https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

Open Access Earth System Science Data Discussions

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

Rohit Mukherjee[1, *], Frederick Policelli[2, *], Ruixue Wang[1], Beth Tellman[1], Prashanti Sharma[1], Zhijie Zhang[1], and Jonathan Giezendanner[1]

[1]University of Arizona, Tucson, USA
[2]NASA Goddard Space Flight Center, Maryland, USA
[*]These authors contributed equally to this work.

Correspondence: Rohit Mukherjee (rohitmukherjee@arizona.edu)

Abstract. Effective monitoring of global water resources is increasingly critical due to climate change and population growth. Advancements in remote sensing technology, especially in spatial, spectral, and temporal resolutions, have revolutionized 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 in measurement principles between optical (Sentinel-2) and radar (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 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 PlanetScope imagery. Our surface water extent dataset comprises of 90 images, each with a size of 1024x1024 pixels, which were sampled using a stratified random sampling strategy. We covered all 14 biomes, and also 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 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 satellite products and methods to gain 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.

1 Introduction

Mapping surface water is becoming increasingly significant due to the impact of climate change, as several regions face the prospect of droughts (Dai, 2013) and floods (Tellman et al., 2021). Timely, accurate, and reliable monitoring of sur-

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

face water for better management, conservation, and risk reduction practices has been a growing challenge for researchers. Remotely sensed satellite data products have provided a unique vantage point for measuring surface water (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 methods for monitoring surface water using multiple satellite products (Markert et al., 2020; Pekel et al., 2016). Among these methods, machine learning and deep learning algorithms have become increasingly popular in recent years as they are able to take advantage of large volumes of satellite products (public and commercial) to accurately map the earth's surface (Isikdogan et al., 2017; Martinis et al., 2022; Wieland et al., 2023).

Even though there are several satellite water products based on multiple sensors, their effectiveness is not consistent across all conditions. Every satellite data product has several trade-offs between spatial, spectral, and temporal resolutions (Wulder et al., 2015). Higher spatial resolution satellite products, e.g. PlanetScope (PS), often produce more accurate maps than lower resolution Sentinel-2 (10 meter) or Landsat 8 (30 meter) (Acharki, 2022). Similarly, radar and optical sensors measure surface water properties using different measurement principles leading to variations in accuracy and suitability (Martinis et al., 2022) even at similar spatial resolutions. 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 no single algorithm that is guaranteed to perform well in all situations. Li et al. (2022) compared widely used methods for surface water detection

Evaluating these satellite products and surface water methods using independent validation datasets is crucial to increase trust (Bamber and Bindschadler, 1997). However, these datasets are expensive to build and hence valuable, while existing datasets might not be suitable for all needs. For example, BigEarthNet (Sumbul et al., 2019) contains close to 600,000 multi-labeled Sentinel-2 image patches where 83,000 of which contain water bodies. This dataset will confirm the presence of water within a patch, but will not identify each pixel. For a more high-resolution (1m) large-scale solution, the Chesapeake Conservancy Land Cover dataset (Robinson et al., 2019) contains a water class labeled per pixel. LandCoverNet (Alemohammad and Booth, 2020) contains global data based on 10m Sentinel-2 data also containing a water class. Apart from surface water, mapping floods have been a strong research direction with Sentinel-1 (S1) based NASA Flood Extent Detection dataset (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 both optical (Sentinel-2) and radar (S1). Although these datasets are 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 resolution options, they are not globally distributed. Further, the ephemeral nature of floods requires a specific model for accurate detection even though flood water is technically surface water (Bonafilia et al., 2020). Wieland et al. (2023) developed a semi-automated global binary surface water reference dataset that contains 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.

To establish the effectiveness and robustness of a product, there needs to be multiple independent evaluations 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 the real world (Paullada et al., 2021). Similarly, hand-labeled datasets include subjectivity

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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 the issue of data leakage (Vandewiele et al., 2021). Data leakage is experienced when researchers involve their validation set as a part of their training process which leads to an overfit model. Multiple and independent validation datasets are therefore required for thorough evaluation and increased trust in remote sensing-based surface water products and methods.

In this research, we present our hand-labeled high-quality globally sampled, high-spatial-resolution dataset based on 90 samples of 3m PS images, each of size 1024x1024 pixels. Our work builds upon existing satellite-based remotely sensed datasets for surface water extent validation. Our motivation for this work is to provide a higher resolution independent dataset to evaluate surface water products based on publicly available medium-resolution datasets, such as 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, spatial generalizability, and robustness 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.

