Sentinel-1-Based Water and Flood Mapping: Benchmarking Convolutional Neural Networks Against an Operational Rule-Based Processing Chain

Max Helleis, Marc Wieland, Christian Krullikowski, Sandro Martinis, and Simon Plank

***Abstract*—In this study, the effectiveness of several convolutional neural network architectures (AlbuNet-34/FCN/DeepLabV3+/U-Net/U-Net++) for water and flood mapping using Sentinel-1 amplitude data is compared to an operational rule-based processor (S-1FS). This comparison is made using a globally distributed dataset of Sentinel-1 scenes and the corresponding ground truth water masks derived from Sentinel-2 data to evaluate the performance of the classifiers on a global scale in various environmental conditions. The impact of using single versus dual-polarized input data on the segmentation capabilities of AlbuNet-34 is evaluated. The weighted cross entropy loss is combined with the Lovász loss and various data augmentation methods are investigated. Furthermore, the concept of atrous spatial pyramid pooling used in DeepLabV3+ and the multiscale feature fusion inherent in U-Net++ are assessed. Finally, the generalization capacity of AlbuNet-34 is tested in a realistic flood mapping scenario by using additional data from two flood events and the Sen1Floods11 dataset. The model trained using dual polarized data outperforms the S-1FS significantly and increases the intersection over union (IoU) score by 5%. Using a weighted combination of the cross entropy and the Lovász loss increases the IoU score by another 2%. Geometric data augmentation degrades the performance while radiometric data augmentation leads to better testing results. FCN/DeepLabV3+/U-Net/U-Net++ perform not significantly different to AlbuNet-34. Models trained on data showing no distinct inundation perform very well in mapping the water extent during two flood events, reaching IoU scores of 0.96 and 0.94, respectively, and perform comparatively well on the Sen1Floods11 dataset.**

***Index Terms*—Convolutional neural networks, data augmentation, semantic segmentation, Sen1Floods11, Sentinel-1, Sentinel-2, surface water monitoring.**

I. Introduction

THE demand for reliable and robust crisis information after catastrophic disasters has substantially grown in the past decades [1]. Earth observation satellites are increasingly used to obtain reliable large-scale crisis information, owed to their ability of being almost independent of the underlying terrain, the possibility of acquiring data over large areas in a short amount of time and their different sensor systems. To provide the required information in a timely manner, international initiatives such as the International Charter "Space and Major Disasters" have been founded, linking space agencies from all over the world and allowing a rapid disaster response by sharing the available satellite resources. Flood events make up one-third of all recorded natural disasters in the past century [2] and were related to approximately 52% of all activations of the International Charter "Space and Major Disasters" between the years 1999 and 2013 [1]. They are usually not localized but affect large regions simultaneously and the atmospheric conditions often prevent observation using optical or multispectral sensor systems. These preconditions make synthetic aperture radar (SAR) an ideal sensor system to be used in flood emergency situations. SAR is a side-looking imaging radar system that utilizes microwave radiation to create images of the surface of the Earth. It is an active instrument that requires no illumination from the sun and can penetrate cloud cover, thus it is often used for rapid mapping purposes during flood events [1].

To derive the flood water extent from SAR data with a high accuracy, sophisticated data analysis steps are required. Visual scene interpretation and flood extent mapping are possible, but have several disadvantages. The areas usually affected by flood events are very large, which renders visual interpretation a very time-consuming task. Furthermore, the results depend on the skills and the subjective perception of an image interpretation expert, which poses a problem for reproducibility [3]. Hence, several semi-automatic and automatic approaches for flood extent mapping have been developed, often relying on thresholding as their core concept. Thresholding is a common technique used to classify every pixel of a SAR image into the classes water or non-water. A pixel intensity threshold value that separates the two classes is chosen. Every pixel with an intensity value below that threshold is classified as water, all other pixels are classified as nonwater. The quality of the classification is strongly dependent on the contrast between water and nonwater pixels, which poses a problem for certain land cover types or water conditions (i.e., rough water surfaces, urban areas, sand patches) and can require additional thresholds, for example, to account for flooded vegetation. Several approaches for the determination of these

Manuscript received August 11, 2021; revised January 6, 2022 and January 24, 2022; accepted February 4, 2022. Date of publication February 16, 2022; date of current version March 7, 2022. This work was supported in part by the German Federal Ministry of Education and Research (BMBF) as part of the project "Künstliche Intelligenz zur Analyse von Erdbeobachtungs- und Internetdaten zur Entscheidungsunterstützung im Katastrophenfall" (AIFER) under Grant 13N15525, and in part by the Helmholtz Artificial Intelligence Cooperation Unit as part of 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: helleis.max@gmail.com; marc.wieland@dlr.de; christian.krullikowski@dlr.de; sandro.martinis@dlr.de; simon.plank@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2022.3152127

thresholds exist. While manual methods use the histogram of the scene or manual trial-and-error approaches to derive a working threshold [4]–[9], automatic approaches [10]–[13] work independently of a human operator and are therefore preferable for emergency situations [1].

The fully automated Sentinel-1 flood processing chain (S-1FS) presented by Twele *et al.* [14] is an adapted version of the TerraSAR-X flood processor presented by Martinis *et al.* [15]. It is able to generate flood extent maps based on systematically acquired Sentinel-1 Level-1 ground range detected (GRD) VV-polarized data by combining automatic intensity thresholding, fuzzy-logic based refinement using topographic information from the Shuttle Radar Topography Mission (SRTM) and an exclusion mask based on the height above nearest drainage (HAND) [16], [17] layer. Field studies have shown that the processor is able to achieve Kappa scores of 0.879 and higher, although special conditions like strong winds (and therefore rough water surfaces), sand, or flooded vegetation can lead to classification errors [14]. The S-1FS is in operational use at the Center for Satellite Based Crisis Information (ZKI) at the German Aerospace Center and is part of the ensemble algorithm of the new systematic flood monitoring product of the Copernicus Emergency Management Services [18].

While various rule-based processing algorithms for water mapping and flood detection using SAR systems have been proposed and deployed in the past decades using a variety of space-borne sensor systems, convolutional neural networks (CNNs) have seen a rapid development in recent years. CNNs are a branch of artificial neural networks that have been shown to be capable of outperforming traditional image processing techniques in various fields ranging from image classification through object detection to image segmentation and have already been successfully applied to various problems in the Earth observation domain, including flood and water mapping using either multispectral or SAR data (e.g., [19]–[24]).

The idea of using artificial neural networks for water mapping and flood detection in SAR imagery had already been proposed and tested by Skakun [25] in 2010. Using Self-Organizing Kohonen Maps and data from ERS-2, Radarsat-1 and Envisat acquired over India, China and Eastern Europe he showed the feasibility of using artificial neural networks for flood detection. Gong *et al.* [26] used a restricted Boltzmann machine in 2016 on multitemporal Radarsat data over Canada to detect change in farmland, water bodies, and the coastline, which was adopted by Bayik *et al.* [27]. In 2017, Liu *et al.* [28] were the first to apply CNNs to directly detect flooded areas in multitemporal Radarsat and ERS-1 data over Canada and Switzerland. Also in 2017, Xu *et al.* [29] used the highly successful AlexNet with parameters pretrained on ImageNet to detect sea ice and open water in Radarsat images acquired over the Golf of Lawrence in Canada. Kang *et al.* [30] emphasized in 2018 the importance of using different scenes for the creation of the training, test and validation datasets to prevent the occurrence of spatial autocorrelation in the validation and test dataset. By using data of three flood events in China acquired by Gaofen-3 and a U-Net-based model architecture, they evaluated the impact of choosing different activation functions as nonlinearities, concluding that the best results are achieved by using the Rectified Linear Unit activation function. In 2019, Liu *et al.* [31] used multitemporal Sentinel-1 data of hurricane Harvey in Houston, Texas, and their own modification of U-Net to evaluate the impact of choosing different polarizations (VV, VH, or VV-VH) as input data and showed that using either dual polarization or VH polarization achieves the best results. Also, in 2019, Li *et al.* [20] used TerraSAR-X data of the same flood event in Houston to assess the roles of various SAR information types (i.e., intensity and coherence, uni- and multitemporal) if used as input features for a CNN. They concluded that multitemporal intensity data is the most important feature type, although adding multitemporal coherence information can further increase the classification accuracy. Furthermore, they proposed a temporal ensembling active self-learning CNN architecture to mitigate the effect of limited training samples. By using two models of identical architecture (student and teacher model), where the student model is trained on the training set and the teacher model's weights are obtained by taking the running mean of the student's weights, they showed that unlabeled samples can be classified and added to the training set during training, which yielded an increase of the Kappa score of about 7%. More recently, in 2020, Nemni *et al.* [22] tested a variety of CNN architectures for water mapping and flood detection using Sentinel-1 images and flood maps from the UNOSAT dataset [32]. Where former approaches were focused on few local regions and often just a single region, the UNOSAT dataset contains Sentinel-1 scenes in VV polarization from nine countries, all located either in Southeast Asia or along the eastern coast of Africa. The evaluation of different model architectures showed that a U-Net and a U-Net using a ResNet (RN) encoder pretrained on ImageNet performed quite well in detecting water areas and actual flood events. Also, in 2020, Bonafilia *et al.* [21] assessed different methodologies for the creation of reference water masks to be used in supervised learning. The used dataset contained Sentinel-1 VH data from 11 flood events as well as from globally distributed open water areas, sampled from the surface water dataset of the European Commission Joint Research Center [33]. Their study states that a classical histogram thresholding approach yields the best results for the segmentation of permanent water bodies, although it must be stated that model optimization was not in the scope of their study. Surprisingly, they observe that models trained using reference water masks created by simple histogram thresholding approaches tend to achieve better results compared to models trained on manually created (hand labeled) reference water masks for permanent water surfaces. Furthermore, they report that models trained solely on scenes showing permanent water bodies perform poorly if evaluated on scenes showing actual flood events. Pai *et al.* [34] applied a U-Net on the task of land and water mapping using Sentinel-1 imagery, reporting an excellence performance and showed that artificially extending the dataset using General Adversarial Networks can further improve the results. In 2021, Muñoz *et al.* [35] combined multispectral data with SAR imagery and terrain information using a convolutional data fusion network and reported good performance for the case of using only dual polarized VV-VH SAR data. Katiyar *et al.* [23] used the Sen1Floods11 dataset published

by Bonafilia *et al.* [21] to further test different U-Net-based CNN architectures, reporting that their CNNs do outperform a simple thresholding classifier regarding permanent water bodies. Furthermore, they confirm that using VV-VH data leads to the best results for general surface water mapping. Last but not least, Bai *et al.* [24] also used the dataset produced by Bonafilia *et al.* [21] to test BASNet [36], an enhanced U-Net with a module to refine the residuals of the final feature maps. Furthermore, they assess various loss functions and investigate the impact of fusing Sentinel-1 and Sentinel-2 data, applying data augmentation and using various model architectures.

