers 24.9% of the total surface area, ~~while~~ low-confidence water covers ~~1.3%~~1.2%, and the rest (~~74.6~~73.9%) is non-water (Fig ~~-~~5). ~~We have a well-distributed number of labeled water pixels per sample with an emphasis on images with lower percentages of water pixels since labels with a higher percentage of surface~~ The distribution of water ~~are relatively easy to delineate. We had a preference towards labeling more heterogenous landscapes but also included labels with~~

~~a lot of water to test the limits of satellite data products and mapping methods. Our labels covered~~ pixel percentages for each individual label, as displayed in Figure 6, demonstrates that most labels contain less than 50% water pixels by design, with the mean water surface area per label being 26.10 km^2. This focus on having more non-water area enables better delineation of water boundaries, as the water class itself tends to be more homogeneous and therefore less complex from both labeling and mapping perspectives.

As mentioned previously, our labeled dataset covers water surface areas across different biomes (Table 1). The mean percentage of water content per label varies substantially between biomes, from a low of 5.29% for Mediterranean Forests, Woodlands & Scrub to a high of 42.95% for Temperate Grasslands, Savannas & Shrublands. This demonstrates the diversity of landscapes and water coverage captured in our dataset. In total, our dataset provides 2609.78 km^2 of labeled water surface area, covering a variety of landscapes ~~–~~ such as rivers passing through urban regions, braided rivers in ~~the delta~~deltas, rivers passing through forests and agricultural fields, ~~including~~ and waterbodies in plain and mountainous regions. The diversity and representativeness of our dataset make it a valuable resource for testing the limits and robustness of satellite data products and mapping methods.

2.5 Dataset Structure

All 100 labels are in the GeoTIFF format with the UInt8 data type and a single band. Each pixel can contain 4 possible values: 0 (nodata), 1 (non-water), 2 (low-confidence water), and 3 (water). The labels are in the WGS84 (EPSG:4326) coordinate reference system. Each label has a corresponding PlanetScope image used for labeling in Labelbox. The PlanetScope images are also in the WGS84 (EPSG:4326) CRS and contain three spectral bands (red, green, and blue) in true color composite. Based on our PS image release agreement with Planet, we converted the original surface reflectance values to byte format with possible pixel values between 0 and 255, instead of UInt16.

The label files are named using the following convention: 'SIDX_YYYYMMDD.tif', where 'SIDX' is the unique sample ID (X ranging from 1 to 100) and 'YYYYMMDD' represents the date of the PlanetScope image used for labeling. The corresponding PlanetScope images follow the naming convention: 'SIDX_PSID.tif', where SIDX is the same as the label, but PSID is the original SuperDove PlanetScope image ID, allowing for the retrieval of the original surface reflectance values, provided there is access.

Our dataset is organized using the Spatio-temporal Asset Catalog (STAC) format, which is a standardized way to describe and catalog geospatial data. The STAC format provides a clear and consistent structure for storing and accessing the labels and their corresponding PlanetScope images, along with relevant metadata.

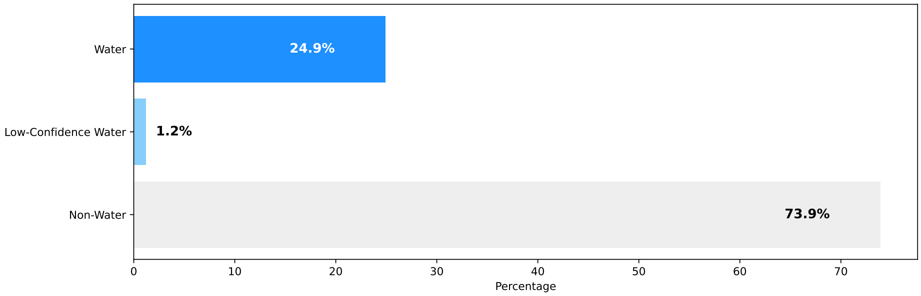

Figure 5. Class distribution across Labels (non-water, low-confidence water, and water) for all chips. Non-water class shares the largest percentage as it encompasses the water class. Low-confidence water pixels are only a minor percentage.

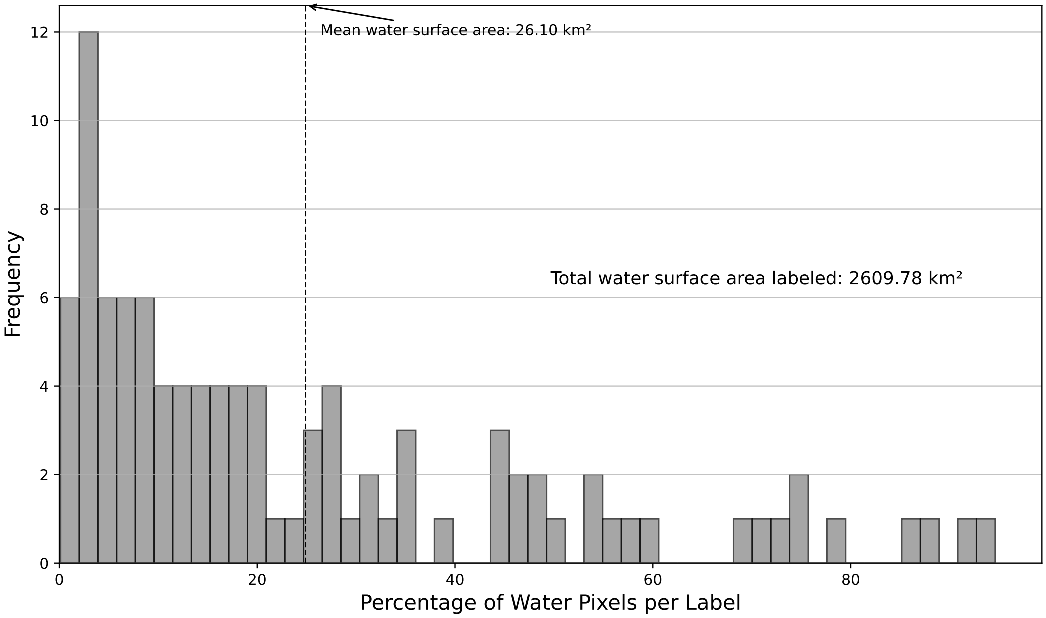

Figure 6. Distribution of water pixels per sample. The figure shows the percentage of water pixels within one sample. Most samples contain less than 50% of water by design, as the focus is to delineate the boundaries since the water class is more homogeneous, therefore, less complex.