2 Data Preparation

2.1 Sampling

Our objective was to build a dataset that is close to a true representation of the true distribution of surface water features. A representative dataset will enable testing surface water extent products for their spatial generalizability in addition to their accuracy. Since achieving a true representation is a nearly impossible task (Paullada et al., 2021), we approached this problem by sampling from different biomes as defined by Olson et al. (2001), as they have multiple climate and land conditions 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. First, we created a buffer of 2 km around shapefiles by Fund (2005) defining global rivers and lakes . Next, we clipped these buffers with the shapefiles of each of the 14 biomes. Then, we randomly placed 50 points on each of the biomes and randomly selected at least 5 of them as samples. Finally, to ensure that we address several contexts in which surface water exists, we randomly selected some samples within urbanized regions (Patterson and Kelso, 2012), braided rivers, and shorelines. Table 1, shows the number of samples for each biome. Figure 1 shows the spatial distribution of the samples globally. 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 of our labels are from rivers, while the rest are lakes.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

Open Access Earth System Science Data Discussions

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

Table 1. Number of samples per biome selected from our stratified random sampling for a total of 90 samples.

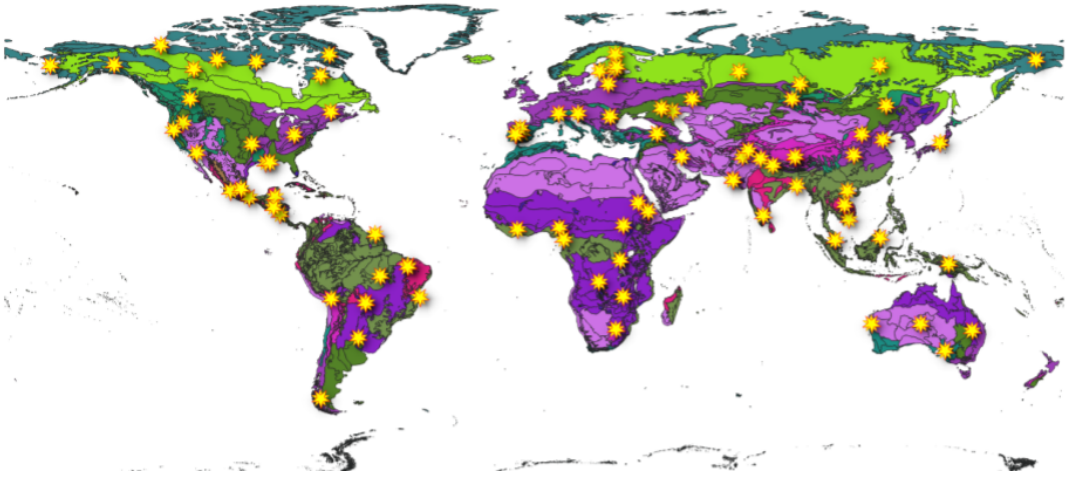


Figure 1. Location of surface water labels sampled globally. The labels have been sampled to be representative of 1) the diverse global biomes (c.f. table 1) and 2) the global spatial distribution.

2.2 Data Processing

After sampling 90 locations based on various criteria, we downloaded 8-band SuperDove PS imagery from the years 2021 and 2022 using our NASA Commercial Smallsat Data (CSDA) Program access. As our objective is to evaluate most medium resolution satellite sensors, including S1 we ensured that the loss of Sentinel-1B satellite did not create a large temporal gap

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

between the label and the last available scene from the satellite. Therefore, for locations found to be only covered by Sentinel-1B satellite and not by 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 perennially frozen water. If a location contained seasonal ice, we substituted that PS image with an image from summer when water was not frozen. From each larger PS scene, we selected an image of 1024x1024 pixels covering an area equal to 9.4 square km. We decided on keeping the labels large enough to ensure that the corresponding spatially coincident medium-resolution imagery from Landsat or Sentinel had sufficient pixels in the image for comparison. For example, 30m Landsat image corresponding to our labels will be 100x100 pixels, while a Sentinel at 10m will be 376x376, where Landsat and Sentinel have enough spatial context.

Figure 2. PlanetScope images selected for labeling are shown in False Color Composite (near infrared, red, and green). Left: Vilyuy River, Sakha Republic, Russia and Right: Tagus River, Toledo, Spain.

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.