While these studies show clearly that water detection using SAR data and CNNs is generally feasible, it is important to quantify how well these methods perform with respect to a current state-of-the-art operational rule-based (flood) water processor (S-1FS) on a global scale, to determine if CNNs can increase the quality of automatically generated flood maps in rapid mappings scenarios compared to the S-1FS. A higher quality of automatically produced maps directly translates to a higher timeliness of the final rapid mapping product and faster access to critical information for disaster relief services.

To answer this overarching question, this study focusses on the following objectives.

1) The assessment of the performance of several state-of-the-art CNN architectures for (flood) water segmentation in Sentinel-1 data with respect to the S-1FS.
2) An evaluation of the use of a linear combination of the distribution-based weighted cross entropy loss function and the region-based Lovász loss function.
3) An impact assessment of various geometric and radiometric data augmentation techniques for Sentinel-1 data in the context of (flood) water mapping and, finally, a transferability assessment.
4) Can CNNs trained for water mapping using a large-scale, globally distributed dataset be used for mapping inundated areas associated with flood events?

II. Data

Flood related disasters are not restrained to any particular region of the world. For effective disaster relief functionality, i.e., rapid mapping activities after a flood event, any flood detection algorithm should therefore be flexible enough to operate on data from any given global region. The main differentiator between water and land areas in SAR data is the contrast in signal return intensity caused by the two types of land cover. While the interaction of the SAR signal with calm water surfaces is mainly characterized by specular reflection (i.e., weak signal return), the rougher land surfaces cause mostly diffuse surface or volume scattering, which lead to a higher signal return intensity. However, depending on the region, land areas can inhibit surface properties that cause the radar return intensity to weaken (e.g., smooths surfaces like sand patches), whereas varying environmental conditions can increase the roughness of the water surface to the order of the signal wavelength, which increases the backscatter intensity. These effects can decrease the contrast and lead to over- and underestimation of the water surface [1].

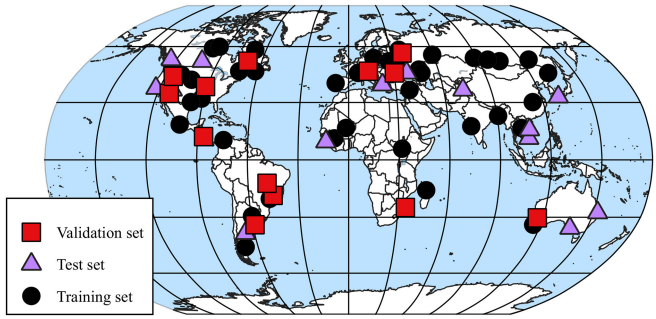

Fig. 1. Locations of all samples contained in dataset I (Map: Made with Natural Earth).

To train a robust classifier and evaluate its performance and generalization capacity for global deployment, it is therefore important to capture the distribution of natural environments by including as many different environmental conditions in the dataset as possible. For this study, three datasets are generated: Dataset I is used to train and evaluate all models and contains a wide range of Sentinel-1 scenes. Dataset II contains data from two observed flood events in Peru and China and is used to further assess the generalization capacities of the models, i.e., how well they can detect water on scenes showing a mix of land surfaces, open water surfaces and temporarily inundated areas. Dataset III is used to compare our method and the S-1FS to other published studies. For all three datasets, the Sentinel-1 data was processed identically.

A. Dataset I

We selected 67 globally distributed, dual polarized Sentinel-1 Level-1 GRD (interferometric wide swath) scenes using a stratified random sampling scheme based on the terrestrial ecoregions [37], land use, and land cover, following the procedure described by Wieland *et al.* [19]. This is done to ensure that the inherent topographic, land cover, and land use variation of the selected samples is maximized throughout the dataset, so that the models can ideally learn all spatial and radiometric contexts in which surface water bodies occur globally. The locations of the scenes are shown in Fig. 1. The Sentinel-1 scenes are geometrically corrected and radiometrically calibrated following the procedure described by Twele *et al.* [14] to ensure that the data can be ingested and processed by the S-1FS. The maximum allowed temporal gap between the Sentinel-1 and the Sentinel-2 acquisition for a given location is set to 30 days. If no nearly cloud free Sentinel-2 acquisition is found that falls within this limit a different location is chosen, based on the aforementioned criteria. The Sentinel-2 scenes are then used to create reference water masks by Normalized Difference Water Index (NDWI) thresholding using a threshold derived by Otsu's method [38]. In the next step extensive manual quality checks and corrections are conducted to fit the reference water masks to the Sentinel-1 data. This step is necessary to correct any misclassifications caused by the NDWI thresholding and by the temporal gap

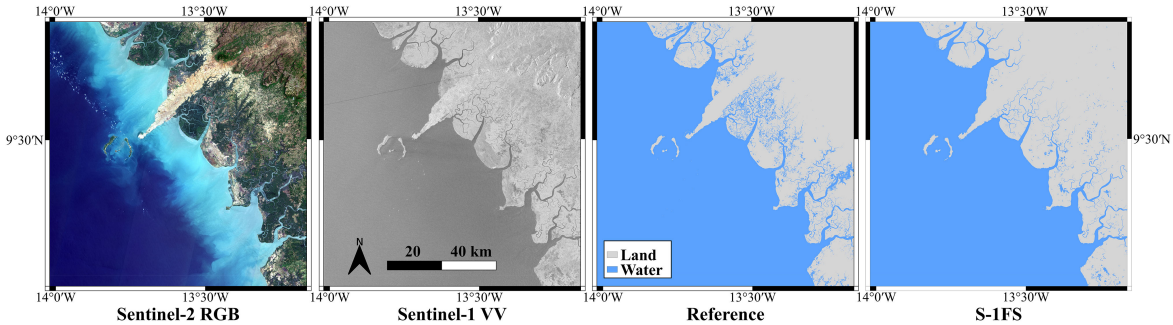

Fig. 2. Example of a scene from dataset I: Sentinel-2 data, Sentinel-1 data, the corresponding reference mask, and the output of the S-1FS.

between the acquisitions of Sentinel-1 and Sentinel-2 (e.g., due to tidal water surface extent variations). Valid pixel masks are created which indicate the locations of clouds in the Sentinel-2 data, i.e., they indicate pixels where the corresponding ground truth is not reliable. Furthermore, the valid masks mark no-data areas near the borders of Sentinel-1 scenes, which occur if the footprint of the Sentinel-2 scene and the Sentinel-1 scene are not perfectly aligned. All Sentinel-1 scenes are processed by the S-1FS. The resulting water segmentation masks are used for later performance comparison between the CNN models and the S-1FS. Since the S-1FS uses a binary exclusion mask based on the HAND index [39] to reduce water-lookalikes by taking the hydrologic-topographical setting into account [14], the same exclusion mask is added to the dataset for every scene but only used during testing. Fig. 2 shows an example of the data from dataset I (scene 37).

B. Dataset II—Case Studies Peru and China

The data contained in dataset I does not cover any distinct flood events. We therefore create a second dataset II to be able to assess the transferability of the representations the models learn from dataset I to data showing actual flood events. This dataset contains data and reference water masks from two flood events, which are adopted from and described in detail by Wieland and Martinis [19]. For this study, we added Sentinel-1 data to the dataset and used the existing reference water masks.

China, June 2016: A flood event caused by heavy monsoon rainfalls. The reference mask is based on a RapidEye image acquired on the June 23, 2016. Since the earliest available Sentinel-1 acquisition was on the July 5, 2016, minor manual corrections to the reference mask had to be made to correct for changes in the water surface extent due to the temporal gap between the acquisitions.

Peru, March 2017: A flood event caused by a strong local El Niño. The reference mask is based on a RapidEye image acquired on the April 1, 2017. Since a Sentinel-1 acquisition was available from the same day, no corrections had to be applied to the reference mask.

C. Dataset III—Sen1Floods11

Sen1Floods11 is a dataset containing Sentinel-1 and Sentinel-2 data from various flood events, which has been recently used in a number of studies (e.g., [23], [24]). Details on the dataset can be found in Bonafilia *et al.* [21]. To be able to ingest the data with our processing pipeline, we rebuild the Sentinel-1 data of the Sen1Floods11 dataset using the provided metainformation and the original Sentinel-1 GRD data as provided by the European Space Agency. This enables us to preprocess the data using our own preprocessing pipeline, as described by Twele *et al.* [14], and make it compatible to our trained models and the S-1FS. For this study, only the 90 hand-labeled tiles from the test set, as specified in the Sen1Floods11 dataset, are used.

