This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

S1S2-Water: A global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 satellite images

Marc Wieland, Florian Fichtner, Sandro Martinis, Sandro Groth, Christian Krullikowski, Simon Plank, Mahdi Motagh

***Abstract*—This study introduces the *S1S2-Water* dataset – a global reference dataset for training, validation and testing of convolutional neural networks for semantic segmentation of surface water bodies in publicly available Sentinel-1 and Sentinel-2 satellite images. The dataset consists of 65 triplets of Sentinel-1 and Sentinel-2 images with quality checked binary water mask. Samples are drawn globally on the basis of the Sentinel-2 tile-grid (100 x 100 km) under consideration of predominant landcover and availability of water bodies. Each sample is complemented with metadata and Digital Elevation Model (DEM) raster from the Copernicus DEM. On the basis of this dataset we carry out performance evaluation of convolutional neural network architectures to segment surface water bodies from Sentinel-1 and Sentinel-2 images. We specifically evaluate the influence of image bands, elevation features (slope) and data augmentation on the segmentation performance and identify best-performing baseline-models. The model for Sentinel-1 achieves an Intersection Over Union of 0.845, Precision of 0.932 and Recall of 0.896 on the test data. For Sentinel-2 the best model produces an Intersection Over Union of 0.965, Precision of 0.989 and Recall of 0.951 respectively. We also evaluate the performance impact when a model is trained on permanent water data and applied to independent test scenes of floods.**

***Index Terms*— Convolutional Neural Networks; Reference dataset; Semantic segmentation; Sentinel-1; Sentinel-2; Surface Water Monitoring**

I. Introduction

Monitoring and understanding the spatio-temporal dynamics of surface water bodies with seamless geographical coverage and over large areas is important for scientists, the economy and political decision makers. Compared to conventional hydrological monitoring systems that rely on networks of point-wise rain and stream gauging stations, satellite remote sensing can provide complementary information on surface water extent over large geographical areas, at high temporal frequency and low cost. Satellite-based water monitoring can contribute relevant information on hydrology in ungauged areas, support the development of infrastructure projects or help to understand the impacts of changes in hydrology on environment, economy and human health at different spatial and temporal scales. In particular, being able to detect and quantify spatio-temporal anomalies, such as flooding or hydrological droughts is essential to support measures across all phases of the disaster risk management cycle.

With the launch of the Sentinel-1 (April 2014, April 2016) and Sentinel-2 (June 2015, March 2017) satellites large-scale monitoring of land surface dynamics at high spatial resolution (~10-20 m Ground Sampling Distance (GSD)), temporal revisit period (~2-6 days depending on geographical location) and with large swath width (>250 km) became possible. In contrast to previous satellite missions, data of the Copernicus missions are acquired systematically over the entire global landmass and are openly available. The Sentinel-1 satellites carry a Synthetic Aperture Radar (SAR) C-band sensor acquiring data in VV and VH polarization amongst others with a capability of penetrating through clouds and acquiring images during day and night. The Sentinel-2 satellites carry a Multi-Spectral Instrument (MSI) that collects data in the visible, near- and short-wave-infrared across 13 spectral bands. Segmentation of water bodies in SAR data is a difficult task and misclassifications are largely related to similar backscatter reflectance patterns between different land cover classes. False positives are in particular caused by confusion with other low backscattering targets, such as sand, concrete or salt pans. False negatives can be caused by strong winds, which increase the surface roughness of water bodies and prevent specular reflection [1]. Multi-spectral satellite images, on the contrary, are influenced by atmospheric effects and the presence of clouds and cloud shadows, which may obstruct objects of interest and introduce bias to further image analysis [2]. Similar spectral response characteristics of shadow and water pixels are a common source for misclassifications. Independent of the observing sensor system, automated water segmentation from satellite images is a challenging task

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project "Künstliche Intelligenz zur Analyse von Erdbeobachtungs- und Internetdaten zur Entscheidungsunterstützung im Katastrophenfall" (AIFER) under Grant 13N15525, and by the Helmholtz Artificial Intelligence Cooperation Unit through the project "AI for Near Real Time Satellite-based Flood Response" (AI4FLOOD) under Grant ZT-IPF-5-39. Corresponding author: Marc Wieland.

The authors are with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), D-82234 Oberpfaffenhofen, Germany (e-mail: marc.wieland@dlr.de; florian.fichtner@dlr.de; sandro.martinis@dlr.de; sandro.groth@dlr.de; christian.krullikowski@dlr.de; simon.plank@dlr.de) and the German Research Centre for Geosciences (GFZ), D-14473 Potsdam, Germany (e-mail: motagh@gfz-potsdam.de). Mahdi Motagh is also with the Leibniz University Hannover, D-30167 Hannover, Germany (e-mail: motagh@ipi.uni-hannover.de).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

because of significant variations in the reflected signal, size and shape of water bodies. Sediment, debris, ice, vegetation, infrastructure, boats or other vehicles interact with the water surface or sub-surface and modify the appearance of water bodies in the satellite images. These conditions are, moreover, highly dynamic and dependent on the geographical location and acquisition time. The fractal geometry of the land-water border further increases the complexity of the segmentation task, since the definition of a hard class boundary may be subjective and strongly dependent on the image resolution.

during 11 flood events (*Sen1Floods11*). They report loss in accuracy when models trained on images depicting only permanent water bodies are being applied to images showing flood events. The *Sen1Floods11* dataset has further been used by Bai et al. (2021) [15] to test a modified U-Net architecture and to examine the influence of data augmentation as well as fusing Sentinel-1 and Sentinel-2 data.

Despite an increasing availability of input data due to the systematic acquisitions of the Sentinel satellites, we observe a general lack of annotated reference data in remote sensing and

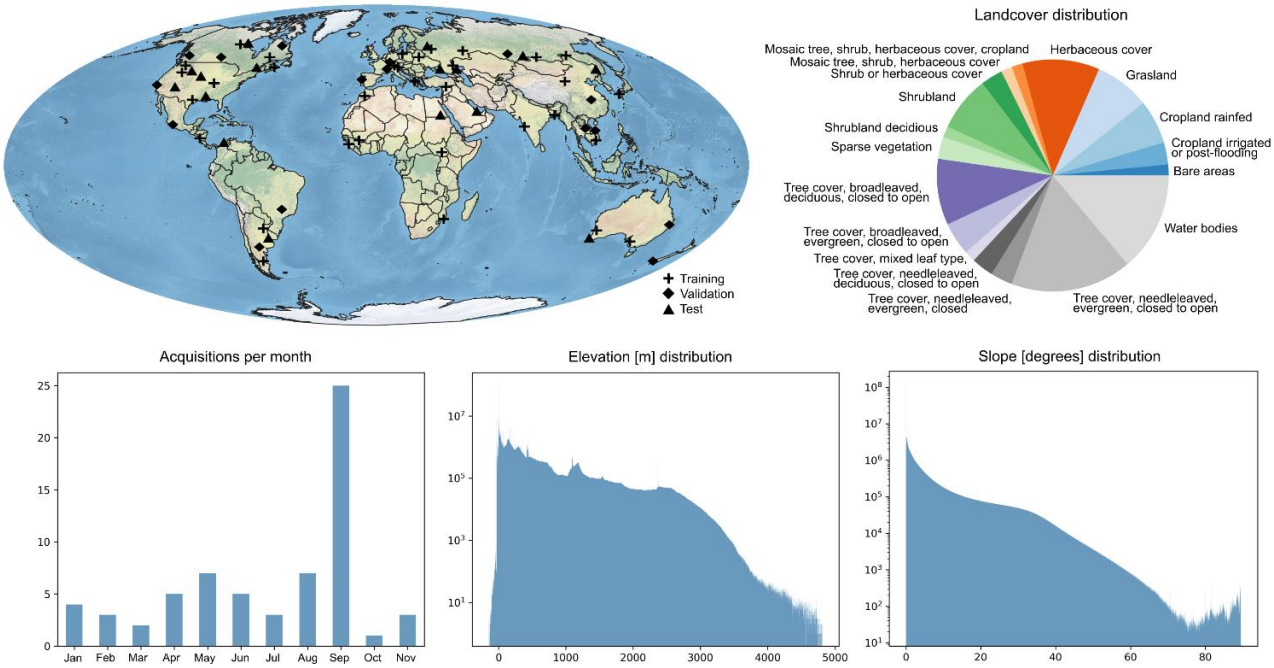

Fig. 1. Spatio-temporal distribution of samples in the S1S2-Water dataset across geographies, months of the year, landcover [19], elevation and slope.

Surface water monitoring is an important topic in remote sensing and detailed reviews can be found in Huang et al. (2018) [3] and Bentivoglio et al. (2022) [4]. While rule-based algorithms have long dominated water mapping studies with multi-spectral and SAR satellite sensors [5]–[7], convolutional neural networks (CNNs) have seen a rapid development in recent years [8]–[11]. Liu et al. (2019) [12] use a modified U-Net architecture to analyze bi-temporal and dual-polarized Sentinel-1 images of floods caused by Hurricane Harvey in the U.S. in 2017. They evaluate the impact of different polarizations as input bands to train a model and show that using VV-VH polarization or VH polarization achieves better results than VV polarization alone. Nemni et al. (2020) [13] compare variations of CNNs based on the U-Net architecture for their performance to segment water in Sentinel-1 images. They developed a dataset of 15 Sentinel-1 images around flood events in eight countries of Eastern-Africa and South-East-Asia. The dataset consists of VV-polarized images and corresponding quality-checked flood extent masks from UNOSAT analyses. Bonafilia et al. (2020) [14] introduce a dataset of VV-VH polarized images from Sentinel-1 acquired in particular for the task of water segmentation. Limitations in geographical coverage reduce the significance of results with respect to global applicability of trained models. To this regard, Mateo-Garcia et al. (2021) [16] developed the *WorldFloods* dataset that consists of Sentinel-2 images with corresponding water masks for 119 globally distributed flood events. They compiled flood water masks from rapid mapping activities carried out by various disaster response organizations. The authors state that the quality of their dataset is varying and label noise is widely present by temporal misalignments between Sentinel-2 image and water mask. This is attributed to the fact that most water masks have been derived semi-automatically from SAR images during rapid mapping. The Sentinel-2 images of the dataset are, moreover, not the source of the water masks, causing temporal misalignments between images and masks. Especially in highly dynamic flood situations, even differences of a few days between acquisitions may potentially introduce large degrees of misalignments and deteriorate the quality of a dataset. The same issues are relevant for the *OMBRIA* [17] dataset, which is directly compiled from rapid mapping

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

products of the Copernicus Emergency Management Service (CEMS). The aforementioned *Sen1Floods11* dataset [14] is released openly and covers larger geographical areas. However, Helleis et al. (2022) [18] indicate issues of this dataset with regards to spatial autocorrelation and expressiveness of the samples. Other relevant reference datasets for Sentinel-1 and / or Sentinel-2 include Zhu et al. (2019) [19], Alemohammad and Booth (2020) [20] and Hong et al. (2021) [21]. These datasets contain a water class but were not specifically created for the task of water segmentation. Therefore, the applied sampling scheme does not consider intra- and inter-class variability of the water class. Furthermore, their geographical coverage is limited and annotation efforts are of varying quality. The *Sen12ms* dataset [22] contains over 180,000 samples (each 256 x 256 pixels in size) with triplets of Sentinel-1, Sentinel-2 image patches and corresponding MODIS land cover data annotations. The dataset has global coverage and samples are well distributed in space and time. Due to its general focus on image fusion and landcover classification it, however, does not specifically account for variability of a water class. Additionally, its annotations are from an independent source and at much coarser native resolution (500 m) than the image data (10-20 m), which renders them not applicable for fine grained water segmentation. The *Sen12-Flood* [23], [24] is another dataset that provides Sentinel-1 and Sentinel-2 image patches for flood events. It is based on the *MediaEval* [25] dataset and provides image-level labels ("image contains a flood or not") rather than pixel-level annotation masks of flooded areas.