We employed labelers with experience in analyzing and labeling surface water in satellite imagery, and performed independent quality checks. To assist the labelers, we created a true-color composite (TCC) and a false-color composite (FCC) with 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.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

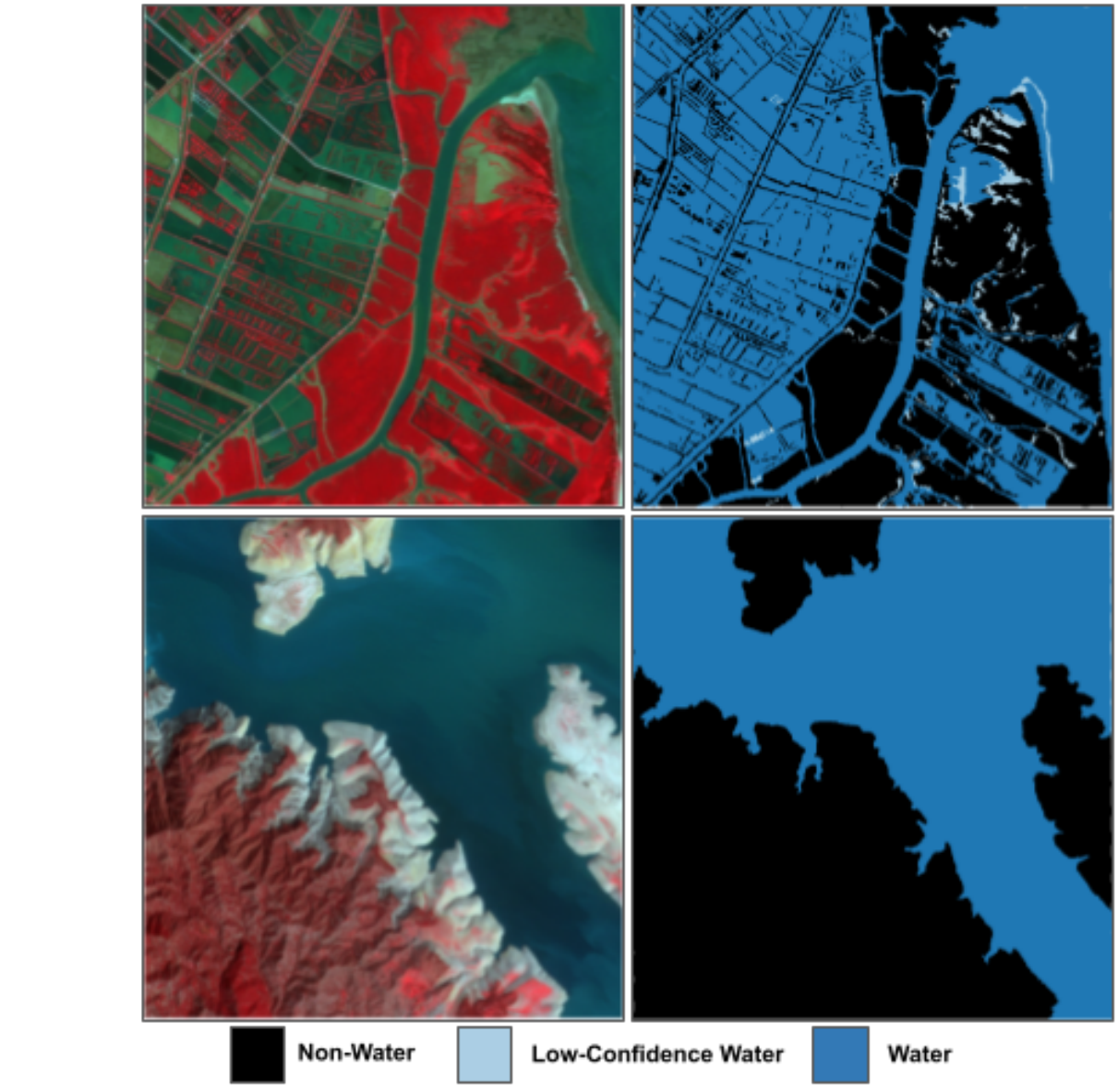


Figure 3. Examples of PlanetScope imagery and corresponding labels (Top Row: Dong Tranh River, Ho Chi Min City, Vietnam, and Bottom Row: Siran River, Pakistan). The images are labeled with three categories: 1) non-water, 2) low-confidence water, and 3) water. The low-confidence water category marks pixels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

During labeling, we encountered several cases where the presence of water was uncertain. However, whenever there was confusion in the PS imagery, we cross-referenced with the higher-resolution basemaps provided by Bing and Google. For features that were still not resolved, we introduced a 'low-confidence water' category. In total, we have three classes - 'water',

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

'low-confidence water', and 'non-water'. However, during evaluation, the low-confidence water class can be ignored since these corresponding features are not confidently identified, hence kept separate from the evaluation process. After labeling, we performed quality controls on each of the labels to ensure the accuracy of our labels. In total, combining time for labeling and quality control, we spent approximately 2 hours for each image, equalling 180 hours of work and leading to 204 square km of labeled surface water from a total surface area of 850 square km. Each label is provided a sampled ID (SID) from 1 to 90 and contains the date (YYYYMMDD) of the PS image from which it was labeled.

