Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1

Derrick Bonafilia[1]
Cloud to Street[1]
www.cloudtostreet.info

Beth Tellman[1,2]
Earth Institute, Columbia University[2]
beth@cloudtostreet.info

Tyler Anderson[1]
tyler@cloudtostreet.info

Erica Issenberg[1]
erica@cloudtostreet.info

Abstract

Accurate flood mapping at global scale can support disaster relief and recovery efforts. Improving flood relief efforts with more accurate data is of great importance due to expected increases in the frequency and magnitude of flood events due to climate change. To assist efforts to operationalize deep learning algorithms for flood mapping at global scale, we introduce Sen1Floods11, a surface water data set including raw Sentinel-1 imagery and classified permanent water and flood water. This dataset consists of 4,831 512x512 chips covering 120,406 km^2 and spans all 14 biomes, 357 ecoregions, and 6 continents of the world across 11 flood events. We used Sen1Floods11 to train, validate, and test fully convolutional neural networks (FCNNs) to segment permanent and flood water. We compare results of classifying permanent, flood, and total surface water from training a FCNN model on four subsets of this data: i) 446 hand labeled chips of surface water from flood events; ii) 814 chips of publicly available permanent water data labels from Landsat (JRC surface water dataset); iii) 4,385 chips of surface water classified from Sentinel-2 images from flood events and iv) 4,385 chips of surface water classified from Sentinel-1 imagery from flood events. We compare these four models to a common remote sensing approach of thresholding radar backscatter to identify surface water. Results show the FCNN model trained on classifications of Sentinel-2 flood events performs best to identify flood and total surface water, while backscatter thresholding yielded the best result to identify permanent water classes only. Our results suggest deep learning models for flood detection of radar data can outperform threshold based remote sensing algorithms, and perform better with training labels that include flood water specifically, not just permanent surface water. We also find that FCNN models trained on plentiful automatically generated labels from optical remote sensing algorithms perform better than models trained on scarce hand labeled data. Future research to operationalize computer vision approaches to mapping flood and surface water could build new models from Sen1Floods11 and expand this dataset to include additional sensors and flood events. We provide Sen1Floods11, as well as our training and evaluation code at: https://github.com/cloudtostreet/Sen1Floods11.

1. Introduction

Floods cause more damage than any other disaster. Today floods account for almost half of all weather-related disasters over the last two decades, affecting 2.3 billion people [11]. This high cost of natural disasters pushes 26 million people into poverty every year, causing setbacks to development as government budgets are stretched and people without financial protection are forced to sell assets [22]. Sea level rise, a changing climate, urbanization, and demographic change all contribute to current and future predicted increases in floods [6, 22, 29]. Improved response to mitigate and manage flood risk can be greatly assisted by satellite observations, which can be used in the mitigation, response, and recovery portions of the disaster cycle [1]. Mapping past flood events via satellite in data poor areas has, for example, aided in refugee relocation in the Republic of Congo in flood prone areas [59]. Near real-time flood information from satellites could optimize emergency vehicle routing, saving millions of dollars [40]. Satellite flood observations could also be used to develop new types of affordable insurance programs to provide financial protection for vulnerable populations [17].

Methods to detect inundation have been developed for dozens of satellite sensors at different spatial resolutions, temporal frequencies, and optical versus radar signal capabilities. Typical approaches to map inundation with MODIS (Moderate Resolution Imaging Spectrometer, 250m spatial resolution) exploit highly absorptive capacities of water in

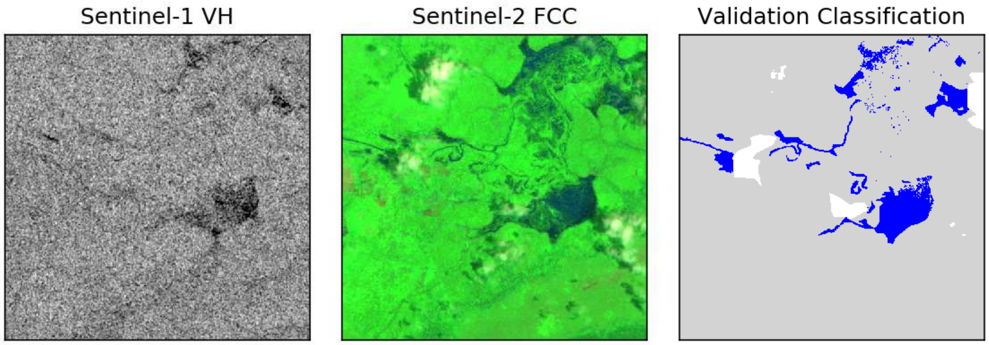


Figure 1. Example of hand labeled validation data.