D. Training/Validation/Test Split

Dataset I is split into three sets on the level of the scenes to prevent the occurrence of spatial autocorrelation between the sets [30]. About 67 Sentinel-1 scenes are distributed to the training, validation, and test sets using a random 60/20/20 split. In a next step, we ensure that every biome type [37] is represented at least once in every set and move scenes between sets if necessary. This is done to ensure that the data distributions of the training, validation, and test datasets are as similar as possible. Skipping this step could lead to a bias in the datasets toward specific environmental conditions or land cover types, since data from different biome types can inhibit very different radiometric properties, which would in turn negatively affect the generalization capabilities of the trained models. If less than three scenes are available for a certain biome type, they are all moved to the training set with the exception of the only scene of biome type "Montane grasslands and shrublands" (scene 61), which is left in the test set so evaluate the ability of the trained models to generalize to a biome they have never seen before. Only the biome "Tropical and subtropical coniferous forests" is missing completely in the dataset.

III. Method

A. Model Selection

Several studies have shown that the U-Net architecture [40] is able to deliver state-of-the-art results in water segmentation tasks using either Multispectral (e.g., [19], [41]) or SAR data (e.g., [22], [23], [30], [34]). After initial tests, a slightly modified version of the U-Net architecture is chosen as the base model, which has been shown to provide better segmentation results

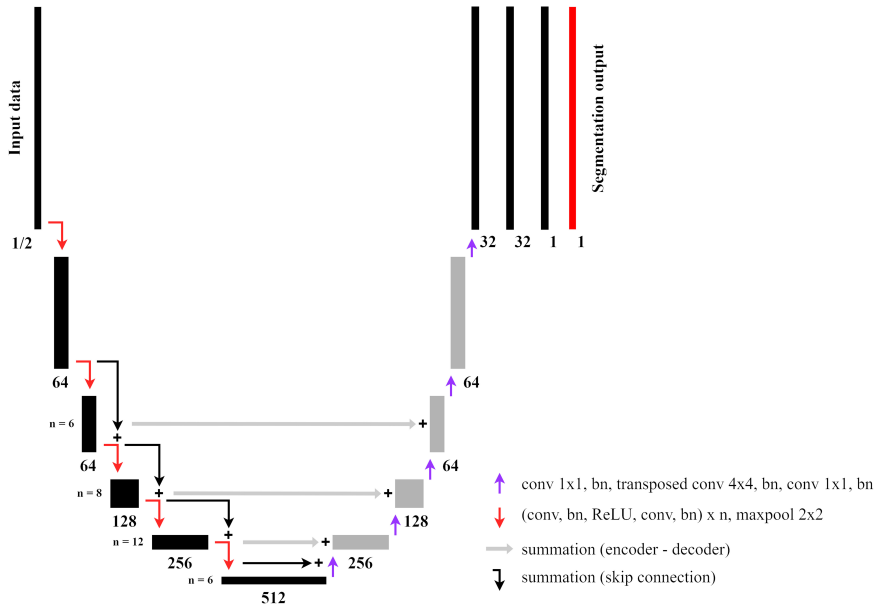

Fig. 3. Schematic architecture of AlbuNet-34 [42] which builds upon the common U-Net architecture. The encoder is replaced by a ResNet-34 and the information flow from the encoder to the decoder is accomplished by summation, where the original U-Net concatenates the feature maps, which reduces the number of trainable parameters.

than U-Net while using fewer parameters: AlbuNet-34 (AN-34) [42]. The main differences between AN-34 and the standard U-Net architecture are as follows. First, the encoder is replaced by a ResNet-34 encoder to incorporate the concept of residual learning, which leverages information flow by adding skip connections between the layers. These skip connections pass the activations from shallow layers to deeper layers and allow the model to easily fit the identity mapping, if necessary [43]. Second, where U-Net uses concatenation to let information flow from the encoder to the decoder, AN-34 uses a simple summation operation, i.e., the activations from the encoder are added to the activations of the decoder, which reduces the number of trainable parameters from 31×10^6 in the original U-Net architecture to 22.7×10^6 in AN-34. Fewer parameters increase the training speed and reduce the required time for inference. Especially the latter is considered as very important: reduced inference time translates to faster map delivery times in an emergency scenario. Fig. 3 shows the schematics of the AN-34 architecture. For a detailed description, refer to [42].

Similar to Bai *et al.* [24], we choose DeepLabV3+ [44] (DL) as a second well-known model architecture to evaluate if a better segmentation can be achieved by employing different architectural concepts. DL applies atrous spatial pyramid pooling by using diluted convolution kernels with different rates to extract information on multiple scales simultaneously. Furthermore, depth-wise separable convolution is applied to decrease computation complexity. To compare AN-34 to different modifications of U-Net, we additionally select the U-Net-ResNet-34 (UN-34) [45] and U-Net++ (UN++) architectures [46]. UN-34 is a modification of the standard U-Net that uses ResNet-34 as an encoder but keeps the concatenation operation compared to AN-34. UN++ is a more extensive modification of U-Net that aims to overcome the problem of finding the right network depth for a segmentation task by fusing multiple U-Nets of varying depths, all sharing the same encoder. Furthermore, its skip connections are designed in a way that allows fusion of feature maps from different scales in the encoder, with the goal to let the network decide how feature maps from various depths should be fused to obtain the best results. The hypothesis is that the architectural differences in the UN++ and DL architectures may allow the models to have a higher level of context-awareness and hence lead to better segmentation results in areas prone to misclassifications, e.g., arid regions or smooth surfaces like airport runways. Finally, we train a fully convolutional network [47] with a ResNet-50 (FCN-50) encoder to compare if any significant difference in performance can be seen between architectures with and a symmetric encoder-decoder structure and the FCN architecture. Refer to [40], [44], [46], and [47] for detailed descriptions of these model architectures. We use the DL, UN-34, and UN++ implementations of Yakubovskiy [45] and the FCN-50 implementation provided by PyTorch.

B. Loss Function

Loss functions are used to measure how well the predictions of a model match the reference data during training [48]. We investigate the effect of using members of two different families of loss functions to train our models: the distribution-based weighted cross entropy loss and the region-based Lovász loss [49]. Distribution-based loss functions describe the similarity of the probability distribution of the predictions of the model and the training data, whereas region-based loss functions maximize the overlap between the predicted segmentation mask and the reference data [48]. The majority of previous studies related to water mapping with convolutional neural networks and SAR data focused on the cross-entropy function. Bai *et al.* [50]

evaluated combinations of various loss functions, including the focal loss and showed that they can lead to better performance compared to the standard cross entropy loss on the segmentation of temporary water bodies, while also reporting decreased performance for permanent water bodies on the Sen1Floods11 test dataset. Since both types of water surfaces are equally important to provide accurate maps in rapid mapping scenarios, we decided to test the combination of the distribution-based weighted cross entropy loss and the Lovász loss. The weighted cross entropy is often used for imbalanced datasets, if one or more classes are underrepresented. It uses the inverse occurrence of a class as a weight to penalize the class specific loss and therefore prevents the model from focusing solely on the dominant class. In the training data of dataset I, there are 2.59 times more land pixels than water pixels, hence 2.59 is used as the cross-entropy weighting factor. The region-based Lovász loss function allows a direct optimization of the intersection-over-union score, which is a commonly used and very intuitive metric to measure the performance of models for semantic segmentation [51]. We follow Rakhlin *et al.* [52] and Bai *et al.* and implement a weighed combination of both loss functions:

$$L = (1 - \alpha_{\text{Loss}}) * L_{\text{CE}} + \alpha_{\text{Loss}} * L_{\text{Lov'asz}}.$$

The weighting coefficient α_{Loss} is used to weigh the two loss functions. L_{CE} denotes the weighted cross entropy loss and $L_{\text{Lov'asz}}$ denotes the Lovász loss.

C. Data Augmentation

To investigate the effect of different commonly used data augmentation techniques on the generalization capabilities of the models individually, we apply six types of data augmentation: Left-right flip (DA-Flip) flips an image tile around its vertical axis. Rotation (DA-Rot) rotates an image tile in 90° increments in a random direction. Intensity augmentation (DA-Int) changes the intensity values of all pixels in an image tile by adding or subtracting a fraction of the standard deviation of the intensity values of all water pixels in the training set. Zoom (DA-Zm) enlarges an image tile by a random factor between 0% and 10%. Following the ideas of Ding *et al.* [53] and Rusak *et al.* [54], we test two approaches for generating speckle: Speckle Gamma (DA-SpG) samples from a Gamma-distribution, Speckle Normal (DA-SpN) samples from a normal distribution. We resample the image tile by a factor of 2, add speckle noise and perform a simple averaging operation, where a 2×2 window with stride 2 is used to compute the average of four neighboring pixels, resulting in an image tile with the same size as the original tile. The resampling and averaging steps are done to imitate the multilooking process that is commonly applied to SAR images to reduce speckle noise.

D. Pre- and Postprocessing

We split the scenes contained in the test and validation set into tiles with a size of 256×256 pixel. To ensure that the input data is zero-centered and has unit variance, we standardize every tile during training and inference with the mean and standard deviation of the intensity values of the training set of dataset I, a common preprocessing step when training a neural network [55], [56]. During testing, we split the Sentinel-1 scenes into tiles using an overlap of 0.3 and feed the tiles to the trained models. The resulting segmentation maps are then recombined to the original scene size by using a tapered cosine function as described by Wieland und Martinis [19] to reduce the prediction errors close to the tile borders. This procedure is applied to the test scenes from datasets I, II, and III. Since only precut tiles are provided in the original dataset III, we extract the corresponding regions from our segmentation results using the bounding boxes of the original Sen1Floods11 reference masks and align the data.