2.4 Dataset Analysis

We labeled a total of 90 1024x1024 PS images at 3m, where water corresponds to 24% of all the surface area, while low-confidence water covers 1.3% and the rest (74.6%) is non-water (Fig 4). 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 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 a variety of landscapes - rivers passing through urban regions, braided rivers in the delta, rivers passing through forests and agricultural fields, including waterbodies in plain and mountainous regions.

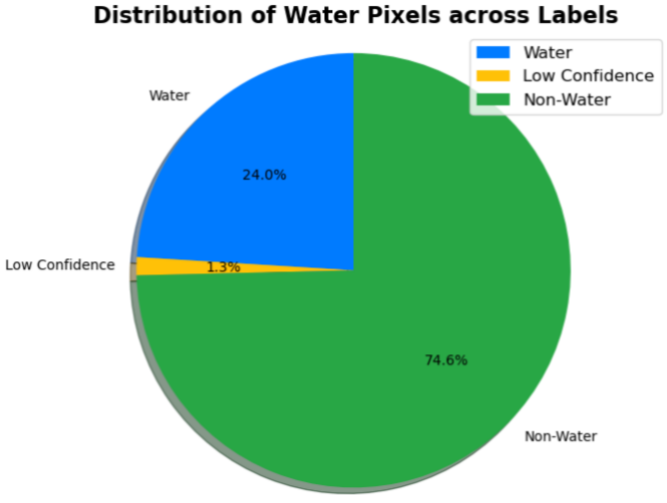


Figure 4. Class distribution across Labels (non-water, low-confidence water, and water) for all chips. Non-water class shares the largest percentage as it encompasses the water class. Low-confidence water pixels are only a minor percentage.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

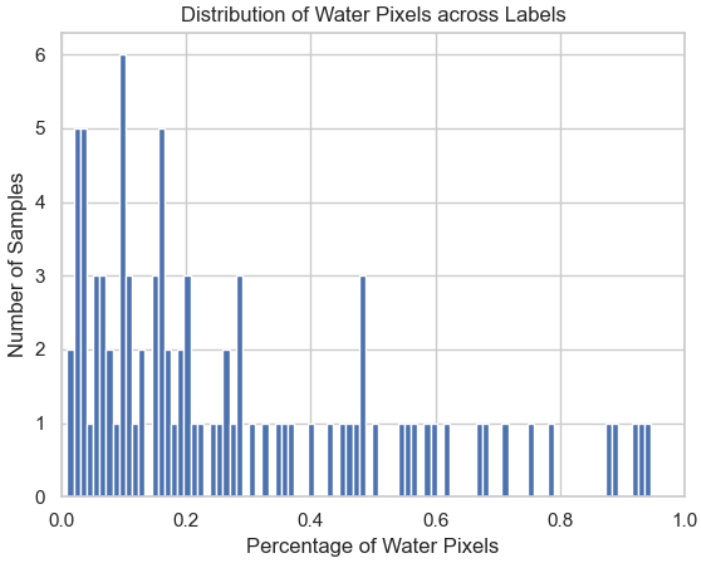


Figure 5. Distribution of water pixels per sample. The figure shows the percentage of water pixels within one sample. Most samples contain less than 50% of water by design, as the focus is to delineate the boundaries since the water class is more homogeneous, therefore, less complex.

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 water maps using our method using S1 imagery and evaluated these water maps 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 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.

3.1 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., in preparation). After pre-processing, as per Twele et al. (2016) and using code implemented by the Alaska Satellite Facility (Meyer et al., 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

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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 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. 6), and moving the threshold until the expected percentage of false positives is equal to the expected percentage of false negatives (within +/- 0.1 dB).

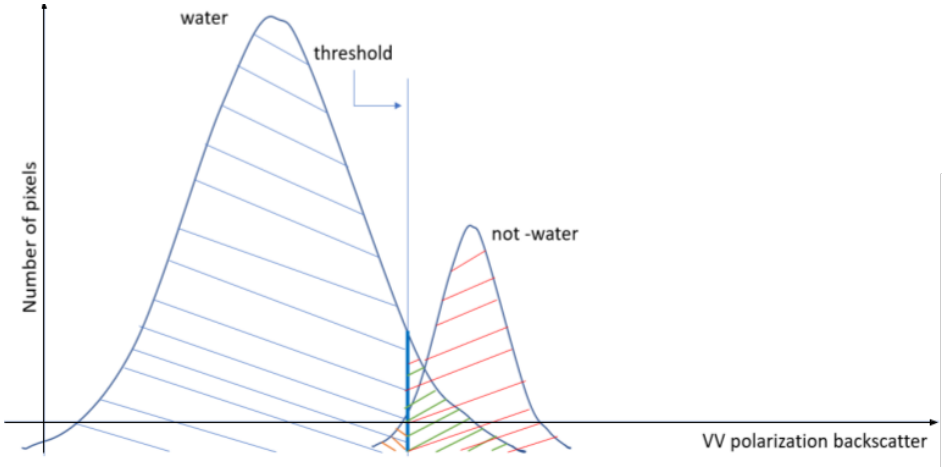


Figure 6. Visualization of the thresholding algorithm based on the Sentinel-1 VV band: a threshold value is chosen to identify the valley in the bimodal distribution to differentiate between the water and non-water classes. In this given schematic representation, the number of water pixels dominates.