short wave infrared spectrum (SWIR) relative to other objects (1628-1652nm, MODIS Band 6) [56, 26, 3], or use the near infrared (NIR) spectrum (841-875nm, MODIS Band 2) relative to the visible spectrum (621-670nm MODIS Band 1) [5] [19]. These approaches can provide water detection at a daily time step globally, although interpolation may be required when clouds obscure clear views [28, 27].

Inundation can be identified in medium resolution sensors such as Landsat and Sentinel-2 using similar approaches as MODIS algorithms. Landsat algorithms use band thresholding, normalized differencing, or more complex combinations of SWIR (1560- 1660nm, Landsat 8 Band 6) and NIR (630-690nm, Landsat 8 Band 4) with other bands [20, 58, 7, 14, 13]. Pekel et al (2016) [43] mapped SWIR, NIR, and NIR+Green+Blue (525-600 and 450-515 for bands 2 and 3, respectively in Landsat 8) onto HSV (Hue, Saturation, and Value) to estimate global surface water extent at a monthly time step since 1984. Algorithms to map surface water for Sentinel-2 similarly rely on water's absorption in SWIR (1539-1681 nm, Band 11) and NIR (768-796 nm, Band 8) [16]. Both Landsat and Sentinel-2 suffer from misclassifications of water and cloud shadows, which both have low reflectance values in SWIR and NIR. Moderate resolution optical sensors can resolve inundation at a higher resolution than MODIS (30m and 10m for Landsat and Sentinel-2 respectively) but suffer from less frequent revisit time (around 3 days when connecting Landsat and Sentinel-2) [27, 55].

SAR (Synthetic Aperture Radar) sensors can be important for flood detection due to their ability to detect through clouds. SAR sensors such as Sentinel-1 have been used to map inundation by identifying water, which has typically has lower backscatter values relative to other features (in VV, HH, VH, and HV bands). Water is identified by thresholding backscatter values on a single image [37, 39], the difference in backscatter between two images [21, 47], or variance of backscatter in a time series [10, 12]. Inundated vegetation and flooding in urban areas may present an increase in backscatter during flood events due to a "double bounce" effect [38, 8]. Urban flood damage has been approximated using the loss of interferometric signal coherence (phase information) of SAR sensors between two time periods [9]. Nearly all of these methods, however, rely on single threshold use to determine flood versus non-flood areas, which results in "speckle" in images due to strong overlap between water and non-water classes [49]. This "noise" often results in image cleaning techniques such as spatial filtering using neighborhood focal functions, refined Lee filtering [30] region growing, or object oriented classification, but often at the expense of removing small or isolated streams or water bodies distant from the main flood channel [49].

Advances in deep learning and computer vision are already shaping a new era in remote sensing of the Earth's surface [61]. Deep learning approaches to identify clouds and cloud shadows are far outperforming physically based algorithms to identify these features [60, 53]. Machine learning techniques such as random forests and support vector machines are now common in remote sensing land use and land cover classifications due to their superiority over previous methods [2]. Deep learning approaches, especially convolutional neural networks, have proven even more accurate in land cover classification [52], and their use has increased dramatically since 2015 [36].