Based on the need for globally applicable, validated water segmentation methods and the observed lack of appropriate benchmark datasets for this task, we introduce *S1S2-Water* – a global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 satellite images. This multi-modal benchmark dataset consists of 65 triplets of Sentinel-1 and Sentinel-2 images with quality checked binary water masks. Tiling it into non-overlapping 256 x 256 pixels patches results in 100,000+ samples per sensor. Each sample is complemented with detailed metadata and Digital Elevation Model (DEM) raster from the Copernicus DEM [26]. Samples are distributed globally on the basis of the Sentinel-2 tile-grid (100 x 100 km) and under consideration of landcover and availability of water bodies. On the basis of *S1S2-Water* we carry out performance evaluation of convolutional neural network architectures to segment water bodies from Sentinel-1 and Sentinel-2 images. We specifically evaluate the influence of image bands, elevation features (slope) and data augmentation on the segmentation performance and identify best-performing baseline-models for each sensor. We also evaluate the performance impact when a model is trained on normal water data (*S1S2-Water*) and applied to independent test scenes of flood events. This aims at testing the applicability of *S1S2-Water* for normal water monitoring as well as flood mapping and allows to identify limitations and future improvements to the dataset. The dataset is released openly alongside this publication [27].

Section II provides an overview of requirements for remote sensing reference datasets and introduces the *S1S2-Water* dataset. Section III describes the methods used to compile the *S1S2-Water* dataset and explains the experimental setup of the performance evaluation of convolutional neural network architectures. Results are presented in section IV, discussed in section V and concluded in section VI.

II. Data

Reference data for the analysis of Earth observation data should primarily be adapted to the requirements of practical application and be geared to the properties of interpretation algorithms. Essentially, the creation of reference data in this context is aimed at training, testing and validating machine learning methods for the semantic segmentation of water bodies. Accordingly, the reference dataset should consist of sufficient and as accurately annotated samples as possible, which explicitly cover the expected challenges in practical application scenarios. Based on these assumptions and following the guidance of building remote sensing benchmark datasets outlined by Long et al. (2021) [28], the following requirements were considered when constructing the *S1S2-Water* dataset.

- **Diversity**: Diversity means that samples of the same class should cover as wide a spectrum as possible in terms of appearance, size, shape and orientation. Differences in background (e.g., water surrounding landcover), scale, acquisition time and geographical context should also be considered.
- **Richness**: While diversity emphasizes the semantic otherness of samples, richness of a dataset refers to the variation of image features. These include lighting, background, occlusion, radiometry, pixel resolution and sensor characteristics, amongst others. In this context, the selection of time periods, geographical coverage and acquisition sensors is influenced by the use case.
- **Scalability**: Scalability refers to the ability to change and expand a dataset. It should be possible to easily integrate newly annotated images into an existing dataset and to update dataset parameters such as the classification scheme. For this reason, data formats and the storage structure of annotations and images as well as metadata are important to control the scalability of a data set.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

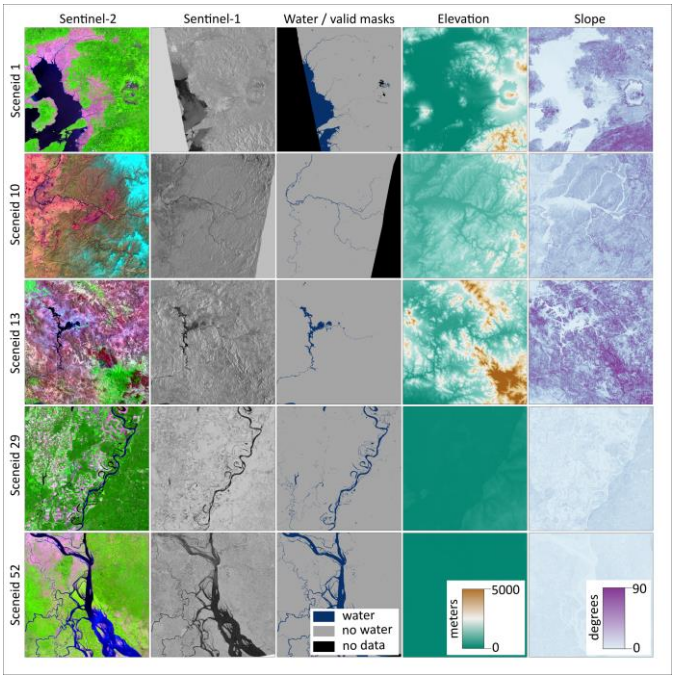

Fig. 2. Examples of *S1S2-Water* samples with Sentinel-1 (VV polarization) and Sentinel-2 (SWIR-NIR-Green) satellite images, associated water / valid masks as well as elevation and slope information. Scenes cover an area of 100 x 100 km.

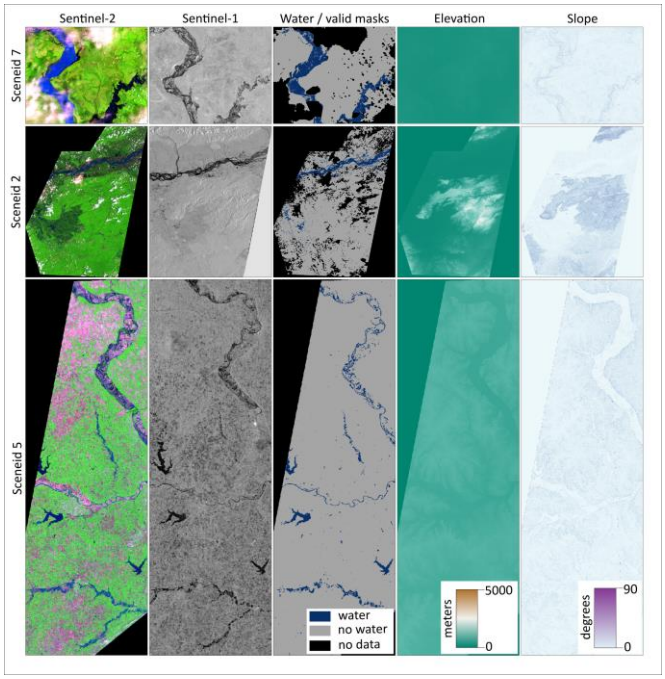

Fig. 3. Examples of *S1S2-Flood* samples with Sentinel-1 and Sentinel-2 satellite images as well as associated water / valid masks, elevation and slope information. Scenes are of varying size and coverage details are provided in Table I.

A. S1S2-Water dataset

The *S1S2-Water* dataset follows recent guidelines for the construction of remote sensing benchmark datasets as much as possible. A stratified random sampling was performed to be representative of a variety of climatic, atmospheric and land cover conditions, to cover different seasons and to ensure that samples always include water areas in addition to other land cover classes (Figure 1). In addition, a fixed division into training, validation and test splits is defined to ensure the transparency and repeatability of experiments. The dataset follows Open Source standards regarding data formats and structure. Each sample consists of Sentinel-1 Ground Range Detected (GRD) Interferometric Wide (IW) swath data and Sentinel-2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata (Figure 2). Each sample is aligned to the Sentinel-2 tile-grid and covers an extent of 100 x 100 km with a pixel-spacing of 10 m.

The final dataset consists of 65 samples (each 100 x 100 km in size) that are spread across 29 countries and cover an area of approximately 650,000 km². The samples cover 18 pre-dominant landcover types and show a wide distribution across elevation and slope. Images have been acquired between 2018-05-21 and 2020-11-26 and are distributed across nearly all months of the year (no samples are available for December). The difference between Sentinel-1 and Sentinel-2 acquisitions per sample is 1 day in average with a standard deviation of 4 days.

We decided against tiling the samples into smaller patches, but provide a Python package along with the data to do so instead. This way, the user has greater flexibility to prepare the samples according to requirements of the desired network architecture and hardware setup. A tiling of *S1S2-Water* into non-overlapping 256 x 256 pixels patches, results in a total of 50,000+ patches for training and 25,000+ patches for validation and testing respectively. Class distribution in the training data is 1 (“water”) to 7.5 (“background”). Dataset and preparation package are released openly alongside this publication [27].

B. Independent flood water dataset

The *S1S2-Water* dataset covers normal water bodies and does not specifically consider anomalously flooded areas. To evaluate the performance impact when a model is trained on normal water data (*S1S2-Water*) and applied to floods, we developed an independent reference dataset *S1S2-Flood* that covers 12 major flood events across the globe (Table I). Similar to *S1S2-Water*, each sample of this *S1S2-Flood* dataset consists of Sentinel-1 GRD and Sentinel-2 L1C images with associated quality-controlled binary water mask annotations, elevation and slope layers as well as metadata (Figure 3). Maximum time difference between acquisition dates of Sentinel-1 and Sentinel-2 images has been limited to one day to ensure spatial and temporal consistency of the flood masks. We used the same procedure as for *S1S2-Water* samples to annotate the satellite images. First, water bodies are roughly identified using a threshold procedure based on NDWI and then manually refined for each sensor specifically in multiple iterations using extensive quality checks and corrections.

We did not rely on the available rapid mapping products of the Copernicus Emergency Management Service (CEMS) to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

ensure consistency between input images and target masks. Most flood delineation products of the CEMS are derived from very high-resolution optical or SAR satellite images. Using their products directly as annotations masks alongside Sentinel-1 or Sentinel-2 images could cause significant amounts of label noise related to geometric or temporal differences between annotation source and reference image. CEMS flood delineation products that were derived directly from Sentinel-1 or Sentinel-2 images could potentially be very valuable samples. However, these products seem to be extracted in a semi-automated manner with unclear levels manual improvements and quality checks. They are, moreover, produced in a rapid mapping context where production time is critical and usually of higher priority than annotation quality. This means that even these samples should be treated with care and would require additional quality and consistency checks before including them in a reference dataset like *S1S2-Flood*.

TABLE I
OVERVIEW OF FLOOD EVENTS AND RESPECTIVE SOURCE IMAGES IN THE S1S2-FLOOD DATASET.