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 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.

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

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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.

During our evaluation of the resulting water maps (discussed below), 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. 7). Equal false positives and negatives would allow the model to produce more balanced water maps that do not tend to overpredict or underpredict.

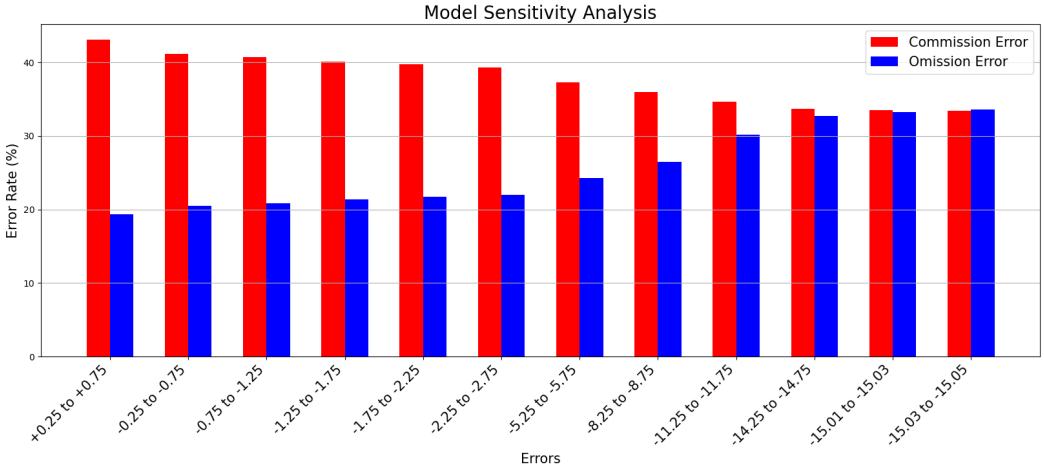


Figure 7. Sensitivity analysis of the model parameters to produce equal false positives and false negatives.

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 average the percent false positives were equal to the percent false negatives. All pixels with VV polarization backscatter below the resulting threshold are provisionally labeled as water and the rest of the pixels are provisionally labeled as non-water. Similarly to the work by (Twele et al., 2016) and Meyer et al., in preparation (they however, use a fuzzy-logic-based approach) any pixels with a slope greater than 15 degrees or Height Above Nearest Neighbor (HAND; Nobre et al., 2011) above 15m are then labeled as not water.

3.2 Evaluating Sentinel-1 imagery and the Equal Percent Solution (EPS)

We downloaded the S1 images that were spatially and temporally coincident with our labels. All the 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.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

Open Access Earth System Science Data Discussions

We analyzed the distribution of pixel values of S1's VV band across our labeled water pixels (Fig. 8). 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 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 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 S1 images are selected. Analyzing these spurious S1 pixels from within the water class, we made several observations (Fig 9) - the presence/absence of temporary objects such as ships on waterbodies is expected due to temporal differences between 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.

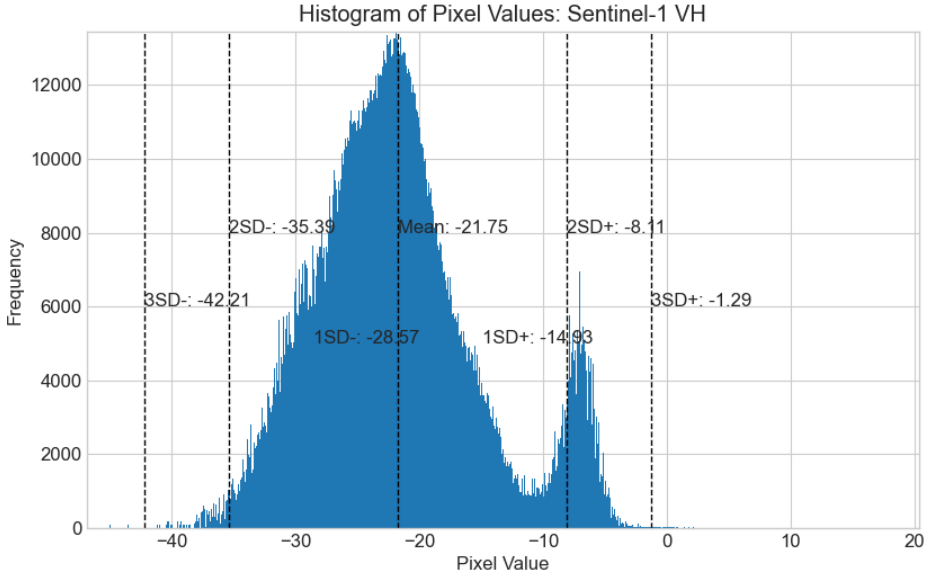


Figure 8. Distribution of pixel values in VV polarisation across Sentinel-1 imagery corresponding to the water class in the hand labels