However, most algorithms are applied to urban and vegetated land cover, with very few examples of surface water detection. One notable example of water detection using fully convolutional neural networks (FCNNs) on Landsat imagery trained on a permanent water dataset derived from MODIS outperformed both MNDWI thresholds and Multi-Layer Perceptron models at global scales [25]. CNNs have been used in at least three other examples, to our knowledge, to map floods specifically, all focused on examples for Hurricane Harvey in Houston. One study used CNNs to segment flooding areas in very high resolution UAV imagery [44], a second used a CNN to fuse Sentinel-1 and Sentinel-2 imagery to identify flooded buildings [45], and a third trained a CNN on Terra-Sar X (3m resolution commercial radar, HH polarization) using 35cm NOAA aerial photos for Hurricane Harvey [31]. While these models performed well on Hurricane Harvey examples, they were not trained on a large sample of floods, which can have unique signals in radar and optical data in urban areas, distinct soil types, or different kinds of flood disasters (e.g. flash floods, storm surge, or pluvial urban floods). Extant labeled training data for higher resolution public sensors like Sentinel-1 and 2 (e.g. SEN12MS) may include some permanent water data, but it represents a very small part of the dataset, and does not include any flood events [48].

Public datasets have spurred advances in several areas of computer vision. For image classification and object detection, the ImageNet training dataset has provided large scale data to researchers around the world, and its test dataset has documented the performance improvements from advances in computer vision [46]. For object detection and instance segmentation, Microsoft COCO has served a similar role [32], and for semantic segmentation, PASCAL VOC has done the same [18]. In part by removing the overhead of collecting data and providing standardized benchmarks, these datasets have been closely tied with many recent computer vision advances.

However, no such training data set exists focused on water and flood contexts. In order to examine the potential of CNN to recognize flood events spanning the variation of urban, rural, and geographic contexts, we produced and are releasing publicly Sen1Floods11 which includes both flooding and permanent water. The main contribution of this paper is a dataset for training and validation of deep learning algorithms for flood detection for Sentinel-1. The dataset provides global coverage of 4,831 chips of 512 x 512 10m pixels across 11 distinct flood events, covering 120,406 sq km. 4,370 chips were automatically labeled using simple remote sensing classification algorithms to be used as weakly supervised training data. Another 446 chips were hand labeled and are used as high quality training, testing, and validation data. Details are provided in section 2.

A second contribution of this paper is to use this dataset to explore four research questions in order to improve flood detection efforts and move towards operationalizing CNNs for global flood mapping. We don't purport to solve these questions, but rather to explore some of the avenues of research into which this dataset lends insight. The research questions are:

1. Do we need hand-labeled training data to train CNNs to detect flood water or can we use weakly supervised training data derived from remote sensing water detection algorithms?

2. Which imagery sources and algorithms provide the best labels for weakly supervised training?

3. What is the impact on model performance when a CNN is trained on permanent water data only as compared to training data that included flood events?

4. Do CNNs identify flood and/or permanent water in radar data more accurately than conventional remote sensing methods such as backscatter thresholding?

Section 3 describes the methods, and section 4 reports the results of models designed to explore the above research questions. Section 5 discusses how model results could shape the research agenda to improve upon initial models and training data presented here.

2. Datasets

This section describes the training and validation data created for this paper and the methodology used for creation and sampling.

2.1. Sampling Permanent Water Data

The JRC (European Commission Joint Research Centre) surface water dataset [43] was used as one source of training data. This dataset provides monthly observations of surface water at 30m resolution using the Landsat satellite. In order to provide an accurate label of water and non-water, water samples were derived from the transition layer, which identifies "permanent" water as pixels that were observed to have detected water presence at both the beginning (1984) and the end (2018) of the dataset. The "not water" label was used for any pixels that were never observed as water. For permanent water, the JRC dataset has commission and omission error rates of less than 1 percent. Its accuracy on all other classes of water (seasonal, ephemeral, or changing water classes) is much lower [43]. We therefore use the permanent water class as true positive examples and treat the other water classes as pixels to ignore (since we cannot be sure if these are true positive examples). We could not use seasonal or ephemeral water as flood training data; other studies have found that the JRC surface water product underpredicts flooding in part because water must be observed in a cloud free pixel, not highly sedimented, and often must be present for up to one month (or in multiple Landsat images) to be included in the JRC dataset [50].