3 ~~Evaluating surface water mapping method using our hand-labeled dataset~~

~~To demonstrate the use of our dataset, we evaluate a novel surface water mapping method called the Equal Percentage Solution (described below)based on S1 imagery. First, we generated watermaps using our method using S1 imagery and evaluated these watermaps against our labels . Apart from evaluating the maps quantitatively, we also visually compared the water maps with the original PS data to understand the possible reasons behind the performance.S1 radar imagery has the advantage of not being~~

Biome	Mean Water Content per Label %
Boreal Forests & Taiga	22.48
Deserts & Xeric Shrublands	18.96
Flooded Grasslands & Savannas	27.45
Mangroves	40.75
Mediterranean Forests, Woodlands & Scrub	5.29
Montane Grasslands & Shrublands	23.71
Temperate Broadleaf & Mixed Forests	19.48
Temperate Conifer Forests	6.55
Temperate Grasslands, Savannas & Shrublands	42.95
Tropical & Subtropical Coniferous Forests	16.80
Tropical & Subtropical Dry Broadleaf Forests	20.71
Tropical & Subtropical Grasslands, Savannas & Shrublands	11.96
Tropical & Subtropical Moist Broadleaf Forests	27.39
Tundra	30.77

Table 1. Mean percentage of water content per label across different biomes. The table shows the average proportion of water pixels within the labeled samples for each biome, highlighting the variability in water coverage across diverse ecological regions.

~~significantly affected by clouds which increases its overall data availability compared to optical sensors such as Landsat 8/9 and Sentinel-2. We downloaded S1 data using Alaska Satellite Facility's Vertex platform and HyP3 package.In the next two subsections, we explain our surface water mapping method, then we analyze the S1 data and evaluate the results.~~

3 Evaluating surface water mapping methods using our hand-labeled dataset

3.1 ~~The Equal Percentage Solution~~

~~Our approach to water mapping with ESA S1 SAR scenes begins with downloading the VV-polarization scenes from the Alaska Satellite Facility (ASF) and pre-processing the VV scenes (Meyer et al.~~ We evaluated two surface water mapping methods based on an optical and a radar satellite imagery product to demonstrate the use of our validation dataset. We used standard metrics for classification - Precision, Sensitivity, Specificity, F1, IoU, and Accuracy for evaluating the two surface water maps. We measured their performance across each biome and their overall performance.

3.1 Performance of Sentinel-2 based Dynamic World on detecting surface water

Dynamic World (DW) is a land use land cover product from Google that utilizes a deep learning model trained on their own labeled dataset. The product includes 9 classes, ~~in preparation). After pre-processing, as per Twele et al. (2016) and using~~

~~code implemented by the Alaska Satellite Facility (Meyer et al.~~ including water, ~~in preparation), we divide the scene into 200x200 pixel subscenes and order the subscenes according to their backscatter variability. High variability in subscenes is an indication that multiple land cover classes are likely in the subscene, with a significant probability that these will include both water~~and ~~not-water classes. Next, (as per Meyer et al., in preparation), we eliminate the subscenes with the highest and lowest 5-percentile variability; this is done to mitigate the possibility of selecting scenes with a high level of anomalies. Uniquely to our algorithm (to the best of our knowledge) we then filter~~produces a map for every Sentinel-2 image. Each Sentinel-2 image is post-processed and cloud-removed. We downloaded Sentinel-2 images within 3 days of each of the ~~selected sub-scenes to include only those exhibiting bi-modal behavior and then model the distribution of each resulting sub-scene as a bi-modal Gaussian distribution. From this model, we derive a threshold at which the percentage of false positives and false negatives are expected to be equal. This is done by selecting an essentially random threshold (-15.5 dB) as a starting place, calculating the percentage of expected false positives and false negatives (see Fig. ??), and moving the threshold until the expected percentage of false positives is equal to the expected percentage of false negatives (within +/- 0.1 dB)~~ 100 labeled PlanetScope images. We also applied a Not-a-Number (NaN) filter, ensuring that images with at least 90% valid pixels are considered. After applying the temporal and NaN filters, there were 53 corresponding Sentinel-2 based DW maps out of our 100 labels. From each DW map, we extracted the first band, which contains the water class. Each DW class contains continuous values between 0 and 1, where 1 denotes the highest confidence in the model prediction. We converted the continuous values to binary, thresholding at 0.3. The water class is one of the least confused classes in the DW product, so mixed pixels are less likely. Finally, we evaluated DW on our labels. Note that for evaluation, we converted the low-confidence water class to water. We finally resampled the DW water class to match the resolution of the labels at 3-meters using nearest neighbor interpolation before evaluating. Note that for evaluation, we merged the low-confidence water class with water. Therefore, labels were either 0 (non-water) or 1 (water).