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

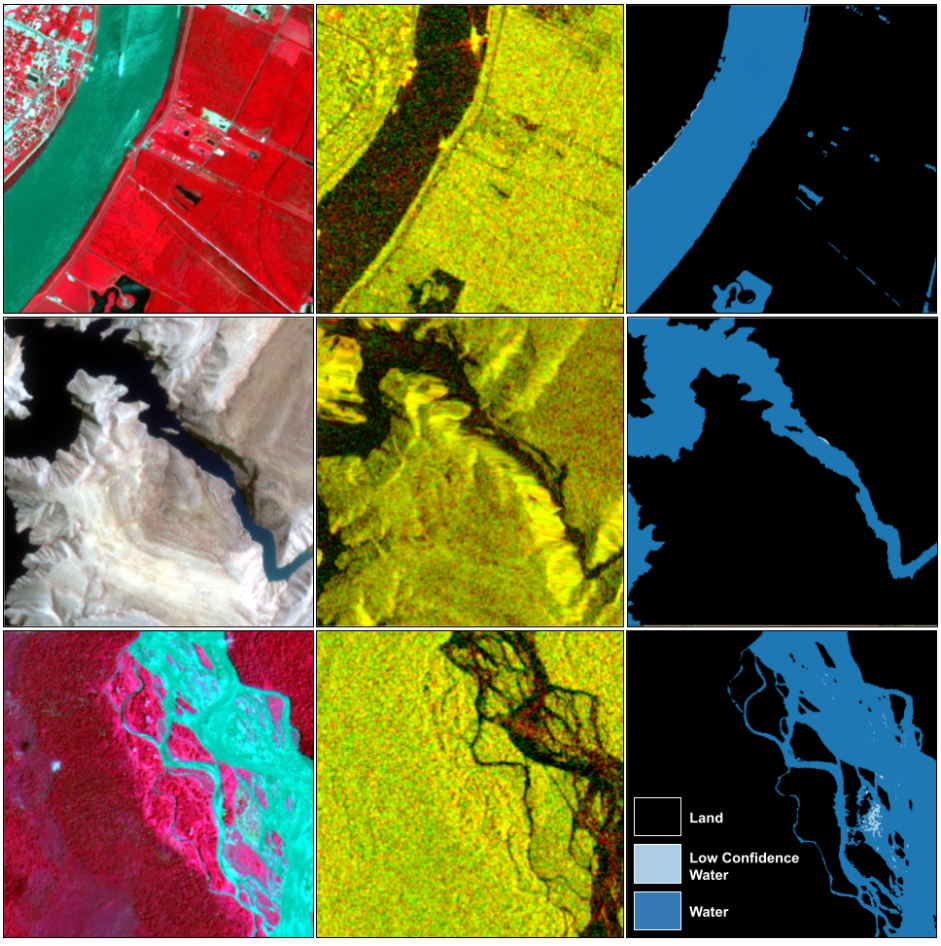


Figure 9. Differences between PlanetScope (all in False color composite), Sentinel-1, and Hand-Labels. First row (New Orleans, Lousiana, USA): differences due to passing water vehicles. Second (Tekeze river, Ethiopia): narrower river width in Sentinel-1 compared to PlanetScope. Third (Papua, Indonesia): Sentinel-1 did not pick up muddy waters in the river channel.

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 10). 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 9. S1 is more effective for identifying calm waters, but not for turbulent and shallow waters. Our method takes advantage of the S1 properties to detect calm surface water while missing rougher surfaces. The low