TABLE I
DESCRIPTION OF THE USED PERFORMANCE METRICS [56], [57]

Metric	Equation	Description
Accuracy	$\frac{tp + tn}{tp + tn + fp + fn}$	Measures the overall accuracy of a classifier
Precision	$\frac{tp}{tp + fp}$	Fraction of correctly detected water pixels
Recall	$\frac{tp}{tp + fn}$	Fraction of detected water pixels compared to all water pixels
F1	$\frac{2 * recall * precision}{recall * precision}$	Harmonic mean of precision and recall
IoU	$\frac{tp}{tp + fp + fn}$	Intersection-over-Union. Measures the ratio of the intersection between the reference mask and the segmentation mask over their union [51]
Kappa	See [58]	Compares the classification result to one achieved by complete random classification

E. Performance Metrics

To evaluate the performance of the trained models, several common performance metrics based on the true positives tp, true negatives tn, false positives fp, and false negatives fn are used (see Table I).

We compute the performance metrics for every scene in the test set individually in a first step. Then, we compute the weighted averages for all metrics for all scenes, based on the number of valid pixels per scene as defined by the valid pixel mask. By doing so, we can evaluate the performance of the models on the level of the individual scenes as well as the average performance on the complete test set. To make the results of the trained models comparable to the results of the rule-based S-1FS, we must apply the HAND exclusion mask to the model output before computing the metrics, i.e., pixels marked as non-flood-prone by the HAND exclusion mask are set to value 0 (class land) in the output segmentation masks of the models. However, the HAND exclusion mask is not used for the evaluation of the CNN model for the Sen1Floods11 dataset, to allow a more direct comparison to the literature. Our implementation of the used metrics is published as an open-source Python package [54].

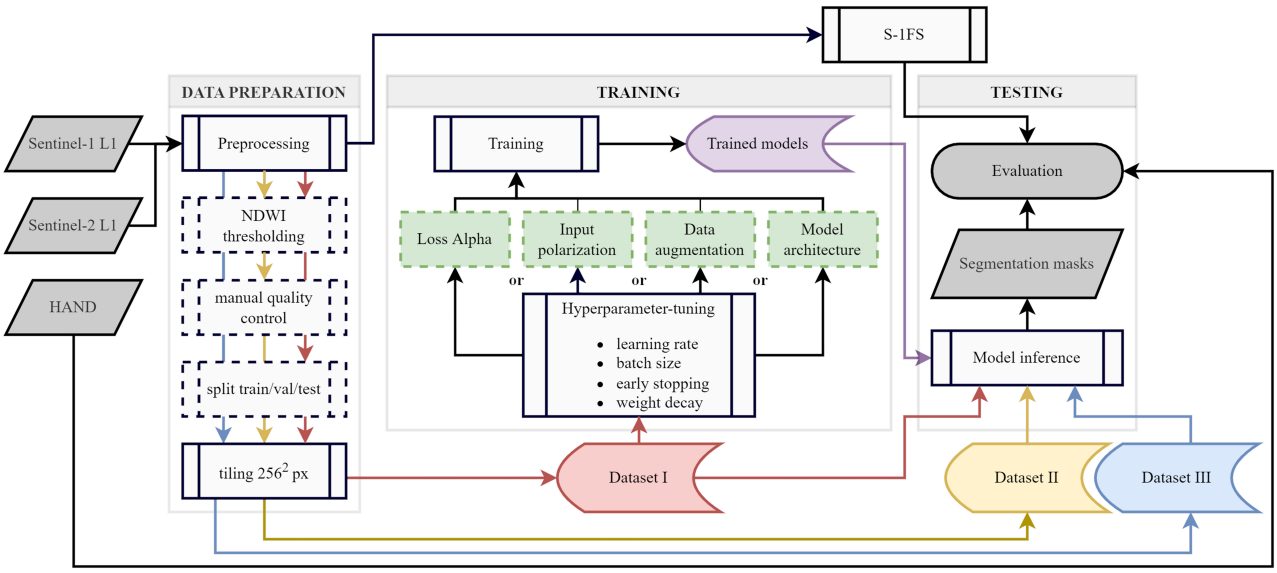

Fig. 4. General processing workflow used for this study. Dataset I is used for training and testing of the models. Datasets II and III are solely used for testing. Dataset I contains Sentinel-1 and Sentinel-2 data. Dataset II contains only Sentinel-1 data and the reference water mask derived from RapidEye data by Wieland and Martinis [19]. Dataset III contains only the Sentinel-1 data that was used to recreate the Sen1Flood11 dataset by Bonafilia *et al.* [21] to work with our preprocessing pipeline, as well as the provided reference water masks.

TABLE II
HYPERPARAMETERS USED FOR TRAINING

Hyperparameter	Value
Initial learning rate	1.0e-4
Learning rate decay	0.5 / 10 epochs
Loss weight (cross entropy)	2.59 (water class)
Max. number of epochs	50 (100 for DL)
Batch size	16
Optimizer	Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [61]
Weight decay	0.001

TABLE III
EXPERIMENT SETUP

Experiment	Input features	α_{Loss}	DA	Model
Polarization	**VV, VH, VV-VH**	0.0	-	AN-34
Loss function	VV-VH	**0, 0.25, 0.5, 0.75, 1.0**	-	AN-34
Data Augmentation	VV-VH	0.5	**DA_{flip}, DA_{int}, DA_{rot}, DA_{SpG}, DA_{SpN}, DA_{zm}**	AN-34
Models	VV-VH	0.5	DA_{int}	**DL: ResNet-34, -50, -101; UN/ UN++: ResNet-34; FCN: ResNet-50**

Note: Bold text indicates which parameter is varied for a given experiment.

F. Setup and Hyperparameters

All models are trained using the hyperparameters specified in Table II, which have been obtained through initial experiments. The learning rate is halved if the Intersection-over-Union (IoU) score stagnates for three epochs. Early stopping is applied if the IoU score does not increase for ten epochs. Random seeds are set to 1 for Numpy, PyTorch, and Python. CUDA deterministic mode is activated to allow reproducible results. All model parameters are initialized using He initialization [60]. During training the best model state is determined by the highest validation IoU score and stored for evaluation. The general processing workflow used in this study is summarized in Fig. 4. All experiments are conducted on a machine running Ubuntu 18.04 with and Intel Xeon W-2133 CPU and a Nvidia Quadro P4000 GPU.

G. Experiment Setup

All experiments are listed in Table III in chronological order. For every experiment one training parameter is varied, while all other parameters remain constant.

To evaluate how the models trained on data showing no explicit flood events generalize to data obtained during a flood event, we further test the best performing model from dataset I and the S-1FS on datasets II and III**.**

IV. RESULTS

A. Results Dataset I

The results for all experiments are shown in Tables IV to IX (bold values indicate the highest scores for a given metric). The average testing scores in Table IV show that the model trained on VV polarized data persistently yield the lowest testing

TABLE IV
AVERAGE TESTING SCORES FOR THE MODELS TRAINED USING VARYING POLARIZATIONS AS INPUT FEATURES

Metric	Polarization (Pol)			S-1FS
	Pol-VV	Pol-VH	Pol-VVVH	
Acc.	93.7	**98.8**	98.6	97.7
Rec.	85.5	75.5	**86.0**	81.4
Prec.	71.0	**83.9**	79.3	78.9
F1	72.1	77.2	**79.5**	75.2
IoU	63.6	68.9	**71.1**	65.7
Kap.	70.7	76.2	**78.7**	73.8

TABLE V
AVERAGE TESTING SCORES FOR THE MODELS TRAINED USING DIFFERENTLY WEIGHTED COMBINATIONS OF THE WEIGHTED CROSS-ENTROPY LOSS AND THE LOVÁSZ LOSS FUNCTIONS

	Experiment Loss function					S-1FS
	Alpha					
Metric	0.00	0.25	0.50	0.75	1.00	
Acc.	98.6	98.7	**98.8**	98.7	98.6	97.7
Rec.	86.0	**86.4**	86.1	86.3	86.3	81.4
Prec.	79.3	81.6	**82.8**	82.4	79.8	78.9
F1	79.5	80.6	**81.3**	81.1	79.6	75.2
IoU	71.1	72.5	**73.0**	72.9	71.5	65.7
Kap.	78.7	79.8	**80.5**	80.4	78.9	73.8

TABLE VI
AVERAGE TESTING SCORES OF THE MODEL TRAINED USING DIFFERENT TYPES OF DATA AUGMENTATION

	Experiment Data Augmentation (DA)							S-1FS
Metric	None	Flip	Int	Rot	SpG	SpN	Zm	
Acc	98.8	98.6	98.9	98.6	98.9	**98.9**	98.7	97.7
Rec.	86.1	85.7	85.7	86.1	86.1	**86.4**	86.2	81.4
Prec.	82.8	82.0	**84.0**	81.6	83.1	82.9	82.3	78.9
F1	81.3	80.8	**81.9**	79.9	81.6	81.8	81.1	75.2
IoU	73.0	72.4	**73.7**	72.1	73.3	73.4	72.9	65.7
Kap.	80.5	80.0	**81.2**	79.3	80.8	81.2	80.2	73.7