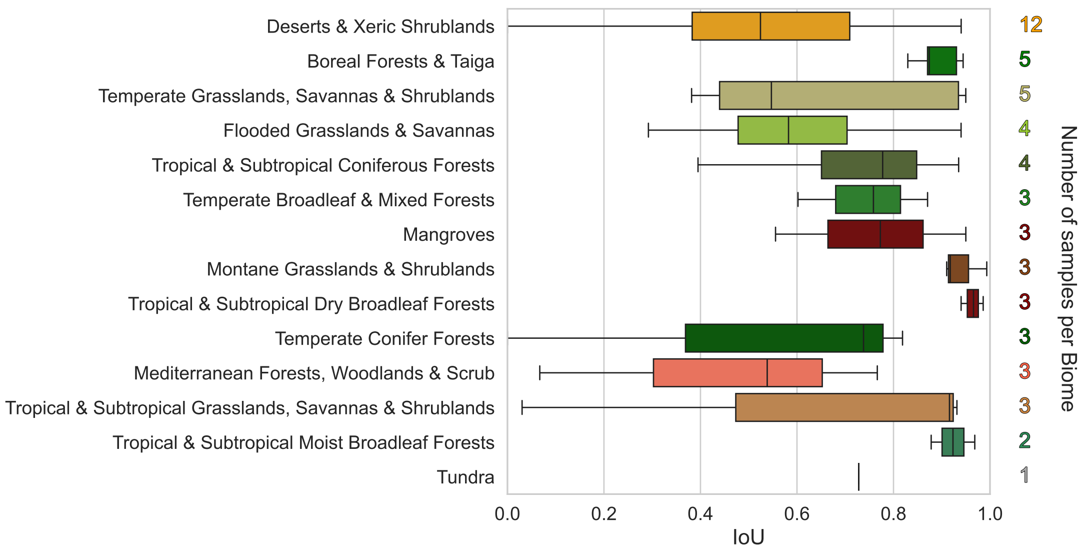

Figure 7. ~~Visualization~~ Intersection over Union (IoU) performance of the ~~thresholding algorithm based~~ Dynamic World (DW) water class across different biomes. The number of samples per biome is shown on the ~~Sentinel-1 VV band: a threshold value is chosen to identify the valley~~ right of each bar. Higher IoU scores suggest better performance in ~~the bimodal distribution to differentiate between the~~ detecting surface water~~and non-water classes~~. ~~In this given schematic representation,~~ The error bars represent the ~~number~~ standard deviation of ~~water pixels dominates~~IoU scores within each biome.

~~If the threshold is between -38.332 dB and -10.278 dB, we add this threshold to a list of candidate thresholds for the entire scene. The upper and lower boundary values for the threshold were established by evaluating the statistics of the S1 pixels corresponding to near-coincident optical data hand-labeled as water in~~ Figure 9 illustrates the performance of the ~~evaluation process (discussed more below). For those pixels evaluated as water, a bimodal, near-Gaussian distribution of the backscatter values was found. On closer inspection, it was found that the higher mode corresponded frequently with rough water , but also other anomalies mentioned below. A backscatter value from a single polarization band cannot be used to distinguish such features from land, so we used the statistics (specifically 3 standard deviations) for the lower distribution to set upper and lower bounds for the subscene thresholds.~~ water class in the Dynamic World product across different biomes using IoU. IoU provides an assessment of the overlap between the predicted and ground truth water pixels, with higher values indicating better performance. The number of samples per biome varies, with some biomes having more representative data than others. For biomes with a larger number of samples, such as Deserts & Xeric Shrublands and Boreal Forests & Taiga, the IoU scores provide a more robust evaluation of the DW water class performance. Despite the variations in sample size, notable differences in performance can be observed among the biomes. It is important to note that the IoU metric is influenced by the amount of water present in each label. Higher water percentage often leads to higher IoU. However, our dataset has an average of 26.1% surface water pixels, providing a balanced assessment of the DW water class performance.

~~As per Twele et al. (2016), we then repeat this process starting with the subscenes meeting the requirements above with the highest variability and proceeding with subscenes having successively lower variability. When a maximum of 10 candidate thresholds have been found, we take their median value (similarly to Twele et al. (2016) and Meyer et al. , in preparation). This median threshold is then used as the candidate upper threshold for water identification in the full scene. Similar to ASF, if all subscenes meeting the above requirements have been inspected and 10 candidate thresholds are not found, but a minimum of 5 are found, we still consider this a representative sample and use their median value as the threshold for the full scene. If less than 5 candidate thresholds are found, a default value of -15.5 dB (as per Meyer et al., in preparation) is used.~~