2.2. Flood Event Data

A second source of data came from 11 flood events identified from a global database of flood event areas from the Dartmouth Flood Observatory [4]. Event selection required that the flood event had coverage from Sentinel-1, as well as coincident Sentinel-2 imagery on the same day or within 2 days of the Sentinel-1 image. Two events in Cambodia and

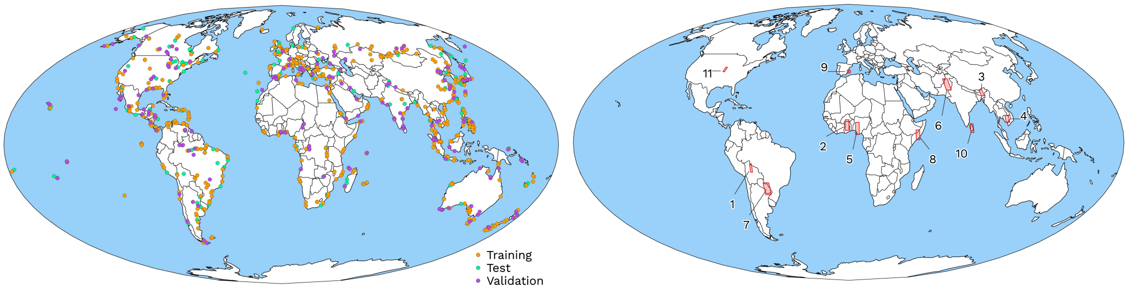
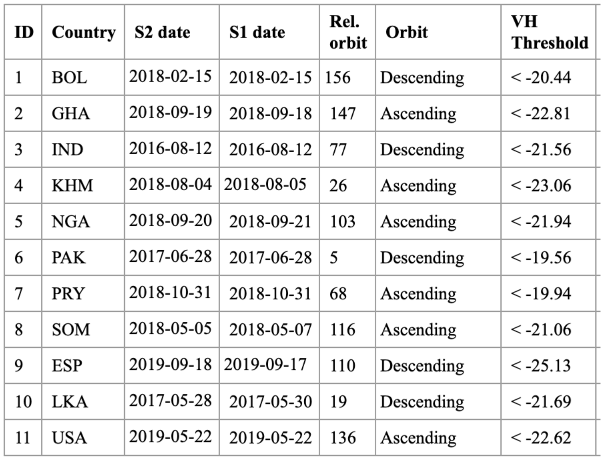


Figure 2. Locations from where permanent water data was sampled: Left. Locations from where flood event data was sampled: Right.



ID	Country	S2 date	S1 date	Rel. orbit	Orbit	VH Threshold
1	BOL	2018-02-15	2018-02-15	156	Descending	< -20.44
2	GHA	2018-09-19	2018-09-18	147	Ascending	< -22.81
3	IND	2016-08-12	2016-08-12	77	Descending	< -21.56
4	KHM	2018-08-04	2018-08-05	26	Ascending	< -23.06
5	NGA	2018-09-20	2018-09-21	103	Ascending	< -21.94
6	PAK	2017-06-28	2017-06-28	5	Descending	< -19.56
7	PRY	2018-10-31	2018-10-31	68	Ascending	< -19.94
8	SOM	2018-05-05	2018-05-07	116	Ascending	< -21.06
9	ESP	2019-09-18	2019-09-17	110	Descending	< -25.13
10	LKA	2017-05-28	2017-05-30	19	Descending	< -21.69
11	USA	2019-05-22	2019-05-22	136	Ascending	< -22.62

Figure 3. Flood Event Metadata

Spain were added in order to add more geographic dispersion and heterogeneity to the data. We provide an overview of these events in Figure 3. These two events were selected by a search through news events of flooding in the regions, before verifying that coincident imagery existed. The orbit direction and size of overlapping Sentinel-1 and Sentinel-2 area varies by event. In total, 5 events had coincident imagery, 4 had imagery within 1 day of each other, and 2 had imagery within 2 days of each other. The ground resolution of the imagery is sampled to 10 meters on all bands, with 2 bands (VV and VH) for Sentinel-1 and 13 bands for Sentinel-2. Locations for flood events are shown in Figure 2.