ID	Location	Date Sentinel-1	Date Sentinel-2	Coverage (km²)	Split
1	Greece, Thessaloniki	2018-02-28	2018-02-28	21,500	Test
2	India, Bishwanath	2016-08-12	2016-08-12	16,800	Test
3	Spain, Murcia	2019-09-17	2019-09-18	1,600	Test
4	Sweden, Mora	2018-05-10	2018-05-10	5,200	Test
5	USA, Kansas City	2019-05-22	2019-05-22	9,500	Test
6	Vietnam, Vinh	2017-08-11	2017-08-12	9,500	Test
7	Ghana, Janga	2018-09-18	2018-09-19	1,400	Train
8	Paraguay, Asunción	2018-10-31	2018-10-31	17,000	Train
9	Greece, Kavala	2018-06-29	2018-06-30	2,700	Train
10	Ethiopia, Kelafo	2018-05-07	2018-05-08	155	Train
11	Kenya, Merti	2018-05-04	2018-05-04	10,500	Val
12	Italy, Parma	2017-12-12	2017-12-13	2,500	Val

A fixed random split into training, validation and test samples is finally applied. Due to the limited number of available samples and to increase the significance of the results, a larger preference (in terms of number of samples) is given to the test split. Therefore, the *S1S2-Flood* dataset should be considered as an independent test dataset. Small training and validation splits are separated only for transfer learning experiments.

III. METHOD

In the following, we describe methods applied to compile the *S1S2-Water* dataset, including the design of the sampling scheme, image preparation and annotation procedures as well as data format and structure considerations. To test the applicability of *S1S2-Water* for normal water monitoring and flood mapping we carry out performance evaluation of convolutional neural network architectures. All experiments are carried out separately for Sentinel-1 and Sentinel-2. First, we compare the performance of different network architectures and choose a baseline model for each sensor. Further, we evaluate the influence of different input image bands and the effects of data augmentation methods on the performance.

A. Sampling scheme

A reference dataset for machine learning should represent the real probability distribution of the data, which is to be expected in the operational use of the developed methods. Since the number of samples in a dataset is always limited, care must be taken when selecting the samples so that they are representative for the task. Simple random samples often cannot meet this requirement. It is important to cover both the distribution of the target class and the variance of environmental parameters as good as possible. A suitable tool for this is the stratified random sample, which makes the selection of samples dependent on additional information about desired properties. For the *S1S2-Water* dataset we use the global Sentinel-2 tile-grid with 100 x 100 km grid cells as geographical reference grid from which we choose the samples. For each grid cell we compute the pre-dominant landcover type from a global landcover map of the ESA CCI Land Cover project with a native spatial resolution of 300 m [29]. Since we want each sample to contain a minimum percentage of water, we remove grid cells that do not cover any water bodies. We use the Global Surface Water maximum water extent layer developed by [30] as indicator for the presence of water. The remaining grid cells contain water bodies, have an assigned pre-dominant landcover type and form the strata for our stratified random sampling. During sampling we also apply a minimum distance constraint of 370 km, which equals twice the swath width of a Sentinel-2 scene, to avoid choosing neighboring grid cells.

Developing a supervised machine learning model requires a reference dataset, which is commonly split into training, validation and test data. Each of these parts has a specific function and is assigned to a phase of model development. The interaction of the three data splits requires at least that they should come from the same data-generating process, the data should follow the same probability distribution and be independent as well as randomly distributed. A fixed split of the data is also important for model comparison. Only when different models are evaluated using the same test data, their performance can be compared. Moreover, for geospatial data it is crucial to minimize the spatial autocorrelation between samples. Spatial autocorrelation describes the assumption that

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

the similarity of the properties of two samples increases with decreasing spatial distance. In practice, this means that one should avoid tiling individual large satellite scenes and dividing the resulting smaller patches into the different data splits, since neighboring tiles could end up in the training and test splits. Therefore, we decided to apply a fixed train, validation and test splitting at the level of individual satellite scenes that have themselves already been sampled under consideration of a minimum distance constraint as described above.

B. Data format and structure

The *S1S2-Water* dataset exposes all samples as Spatio Temporal Asset Catalog (STAC) items to make spatio-temporal searches and attribute filtering easy and according to Open Source standards [31]. STAC defines a basic metadata format for geodata with the aim of unifying access to and querying geodata from a wide range of providers and enabling it with the help of a single Application Programming Interface (API). Metadata items in *S1S2-Water* contain information about the sample footprint and spatial reference system as well as properties, such as landcover, split (train, validation, test) or source-ids of the Sentinel-1 and Sentinel-2 scenes that have been used to generate the sample. The STAC item further contains references to the assets which are the actual raster data layers (Sentinel-1 and Sentinel-2 images, water masks, valid pixel masks, as well as elevation and slope layers).

Raster data layers are stored and made available as Cloud Optimized GeoTIFF (COG) [32]. While with standard GeoTIFF format it is necessary to download raster data completely, even if the section to be examined only occupies a small part of the raster file, the COG format allows the targeted download of the desired image section. The conversion to the COG format is based on two general steps: On the one hand, the raster data is internally divided into tiles, which can later be made available directly via an HTTP request. On the other hand, one or more low-resolution variants of the grid are created, which enable the data to be provided more quickly, e.g. for graphical representation. The combination of the STAC and COG standards with an object storage service optimized for data storage on cloud platforms such as AWS Simple Storage Service (S3) therefore enables the exploration of the data sets with a uniform API and the targeted streaming of the desired bytes, thus laying the basis for a cloud-based processing.

C. Data preparation and annotation

Figure 4 gives an overview of the processing steps used to prepare the *S1S2-Water* samples. The Sentinel-1 images are geometrically corrected and radiometrically calibrated according to the method described by Twele et al. (2016) [33]. Both VV and VH polarizations are used and stored in separate image channels. All images are cropped to the spatial extent of the sample. In case multiple Sentinel-1 images are required to cover a sample, they are stitched together prior to cropping. For each sample, additional Sentinel-2 scenes are acquired within a maximum time difference of 21 days (the average time difference between acquisition times is 1 day) to the Sentinel-1 image. The choice of a respective Sentinel-2 image is further constrained by a filter on the cloud-cover (5 %). Once acquired, the Sentinel-2 image is resampled to a uniform resolution and stored in separate image channels. Based on findings from previous work and as described in [11], only the spectral channels blue (B), green (G), red (R), near infrared (NIR), shortwave infrared I (SWIR1) and II (SWIR2) are used.

Multi-spectral Sentinel-2 scenes are usually easier to interpret compared to SAR images of Sentinel-1. Therefore, we used Sentinel-2 images as starting point for the annotation of water masks. These distinguish the classes "water" from "no water" (background). We use a semi-automated annotation procedure, in which water bodies are first roughly identified using a threshold procedure based on a Normalized Difference Water Index (NDWI) and then refined using extensive manual quality checks and corrections. The manual editing is performed in multiple iterations by three independent image interpretation experts to ensure the quality and consistency of the resulting reference masks. Within each iteration, water masks are visually compared with the reference images by the expert and in case of discrepancies are corrected accordingly. Water masks are adjusted to the images of each sensor specifically. This is required since in some areas even minor differences in the acquisition time can result in major discrepancies of the observed water extent. For example, in highly dynamic water regimes (e.g., tidal areas or during floods) the water extent can significantly change in short time and thus even images acquired at the same day but at different hours may not be comparable. In addition to the water masks, valid pixel masks are created to identify areas where the reference data is of insufficient quality or where no image data is available (e.g., due to partial cloud cover). Each sample is also supplemented by metadata and a DEM. Both elevation and slope information are provided.

D. Model hyperparameters and training setup

All models are trained using the same values for initial learning rate (1e-3) and weight decay (1e-2), which have been obtained through initial experiments. Model parameters are initialized as described in He et al. (2015) [34]. For the optimization algorithm we select AdamW with $\beta1=0.9$ and $\beta2=0.9$ [35]. We use a weighted combination of cross-entropy and Lovász loss that optimizes the IoU score during backpropagation. Following the results of Helleis et al. (2022) [18] we equally weight cross-entropy and Lovász loss. To account for class-imbalance we, moreover, weight cross-entropy according to the positive class distribution in the training data. We train the networks in batches of 64 until convergence for a maximum number of 100 epochs. The learning rate is step-wise reduced by a factor of 0.1 if no improvement is seen for three epochs. If no improvement is observed for nine consecutive epochs, we apply early stopping. During training the best model state is determined by the lowest validation loss, which is also stored for evaluation. All experiments are performed on a machine

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

running CentOS 8.2 with two NVIDIA Tesla K80 GPUs and Intel Xeon E5-2697A CPUs.

We split the scenes contained in the training and validation sets into non-overlapping tiles with a size of 256 x 256 pixels. The input image feature space in all experiments is standardized to zero mean and unit variance. Mean and standard deviation have been computed on the training dataset and are applied to the validation and testing sets. During testing, we split the test image scenes into overlapping tiles of 256 x 256 pixels before feeding them to the trained models for inference. The resulting segmentation maps are then recombined to the original scene size and a tapered cosine window function is used to smoothly blend overlapping tiles together with the aim to reduce prediction errors close to the tile borders [11].

E. Performance metrics

To assess the segmentation performance, we report standard accuracy metrics Intersection over Union score (IoU), Precision (Prec) and Recall (Rec). Overall accuracy is not reported in this study due to the imbalanced class distribution of the datasets under evaluation, which would render this metric unreliable and misleading. Prediction probabilities are converted into binary segmentation masks by applying a threshold of 0.5 on the probabilities. A pixel belongs to the water class if its probability >= 0.5, otherwise it belongs to the background class. We compute the performance metrics for every scene in the test set and further compute the weighted averages across all scenes. Our implementation of the used metrics is published as an open-source Python package [36]. Given the overall dataset size and the number of experiments in this study, cross-validation is not considered because of the computational overhead of running each experiment multiple times. Instead, we fix a random seed (4) for Numpy, PyTorch, and Python and apply it to all components of the experimental setup that involve randomness (e.g., data shuffling, weight initialization, etc.). Deterministic GPU floating point calculations are enforced to ensure that results are reproducible and comparable. We further report model throughput measured in megapixel per second (mp/s). Measurements per experiment are averaged across five prediction runs on 5,000 tiles with shape (256, 256, n), with n being the number of image bands.

F. Baseline model (BM)

The models that we test in this study are based on the widely-used U-Net architecture [37], which has proven to deliver highly accurate results for water segmentation tasks in high-resolution satellite images at relatively low computational complexity [11], [18]. In combination with the U-Net architectures we compare different encoders, namely MobileNet-V3, ResNet-50, EfficientNet-B0 and EfficientNet-B4. We selected these for our experiments since they show a good trade-off between number of model parameters and Imagenet Top-1 accuracy [38]. The hypothesis is, that more complex models may allow the model to have a higher level of image understanding, which may positively impact on the segmentation performance and in particular on the generalization ability.

G. Input bands (IB)

For Sentinel-2 we use only spectral bands that are available across different satellite sensors (e.g., Landsat OLI) to ensure a high degree of transferability of the trained models. Water shows low reflectance in the NIR and SWIR wavelengths as it absorbs more energy, while non-water generally has a higher reflectance. This leads to a high contrast in reflectance values between water and non-water landcover classes in the NIR and SWIR spectral bands compared to the visible R, G and B bands. Moreover, atmospheric aerosols have a stronger effect on shorter wavelengths and affect mainly the visible spectrum. This means that particularly the SWIR spectral bands show improved atmospheric transparency and hence can support in the distinction between landcover and atmospheric effects like clouds or haze. To this regard, we test the influence of the NIR and SWIR spectral bands on the water segmentation performance. Additionally, we also consider slope information derived from the freely and globally available Copernicus DEM. We compute slope in degrees at the 30 m native spatial resolution of the DEM and resample it to match the respective sample image resolution of 10 m. We apply different band combinations separately on the training dataset to test their individual contribution to the segmentation performance. Similar for Sentinel-1, we test the influence of different available polarizations (VV and VH) and their combination with slope information.