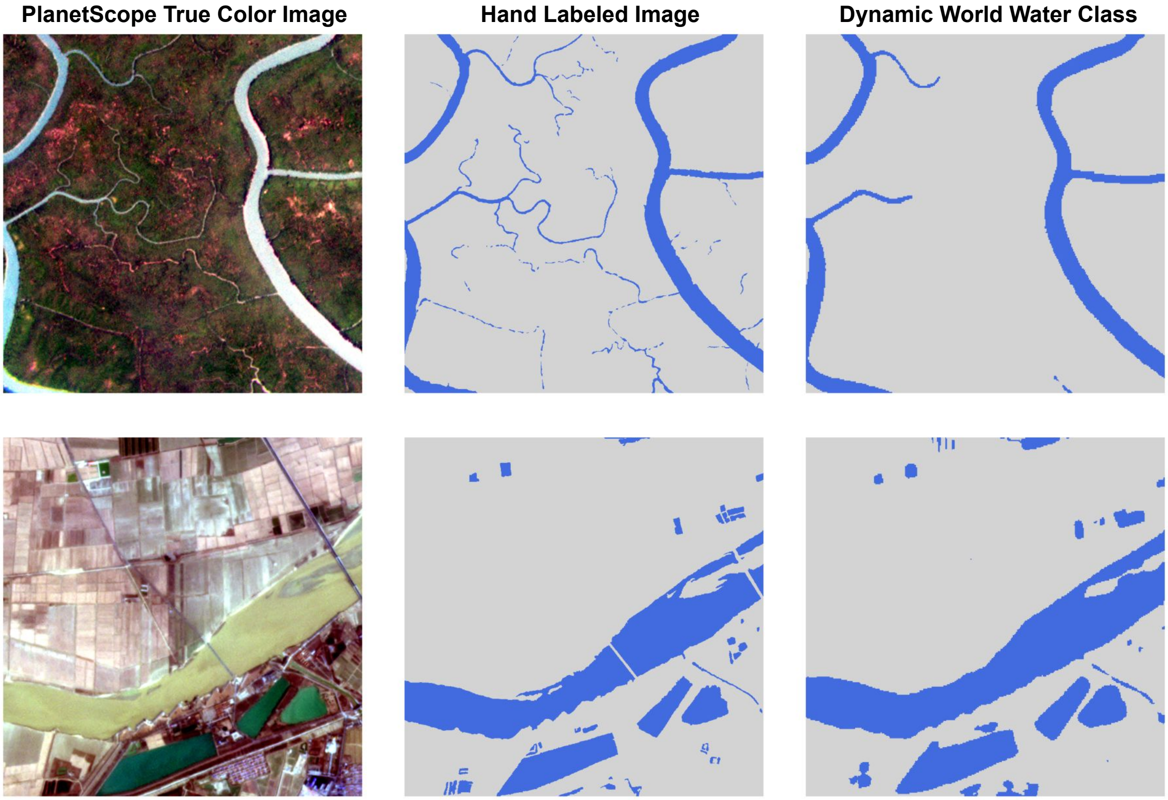

Figure 8. Comparison of a PlanetScope true color image (left), the corresponding hand-labeled image (middle), and the Dynamic World water class prediction (right). Top: Sundarban National Park, Bangladesh (SID01), Bottom: Shandong, China (SID13).

~~During our evaluation of the resulting water maps (discussed below)~~ Figure 8 provides a visual comparison of the Dynamic World water class predictions with the hand labels for two locations: Sundarban National Park, Bangladesh (SID01) and Shandong, China (SID13). The DW product appears to capture the majority of the water pixels accurately, ~~the percentage of false positives was found to not be the same as the percentage of false negatives,~~ however, ~~when we modified the algorithm to require the percentage of false positives to be equal to the percentage of false negatives minus an offset percentage, the evaluation results trended toward equal false positives and false negatives (see Fig. ??) . Equal false positives and negatives~~

~~would allow the model to produce more balanced water maps that do not tend to overpredict or underpredict~~it misses the narrow rivers (SID01) and the incorrectly ignores two bridges (SID13).

Table 2. Performance metrics for the Dynamic World (DW) water class evaluated on our hand-labeled dataset. The table presents the mean and standard deviation of various metrics. IoU denotes Intersection over Union. Higher values indicate better performance.

~~Sensitivity analysis of the model parameters to produce equal false positives and false negatives.~~

Metric	**Mean**	**Std Dev**
Precision	0.8812	0.2301
Sensitivity	0.7745	0.2830
Specificity	0.9656	0.0888
F1 Score	0.7970	0.2623
IoU	0.7216	0.2763
Accuracy	0.9529	0.0542

~~Based on these results, we incorporated an offset of 15.3 percent +/- 0.1 percent offset between the false positives and the false negatives into our algorithm. Requiring the percent false positives to be equal to the percent false negatives less this offset in the algorithm described above led to evaluated results for which on averagethe percent false positives were equal to the percent false negatives. All pixelswith VV polarization backscatter below the resulting threshold are provisionally labeled as waterand the rest of the pixels are provisionally labeled as non-water. Similarly to the work by (Twele et al., 2016) and Meyer et al.~~ Table 2 summarizes the performance metrics for the Dynamic World water class evaluated on our hand-labeled dataset. The mean precision of 0.8812 indicates that, on average, ~~in preparation (they however, use a fuzzy-logic-based approach) any pixelswith a slope greater than 15 degrees or Height Above Nearest Neighbor (HAND; Nobre et al., 2011) above 15m are then labeled as not water~~88.12% of the pixels predicted as water by DW are actually water in our ground truth labels. The mean sensitivity (recall) of 0.7745 suggests that DW correctly identifies 77.45% of the water pixels in our labels. The high mean specificity (0.9656) indicates that DW accurately classifies non-water pixels, with minimal misclassification as water. The F1 score, which is the harmonic mean of precision and recall, has a mean value of 0.7970, indicating a good balance between the two metrics. The mean IoU of 0.7216 signifies that, on average, there is a 72.16% overlap between the predicted and ground truth water pixels. Lastly, the mean accuracy of 0.9529 shows that DW correctly classifies 95.29% of the pixels overall, including non-water pixels. However, the high standard deviation indicates that there is a large variability in performance for almost all metrics except Specificity and Accuracy, since they take into account the non-water pixels.

3.2 ~~Evaluating~~ Performance of Sentinel-1 ~~imagery and the Equal Percent Solution (EPS)~~based deep learning model

We ~~downloaded the~~ evaluated the performance of a deep learning model (Paul and Ganju, 2021) for inundation mapping that uses S1 ~~images that were spatially and temporally coincident with our labels. All the~~ radar imagery. This deep learning model was the competition winner at the NASA IMPACT challenge for flood detection challenge. Unlike Dynamic World which contained a surface water class, this method focuses on flood or more specifically inundation class. Technically, our hand-labeled dataset also labels inundation although our labels did not focus on capturing flooding. Therefore, we are not directly comparing S1 ~~images were within 3 days of the labeled PS images. Since we are focusing on permanent surface water in this dataset, we do not expect a significant difference between the S1 scenes and the PS-derived labels with respect to surface water extent~~IMPACT flood model against the Dynamic World water class.

We ~~analyzed the distribution of pixel values of~~ processed radiometrically corrected S1 ~~'s VV band across our labeled water pixels (Fig. ??). We found a bimodal distribution with a larger peak at lower backscatter values and another smaller peak at higher backscatter values. We were expecting a normal distribution as these pixels are all related to the water class. Therefore, we investigated the source of this anomaly by isolating the~~ imagery from Alaska Satellite Facility (ASF)'s data repository using the Hyp3 API. S1 ~~pixels with high backscatter values from this smaller peak. Since the water classes were identified using PS imagery at a higher 3m compared to 10m, we were careful to avoid misregistration errors and include non-water samples. To avoid this, we resampled the~~ imagery was searched for each label 3 days before and after the labeled date. We clipped the S1 ~~images from 10m to 3m using the nearest neighbor algorithm. Next, we created a negative 20m buffer around the water classes in each label to ensure only water pixels from~~ scenes based on the labels and then we applied the trained model to these clipped S1 ~~images are selected . Analyzing these spurious~~ scenes using the trained model. We then evaluated the predictions from the deep learning model on our labels after resampling the imagery to match the resolution of the higher resolution labels using nearest neighbor interpolation. 72 S1 ~~pixels from within the water class, we made several observations (Fig ??) - the presence/absence of temporary objects such as ships on waterbodies is expected due to temporal differencesbetween PS and S1, S1 has significantly different signals for muddy and turbulent waters, further, S1 captured narrower river channels compared to PS in hilly regions possibly due to signal distortion. Additionally, in one case, stray signals from nearby buildings distorted the nearby river and affected the water pixels. These anomalous signals highlight the issues behind detecting water using S1 imagery.~~ images were selected for this evaluation. Note that for evaluation, we converted the low-confidence water class to water. Therefore, labels were either 0 (non-water) or 1 (water).