Preprocessing To preprocess the data, Google Earth Engine was used to filter imagery for each event using DFO start and end dates. Sentinel-1 and Sentinel-2 imagery for each flood event was stacked to a single image, composed of the intersection of scenes from each sensor. Reference flood maps from Sentinel-1 and Sentinel-2 were added to the image stack. Reference Sentinel-1 flood maps were derived using a dynamic threshold on the VH band. A 1 km x 1 km grid was created across the VH band, with high variance grids used to build a histogram. Otsu's thresholding was then utilized on the histogram, which maximizes interclass variance between flooded and non-flooded pixels [41, 14]. This threshold is then applied across a focal mean smoothed VH band to reduce speckle, resulting in a binary flood map. Reference Sentinel-2 flood maps were created after calculating Normalized Difference Vegetation Index (NDVI=(B8-B4)/(B8+B4), B=band) and Modified Normalized Difference Water Index (MNDWI= (B12-B3)/(B12+B3), B=band) bands, and applying an expert defined threshold of 0.2 and 0.3 respectively [57, 16]. Clouds were identified using a blue band reflectance threshold of less than 0.2. Cloud shadows were removed by projecting the shadows based on potential cloud heights, the solar azimuth angle, and solar zenith angle [15]. Cloud shadows have similar spectral signatures as floods, and thus are crucial to mask for accurate optical flood mapping.

Chip Creation and Sampling For each event, a smaller subset of the imagery was selected by remote sensing analysts in order to sample regions predominantly affected by flooding. The resulting subsets were divided further into 512 x 512 pixel non-overlapping chips. Chips with majority cloud cover were removed from the selection process. Sentinel-2 was used as a base water classification (see above) due to its ability to better map small water bodies, robustness in desert areas, and higher on average accuracies in clear conditions during rainy season with multiple flood events (80-83%) when compared to Sentinel-1 (66-73%) [51]. The area of Sentinel-2 water was calculated for each chip, and a stratified sample of 446 chips was selected across all events to be hand labeled for validation. The sample was stratified such that 75% of sampled chips (336) were chips that contained more than 0.02 sq km of water in the Sentinel-2 classification, and 25% of the chips (112) contained 0.02 sq km of water or less. The remaining 4,385 chips that were not selected for hand labeling, but contained water, were exported with only the Sentinel-1 and Sentinel-2 flood classifications and were not quality controlled. In total 4,831 non-overlapping chips covering 120,406 sq. km were created, with 4,385 exported for training, and 446 hand-labeled and exported for training, validation and testing.

Hand Labeling In order to create validation chips, a GUI was created in Google Earth Engine (see Figure 6) for trained remote sensing analysts to hand label water areas. Analysts had access to the Sentinel-1 VH band, two false color composites from Sentinel-2 which highlight water features (RGB: B12, B8, B4 B8, B11, B4), and the reference water classification from Sentinel-2. Using the reference Sentinel-2 water classification, analysts then marked areas to remove from the water classification, add to the water classification, and marked areas they could not confidently identify as "no data". The reference Sentinel-2 classification was then updated using the analyst labeled polygons, and a final water classification for validation was exported for each chip (see Figure 1).

Training, Validation, and Testing Data All hand-labeled data from Bolivia is held out for a distinct test set to evaluate the accuracy of trained models on flood events on which they have not been trained. In this vein, all Sentinel-2 classification maps for Bolivia are also withheld from the training and validation sets, since in the event of a real flood event there most likely would not be coincident Sentinel-2 imagery. However, we do include Sentinel-1 based flood maps for weakly supervised training data for Bolivia, since this data does not require coincident imagery or hand labels to be available.

Apart from the special handling of Bolivia data, the hand labeled data is split into training, validation, and testing data with a random 60-20-20 split and all of the non-hand-labeled Sentinel-1 and Sentinel-2 data is used for weakly supervised training data.

3. Methods

3.1. Models

This section describes the models trained on different training datasets. These models are designed to compare the performance of FCNNs across four dimensions: i) FCNNs trained on optical versus radar data, ii) FCNNs trained on hand-labeled data versus weakly supervised data, iii) FCNNs trained on permanent water only versus flood and permanent water and iv) FCNN performance compared to thresholding backscatter, a common remote sensing algorithm.