H. Data augmentation (DA)

Satellite images are affected by changes in landcover, atmospheric conditions, seasonality and other scene and image properties such as sun elevation or radiometric resolution. Due to this very large variability of influencing factor, even large reference datasets may not cover all possibilities that may occur in real-world applications. To this regard, data augmentation enables a network to learn invariance to changes in the augmented domains to a degree that may go beyond what is present in the raw training image. In this experiment we test the influence of different data augmentation techniques on the segmentation performance. Specifically, we apply random contrast, brightness, scale and image flipping. Factors are randomly applied within predefined ranges to contrast [-0.1, 0.1], brightness [-0.1, 0.1] and scale [0.9, 1.1]. Flipping is performed in left-right direction, which flips an image around its vertical axis. We do not apply rotation augmentation since it may introduce unrealistic SAR image characteristics [39]. All augmentations are applied with equal probability. We apply the different augmentation methods separately on the training dataset to test their individual influence. In an additional test we combine all augmentations and test their joint influence on the segmentation performance during training. We also perform an experiment with test-time augmentation, in which we predict on the original input image and five randomly augmented versions of it. The final prediction is then based on the averaged class probabilities.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

I. Transfer model (TM)

In this experiment we aim at answering the question, whether a model trained on images depicting normal water can be transferred to flood images. Compared to normal water, flood water is mostly characterized by higher contents of sediment and debris. Flooded vegetation, infrastructure or vehicles may further interact with the water surface and modify the reflectance characteristics of the target class in the satellite images. As baseline, we evaluate the performance of sensor-specific models that have been trained solely on normal water (*S1S2-Water*) and apply them to the test images of our independent flood water dataset (*S1S2-Flood*). We then compare these results to models that have been trained on a joint dataset that includes normal and flood water images (*S1S2-Water + S1S2-Flood*). Starting from pre-trained weights can improve performance, because low-level features that are

TABLE II

RESULTS FOR DIFFERENT BASELINE MODELS. MODELS HAVE BEEN TRAINED FOR EACH SENSOR SPECIFICALLY WITH ALL AVAILABLE IMAGE BANDS AND ADDITIONAL SLOPE INFORMATION.

ID	Decoder	Encoder	Throughput (mp/s)	Sentinel-1			Sentinel-2		
				IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
BM-0	U-Net	ResNet-50	9	.842	.955	.873	.927	.989	.937
BM-1	U-Net	Mobilenet-V3	17	.842	.954	.873	.911	.983	.928
BM-2	U-Net	EfficientNet-B0	12	.844	.950	.882	.940	.989	.951
BM-3	U-Net	EfficientNet-B4	7	.850	.950	.885	.922	.987	.935

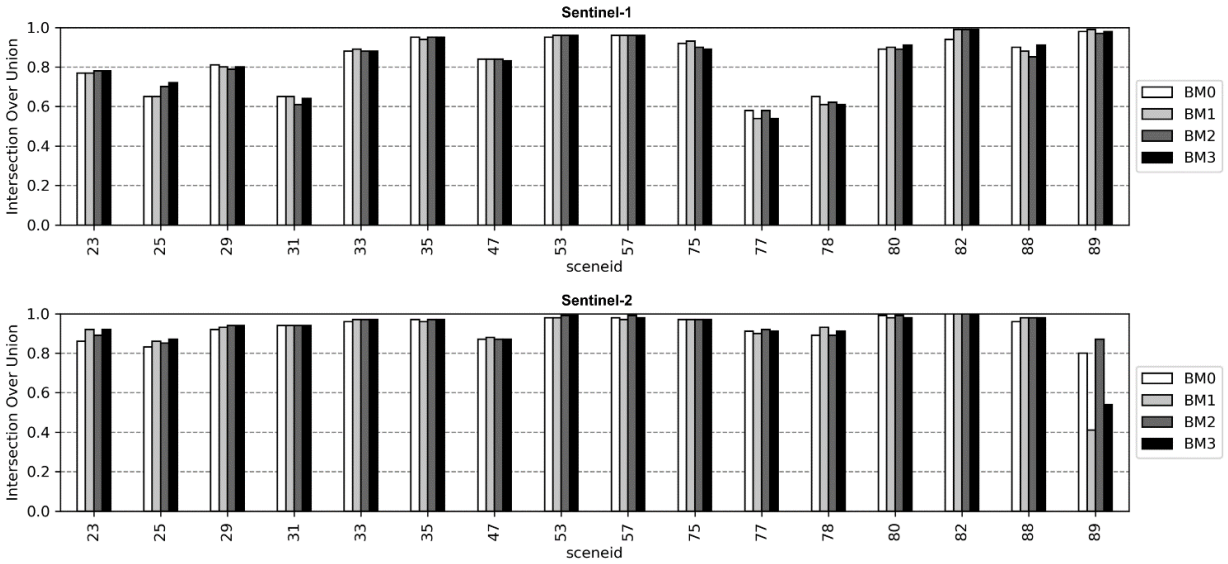

Fig. 5. Comparison of baseline models for each test scene.

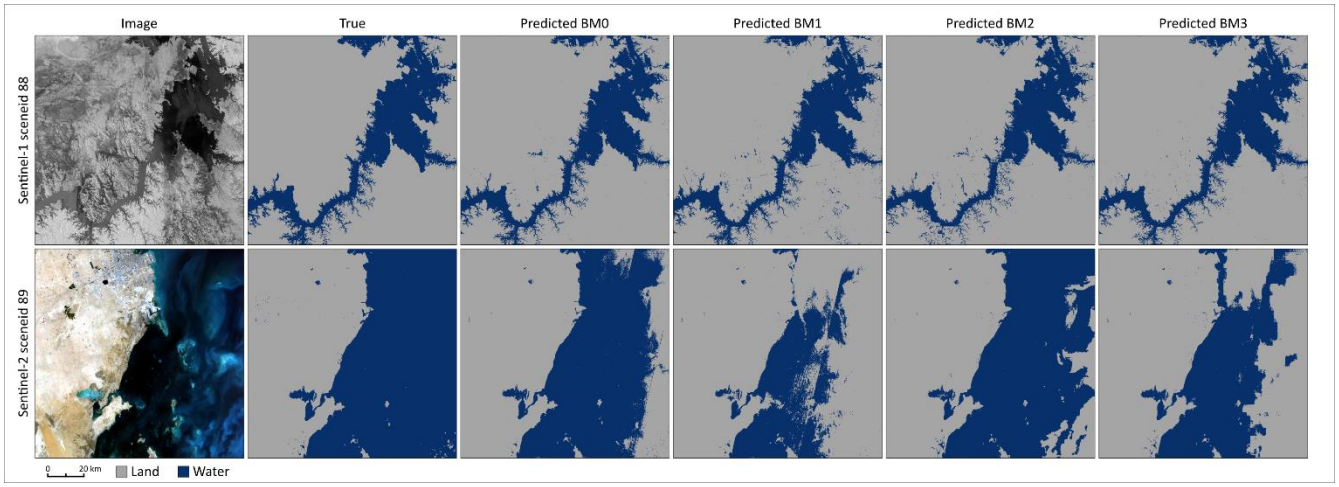

Fig. 6. Comparison of baseline model predictions for selected test scenes. Each scene covers an area of 100 x 100 km.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

being learned in early network layers are similar across image domains. Therefore, we further test models that have been pre-trained on *S1S2-Water* and fine-tune them on *S1S2-Flood*.

IV. Results

Table II shows results for different baseline models (BM). Models have been trained for each sensor separately with all available image bands and additional slope information. No data augmentation has been performed at this stage. It can be seen that the tested encoders have a significant impact on the inference speed, with the more complex models like ResNet-50 or EfficientNet-B4 having much lower throughput compare to Mobilenet-V3 and EfficientNet-B0. In terms of accuracy, however, the differences between tested baseline models are rather small. The Sentinel-1 models perform equally well, with EfficientNet-B4 showing a marginally higher IoU across all test images. Differences between models are slightly more visible for Sentinel-2. In this case EfficientNet-B0 performs better than the other baseline models. In general, we can observe higher test scores for Sentinel-2 compared to Sentinel-1. This difference between sensors is largely attributed to higher recall values for Sentinel-2. Sentinel-1 models, therefore, seem to miss more water pixels and hence produce large amounts of false negatives. Considering throughput and test scores as equally weighted decision criterions, we select EfficientNet-B0 as the preferred baseline model for further experiments.

TABLE III
Results for different Sentinel-1 input bands and additional slope information.

ID	Input bands	Throughput (mp/s)	Sentinel-1		
			IoU	*Prec*	*Rec*
IB-0	VV	12	.677	.771	.846
IB-1	VH	12	.724	.838	.806
IB-2	VV-VH	12	.772	.869	.862
IB-3	VV-VH-SLOPE	12	.844	.950	.882

Figure 5 compares the baseline models individually for each test scene. Larger variance across scenes can be observed for Sentinel-1 compared to Sentinel-2. While the variance of IoU values for different models within scenes is generally relatively small, a few scenes seem to dominate the overall metrics. Figure 6 shows two examples of scenes with larger discrepancies between different models. Sentinel-1 scene 88 depicts an arid landscape with large patches of bare soil and sand (pre-dominant landcover type "bare areas"). The more complex models (BM-0 and BM-3) show better performance in this setting (0.901 IoU for BM-0 and 0.913 IoU for BM-3), which is particularly challenging for water detection in SAR images, compared to BM-1 and BM-2 (0.878 IoU for BM-1 and 0.851 IoU for BM-2). Similarly, the complex water situation in scene 89 (pre-dominant landcover type "water bodies") with several shallow sandbanks and partially high turbidity seems to strongly influence the performance of the Sentinel-2 baseline models. Here BM-0 and BM-2 show better performance (0.806 IoU for BM-0 and 0.872 IoU for BM-2) compared to BM-1 and BM-3 (0.409 IoU for BM-1 and 0.537 IoU for BM-3). The overall performance on this scene, however, leaves room for improvement.

TABLE IV
Results for different Sentinel-2 input bands and additional slope information.

ID	Input bands	Throughput (mp/s)	Sentinel-2		
			IoU	*Prec*	*Rec*
IB-0	R-G-B	12	.830	.965	.854
IB-1	R-G-B-NIR	12	.919	.989	.929
IB-2	R-G-B-NIR-SWIR1	12	.921	.991	.934
IB-3	R-G-B-NIR-SWIR2	12	.921	.994	.928
IB-4	R-G-B-NIR-SWIR1-SWIR2	12	.922	.994	.933
IB-5	R-G-B-NIR-SWIR1-SWIR2-SLOPE	12	.940	.989	.951
IB-6	R-G-B-NIR-SLOPE	12	.945	.985	.959

TABLE V
Results for different data augmentation techniques.