TABLE VII
AVERAGE TESTING SCORES OF ALL TESTED MODEL ARCHITECTURES

Metric / Model	Experiment Models					
	Acc.	Rec.	Prec.	F1	IoU	Kap.
AN-34	**98.9**	85.7	**84.0**	81.9	73.7	81.2
FCN-50	98.8	85.3	83.6	81.5	73.1	80.7
UN-34	**98.9**	84.4	86.2	**82.3**	**74.0**	**81.6**
UN-34++	98.8	85.8	83.2	81.5	73.2	80.7
DL-34	98.8	**86.2**	83.4	81.7	73.3	81.0
DL-50	98.7	**86.2**	81.1	79.9	71.7	79.0
DL-101	98.6	84.9	82.0	79.3	71.7	78.7
S-1FS	97.7	81.4	78.9	75.2	65.7	73.8

TABLE VIII
PERFORMANCE METRICS FOR THE FLOOD EVENTS IN CHINA AND PERU IN DATASET II (AN-34 WITH INTENSITY AUGMENTATION AND THE S-1FS)

	Dataset II – Case studies			
	Peru		China	
Metrics	AN-34 (Int)	S-1FS	AN-34 (Int)	S-1FS
Acc.	**97.6**	97.3	97.2	**97.5**
Rec.	**98.1**	97.8	**96.8**	94.9
Prec.	**97.9**	97.7	96.5	**99.1**
F1	**98.0**	97.7	96.7	**97.0**
IoU	**96.1**	95.6	93.5	**94.1**
Kap.	**95.0**	94.3	94.2	**94.8**

TABLE IX
IOU AND MIOU SCORES FOR DATASET III FOR THE BEST PERFORMING MODEL FROM DATASET I AND THE S-1FS

Metric	Dataset III – Sen1Floods11			
	All water			
	AN-34 (Int)	S-1FS	FCNN* [21]	BASNet** [24]
IoU	49.7	**54.9**	n/a	53.9
mIoU	34.7	31.0	40.8	**42.9**

Note: *Sentinel-2 weak, **Augments + Focal Loss (as named in the publication). The results from Bonafilia *et al.* [21] and Bai *et al.* [24] are shown for reference.

scores. The S-1FS, which also ingests only VV polarized data, performs better than the corresponding model in every metric. However, the models trained using either VH polarized data or both polarizations reach significantly higher F1, IoU and Kappa scores than the S-1FS.

The reported scores for the models trained using differently weighted combinations of the weighted cross entropy loss and the Lovász loss function in Table V show that a weighting coefficient of α equal 0.5 yields the best performing model for all metrics except recall.

The scores for the models trained using different kinds of data augmentation are shown in Table VI. Intensity augmentation yields the model with the best precision, F1, IoU, and Kappa score, whereas Speckle simulation using a normal distribution yields the highest accuracy and recall.

The scores obtained by the model based on the DL architecture with differently sized ResNet encoders (FCN-50, UN-34, and UN++ with a ResNet-34 encoder) are shown in Table VII. For DL, using a ResNet-34 encoder leads to the highest scores in all metrics. UN-34 yields the highest F1, IoU, and Kappa scores.

Fig. 5 lists the IoU scores for the best trained models from each experiment for all scenes in the test set of dataset I. The S-1FS yields the lowest IoU scores for 8 out of 14 scenes.

B. Results Dataset II—Case Studies Peru and China

Table VIII lists the scores obtained by the best AN-34 model (DA-Int) for the two flood events contained in dataset II, as well as the results of the S-1FS. It can be clearly seen that there is no significant difference in performance between AN-34 and the S-1FS for these two events.

C. Results Dataset III—Sen1Floods11

Table IX lists the IoU score and the mean IoU (mIoU) scores obtained by the best performing model from dataset I and the S-1FS on the Sen1Floods11 test set. The scores are computed following the procedure described by Bai *et al.* [24].

V. DISCUSSION

A. Polarization

The segmentation results depend strongly on the choice of polarization we use for training and inference. Contrary to the S-1FS, which tends to produce higher recall, precision and kappa scores if VV polarized data is used for inference [14], the models trained using either VV-VH or only VH polarized input data

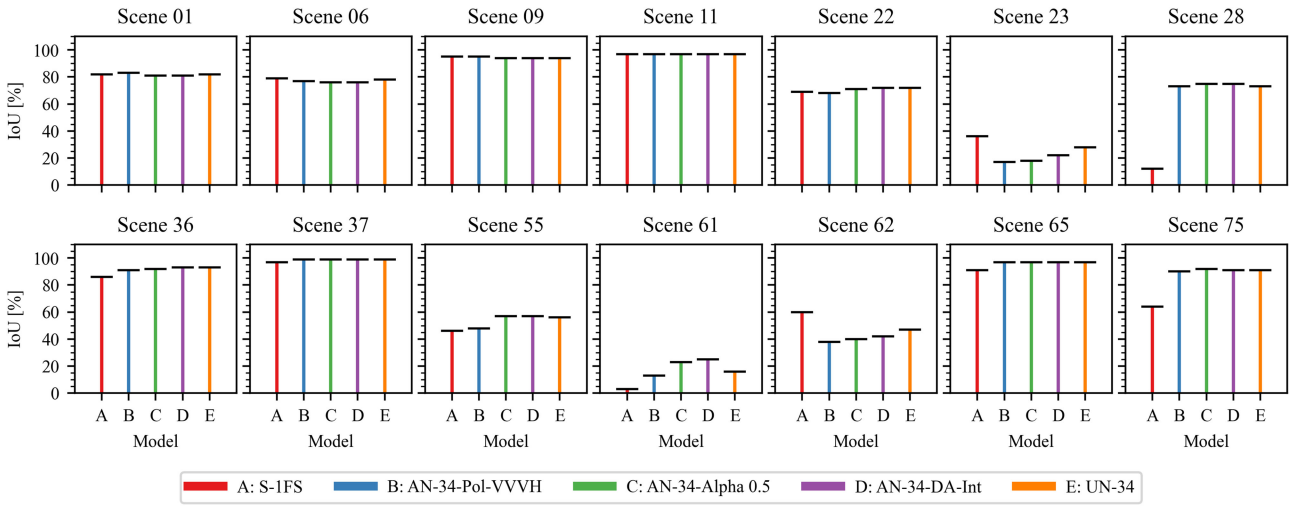

Fig. 5. IoU scores of the best trained model from every experiment and the S-1FS for all scenes from the test set of dataset I. The letter A denotes the S-1FS, B denotes the experiment on polarization, C the experiment on loss functions, D the experiment on data augmentation and E the experiment using different model architectures (scene numbers according to dataset definitions).

outperform the model trained on VV polarized data by a considerable margin in every metric, confirming the observations of Liu *et al.* [31] and Katiyar *et al.* [23]. Fig. 6 shows two extreme examples from the test set where the choice of input polarization has a particularly strong effect on the segmentation results. Scene 62 shows a small group of lakes in an otherwise extremely arid environment. Probably due to the low contrast in backscatter intensity in the VV band, the model trained using only VV polarized data drastically overestimates the water surfaces. The models using either single polarized VH or dual polarized VV-VH data produce a much more accurate segmentation mask, benefiting from the stronger backscatter intensity contrast in the VH band, even though in this particular scene there are still many misclassifications, especially if VV-VH data is used. Scene 65 is more representative of the overall results. Again, using only VV polarized input data leads to a significant number of false positives (similar to the S-1FS) and a high number of false negatives at the water-land boundary close to the coast and on the open water surface. The contrast between land and water in the VV band is quite low, probably due to strong surface winds, which are known to decrease the backscatter intensity contrast between water and land surfaces more strongly in VV than in VH polarized data, which is caused by the higher sensitivity of VV polarized electromagnetic waves to waves and ripples [1], [62]. Hence, the models using either VH or VV-VH polarized input data achieve IoU scores of 0.94 and 0.97, respectively, compared to the 0.84 yielded if only VV data is used and 0.91 reached by the S-1FS.

B. Loss Function

The proposal to use a weighted combination of distribution-based and region-based loss functions by Rakhlin *et al.* [52] and Bai *et al.* [24] also leads in our case to better results than using either of the two loss functions on their own and can be recommended for the combination of the weighted cross entropy loss and the Lovász loss. A weighting coefficient of 0.5 leads to a 2% higher IoU score compared to using only the weighted cross entropy loss and a 1.5% higher IoU score compared to using only the Lovász loss. This is consistent with the results of Bai *et al.*, who reported good performance using a combination of the distribution based focal loss [50], the Structural SIMilarity loss [63], and the region-based IoU loss [64]. Hence, the combination of distribution and region-based loss functions appears to generally be of benefit in the case of water and flood mapping using SAR data and should be employed when training CNNs for this task.