We train and test four FCNNs on each of the training datasets described above and compare them to a backscatter thresholding algorithm. The backscatter thresholding model used Otsu thresholding on the VH band, as described earlier in the paper. From the 11 flood events, we created three separate training datasets as described in section 2. The three models include i) a weakly supervised training dataset using Sentinel-1 based flood classifications as labels (Sentinel-1 Weak); ii) a weakly supervised training dataset using Sentinel-2 based flood classifications as labels (Sentinel-2 Weak), and iii) the hand-labeled flood classifications maps using corrected Sentinel-2 based flood classifications as label (Hand-Labeled). We trained a fourth CNN on the JRC permanent water dataset, which is identified from Landsat 8 data.

We also evaluate the ability of each model to identify permanent water, flood water, and total surface water. This is important for understanding the transferability of permanent water detection to flood water detection.

3.2. Convolutional Neural Networks and Accuracy Assessment

We use PyTorch to train and test all of our models [42]. As our goal is to provide straightforward baselines for the datasets rather than train the best possible models, we do not perform an exhaustive hyperparameter search. To predict water in each pixel, we use a fully convolutional network with a Resnet50 backbone [33, 23]. We use a batch size of 16 images. To account for the relatively small batch size, we convert all of the Batch Normalization layers to Group Normalization layers [24, 54]. We use the AdamW optimizer with a base learning rate of 5e-4 and a weight decay coefficient of 1e-2 [35]. We also use a cosine annealing learning rate scheduler with warm restarts with an initial T_0 of 1 epoch and T_{mult} of 2 [34]. For data augmentation, we randomly crop the images from 512x512 chips to 256x256 and randomly apply horizontal and vertical flips.

We then perform mean and standard deviation normalization using the mean and standard deviation computed over the hand labeled training dataset ([0.6851, 0.5235], [0.0820, 0.1102]). Intersection over union(IOU) is used to evaluate the models; we report the mean IOU(equal weighting of all chips). We report all evaluation metrics on the flooded water dataset. The flooded water dataset contains permanent and flooded water, so we compute separate metrics for the flooded water pixels and permanent water pixels. We also report omission and commission error rates for comparison to remote sensing literature. Omission rates are calculated as the rate of false negative water detections, and commission rates are the rate of false positive water detections. We monitor convergence and overfitting during training using our validation dataset. We use convergence and overfitting data to decide how many epochs to train. The model checkpoint with the highest mean IOU score on our validation data of 11 flood events as our final model for a given training run. Training and evaluation code is available at the link in the abstract.

Dataset	PW	FW	AW
Sentinel-1 Weak	.2872	.2422	.3092
Sentinel-2 Weak	.3818	**.3389**	**.4084**
Hand-Labeled	.2570	.2421	.3125
Permanent Water	.3391	.1693	.2452
Otsu Threshold-VH	**.4571**	.2850	.3591

Table 1. Performance on the hand-labeled test set of 10 flood events (all besides Bolivia) of models trained on each dataset in terms of Mean IOU for the water class. Results shown on permanent water(PW), flooded water(FW), and all water(FW).

Dataset	PW	FW	AW
Sentinel-1 Weak	.2506	**.3296**	**.3871**
Sentinel-2 Weak	.1946	.2738	.3160
Hand-Labeled	.2300	.2905	.3524
Permanent Water	**.2881**	.2684	.3422
Otsu Threshold-VH	.2859	.3239	.3862

Table 2. Performance on the hand-labeled test set of the flood event in Bolivia of models trained on each dataset in terms of Mean IOU for the water class. Results shown on permanent water(PW), flooded water(FW), and all water(FW).

4. Results

4.1. Traditional Remote Sensing Baseline

The results are in Tables 1, 2, 3, and 4. Validation data results reveal that backscatter thresholding on the VH band is the best performing model to identify permanent water (IOU= 0.4571), outperforming CNNs by a considerable amount, mainly because of lower rates of omission error (0.0540). The backscatter thresholding flood detection modeling was the second highest performing to identify flood water and total surface water area (IOU= 0.2850 and 0.3591, respectively). Overall accuracy (OA) metrics would rank this method as the most accurate for all surface water (0.9389). However, high OA is likely due to underprediction (e.g high omission error for flood water at 0.3756) and the fact that a majority of area in each chip is dry land (e.g. not permanent or flood water) which drives the overall accuracy metric.