ID	Augmen-tation	Sentinel-1			Sentinel-2		
		IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
DA-0	None	.844	.950	.882	.945	.985	.959
DA-1	Left-right flip	.837	.949	.872	.938	.988	.949
DA-2	Scale and crop	.833	.947	.869	.945	.990	.954
DA-3	Brightness and contrast	.834	.945	.876	.933	.989	.944
DA-4	DA-1 + DA-2 + DA-3	**.838**	**.954**	**.877**	**.962**	**.991**	**.974**
DA-5	DA-1 + DA-2	.835	.958	.870	.946	.988	.959
DA-6	DA-1 + DA-3	.834	.945	.877	.937	.987	.950
DA-7	DA-2 + DA-3	.834	.948	.875	.931	.991	.938
DA-8	Test-time	.842	.953	.880	.937	.991	.946

Table III and Table IV show the influence of different input bands and additional slope information on the performance of the selected baseline model BM-2 (U-Net with EfficientNet-B0 encoder). For Sentinel-1 (Table III) it can be seen that VH polarization is superior over VV polarization in discriminating water bodies from background. A combination of VV and VH polarizations can further improve the test

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

scores of the water segmentation by a large margin. A similar improvement is achieved when slope information is added to the input feature space. The main contribution to the overall gain in IoU is attributed to Prec. The additional input bands seem to support the reduction of false positives.

The major contribution to test score improvement in Sentinel-2 images can be explained by adding the NIR spectral band. Compared to using R-G-B bands only, we can achieve an improvement of 0.089 IoU. SWIR bands do not improve the overall test scores, while slope information does. The SWIR bands even seem to have a negative effect on a few scenes (Figure 7). In particular, aforementioned Sentinel-2 scene 89 seems to benefit from removing the SWIR bands from the input feature space. Contrary to Sentinel-1, we can observe that by adding additional input bands the recall improves while precision stays relatively stable. This indicates that NIR spectral band and slope information mainly contribute to reducing false negatives.

Table V summarizes the results for different data augmentation techniques using the best-performing model setups from previous experiments (U-Net with EfficientNet-B0 encoder for both sensors with IB-3 for Sentinel-1 and IB-6 for Sentinel-2). While augmentation during training slightly improves IoU for Sentinel-2, no improvements can be observed for Sentinel-1. The best results are achieved for DA-4, which combines several augmentations (left-right flip, scale and crop, brightness and contrast). Test-time augmentation (DA-8) does not improve results. Since every image is augmented and predicted multiple times before the final prediction can be assigned, inference times of this approach are longer compared to train-time augmentation.

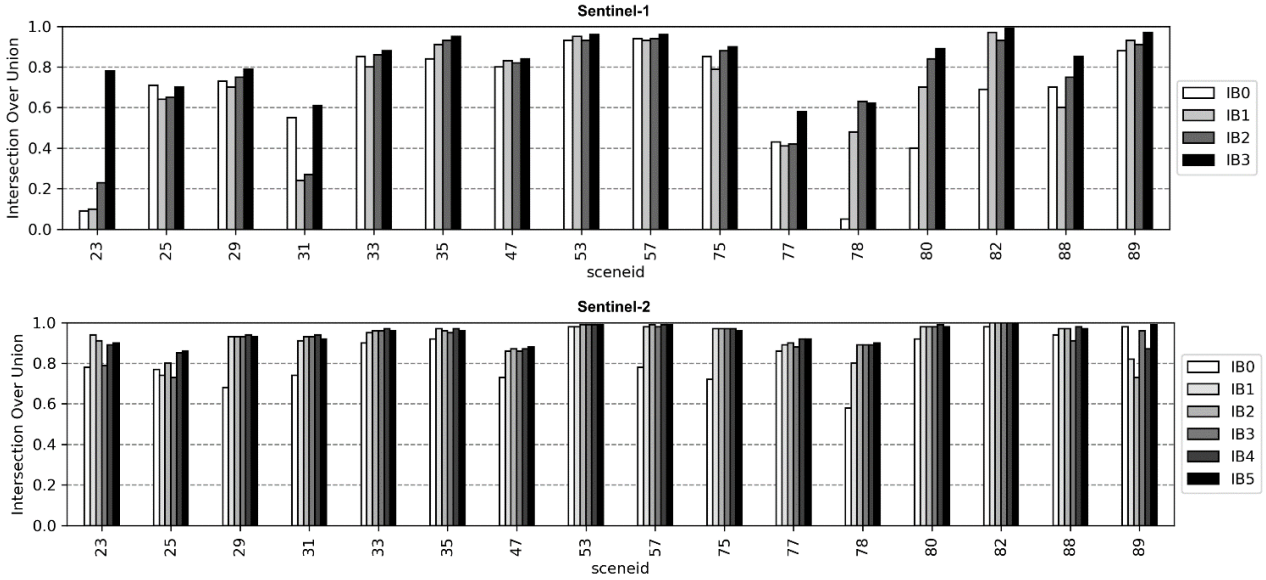

Fig. 7. Comparison of models with different input bands and additional slope information for each test scene.

TABLE VI
RESULTS FOR MODEL TRANSFER BETWEEN NORMAL AND FLOOD WATER SCENES.

ID	Transfer method	Test on *S1S2-Water*						Test on *S1S2-Flood*					
		Sentinel-1			Sentinel-2			Sentinel-1			Sentinel-2		
		IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
TM-0	Train on *S1S2-Water* only (DA-4)	.838	.954	.877	.962	.991	.974	.752	.919	.799	.816	.944	.864
TM-1	Train on *S1S2-Water* with additional flood scenes	**.845**	**.932**	**.896**	**.965**	**.989**	**.951**	**.758**	**.912**	**.809**	**.855**	**.938**	**.908**
TM-2	Fine-tune *S1S2-Water* (DA-4) on additional flood scenes	.841	.948	.877	.931	.985	.948	.739	.911	.789	.862	.940	.913

Table VI shows the results for model transfer between normal and flood water scenes. In this experiment we compare three different training scenarios. Directly applying the sensor-specific models that have been trained solely on normal water (*S1S2-Water*) to the test images of our independent flood water dataset (*S1S2-Flood*) produces reasonable results with IoU of 0.752 for Sentinel-1 and 0.816 for Sentinel-2 (TM-0). In comparison to training and testing on *S1S2-Water*, however, we can observe a clear drop in test scores for Sentinel-1 (-0.086 difference in IoU) and Sentinel-2 (-0.110 difference in IoU). The largest influence on the performance loss for both sensors can be attributed to lower recall values, which indicates that the models trained on *S1S2-Water* tend to

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

under-detect flood water. Training on *S1S2-Water* with additional flood scenes for training and validation splits, we can achieve slightly better results in all test scenarios (TM-1). Improvements on the *S1S2-Flood* test dataset are especially visible for Sentinel-2, for which we can achieve an increase in IoU of 0.039 compared to training on *S1S2-Water* only. We further test models that have been pre-trained on *S1S2-Water* and fine-tune them on *S1S2-Flood* (TM-2). However, fine-

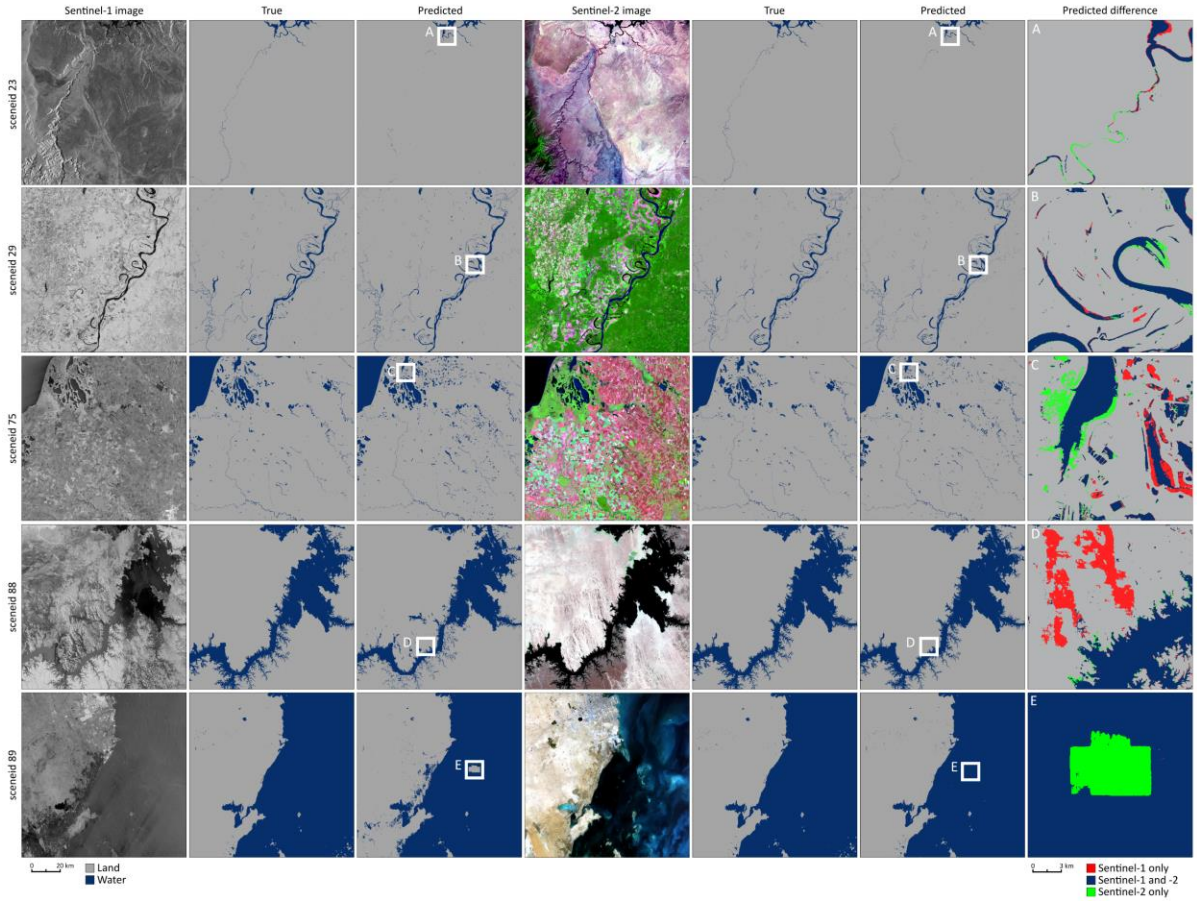

Fig. 8. Examples of best-performing models (TM-1) for Sentinel-1 and Sentinel-2 on test images of S1S2-Water.