~~Distribution of pixel values in VV polarisation across Sentinel-1 imagery corresponding to the water class in the hand labels~~

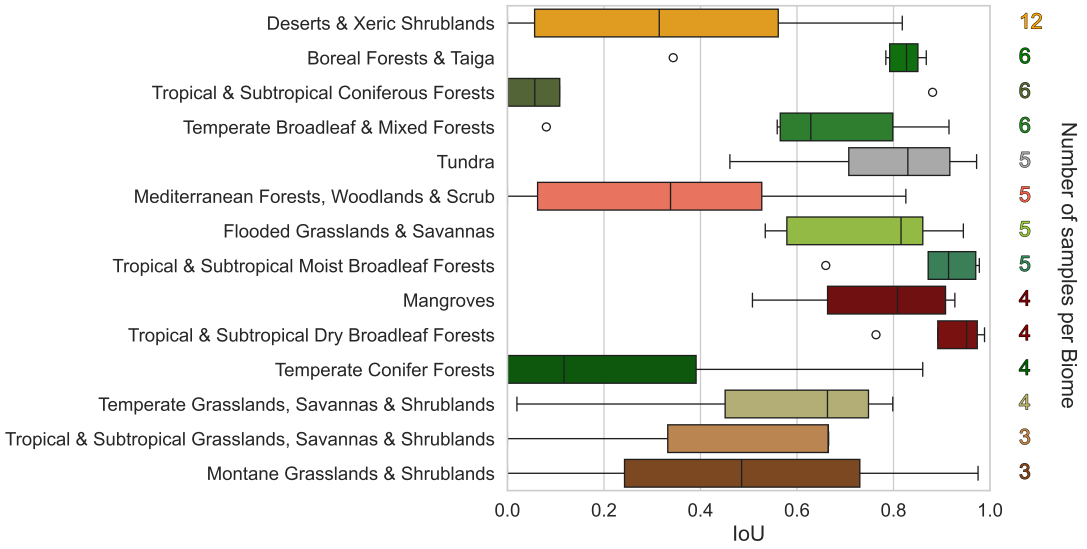

Figure 9. Intersection over Union (IoU) performance of the Sentinel-1 based deep learning model across different biomes. The number of samples per biome is on the right of each bar. Higher IoU scores suggest better performance in detecting surface water. The error bars represent the standard deviation of IoU scores within each biome.

Figure 9 illustrates the performance of the S1-based deep learning model across different biomes using the Intersection over Union (IoU) metric. Performance across biomes has a large variation, with some notable differences. For example, the IMPACT model performed robustly on Tropical & Subtropical Dry Broadleaf Forests, Tropical & Subtropical Moist Broadleaf Forests, Tundra, and Mangroves. Whereas for Tropical & Subtropical Coniferous Forests, Temperate Conifer Forests, and Desert & Xeric Shrublands the model performed less accurately and with large variations. Especially, Mediterranean Forests, Woodlands & Scrub where the model consistently performed poorly. The effectiveness is influenced by the fact that the training dataset of this model is focused on only 5 flood events globally. Therefore, performing accurately on the global surface water dataset is not the objective of this model. Nonetheless, the objective is still detection inundation and the variation in performance provides clues to how such a model can be improved by sampling from biomes or other contexts (urban, river, lake, etc.).

~~Differences between PlanetScope (all in False color composite), Sentinel-1, and Hand-Labels. First row (New Orleans, Lousiana, USA): differences due to passing water vehicles. Second (Tekeze river, Ethiopia): narrower river width in Sentinel-1 compared to PlanetScope. Third (Papua, Indonesia): Sentinel-1 did not pick up muddy waters in the river channel.~~