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

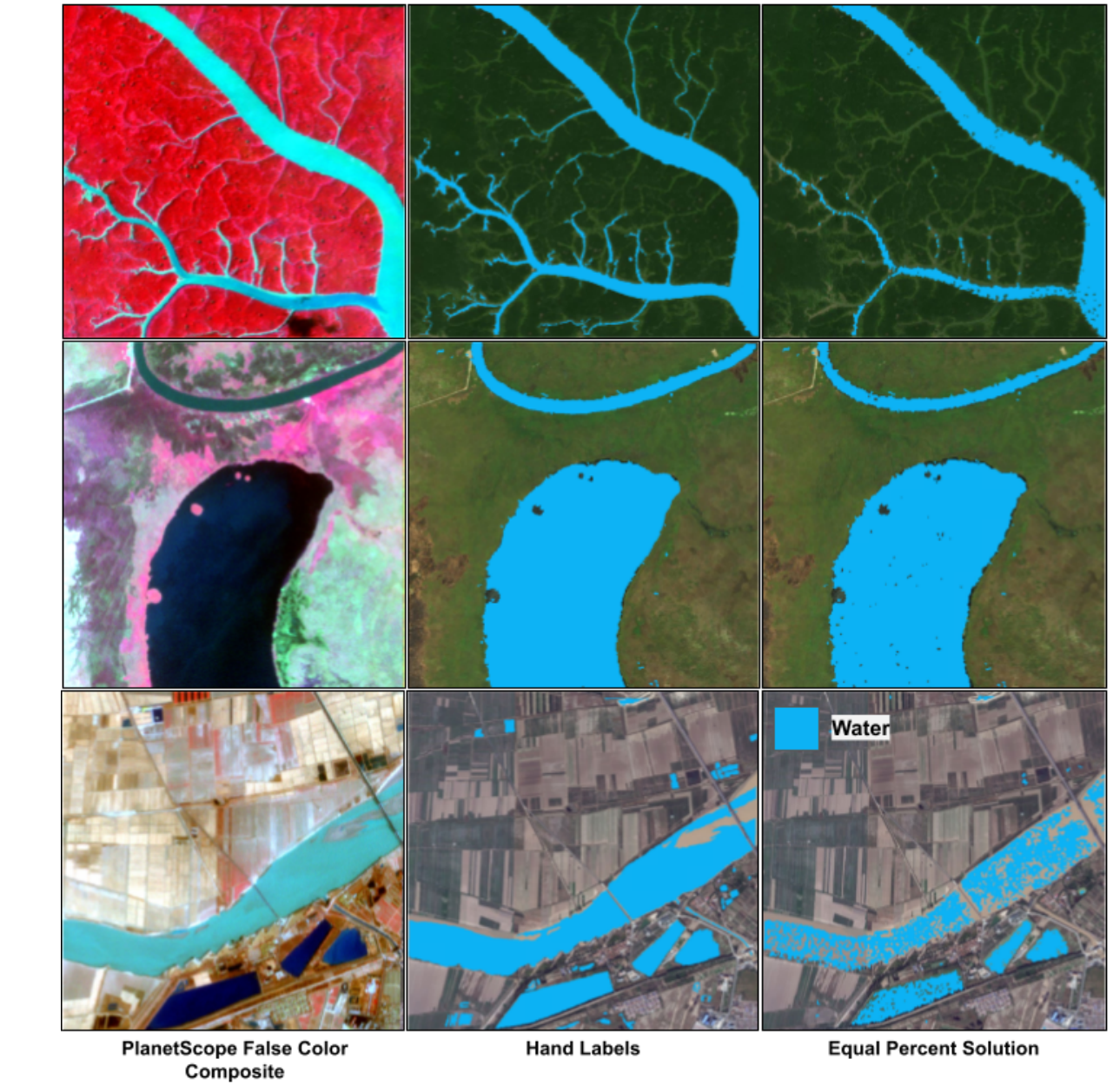
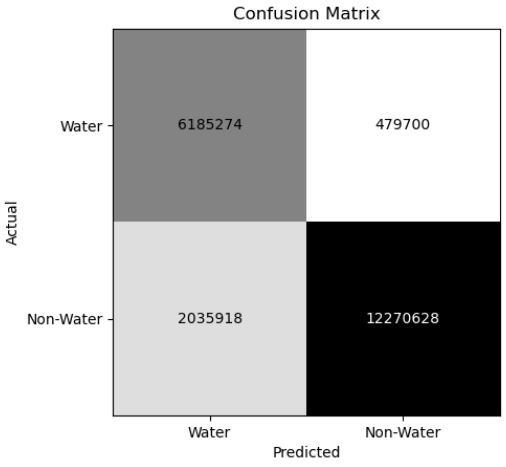


Figure 10. Results from evaluating Equal Percent Solution applied on Sentinel-1. Top Row: Ho Chi Minh City, Vietnam (F1 score: 83.45%). Middle: Sudd, South Sudan (F1 score: 96.89%). Bottom: Shandong, China (F1 score: 75.22%). Note the omissions due to surface roughness in the Equal Percentage Solution.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

Metric	Value (%)
Userś Accuracy	87.90
Producerś Accuracy	65.67
Omission Error	34.33
Commission Error	12.10
F1	72.16
Jaccard	62.47
Overall Accuracy	88.00

(a) Overall metrics from the evaluation expressed in percentages.