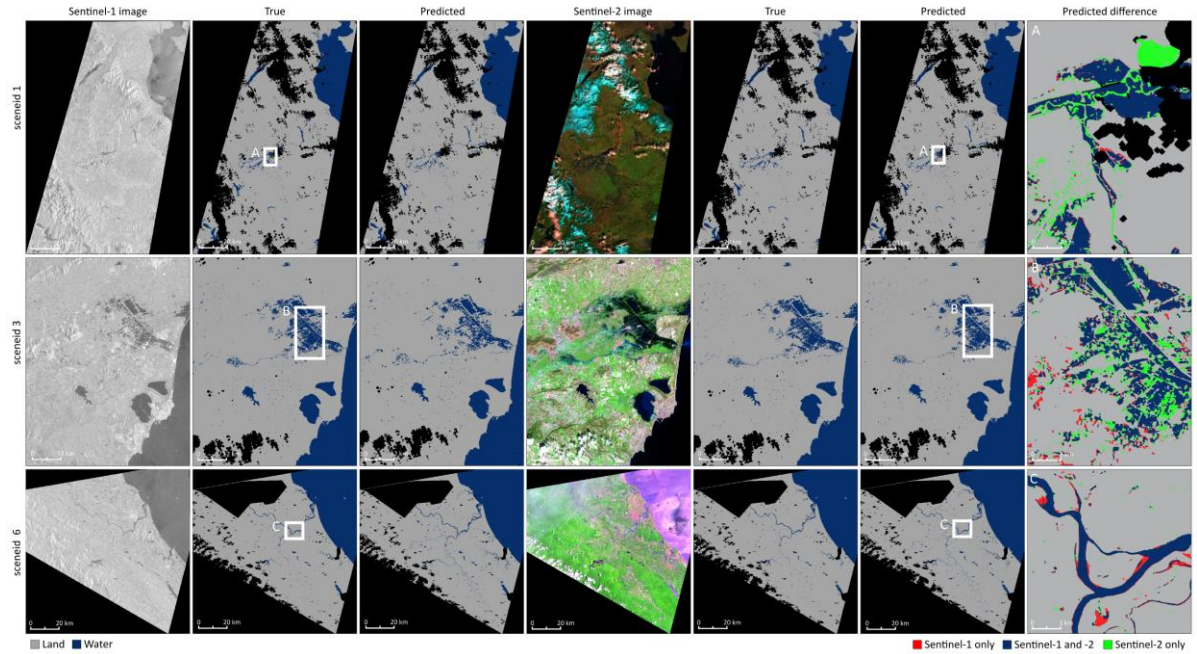

Fig. 9. Examples of best-performing models (TM-1) for Sentinel-1 and Sentinel-2 on test images of S1S2-Flood.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

tuning does not improve the performance on any of the datasets and in this experimental setup training on *S1S2-Water* with additional flood scenes is the preferred option to gain models with better generalization ability.

Figure 8 shows results of the best-performing models (TM-1) for Sentinel-1 and Sentinel-2 test images of *S1S2-Water*. The models can produce highly accurate segmentation masks that are, moreover, largely consistent between Sentinel-1 and Sentinel-2 scenes for the same sample. Discrepancies between sensors occur mainly over bare soil and sand, where the Sentinel-1 model tends to overestimate water while the Sentinel-2 model predicts correctly no water (e.g., sceneid 88). Moreover, water with partial vegetation cover (e.g., sceneid 29) and narrow rivers (e.g., sceneid 23) are better segmented in Sentinel-2 than in Sentinel-1. Wind-induced roughening effects over large open water bodies are known to negatively impact on the detection of water in SAR images as they prevent specular reflection [1]. While our model seems to be able to deal well with such effects over inland water bodies, we could still observe related false negatives in few samples over the open ocean (e.g., sceneid 89). Other discrepancies are likely caused by the difference in acquisition dates for some of the samples that depict highly dynamic surface water environments (e.g., sceneid 75 with 3 days between acquisitions).

Figure 9 shows examples of the best-performing models (TM-1) for Sentinel-1 and Sentinel-2 images tested on *S1S2-Flood* samples. The flood scene over Thessaloniki, Greece (sceneid 1) shows generally good overlap between water segmentations derived from Sentinel-1 and Sentinel-2. Minor discrepancies occur along narrow water channels where Sentinel-2 segments the existing water extent more accurately. However, also false positives can be observed in the Sentinel-2 water masks; these occur especially in areas where cloud-shadow has not been fully detected by the cloud and cloud-shadow mask algorithm as part of the valid mask [2]. Similar to our observations on the *S1S2-Water* test dataset, some discrepancies between sensors are the result of a highly dynamic surface water situation coupled with a time difference between acquisition dates. In the case of floods even shorter differences between acquisitions can cause large variations in the water extent and thus induce discrepancies in segmentation masks between Sentinel-1 and Sentinel-2 (e.g., sceneids 3 and 6 with 1 day between acquisitions).

Figure 10 depicts results of the best-performing models (TM-1) for Sentinel-1 and Sentinel-2 on test images of *S1S2-Water* and *S1S2-Flood* grouped by predominant landcover. With IoU values above 0.8 Sentinel-2 generally performs better than Sentinel-1 across all landcover types. The Sentinel-1 model seems to have issues in scenes that are dominated by “sparse vegetation”, “tree cover, needleleaved, deciduous, closed to open” and “grassland”. A large spread in IoU values can, moreover, be observed in “cropland rainfed” scenes. Relative to the other classes, “Grassland” and “cropland rainfed” scenes also show inferior performance with the Sentinel-2 model.

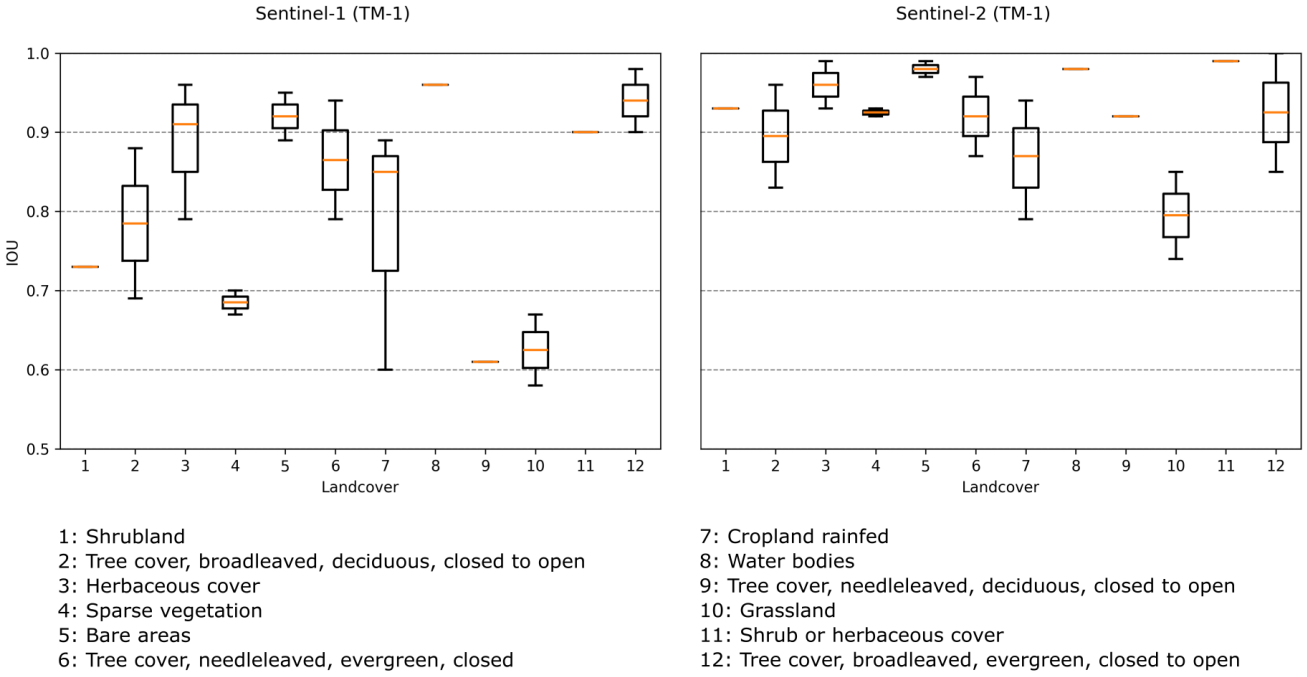

Fig. 10. Results of best-performing models (TM-1) for Sentinel-1 and Sentinel-2 on test images of S1S2-Water and S1S2-Flood grouped by predominant landcover.

Table VII shows performance of the TM-1 model on the test splits of *Sen1Floods11* (Sentinel-1) and *WorldFloods* (Sentinel-2) datasets. A performance decrease on both datasets can be observed when compared to the results on our *S1S2-*

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

Flood dataset. This is largely attributed to lower recall values, which indicates that our model tends to predict less water compared to the reference annotations of *Sen1Floods11* and *WorldFloods*.

TABLE VII
RESULTS OF BEST-PERFORMING MODELS (TM-1) ON SEN1FLOODS11 AND WORLDFLOODS DATASETS.

Test dataset	TM-1		
	IoU	*Prec*	*Rec*
Sen1Floods11 (Sentinel-1)	.556	.824	.631
WorldFloods (Sentinel-2)	.608	.861	.675

V. DISCUSSION

This study introduces a new global reference dataset for semantic segmentation of surface water bodies in single-temporal Sentinel-1 and Sentinel-2 satellite images. Our *S1S2-Water* dataset aims to overcome the observed lack of appropriate remote sensing benchmark datasets for the segmentation of water bodies in multi-modal satellite images. It distinguishes itself from similar available datasets in that it follows recent guidelines for the development of remote sensing benchmark datasets [28] and specifically considers sample diversity and richness by applying a dedicated stratified sampling scheme. Moreover, the dataset aims to be scalable and builds on recent cloud-optimized open-source geospatial standards, namely STAC for metadata handling and COG as raster data format. Compared to other remote sensing reference datasets, we leave critical choices about data preparation (e.g., tile size, band selection) up to the user and provide a Python package to individually prepare data for training, validation and test of machine learning models.

Further differences with respect to the closely related *WorldFloods* [16], *OMBRIA* [17], *Sen1Floods11* [14] and *Sen12ms* [22] datasets are that we specifically adjust annotations between multiple satellite acquisitions of a sample and base the annotation procedure directly on the underlying images rather than independent sources. Spatial and temporal differences between annotation source and reference image can introduce significant degrees of label noise into a dataset. While this can be intended to increase training data with minimal annotation effort and even may support the generalization ability of a trained model [40], it is not acceptable for a representative independent test split used for benchmarking of models. Our experiments highlight the importance of the spatio-temporal alignment between images and annotation masks, especially in highly dynamic flood situations (Figure 9).

Image annotation is time- and resource-consuming even with the semi-automated procedure used in this study. Novel approaches to semi-automated labelling of satellite images as for example proposed by Geiss et al. (2020) [41] could potentially provide more accurate initial segmentation masks and hence reduce the manual interventions required to adjust them during post-processing. Quality checks by several independent operators are, however, essential for producing accurate annotations and more efforts should be spent on providing guidelines and standards for quality control of remote sensing benchmark datasets.