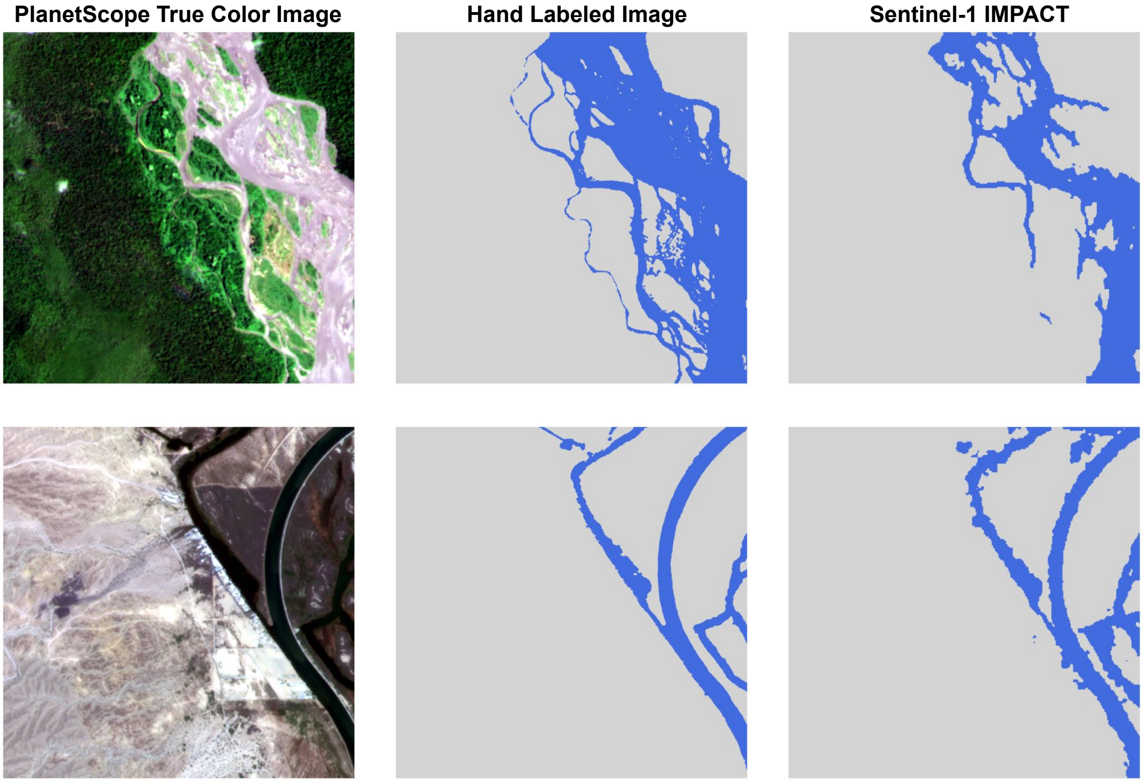

Figure 10. Comparison of a PlanetScope true color image (left), the corresponding hand-labeled image (middle), and the surface water predictions of the Sentinel-1 based deep learning model (right). Top: Nam Dinh, Vietnam (SID33), Bottom: Paymaster Landing, California, USA (SID59).

~~Next, we applied the Equal Percent Solution algorithm to the processed S1 scenes to identify water.EPS achieves a high User's Accuracy at 87.90% but a low Producer's accuracy at 65.67% (Fig **??**) . Additionally, an overall F1 score of 72.16% and an intersection over union (Jaccard)of 62.47%. These scores are influenced by low commission errors and high omission errors by EPS, suggesting that the algorithm misses a lot of water pixels, but rarely misclassifies non-water to water. The second smaller peak observed in the pixel value distribution of S1's VV band over the water class was mostly comprised of turbulent and muddy waters. We know that increased surface roughness produces a different signal in synthetic aperture radar imagery than calm waters leading to high omission errors. This issue of omission with respect to muddy rivers compared to calmer lake waters can be observed in figure~~ **??**~~. S1 is more effective for identifying calm waters, but not for turbulent and shallow waters. Our method takes advantage of~~ Figure 10 provides a visual comparison of the Sentinel-1 based deep learning model's predictions with the ground truth labels for two locations: Nam Dinh, Vietnam (SID33) and Paymaster Landing, California, USA (SID59). The model appears to capture the majority of the water pixels accurately. However, the labels and the corresponding prediction by S1-based model demonstrates the complexity of labeling and identifying water in a meandering braided river (SID33). In case of SID59, the S1 ~~properties to detect calm surface water while missing rougher surfaces. The low commission error in useful in situations such as flooding, where fair allocation of resources is important. Therefore, a method with a high User's Accuracy is more reliable~~model performs well except for the coarser edges of a river in a more arid landscape.

Table 3. Performance metrics for the Sentinel-1 (S1) IMPACT flood detection model evaluated on our hand-labeled dataset. The table presents the mean and standard deviation of various metrics. IoU denotes Intersection over Union. Higher values indicate better performance.

~~Results from evaluating Equal Percent Solution applied on Sentinel-1. Top Row: Ho Chi Minh City, Vietnam (F1 score: 83.45%). Middle: Sudd, South Sudan (F1 score: 96.89%). Bottom: Shandong, China (F1 score: 75.22%). Note the omissions due to surface roughness in the Equal Percentage Solution.~~

Metric	**~~Value (%)~~Mean**	**Std Dev**
~~Userś Accuracy~~ Precision	~~87.90 Producerś Accuracy~~ 0.6547	~~65.67~~ 0.3488
~~Omission Error~~ Sensitivity	~~34.33~~ 0.7485	0.3408
~~Commission Error~~ Specificity	~~12.10~~ 0.8653	0.2309
F1 Score	~~72.16~~ 0.6579	0.3435
~~Jaccard~~ IoU	~~62.47~~ 0.5761	0.3406
~~Overall~~ Accuracy	~~88.00~~ 0.8734	0.1922

~~Overall metrics from the evaluation expressed in percentages. Confusion Matrix showing the true positives, false positives, true negatives, and false negatives from the evaluation. Overall performance metrics from the evaluation of Equal Percentage Solution on Sentinel-1 on our hand-labeled dataset.~~

~~Per label variance in User's Accuracy, Producer's Accuracy, F1 score, and Overall Accuracy from the surface water maps of Equal Percent Solution on Sentinel-1.~~ Table 3 summarizes the performance metrics for the S1 based deep learning model evaluated on our hand-labeled dataset. The metrics exhibit significant variability across the evaluated labels. The S1 IMPACT model generally found it difficult to predict water pixels across several biomes. Apart from the differences in resolution, turbulent water and water located in spatially heterogeneous landscapes are more complicated to detect. Given the cloud free observations, S1 based models can be of considerable benefit for regular monitoring and consistent observations.

4 Limitations

Although our hand-labeled dataset provides a valuable resource for evaluating surface water extent products, it has several limitations that must be considered. First, the spatial resolution of the dataset is limited to 3m, making it more suitable for evaluating lower spatial resolution imagery (> 3m). For higher resolutions (<= 3m), the influence of human labeling errors on the evaluation results is likely to increase. Despite our efforts to cross-reference multiple sources ~~(PS imagery, Bing ,~~ of higher resolution (<1m Bing and Google basemaps) during our labeling process and implement ~~significant~~ considerable quality control, the dataset unavoidably contains biases from our labelers~~and the data used to label. In other words, a~~ , in addition to the biases in the optical PS imagery itself. A model using PS will likely perform the best since PS was the primary source for labeling. Moreover, some features remained unresolved, especially features finer than 3m, leading to the addition of another class called "low-confidence water".