(b) Confusion Matrix showing the true positives, false positives, true negatives, and false negatives from the evaluation.

Figure 11. Overall performance metrics from the evaluation of Equal Percentage Solution on Sentinel-1 on our hand-labeled dataset.

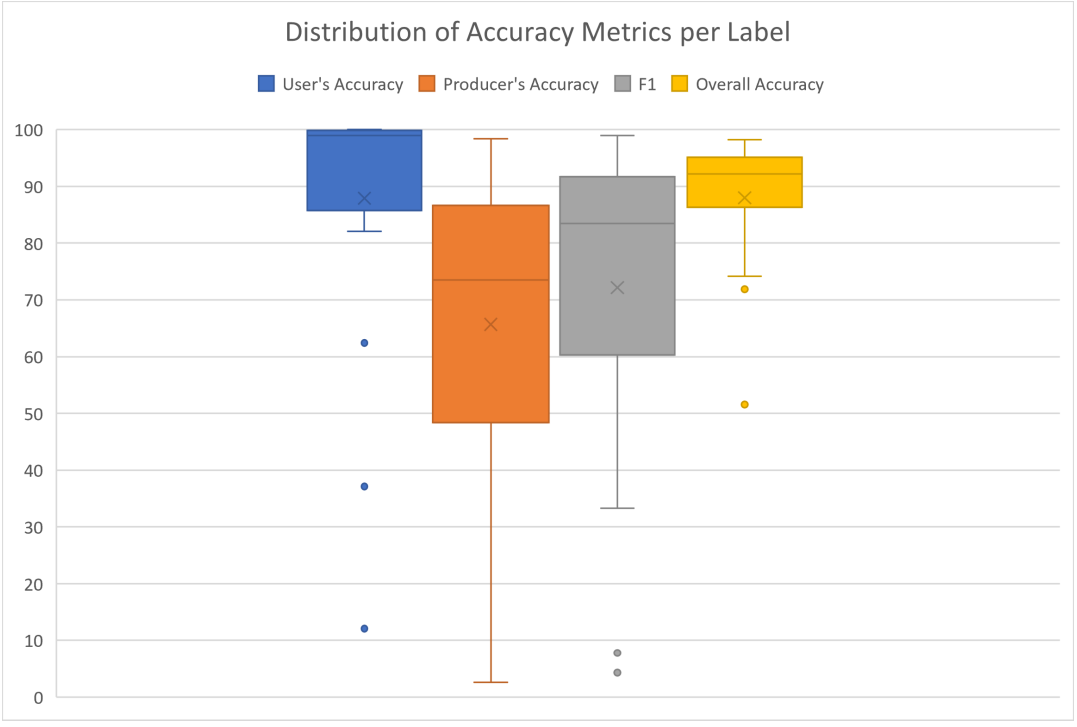


Figure 12. Per label variance in User's Accuracy, Producer's Accuracy, F1 score, and Overall Accuracy from the surface water maps of Equal Percent Solution on Sentinel-1.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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, and Google basemaps) during our labeling process and implement significant quality control, the dataset unavoidably contains biases from our labelers and the data used to label. In other words, a model using PS will likely perform the best since PS was the primary source for labeling. Moreover, some features remained unresolved, 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. 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, our dataset is primarily a validation dataset and does not include the input images of our labels, which are required for training models. Hence, it cannot be used for benchmarking methods. However, this ensures that there is no data leak from the training process, maintaining the integrity of the evaluation process.

5 Discussion and Conclusions

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 is crucial for developing reliable surface water extent products.

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, covering 14 biomes, from multiple continents, urban and rural, lakes, and rivers, including braided rivers and 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 the need for developing and utilizing 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 methods. We believe that the availability of such datasets can facilitate standardized evaluations of data products and surface water extent methods. Ultimately, our dataset

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

contributes to the development of more effective and sustainable water management practices, which are essential for the conservation of our natural resources.

Code and data availability. Awaiting DOI from Radiant Earth ML Hub. Preliminary Access: https://data.cyverse.org/dav-anon/iplant/home/jgiezendanner/Mukherjee_HighResolutionSurfaceWaterLabels_Mai2023.zip. Code can be provided after receiving permission from NASA to release.

Author contributions. R.M., F.P., and B.T. designed the dataset, developed sampling strategy, and structured the paper. R.M. and J.G. wrote the paper. R.W. and P.S. labeled the images. R.M., R.W., P.S., 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]

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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.

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.

Fund, W. W.: 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.

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.

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., 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.

https://doi.org/10.5194/essd-2023-168
Preprint. Discussion started: 30 August 2023

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.

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.

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.

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.