On the basis of *S1S2-Water* we provide a performance evaluation of convolutional neural networks and assess the influence of image bands, elevation features (slope) and data augmentation on the segmentation performance for Sentinel-1 and Sentinel-2 images. On the basis of a U-Net decoder, we compared different encoders with varying complexity to identify a suitable baseline model. While the differences in test scores of the trained baseline models are small, differences in model throughput dominate the final decision for a U-Net with Efficientnet-B0 encoder (Table II). While the average IoU values across all scenes for the tested baseline models are high (IoU >= 0.842 for Sentinel-1 and IoU >= 0.911 for Sentinel-2), test scores drop for single test scenes reveal the need for further model improvements by means of data-focused hyperparameters such as input image bands and augmentations (Figure 5).

Regarding the influence of input image bands, our results for Sentinel-1 confirm the findings of Liu et al. (2019) [12] and Helleis et al. (2022) [18] that jointly using VV and VH polarizations can significantly improve the water segmentation compared to using single polarized data. We observe an improvement of 0.095 IoU compared to using VV polarization alone and 0.048 IoU compared to using VH polarization alone (Table III). In our experimental setup, VH polarization seems to have a larger positive impact on the test scores of the water segmentation than VV polarization. This is contradicting with several studies that focus solely on VV polarization for water segmentation [33], [42]. These studies show that by using solely VV polarization land and water can be distinguished very well. While this is true in particular for smaller water bodies, VV polarization is sensitive to wind-induced roughening effects and hence prone to cause false-negatives over larger open water bodies. VH polarization on the contrary is known to be less sensitive to roughening effects and can aid in reducing false-negatives in such situations. Therefore, our results also underline the theoretical assumption that a combination of both polarizations works best in practice.

For Sentinel-2, combining the NIR spectral band with the R-G-B bands provided an improvement of 0.089 IoU compared to using R-G-B bands alone (Table IV). Adding SWIR bands improved the results only marginally. Theoretically SWIR bands should be less impacted by haze and thin cloud cover as the longer wavelengths show better atmospheric transparency, which may aid in the distinction between landcover and atmospheric effects. The added value of SWIR bands is also confirmed by previous studies of the authors [2], [11]. A possible explanation for the different findings in this study could be that *S1S2-Water* does not contain images with cloud-cover >5 %. Therefore, the benefits of enhanced atmospheric transparency of the SWIR spectral bands are likely not visible in the reported results. Expanding

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

the reference dataset with atmospherically more challenging scenes could provide further insights into this topic and help to better align trained models to real-world applications. Adding additional slope information from an independent DEM could further improve the test scores of water segmentations for both Sentinel-1 (improvement of 0.072 IoU) and Sentinel-2 (improvement of 0.023 IoU) compared to the best image band combinations (Table III and Table IV).

Various scene and image properties, such as the characteristics of the sensor being used, atmospheric conditions, landcover of the background class and appearance of the water class, can affect the target domain. Despite our aim to cover such variations as part of the reference dataset, we decided to test different data augmentation methods (Table V). Since augmentation increases the training sample size and allows to cover a larger range of conditions that may occur during inference, we would expect enhanced performance on the test dataset. However, compared to using no data augmentation (DA-0), random left-right flipping (DA-1), scaling and cropping (DA-2) as well as brightness and contrast augmentation (DA-3) lead to slightly worse test scores if performed separately. Only when the single augmentations are performed together (DA-4) we can observe an increase in test scores by 0.017 IoU for Sentinel-2. This observation is in line with the findings of a previous study of the authors [11] and provide an indication that augmentation can at least partially help to tackle variations related to atmospheric effects in L1C data that is being used in this study and that is not atmospherically corrected. A comparison with atmospherically corrected L2A data would be an interesting topic for a follow-up study in this direction. While test scores for Sentinel-2 improved by applying data augmentation as of DA-4, the ones for Sentinel-1 remained unchanged. As stated by Zhu et al. (2021) [39] and confirmed by Helleis et al. (2022) [18], geometric modifications of SAR images like those provided by Sentinel-1 need to be handled carefully since unrealistic image characteristics are easily introduced. Due to this reason we did not apply rotation augmentation in our experiments and applied scaling only within a very small range. More research into SAR-specific augmentations (e.g., speckle noise augmentation) would be needed to account for more complex variations in the target domain and to investigate their potential for model improvement.

S1S2-Water contains samples that depict normal water situations. It does not specifically contain flood water samples. Nonetheless, we evaluate how well models trained on S1S2-Water can be transferred to flood water scenes (Table VI). Similar to the findings of Bonafilia et al. (2020) [14] our experiments confirm that adding flood water samples to the training data supports model transfer and increases test scores on the independent *S1S2-Flood* dataset. Performance gains compared to using a model trained exclusively on *S1S2-Water* are very small for Sentinel-1 images (improvement of 0.006 IoU) and more substantial for Sentinel-2 images (improvement of 0.039 IoU). Training a model with a joint training dataset that contains normal and flood water samples (TM-1) outperforms fine-tuning a pre-trained model on *S1S2-Water* with additional flood samples (TM-2) on Sentinel-1. On Sentinel-2 images both approaches show the same improvement of test scores. James et al. (2021) [9] analyzed a similar model transfer for water segmentation in Sentinel-2 images. Even though their transfer experiment targeted geographical differences, their results emphasize the added value of retraining with limited samples of the target domain.

In summary, our results illustrate that more efforts should be spend on preparing high-quality reference data and dedicate more research towards data preparation and its incremental improvement. At last, modifications to the training dataset (Table III, Table IV, Table V and Table VI) enabled larger test score improvements compared to changes in the network architecture (Table II). Known limitations of the current version of the *S1S2-Water* dataset include limited availability of samples with complicated atmospheric conditions (e.g., dense cloud cover, haze or smoke), focus on normal water bodies (e.g., flood water bodies are not specifically considered), and underrepresentation of some predominant landcover classes for the background class (e.g., urban and bare areas, snow/ice). Based on these limitations, we can identify the following directions for future improvements of *S1S2-Water*. More samples with clouds and cloud shadows for Sentinel-2 are required to better distinguish between water and shadows. Adding task-specific samples could further improve segmentation accuracy for specific applications like flood mapping. A larger variety of landcover classes for the background class should be considered. Given provided data format and structure as well as the fact that all input data sources that we used to compile *S1S2-Water* are freely available, it should be possible to expand the dataset accordingly in a community-effort.

VI. Conclusions

In this study, we introduced *S1S2-Water* – a new reference dataset for segmentation of surface water bodies in single-tempoSentinel-1 and Sentinel-2 satellite images. The dataset aims to follow recent guidelines for the construction of remote sensing benchmark datasets by applying a stratified random sampling and a fixed division into training, validation and test splits to ensure the representativeness of samples as well as transparency and repeatability of experiments. The dataset follows Open Source standards regarding data formats and structure to support interoperability and scalability. It consists of 65 samples with size 100 x 100 km that are spread across 29 countries and cover an approximate area of 650,000 km².

Based on *S1S2-Water* we compared the performance of several CNN architectures and encoders (U-Net with ResNet-50, Mobilenet-V3, EfficientNet-B0 and EfficientNet-B4) to segment surface water in SAR and multi-spectral satellite images. Across several experiments we identified the superior performance of U-Net Efficientnet-B0 models, which show good generalization ability across varying environmental conditions and produce high accuracies at high throughput in both SAR and multi-spectral images. In this context, not only

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

the choice of model architecture and encoder is relevant, but also the sensor-specific input feature space and the way training data are augmented.

By successfully applying the model to six flood events (*S1S2-Flood*) and independent test splits of other benchmark datasets (*Sen1Floods11* and *WorldFloods*), we could highlight the usefulness of this work for rapid mapping activities to support situational awareness in emergency response. This underlines the findings of previous work of the authors that it is possible to train a model that is able to cope with highly diverse data availability scenarios in disaster situations [43]. However, more work is required to further adapt dataset and model training to specific scene properties that have caused misclassifications in the test images. These include for example adding additional samples for underrepresented landcover (background) classes like “bare areas” or “snow / ice”. To this regard, the *S1S2-Water* dataset should not be considered a static product. It has been designed to support incremental updates as part of a community effort.

The trained models are deployed in systematic water and flood monitoring services based on Sentinel-1 [18] and Sentinel-2 [11]. Ongoing and future works that are based on the *S1S2-Water* dataset include large-scale water monitoring studies [44] and probabilistic SAR-based flood segmentation with adapted Bayesian convolutional neural networks [45]. We hope that the *S1S2-Water* dataset will stimulate further research and development activities across various application domains and will support transparent and reproducible research.

REFERENCES

[1] S. Martinis, J. Kersten, and A. Twele, ‘A fully automated TerraSAR-X based flood service’, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 104, pp. 203–212, 2015, doi: 10.1016/j.isprsjprs.2014.07.014.

[2] M. Wieland, Y. Li, and S. Martinis, ‘Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network’, *Remote Sensing of Environment*, vol. 230, no. 111203, Art. no. 111203, 2019.

[3] C. Huang, Y. Chen, S. Zhang, and J. Wu, ‘Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review’, *Rev. Geophys.*, vol. 56, no. 2, Art. no. 2, 2018, doi: 10.1029/2018RG000598.

[4] R. Bentivoglio, E. Isufi, S. N. Jonkman, and R. Taormina, ‘Deep learning methods for flood mapping: a review of existing applications and future research directions’, *Hydrology and Earth System Sciences*, vol. 26, no. 16, pp. 4345–4378, 2022, doi: 10.5194/hess-26-4345-2022.

[5] L. Landuyt, A. Van Wesemael, G. J.-P. Schumann, R. Hostache, N. E. C. Verhoest, and F. M. B. Van Coillie, ‘Flood Mapping Based on Synthetic Aperture Radar: An Assessment of Established Approaches’, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 722–739, Feb. 2019, doi: 10.1109/TGRS.2018.2860054.

[6] L. Landuyt, F. M. B. Van Coillie, B. Vogels, J. Dewelde, and N. E. C. Verhoest, ‘Towards Operational Flood Monitoring in Flanders Using Sentinel-1’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11004–11018, 2021, doi: 10.1109/JSTARS.2021.3121992.

[7] C. Krullikowski *et al.*, ‘Estimating Ensemble Likelihoods for the Sentinel-1-Based Global Flood Monitoring Product of the Copernicus Emergency Management Service’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 6917–6930, 2023, doi: 10.1109/JSTARS.2023.3292350.

[8] F. Isikdogan, A. C. Bovik, and P. Passalacqua, ‘Surface Water Mapping by Deep Learning’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 11, Art. no. 11, 2017, doi: 10.1109/JSTARS.2017.2735443.

[9] T. James, C. Schillaci, and A. Lipani, ‘Convolutional neural networks for water segmentation using sentinel-2 red, green, blue (RGB) composites and derived spectral indices’, *International Journal of Remote Sensing*, vol. 42, no. 14, pp. 5338–5365, 2021, doi: 10.1080/01431161.2021.1913298.

[10] Y. Li, S. Martinis, M. Wieland, S. Schlaffer, and R. Natsuaki, ‘Urban Flood Mapping Using SAR Intensity and Interferometric Coherence via Bayesian Network Fusion’, *Remote Sensing*, vol. 11, no. 19, Art. no. 19, 2019, doi: 10.3390/rs11192231.

[11] M. Wieland and S. Martinis, ‘A modular processing chain for automated flood monitoring from multi-spectral satellite data’, *Remote Sensing*, vol. 11, no. 9, Art. no. 9, 2019.