While we made an effort to include samples from diverse contexts in which water can be found (urban, lakes, braided rivers, mountainous regions) and multiple biomes covering different seasons, designing a truly representative dataset is not feasible. The stratified random sampling strategy used to create the dataset aims to cover diverse contexts and biomes but may not capture all the variability in surface water appearance across different regions and seasons. Additionally, the dataset only represents a snapshot in time and does not account for temporal changes in surface water extent, which can be significant in some regions due to seasonal variations, human interventions, or flooding. For example, this dataset does not include frozen water bodies.

Therefore, we recommend using evaluations from multiple independent datasets from various sources to achieve further robustness in evaluation. ~~Finally,~~ While our dataset is primarily ~~a validation dataset and~~ designed for validation purposes, it can still be used for fine-tuning pre-trained models. However, it does not include the ~~input~~ original input PlanetScope images of our labels, which are required for training models. ~~Hence, it cannot be used for benchmarking methods. However, this~~ This ensures that there is no data leak from the training process, maintaining the integrity of the evaluation process.

~~Reliable and accurate monitoring of global water resources is crucial for sustainable water management and conservation. Remote sensing technology, with the recent rise in data availability and access to computational resources, has revolutionized our ability to monitor water resources using high-resolution products and advanced machine learning algorithms. Despite the existence of numerous solutions, trust in these products remains a challenge since there is no single perfect product or method for surface water mapping. Hence, identifying the advantages and drawbacks of each of these solutions under different conditions~~ Nevertheless, relying on a single dataset for evaluation has its limitations, and using multiple independent datasets is crucial for ~~developing reliable surface water extent products~~assessing the robustness and generalizability of surface water mapping methods.

5 Discussion and Conclusions

In this study, we have presented ~~globally sampled high spatial resolution hand labels based on 3m PlanetScope imagery which serves as an independent validation dataset for surface water extent mapping methods. Our dataset includes locations of surface water from diverse contexts~~a globally sampled, ~~covering 14 biomes, from multiple continents,~~ high-resolution surface water dataset consisting of 100 hand-labeled images derived from 3-meter PlanetScope imagery. Our dataset covers diverse biomes and contexts, including urban and rural areas, lakes, ~~and rivers, including braided riversand shorelines. Using this dataset, we introduced and evaluated a novel Sentinel-1 algorithm called the Equal Percentage Solution for surface water extent mapping~~

~~. The evaluation process using our hand-labels highlighted the advantages and drawbacks of the satellite imagery product and the method introduced in this study.~~

~~Our study underscores~~ rivers, braided rivers, and coastal regions. The thorough labeling process, which involves cross-referencing multiple data sources and extensive quality control, ensures the reliability of the labels. These characteristics make our dataset a valuable resource for evaluating the performance and robustness of surface water mapping methods across a wide range of landscapes.

By applying our dataset to the S2-based Dynamic World and S1-based NASA IMPACT models, we demonstrated its utility in identifying the strengths and limitations of different satellite imagery products and methodologies. The variability in performance across biomes highlights the importance of using representative validation data to assess the spatial generalizability of mapping methods. Our findings underscore the need for ~~developing and utilizing~~ multiple independent validation datasets to ~~ensure accurate and reliable water resource monitoring . The insights from our evaluation process improve our understanding of the characteristics of the satellite product used in our study and how it influences the effectiveness of~~ comprehensively evaluate surface water products and build trust in their results.

Accurate and reliable monitoring of surface water resources is crucial for sustainable water management, climate change adaptation, and conservation efforts. High-quality validation datasets like ours play a vital role in advancing these goals by enabling the development and assessment of more effective mapping methods. We ~~believe that the availability of such datasets can facilitate standardized evaluations of data products and surface water extent methods. Ultimately, our dataset contributes to the development of more effective and sustainable water management practices, which are essential for the conservation of our natural resources~~ anticipate that our dataset will contribute to improving the accuracy, robustness, and spatial generalizability of surface water mapping products, ultimately supporting better-informed decision-making and more efficient management of our precious water resources in the face of growing global challenges.

Data availability. Our global surface water dataset (Mukherjee et al., 2024) used in this study is available in the CyVerse Data Commons, accessible via https://doi.org/10.25739/03nt-4f29.

Author contributions. R.M., F.P., and B.T. designed the dataset, developed the sampling strategy, and structured the paper. R.M. and J.G. wrote the paper. R.W., P.S., E.A. labeled the images. R.M., R.W., P.S., E.A., and Z.Z. processed the data. Z.Z. uploaded the dataset.

Competing interests. No competing interests are present.

Acknowledgements. This work was supported by the NASA Earth Science Division ACCESS Program [19-ACCESS19-0041]

References

Acharki, S.: PlanetScope contributions compared to Sentinel-2, and Landsat-8 for LULC mapping, Remote Sensing Applications: Society and Environment, 27, 100 774, 2022.

Alemohammad, H. and Booth, K.: LandCoverNet: A global benchmark land cover classification training dataset, arXiv preprint arXiv:2012.03111, 2020.

Bamber, J. and Bindschadler, R.: An improved elevation dataset for climate and ice-sheet modelling: validation with satellite imagery, Annals of Glaciology, 25, 439–444, 1997.

Bijeesh, T. and Narasimhamurthy, K.: Surface water detection and delineation using remote sensing images: A review of methods and algorithms, Sustainable Water Resources Management, 6, 1–23, 2020.

Bonafilia, D., Tellman, B., Anderson, T., and Issenberg, E.: Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 210–211, 2020.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., and Tait, A. M.: Dynamic World, Near real-time global 10 m land use land cover mapping, Scientific Data, 9, 251, 2022.

Chesapeake Bay Program: Chesapeake Bay Land Use and Land Cover (LULC) Database 2022 Edition, U.S. Geological Survey data release, https://doi.org/10.5066/P981GV1L, 2023.