C. Data Augmentation

The testing scores of the models trained using different kinds of data augmentation vary only slightly. Compared to using no data augmentation, random flipping, rotation or zooming lead to slightly worse testing scores. With regard to rotation, this confirms Zhu *et al.* [65], who state that rotating an SAR image leads to unrealistic data, since the azimuth and range direction in SAR imagery are not arbitrary. Due to the close-to-constant orbital altitude of 685 km, introducing zooming might also lead to unrealistic imagery and can be omitted. Left-right flipping of the tiles can be seen, at least to a first order, as a change of the acquisition geometry from the ascending to the descending path or vice-versa. Hence, the observed deterioration of the performance of the models is surprising and cannot be entirely explained with the introduction of unrealistic image geometries. Radiometric data augmentation on the other hand leads to slightly increased testing scores if either intensity augmentation or speckle noise augmentation is used. For speckle noise augmentation, no significant difference between using a normal or a gamma distribution can be observed, hence using a simple normal distribution to augment the data might be sufficient to use, neglecting the actual gamma distribution of speckle noise. Interestingly, Bai *et al.* [24] applied left-right flipping, rotation and image cropping in

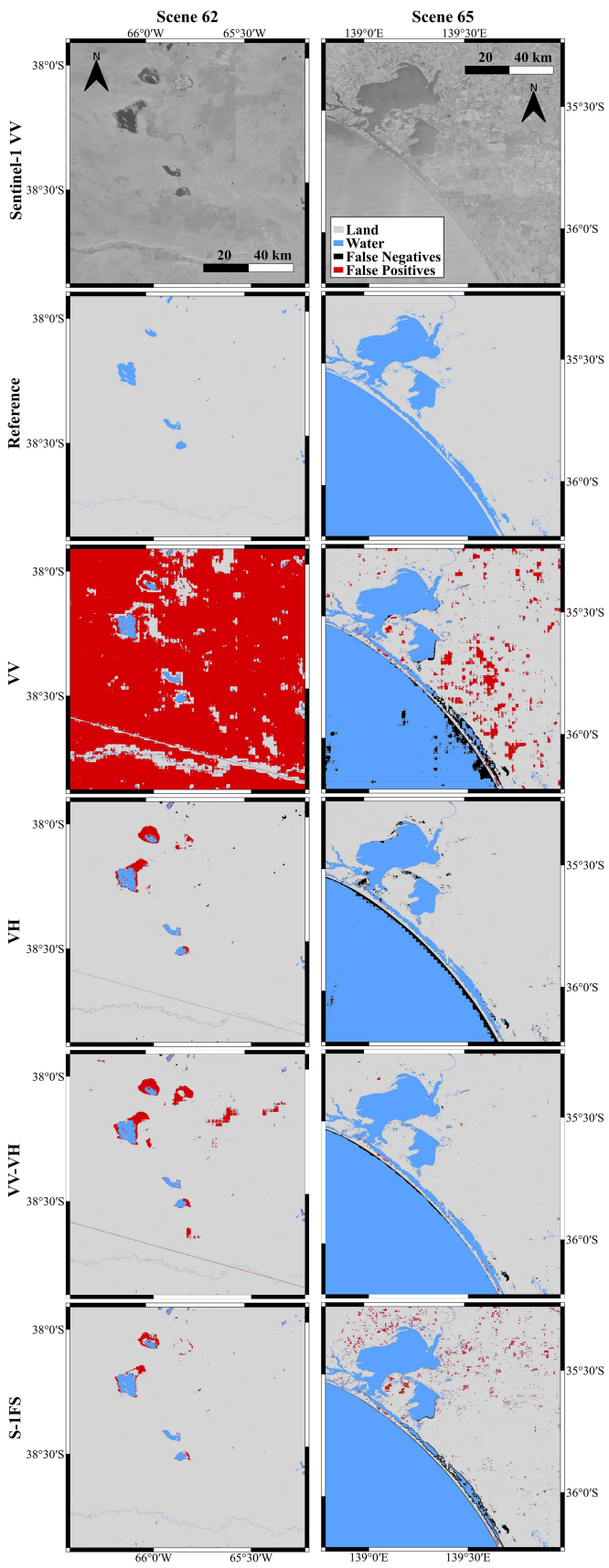

Fig. 6. Segmentation results for scenes 62 (Argentina) and 65 (Australia) for AN-34 models using either VV, VH or VV-VH polarized input data, and the S-1FS.

TABLE X
AMOUNT OF MODEL PARAMETERS AND IOU SCORES FOR ALL TRAINED MODEL ARCHITECTURES AND THE RELATIVE DEVIATION COMPARED TO AN-34

Model	Parameters		IoU	
	Amount [x 10^6]	Rel. dev. [%]	Score [%]	Rel. dev. [%]
AN-34	22.7	0.0	73.7	0.0
UN	24.4	7.5	74.0	0.41
UN++	26.1	15.0	73.2	-0.7
FCN-50	32.9	44.9	73.1	-0.8
DL-34	22.4	-1.3	73.3	-0.5
DL-50	26.7	17.6	71.7	-2.7
DL-101	45.7	101.3	71.7	-2.7

their study and reported a minor increase of the IoU score of 0.031 for detecting all water surfaces. However, this difference in the score can probably be attributed to differences between the datasets and to the unknown contribution of the image cropping augmentation.

D. Model Architecture

The DL, UN++ with a ResNet-34 encoder and the FCN-50 architectures lead to very comparable results to the AN-34 architecture. The results for DL with deeper encoders indicate that deeper models with more layers do not necessarily improve the predictive capabilities of the models. Furthermore, a visual comparison of the segmentation results reveals no systematic differences between the output of AN-34, UN-34, and DL/UN++, whereas the output of FCN-50 tends to produce noisy predictions near image borders in a few cases. The hypothesis that the concept of ASPP incorporated in DL or the multidepth encoders and multiscale feature map fusion incorporated in UN++ help with the correct classification of areas prone to misclassifications, such as airport runways or arid regions, can be discarded, at least within the scope of the data used in this study. This is well aligned with the results from Bai *et al.* [24], who also reported that DL yields results similar to other CNN architectures. The inherent multiscale information flow in AN-32 or UN-32 that is intrinsic in any encoder might be sufficient for this task, hence using a dedicated architecture focusing on multiscale information fusion does not improve the performance. Furthermore, problematic structures such as airport runways or sand patches are naturally heavily underrepresented in the training data. Hence, the models cannot significantly reduce the training loss by focusing on these areas and might tend to ignore them. This would indicate that the model architecture might not be the limiting parameter for obtaining better segmentation results at this stage. To further improve the performance of the models, adding additional training data with a focus on hard examples might be more beneficial than optimizing the model architecture.

Table X lists the amount of trainable model parameters for each architecture, the best achieved IoU score and the relative deviation of both compared to AN-34. Using concatenation instead of summation to fuse the feature maps from the encoder to the decoder, as done in UN-34, yields an increase in IoU score of 0.41% at the cost of requiring 7.5% more trainable parameters.

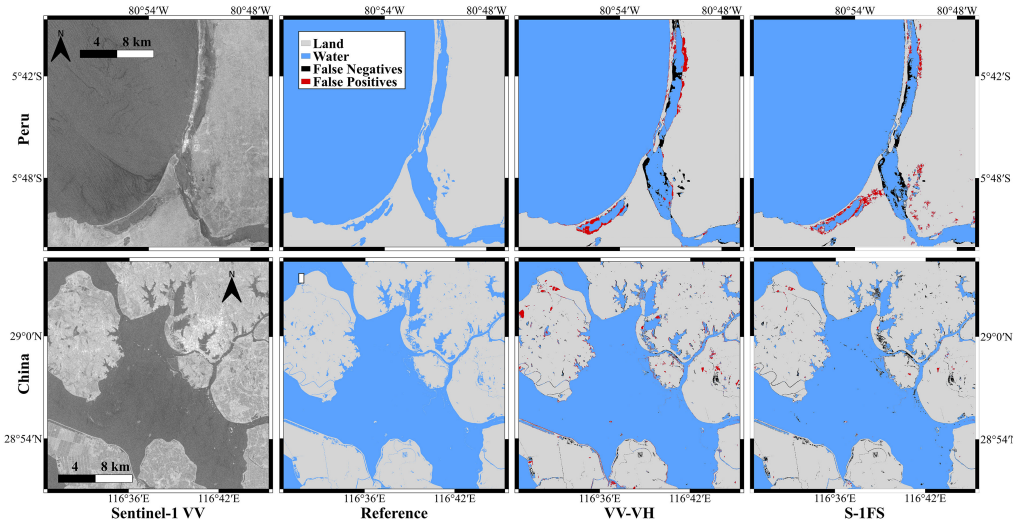

Fig. 7. Segmentation results for the case studies in Peru and China for the best performing AN-34 model (DA-Int) and the S-1FS.

E. Comparison of the S-1FS and CNNs

The scores achieved by the best-performing models on the test set from dataset I show that a CNN-based approach to water mapping using dual polarized VV-VH data can outperform the rule-based S-1FS in general by a considerable margin. All tested model architectures outperformed the S-1FS in every metric, at least if dual polarized VV-VH input data is used. There are certain conditions that present major difficulties to the CNNs and the S-1FS alike (scenes 23, and 61 and 62 in Fig. 5). Scenes 23 and 62 show very arid environments with large areas with low backscatter intensities, whereas scene 61 shows an arid, mountainous environment containing a large fraction of radar shadow. Both, arid environments and radar shadow, are well known to cause problems with water mapping in SAR imagery [1], [66]–[71]. Furthermore, scene 61 is the only scene from the biome "Montane grasslands and shrublands" in the dataset, hence it is an out-of-training-distribution sample for the model. Another aspect that must be considered when comparing both approaches, is the flexibility of the methods. While the S-1FS is a rather complex system that cannot easily be modified to incorporate additional information, it is straight-forward to add new information to the CNN. Further research should assess if this might help with arid or mountainous scenes, i.e., by adding slope, land cover or climatic information and further hard training examples.

F. Transferability to Flood Detection

While the performance metrics show no significant difference in the scores between AN-34 and the S-1FS for the flood events in Peru and China, a closer look at the segmentation masks in Fig. 7 reveals that there are qualitative differences to be noted.

In Peru, the S-1FS detects a number of false positives in the south and south-east regions of the scene, confusing arid land areas and sand patches with inundated areas. While AN-34 produces almost no false positives in these areas, the inundated area in the very south is slightly overestimated by AN-34 and by the S-1FS alike. The inundated strip close to the coast in the northern half of the scene is slightly underestimated by both methods, probably owed to the weak contrast between water and land in the VV band. With the results from dataset I in mind, this is not surprising, as both methods sometimes exhibit problems with the correct segmentation of water bodies in arid regions. In China, S-1FS tends to miss parts of smaller rivers, which are captured with a higher (but not perfect) accuracy by AN-34. Generally, the processor tends to produce more false negatives, whereas AN-34 tends to produce more false positives. However, both methods produce very detailed water extent maps that accurately depict every major water body. This shows that a model trained only on Sentinel-1 data showing no particular flood events can learn a sufficient representation of water surfaces that can be transferred to detect inundated areas during flood events. However, the problematic scenes from dataset I indicate that the robustness of the CNN models must be further improved to reliably work in all conditions.