[12] B. Liu, X. Li, and G. Zheng, ‘Coastal Inundation Mapping From Bitemporal and Dual-Polarization SAR Imagery Based on Deep Convolutional Neural Networks’, *Journal of Geophysical Research: Oceans*, vol. 124, no. 12, pp. 9101–9113, 2019, doi: 10.1029/2019JC015577.

[13] E. Nemni, J. Bullock, S. Belabbes, and L. Bromley, ‘Fully Convolutional Neural Network for Rapid Flood Segmentation in Synthetic Aperture Radar Imagery’, *Remote Sensing*, vol. 12, no. 16, Art. no. 16, 2020, doi: 10.3390/rs12162532.

[14] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, ‘Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1’, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 835–845. doi: 10.1109/CVPRW50498.2020.00113.

[15] Y. Bai *et al.*, ‘Enhancement of Detecting Permanent Water and Temporary Water in Flood Disasters by Fusing Sentinel-1 and Sentinel-2 Imagery Using Deep Learning Algorithms: Demonstration of Sen1Floods11 Benchmark Datasets’, *Remote Sensing*, vol. 13, no. 11, Art. no. 11, 2021, doi: 10.3390/rs13112220.

[16] G. Mateo-Garcia *et al.*, ‘Towards global flood mapping onboard low cost satellites with machine learning’, *Sci Rep*, vol. 11, no. 1, p. 7249, Dec. 2021, doi: 10.1038/s41598-021-86650-z.

[17] G. I. Drakonakis, G. Tsagkatakis, K. Fotiadou, and P. Tsakalides, ‘OmbriaNet—Supervised Flood Mapping via Convolutional Neural Networks Using Multitemporal Sentinel-1 and Sentinel-2 Data Fusion’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2341–2356, 2022, doi: 10.1109/JSTARS.2022.3155559.

[18] M. Helleis, M. Wieland, C. Böhnke, S. Martinis, and S. Plank, ‘Water mapping for flood detection using SAR data and convolutional neural networks’, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2023–2036, 2022.

[19] X. Zhu *et al.*, ‘So2Sat LCZ42: A Benchmark Dataset for Global Local Climate Zones Classification’, *ArXiv*, 2019, doi: 10.1109/mgrs.2020.2964708.

[20] H. Alemohammad and K. Booth, ‘LandCoverNet: A global benchmark land cover classification training dataset’, *arXiv:2012.03111 [cs]*, 2020.

[21] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, ‘Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model’, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021, doi: 10.1016/j.isprsjprs.2021.05.011.

[22] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, ‘SEN12MS - A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion’, in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, 2019, pp. 153–160. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

[23] C. Rambour, N. Audebert, E. Koeniguer, B. Le Saux, M. Crucianu, and M. Datcu, ‘Flood detection in time series of optical and SAR images’, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2-2020, pp. 1343–1346, Aug. 2020, doi: 10.5194/isprs-archives-XLIII-B2-2020-1343-2020.

[24] C. Rambour, ‘SEN12-FLOOD : a SAR and Multispectral Dataset for Flood Detection’. IEEE, Sep. 14, 2020. Accessed: Aug. 31, 2023. [Online]. Available: https://ieee-dataport.org/open-access/sen12-flood-sar-and-multispectral-dataset-flood-detection

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

[25] B. Bischke, P. Helber, C. Schulze, V. Srinivasan, A. Dengel, and D. Borth, 'The Multimedia Satellite Task at MediaEval', in *Proceedings of the MediaEval 2019 Workshop*, 2019.

[26] E. Fahrland, P. Jacob, H. Schrader, and H. Kahabka, 'Copernicus Digital Elevation Model Product Handbook', Airbus, AO/1-9422/18/I-LG, 2020. Accessed: Aug. 10, 2022. [Online]. Available: https://spacedata.copernicus.eu/documents/20126/0/GEO1988-CopernicusDEM-SPE-002_ProductHandbook_I1.00.pdf/082dd479-f908-bf42-51bf-4c0053129f7c?t=1586526993604

[27] M. Wieland, F. Fichtner, S. Martinis, C. Krullikowski, S. Plank, and M. Motagh, 'S1S2-Water: A global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 satellite images (v1.0.0)'. Zenodo, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8314175

[28] Y. Long *et al.*, 'On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4205–4230, 2021, doi: 10.1109/JSTARS.2021.3070368.

[29] 'Land Cover CCI Product User Guide Version 2', European Space Agency, Technical Report CCI-LC-PUGV2, 2017. [Online]. Available: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

[30] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, 'High-resolution mapping of global surface water and its long-term changes', *Nature*, vol. 540, no. 7633, Art. no. 7633, 2016, doi: 10.1038/nature20584.

[31] 'STAC: SpatioTemporal Asset Catalogs'. Accessed: Nov. 24, 2022. [Online]. Available: https://stacspec.org/en/

[32] 'Cloud Optimized GeoTIFF'. Accessed: Nov. 24, 2022. [Online]. Available: https://www.cogeo.org/

[33] A. Twele, W. Cao, S. Plank, and S. Martinis, 'Sentinel-1-based flood mapping: a fully automated processing chain', *International Journal of Remote Sensing*, vol. 37, no. 13, Art. no. 13, 2016, doi: 10.1080/01431161.2016.1192304.

[34] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', *arXiv:1512.03385 [cs]*, 2015.

[35] I. Loshchilov and F. Hutter, 'Decoupled Weight Decay Regularization'. arXiv, 2019. doi: 10.48550/arXiv.1711.05101.

[36] 'ukis-metrics'. German Aerospace Center, Earth Observation Center, Jul. 18, 2022. Accessed: Sep. 29, 2022. [Online]. Available: https://github.com/dlr-eoc/ukis-metrics

[37] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[38] M. Tan and Q. V. Le, 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', *arXiv:1905.11946 [cs, stat]*, 2019.

[39] X. X. Zhu *et al.*, 'Deep Learning Meets SAR: Concepts, models, pitfalls, and perspectives', *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 143–172, 2021, doi: 10.1109/MGRS.2020.3046356.

[40] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, 'Learning Aerial Image Segmentation From Online Maps', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, Art. no. 11, 2017, doi: 10.1109/TGRS.2017.2719738.

[41] C. Geiss, P. Aravena Pelizari, S. Bauer, A. Schmitt, and H. Taubenbock, 'Automatic Training Set Compilation With Multisource Geodata for DTM Generation From the TanDEM-X DSM', *IEEE Geosci. Remote Sensing Lett.*, vol. 17, no. 3, pp. 456–460, 2020, doi: 10.1109/LGRS.2019.2921600.

[42] B. Bauer-Marschallinger *et al.*, 'Satellite-Based Flood Mapping through Bayesian Inference from a Sentinel-1 SAR Datacube', *Remote Sensing*, vol. 14, no. 15, p. 3673, Jul. 2022, doi: 10.3390/rs14153673.

[43] M. Wieland, S. Martinis, R. Kiefl, and V. Gstaiger, 'Learning water body segmentation from very high-resolution satellite and aerial images', *Remote Sensing of Environment*, in review.

[44] A. Schneibel *et al.*, 'User-driven flood response & monitoring information – Key findings of the Data4Human project', in *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, 2022, pp. 46–53. doi: 10.1109/GHTC55712.2022.9911021.

[45] V. Hertel, C. Chow, O. Wani, M. Wieland, and S. Martinis, 'Probabilistic SAR-based water segmentation with adapted Bayesian convolutional neural network', *Remote Sensing of Environment*, vol. 285, p. 113388, 2023, doi: 10.1016/j.rse.2022.113388.

Marc Wieland received the Diploma degree in Geography from the Ruprecht-Karls Universität Heidelberg, Germany, in 2009 and the PhD degree from the Technical University of Berlin in 2013. In 2010 he joined the German Research Centre for Geoscience in Potsdam. In 2015 he moved to Chiba University, Japan to work on statistical pattern recognition in SAR time-series. In 2016 he joined the University of Oxford as postdoctoral researcher. He is currently based at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR) where his research activities focus on machine learning techniques for emergency response.

Florian Fichtner received the BSc degree in Geography at the University of Tübingen, Germany, in 2014 and his MSc degree in Geomatics at Delft University of Technology, the Netherlands, in 2016. He then worked for two years at the IT Consultancy Tensing in the Netherlands before he joined Telefónica NEXT in October 2018. In 2020, he joined the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR). His research activities focus on scaling thematic processing and deriving new products from its results for efficient disaster response.

Sandro Martinis received the Diploma degree in Geography, Physics, and Remote Sensing from the University of Munich, Germany, in 2006. He received the PhD degree from the University of Munich in 2010 working on automatic flood detection using high resolution X-band SAR satellite data at DLR. From 2006-2007, he was a research associate with the University of Munich working on the development of remote sensing-based methods for the monitoring of glacier motions and subglacial volcanic eruptions. Since 2013, he is head of the team "Natural Hazards" at DLR. Since 2016 he is leading the operational activities of Germany's contribution to the International Charter "Space and Major Disasters".

Sandro Groth received his BSc degree in Geography at the Ludwig-Maximilians-Universität Munich, Germany, in 2018 and his MSc degree in Earth Observation and Geospatial Data Analytics at the Julius-Maximilians-Universität Würzburg in 2022. He joined the department of Geo-Risks and Civil Security at the German Remote Sensing Data Center (DFD) of the German Aerospace

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3333969

Center (DLR) in 2022. His research focuses on the development and optimization of large-scale processing methods to investigate hydrological extremes.

Christian Krullikowski received the BSc degree in Geological Sciences at the Freie Universität Berlin, Germany, in 2013 and his MSc degree in Geological Sciences with a specialization in Hydrogeology at the Freie Universität Berlin, Germany, in 2016. In 2017, he joined the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR) in the department Geo-Risks and Civil Security. His research activities focus on thematic processing with rule-based and machine learning methods for efficient disaster response.

Simon Plank received the Diploma degree in Geology from the Technical University of Munich (TUM), Germany, in 2009, the MSc degree in Geographical Information Science and Systems from the Paris Lodron University of Salzburg, Austria, in 2011, and the PhD degree from TUM in 2012. Since 2009, he has been working on InSAR-based deformation monitoring of mass movements. From January 2013 to March 2013, he was a Post-Doctoral Researcher with TUM. In April 2013, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR) where he has been involved in research projects focusing on the development of algorithms and methods for crisis-related information extraction from optical, thermal, and SAR remote sensing imagery.

Mahdi Motagh received the BSc degree in Surveying Engineering and the MSc degree in Geodesy from the University of Tehran, Iran, in 1998 and 2002, respectively, and the PhD degree in Earth Sciences from the University of Potsdam, Germany, in 2007. As of 2007, he was a Postdoctoral Scientist with GFZ German Research Centre for Geosciences, Germany, where he became a Permanent Research Staff in 2011. He currently holds a Professorship in radar remote sensing with GFZ and Leibniz University Hannover (LUH), where he has been leading the research group on radar and optical remote sensing for geohazards. His research activities focus on the use of SAR/InSAR data to investigate processes related to natural and man-made hazards.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/