Cloud to Street, Microsoft, and Radiant Earth Foundation: A Global Flood Events and Cloud Cover Dataset (Version 1.0), https://doi.org/10.34911/rdnt.oz32gz, [Date Accessed], 2022.

Dai, A.: Increasing drought under global warming in observations and models, Nature climate change, 3, 52–58, 2013.

Gahlot, S., Gurung, I., Molthan, A., Maskey, M., and Ramasubramanian, M.: Flood Extent Data for Machine Learning, [Date Accessed]. Radiant MLHub, https://doi.org/10.34911/rdnt.ebk43x, 2021.

Ghayour, L., Neshat, A., Paryani, S., Shahabi, H., Shirzadi, A., Chen, W., Al-Ansari, N., Geertsema, M., Pourmehdi Amiri, M., Gholamnia, M., et al.: Performance evaluation of sentinel-2 and landsat 8 OLI data for land cover/use classification using a comparison between machine learning algorithms, Remote Sensing, 13, 1349, 2021.

Giezendanner, J., Mukherjee, R., Purri, M., Thomas, M., Mauerman, M., Islam, A. K. M. S., and Tellman, B.: Inferring the Past: A Combined CNN-LSTM Deep Learning Framework To Fuse Satellites for Historical Inundation Mapping, pp. 2155–2165, https://doi.org/10.1109/CVPRW59228.2023.00209, 2023.

Isikdogan, F., Bovik, A. C., and Passalacqua, P.: Surface water mapping by deep learning, IEEE journal of selected topics in applied earth observations and remote sensing, 10, 4909–4918, 2017.

Li, J., Ma, R., Cao, Z., Xue, K., Xiong, J., Hu, M., and Feng, X.: Satellite detection of surface water extent: A review of methodology, Water, 14, 1148, 2022.

Markert, K. N., Chishtie, F., Anderson, E. R., Saah, D., and Griffin, R. E.: On the merging of optical and SAR satellite imagery for surface water mapping applications, Results in Physics, 9, 275–277, 2018.

Markert, K. N., Markert, A. M., Mayer, T., Nauman, C., Haag, A., Poortinga, A., Bhandari, B., Thwal, N. S., Kunlamai, T., Chishtie, F., et al.: Comparing sentinel-1 surface water mapping algorithms and radiometric terrain correction processing in southeast asia utilizing google earth engine, Remote Sensing, 12, 2469, 2020.

Martinis, S.: Improving flood mapping in arid areas using Sentinel-1 time series data, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 193–196, IEEE, 2017.

Martinis, S., Groth, S., Wieland, M., Knopp, L., and Rättich, M.: Towards a global seasonal and permanent reference water product from Sentinel-1/2 data for improved flood mapping, Remote Sensing of Environment, 278, 113 077, 2022.

Misra, I., Lawrence Zitnick, C., Mitchell, M., and Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2930–2939, 2016.

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., Lymburner, L., McIntyre, A., Tan, P., Curnow, S., et al.: Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia, Remote Sensing of Environment, 174, 341–352, 2016.

Mukherjee, R., Zhang, Z. J., Policeli, F., Tellman, B., and Wang, R.: Rohit_GlobalSurfaceWaterDataset_2024, https://doi.org/10.25739/03nt-4f29, 2024.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., and Saleska, S.: Height Above the Nearest Drainage–a hydrologically relevant new terrain model, Journal of Hydrology, 404, 13–29, 2011.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al.: Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity, BioScience, 51, 933–938, 2001.

Patterson, T. and Kelso, N. V.: World Urban Areas, LandScan, 1:10 million (2012) [Shapefile], North American Cartographic Information Society, https://earthworks.stanford.edu/catalog/stanford-yk247bg4748, 2012.

Paul, S. and Ganju, S.: Flood segmentation on sentinel-1 SAR imagery with semi-supervised learning, arXiv preprint arXiv:2107.08369, 2021.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A.: Data and its (dis) contents: A survey of dataset development and use in machine learning research, Patterns, 2, 100 336, 2021.

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, 418–422, 2016.

Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., and Datcu, M.: Flood detection in time series of optical and sar images, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 1343–1346, 2020.

Robinson, C., Hou, L., Malkin, K., Soobitsky, R., Czawlytko, J., Dilkina, B., and Jojic, N.: Large Scale High-Resolution Land Cover Mapping with Multi-Resolution Data, in: Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2019.00264, 2019.

Sharma, M., Rasmuson, D., Rieger, B., Kjelkerud, D., et al.: Labelbox: The best way to create and manage training data. software, LabelBox, Inc, https://www. labelbox. com, 2019.

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

Tellman, B., Sullivan, J., Kuhn, C., Kettner, A., Doyle, C., Brakenridge, G., Erickson, T., and Slayback, D.: Satellite imaging reveals increased proportion of population exposed to floods, Nature, 596, 80–86, 2021.

Twele, A., Cao, W., Plank, S., and Martinis, S.: Sentinel-1-based flood mapping: a fully automated processing chain, International Journal of Remote Sensing, 37, 2990–3004, 2016.

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., et al.: Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling, Artificial Intelligence in Medicine, 111, 101 987, 2021.

Wieland, M., Martinis, S., Kiefl, R., and Gstaiger, V.: Semantic segmentation of water bodies in very high-resolution satellite and aerial images, Remote Sensing of Environment, 287, 113 452, 2023.

Wolpert, D. H.: The supervised learning no-free-lunch theorems, Soft computing and industry: Recent applications, pp. 25–42, 2002.

World Wildlife Fund: Global Lakes and Wetlands Database: Large Lake Polygons (Level 1), Online publication, https://www.worldwildlife.org/publications/global-lakes-and-wetlands-database-large-lake-polygons-level-1, 2005.

Wulder, M. A., Hilker, T., White, J. C., Coops, N. C., Masek, J. G., Pflugmacher, D., and Crevier, Y.: Virtual constellations for global terrestrial monitoring, Remote Sensing of Environment, 170, 62–76, 2015.