The results obtained on the hand-labelled Sen1Floods11 test set show that AN-34 achieves results comparable to the methods used by Bonafilia *et al.* [21] and Bai *et al.* [24]. The lower IoU and the mIoU scores are not surprising, since, contrary to the other models, AN-34 was not trained on data from these flood events. While the S-1FS actually performs comparatively very well if the IoU score is computed over the complete test set, it performs worse with regard to the equally weighted mIoU score, which is more strongly influenced by the scores from individual tiles. However, a close inspection of the Sen1Floods11 dataset reveals a few deficiencies that have to be addressed.

1) The data of individual events is distributed into the training, validation, and test sets. As noted by Kang *et al.* [30], this can lead to the occurrence of spatial autocorrelation between the sets and should be prevented. While there is a small set of data from a single flood event in Bolivia that is held separately and not used for training, there are only 15 samples contained in that set, which is hardly a sufficient

sample-size to assess the generalization capacity of newly developed methods. Using data from multiple, completely unseen events would be preferable.

2) Five out of these 15 Bolivia samples are taken from the border area of the underlying Sentinel-1 data. This leads to a sharp edge between image data and no-data areas. Even though these no-data areas are marked as non-valid, the effects of these sharp edges are naturally propagated by the CNNs into the image regions that are evaluated. Hence, these samples inherently measure the ability of the models to cope with these no-data borders, even though they rarely occur if complete scenes are processed during an actual flood event.
3) Out of these 15 samples, only approximately 73% (∼85% for the complete test set) of the pixels are marked as valid, which further hampers the expressiveness.

VI. Conclusion

In this study, we compared the performance in water and flood mapping of a state-of-the-art rule-based Sentinel-1 flood processor (S-1FS) with five CNN architectures and assessed the impact of various hyperparameters on the performance of the trained CNN models.

1) We confirmed the observation by Liu *et al.* [31] and Katiyar *et al.* [23] that VH or VV-VH polarized data is the preferred input feature for the purpose of water mapping using CNNs and Sentinel-1 data.
2) We further showed that a linearly weighted combination of the weighted cross entropy loss function and the Lovász loss function yields better testing results than either of the loss functions on their own. Considering other studies which reported similar results using either different loss functions [24] or data from a different domain [52], our results strengthen the assumption that combining distribution-based and region-based loss functions is beneficial for many segmentation tasks.
3) Our results further indicate that geometric data augmentation methods should be treated with care when working with SAR data for water mapping, whereas radiometric data augmentation in the form of intensity augmentation or speckle noise simulation leads to better testing results.
4) All examined CNN architectures outperform the rule-based S-1FS in the task of water mapping, however, problems can arise in arid or mountainous environments using either method. Since all tested CNN model architectures perform very comparable to each other on a high level, we suggest that more potential for further enhancement of the segmentation performance lies with optimizing the training data itself than with further optimization of the model architectures.
5) Using data from two flood events we showed that CNNs trained solely on data containing no distinct flood events can work very well on data that includes inundated areas.
6) Benchmarking our methods on the Sen1Floods11 dataset leads to results that are comparable to the literature, considering that our models have not trained on data from the events contained in the dataset. However, within the frame of our study, we conclude that the Sen1Floods11 dataset is not suited as a benchmark dataset for flood detection using Sentinel-1 data, due to several qualitative deficiencies.

CNNs appear to be a superior choice for flood and water mapping using SAR data and could replace rule-based systems in operational environments, e.g., for rapid mapping purposes during or after flood disasters or for the monitoring of water surfaces. Even though we assessed only Sentinel-1 data, our methods should be easily transferable to other SAR sensors. To tackle the problems with arid and mountainous environments, further research should assess if additional information in the form of slope or land cover information or the extension of the training set with more scenes from difficult environments can improve the results and increase the robustness of the models. The release of an improved version of dataset I is under preparation [72].

References

[1] S. Martinis, C. Kuenzer, and A. Twele, "Flood studies using synthetic aperture radar data," in *Remote Sensing Handbook Volume III - Remote Sensing of Water Resources, Disasters, and Urban Studies*, P. Thenkabail, Ed. New York, NY, USA: Taylor & Francis, 2015, pp. 145–173, doi: 10.1201/b19321-10.

[2] D. Guha-Sapir, "EM-DAT: The emergency events database," Université catholique de Louvain (UCL) - CRED, Chicago, IL, USA, 2009.

[3] J. Sanyal and X. Lu, "Application of remote sensing in flood management with special reference to monsoon Asia: A review," *Natural Hazards*, vol. 33, no. 3, pp. 283–301, 2004, doi: 10.1023/B:NHAZ.0000037035.65105.95.

[4] P. A. Townsend and S. J. Walsh, "Modeling floodplain inundation using an integrated GIS with radar and optical remote sensing," *Geomorphology*, vol. 21, no. 3, pp. 295–312, 1998. [Online]. Available: https://doi.org/10.1016/S0169-555X(97)00069-X

[5] P. A. Brivio, R. Colombo, M. Maggi, and R. Tomasoni, "Integration of remote sensing data and GIS for accurate mapping of flooded areas," *Int. J. Remote Sens.*, vol. 23, no. 3, pp. 429–441, 2002, doi: 10.1080/01431160010014729.

[6] J.-B. Henry, P. Chastanet, K. Fellah, and Y.-L. Desnos, "Envisat multi-polarized ASAR data for flood mapping," *Int. J. Remote Sens.*, vol. 27, no. 10, pp. 1921–1929, 2006, doi: 10.1080/01431160500486724.

[7] P. Matgen, G. Schumann, J.-B. Henry, L. Hoffmann, and L. Pfister, "Integration of SAR-derived river inundation areas, high-precision topographic data and a river flow model toward near real-time flood management," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 9, no. 8, pp. 247–263, 2007, doi: 10.1016/j.jag.2006.03.003.

[8] M. Lang, P. Townsend, and E. Kasischke, "Influence of incidence angle on detecting flooded forests using C-HH synthetic aperture radar data," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 3898–3907, 2008.

[9] V. Gstaiger, J. Huth, S. Gebhardt, T. Wehrmann, and C. Kuenzer, "Multi-sensoral and automated derivation of inundated areas using TerraSAR-X and ENVISAT ASAR data," *Int. J. Remote Sens.*, vol. 33, no. 11, pp. 7291–7304, 2012, doi: 10.1080/01431161.2012.700421.

[10] S. Martinis, A. Twele, and S. Voigt, "Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-X data," *Natural Hazards Earth Syst. Sci.*, vol. 9, no. 3, pp. 303–314, 2009, doi: 10.5194/nhess-9-303-2009.

[11] G. Schumann, G. Di Baldassarre, D. Alsdorf, and P. D. Bates, "Near real-time flood wave approximation on large rivers from space: Application to the River Po, Italy," *Water Resour. Res.*, vol. 46, no. 5, 2010, doi: 10.1029/2008WR007672.

[12] P. Matgen, R. Hostache, G. Schumann, L. Pfister, L. Hoffmann, and H. H. G. Savenije, "Towards an automated SAR-based flood monitoring system: Lessons learned from two case studies," *Phys. Chem. Earth, Parts A/B/C*, vol. 36, no. 7, pp. 241–252, 2011. [Online]. Available: https://doi.org/10.1016/j.pce.2010.12.009

[13] L. Pulvirenti *et al.*, "Detection of floods and heavy rain using Cosmo-skymed data: The event in northwestern Italy of November 2011," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 3026–3029, doi: 10.1109/IGARSS.2012.6350788.

[14] A. Twele, W. Cao, S. Plank, and S. Martinis, "Sentinel-1-based flood mapping: A fully automated processing chain," *Int. J. Remote Sens.*, vol. 37, no. 13, pp. 2990–3004, 2016, doi: 10.1080/01431161.2016.1192304.

[15] S. Martinis, J. Kersten, and A. Twele, "A fully automated TerraSAR-X based flood service," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, no. 6, pp. 203–212, 2015. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2014.07.014

[16] A. Nobre *et al.*, "Height above the nearest drainage - A hydrologically relevant new terrain model," *J. Hydrol.*, vol. 404, no. 5, pp. 13–29, 2011, doi: 10.1016/j.jhydrol.2011.03.051.

[17] C. Chow, A. Twele, and S. Martinis, "An assessment of the height above nearest drainage terrain descriptor for the thematic enhancement of automatic SAR-based flood monitoring services," *Proc. SPIE*, vol. 9998, Oct. 2016, Art. no. 999808, doi: 10.1117/12.2240766.

[18] P. Salamon *et al.*, "The new, systematic global flood monitoring product of the copernicus emergency management service," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 1053–1056.

[19] M. Wieland and S. Martinis, "A modular processing chain for automated flood monitoring from multi-spectral satellite data," *Remote Sens.*, vol. 11, no. 19, Oct. 2019, Art. no. 2330, doi: 10.3390/rs11192330.

[20] Y. Li, S. Martinis, and M. Wieland, "Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 178–191, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.014.

[21] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 835–845, doi: 10.1109/CVPRW50498.2020.00113.

[22] E. Nemni, J. Bullock, S. Belabbes, and L. Bromley, "Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery," *Remote Sens.*, vol. 12, no. 16, Aug. 2020, Art. no. 2532, doi: 10.3390/rs12162532.

[23] V. Katiyar, N. Tamkuan, and M. Nagai, "Near-real-time flood mapping using off-the-shelf models with SAR imagery and deep learning," *Remote Sens.*, vol. 13, no. 12, Jun. 2021, Art. no. 2334, doi: 10.3390/rs13122334.

[24] 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 Sens.*, vol. 13, no. 11, Jun. 2021, Art. no. 2220, doi: 10.3390/rs13112220.

[25] S. Skakun, "A neural network approach to flood mapping using satellite imagery," *Comput. Inf.*, vol. 29, no. 6, pp. 1013–1024, 2010.

[26] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[27] C. Bayik, S. Abdikan, G. Ozbulak, T. Alasag, S. Aydemir, and F. Balik Sanli, "Exploiting multi-temporal Sentinel-1 SAR data for flood extend mapping," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-3/W4, pp. 109–113, Mar. 2018, doi: 10.5194/isprs-archives-XLII-3-W4-109-2018.

[28] T. Liu, Y. Li, Y. Cao, and Q. Shen, "Change detection in multitemporal synthetic aperture radar images using dual-channel convolutional neural network," *J. Appl. Remote Sens.*, vol. 11, no. 4, pp. 1–13, 2017, doi: 10.1117/1.JRS.11.042615.

[29] Y. Xu and K. A. Scott, "Sea ice and open water classification of SAR imagery using CNN-based transfer learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 3262–3265, doi: 10.1109/IGARSS.2017.8127693.

[30] W. Kang, Y. Xiang, F. Wang, L. Wan, and H. You, "Flood detection in Gaofen-3 SAR images via fully convolutional networks," *Sensors*, vol. 18, no. 9, 2018, Art. no. 2915, doi: 10.3390/s18092915.

[31] B. Liu, X. Li, and G. Zheng, "Coastal inundation mapping from bitemporal and dual-polarization SAR imagery based on deep convolutional neural networks," *J. Geophys. Res., Oceans*, vol. 124, no. 12, pp. 9101–9113, 2019, doi: 10.1029/2019JC015577.

[32] UNOSAT, *UNOSAT Flood Dataset*, 2019. [Online]. Available: http://floods.unosat.org/geoportal/catalog/main/home.page

[33] J.-F. Pekel, A. Cottam, N. Gorelick, and S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, pp. 418–422, Dec. 2016, doi: 10.1038/nature20584.

[34] M. M. M. Pai, V. Mehrotra, U. Verma, and R. M. Pai, "Improved semantic segmentation of water bodies and land in SAR images using generative adversarial networks," *Int. J. Semantic Comput.*, vol. 14, no. 1, pp. 55–69, Mar. 2020, doi: 10.1142/S1793351X20400036.

[35] D. F. Muñoz, P. Muñoz, H. Moftakhari, and H. Moradkhani, "From local to regional compound flood mapping with deep learning and data fusion techniques," *Sci. Total Environ.*, vol. 782, Aug. 2021, Art. no. 146927, doi: 10.1016/j.scitotenv.2021.146927.

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7471–7481, doi: 10.1109/CVPR.2019.00766.

[37] D. Olson *et al.*, "Terrestrial ecoregions of the world: A new map of life on earth," *BioScience*, vol. 51, no. 11, pp. 933–938, 2001, doi: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

[38] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[39] C. D. Rennó *et al.*, "HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia," *Remote Sens. Environ.*, vol. 112, no. 9, pp. 3469–3481, Sep. 2008, doi: 10.1016/j.rse.2008.03.018.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, vol. 9351, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

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

[42] A. Shvets, V. Iglovikov, A. Rakhlin, and A. Kalinin, "Angiodysplasia detection and localization using deep convolutional neural networks," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl.*, 2018, no. 4, pp. 612–617.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[44] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis.*, Cham, Switzerland, 2018, pp.pp. 833–851.

[45] P. Yakubovskiy, "Segmentation models pytorch," *Github Repository*, 2020. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch

[46] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[47] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[48] J. Ma *et al.*, "Loss odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102035, doi: 10.1016/j.media.2021.102035.

[49] M. Berman, A. Rannen, and M. Blaschko, "The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421, doi: 10.1109/CVPR.2018.00464.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[51] M. Berman, A. Rannen, and M. Blaschko, "The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421, doi: 10.1109/CVPR.2018.00464.

[52] A. Rakhlin, A. Davydow, and S. Nikolenko, "Land cover classification from satellite imagery with U-net and Lovász-Softmax loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 257–2574, doi: 10.1109/CVPRW.2018.00048.

[53] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[54] E. Rusak *et al.*, "A simple way to make neural networks robust against diverse image corruptions," in *European Conference on Computer Vision*, vol. 12348, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 53–69, doi: 10.1007/978-3-030-58580-8_4.

[55] E. Kreyszig, H. Kreyszig, and E. J. Norminton, *Advanced Engineering Mathematics*, 10th ed., Hoboken, NJ, USA: Wiley, 2011.

[56] I. Goodfellow and Y. Bengio, *Deep Learning*, 1st ed., Cambridge, MA, USA: MIT Press, 2016.

[57] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: https://doi.org/10.1016/j.ipm.2009.03.002
[58] W. Tang, J. Hu, H. Zhang, P. Wu, and H. He, "Kappa coefficient: A popular measure of rater agreement," *Shanghai Arch Psychiatry*, vol. 27, no. 1, pp. 62–67, Feb. 2015, doi: 10.11919/j.issn.1002-0829.215010.
[59] M. Helleis, F. Fichtner, M. Mühlbauer, and C. Krullikowski, *ukis-metrics*, 2021. [Online]. Available: https://pypi.org/project/ukis-metrics/
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2015, pp. 1026–1034.
[61] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, Dec. 2015, pp. 13–22.
[62] G. Schumann *et al.*, "High-resolution 3-D flood information from radar imagery for flood hazard management," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1715–1725, Jun. 2007.
[63] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402, doi: 10.1109/ACSSC.2003.1292216.
[64] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3458–3466, doi: 10.1109/ICCV.2017.372.
[65] X. Zhu *et al.*, "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 143–172, Dec. 2021.
[66] A. Wendleder, B. Wessel, A. Roth, M. Breunig, K. Martin, and S. Wagenbrenner, "TanDEM-X water indication mask: Generation and first evaluation results," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 6, no. 1, pp. 171–179, Feb. 2013.
[67] A. Bertram, A. Wendleder, A. Schmitt, and M. Huber, "Long-term monitoring of water dynamics in the Sahel region using the multi-SAR-system," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLI-B8, pp. 313–320, Jun. 2016, doi: 10.5194/isprs-archives-XLI-B8-313-2016.
[68] D. O'Grady, M. Leblanc, and D. Gillieson, "Use of ENVISAT ASAR global monitoring mode to complement optical data in the mapping of rapid broad-scale flooding in Pakistan," *Hydrol. Earth Syst. Sci.*, vol. 15, no. 11, pp. 3475–3494, Nov. 2011, doi: 10.5194/hess-15-3475-2011.
[69] R. S. Westerhoff, M. P. H. Kleuskens, H. C. Winsemius, H. J. Huizinga, G. R. Brakenridge, and C. Bishop, "Automated global water mapping based on wide-swath orbital synthetic-aperture radar," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 2, pp. 651–663, Feb. 2013, doi: 10.5194/hess-17-651-2013.
[70] S. Martinis, "Improving flood mapping in arid areas using Sentinel-1 time series data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 193–196, doi: 10.1109/IGARSS.2017.8126927.
[71] S. Martinis, S. Plank, and K. Cwik, "The use of Sentinel-1 time-series data to improve flood monitoring in arid areas," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 583.
[72] M. Wieland, M. Helleis, C. Krullikowski, S. Martinis, and S. Plank, "Data_s1s2_water: A global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 data," 2022.

Max Helleis received the B.S. degree in mechanical engineering from HTW Berlin, Berlin, Germany, in 2017, and the M.S. degree in earth oriented space science and technology with a specialization in remote sensing from the Technical University of Munich, Munich, Germany, in 2020.

In 2021, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He is involved in the operational activities of the International Charter "Space and Major Disasters" and his research focuses on the use of machine learning techniques for emergency response.

Marc Wieland received the diploma degree in geography from the Ruprecht-Karls Universität Heidelberg, Heidelberg, Germany, in 2009, and the Ph.D. degree (Dr.rer.nat.) from the Technical University of Berlin, Berlin, Germany, in 2013, working on exposure estimation from multi-source imaging for rapid seismic vulnerability assessment.

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 a Postdoctoral Researcher. He is currently based at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, where his research activities focus on machine learning techniques for emergency response.

Christian Krullikowski received the B.S. degree in geological sciences and the M.S. degree in geological sciences with a specialization in hydrogeology from the Freie Universität Berlin, Berlin, Germany, in 2013 and 2016, respectively.

In 2017, he joined the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, 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.

Sandro Martinis received the diploma degree in geography, physics, and remote sensing from the University of Munich, Germany, in 2006, and the Ph.D. 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 has been the Head of the team "Natural Hazards" with DLR, Oberpfaffenhofen, Germany. Since 2016, he has been leading the operational activities of Germany's contribution to the International Charter "Space and Major Disasters."

Simon Plank received the diploma degree in geology from the Technical University of Munich (TUM), Munich, Germany, in 2009, the M.S. degree in geographical information science and systems from the Paris Lodron University, Salzburg, Austria, in 2011, and the Ph.D. 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 Postdoctoral Researcher with TUM. In April 2013, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Oberpfaffenhofen, Germany, where he has been involved in research projects focusing on the development of (semi-)automated algorithms and methods for crisis-related information extraction from optical, thermal, and SAR remote sensing imagery.