4.2. Permanent Water

We trained the CNN on permanent water for 200 epochs. It performed relatively to detect permanent surface water (IOU= 0.33912), but was the lowest performing model to detect flood water (IOU=0.1693) and total surface water (IOU=0.2452). This model had the lowest rates of commission error among any other models for all surface water and flood water(0.0633 and 0.0633 respectively), but higher rates of omission error (0.2440 and 0.2340 respectively).

4.3. Flooded Water

We trained for 3 epoch on Sentinel-1 Weak, 200 epochs on Sentinel-2 Weak, and 100 epochs on Hand-Labeled. Epoch training numbers were determined based on how quickly the training appeared to be converging or overfitting based on the training and validation losses. We observe that training starts overfitting after only a few epochs (3) for Sentinel-1 Weak and eventually occurs on Hand-Labeled as well after 100 epochs. Models trained on Sentinel-2 weak do not appear to overfit, but their validation loss does appear to converge and stop decreasing at around 200 epochs. Test results are in Tables 1, 2, 3, 4. Sentinel-2 weak was the best performing model to detect flood water (IOU= 0.3389, OA=0.9277) and all water (IOU= 0.4084, OA=0.9384) while Sentinel-1 weak was the best performing model on the holdout Bolivia test for flood water (IOU= 0.3296, OA=0.9277) and all water (IOU= 0.3871, OA=0.9384). The FCNN model trained on hand-labels was never the best performing model by any metric.

5. Discussion and Conclusion

These initial analyses provide some valuable insights into training deep learning models to predict floods and permanent surface water. The models and results presented in this paper are trained on a relatively small dataset of 11 flood events. However, the results presented here provide preliminary insights towards answering key research questions.

5.1. Hand-labeled training data is not necessary to train FCNNs to detect flood water

The model trained on hand-labeled data does not provide the best results according to IOU or OA metrics on any of our tests. Building this training dataset was much more time and labor intensive than building any of the other datasets where labels were automatically generated with other remote sensing classification algorithms. These results suggest that, for future datasets, this time and energy could be better spent on automatically generating large quantities of data (such as with Sentinel-1 weak, Sentinel-2 weak and Permanent Water) and expanding the number of flood events for training. This is an encouraging result because given the large number of flood detection algorithms available, and the availability of cloud based remote sensing platforms like Google Earth Engine, generating flood maps for many events should not prove to be a difficult task. Notwithstanding the importance of more weakly supervised classification data, hand labeled data remains essential to validate and compare different models. We note that it is likely that models trained on weakly supervised data perform better due to increased quantity of training data rather than increased quality of training data. Our results are consis-

Dataset	**Permanent Water**		**Flood Water**		**All Water**	
	Om	**Comm**	**Om**	**Comm**	**Om**	**Comm**
Sentinel-1 Weak	.0660	.1354	**.1190**	.0997	**.1124**	.0997
Sentinel-2 Weak	.1209	**.0534**	.2684	.0778	.2482	.0778
Hand-Labeled	.0945	.1519	.1352	.1055	.1297	.1055
Permanent	.1485	.1064	.2440	**.0633**	.2340	**.0633**
Otsu Threshold-VH	**.0540**	.0849	.1510	.0849	.1427	.0849

Table 3. Performance on the hand-labeled test set of 10 flood events (all besides Bolivia) of models trained on each dataset in terms of omission and commission error for the water class.

Dataset	**Permanent Water**		**Flood Water**		**All Water**	
	Om	**Comm**	**Om**	**Comm**	**Om**	**Comm**
Sentinel-1 Weak	.3181	.0715	.3950	.0695	.2787	.0695
Sentinel-2 Weak	.0588	.0467	**.2155**	.0414	.1575	.0414
Hand-Labeled	.0669	.0904	.2148	.0813	**.1518**	.0813
Permanent Water	.3427	**.0420**	.5704	**.0340**	.4073	**.0340**
Otsu Threshold-VH	**.0129**	.0508	.3756	.0525	.2480	.0525

Table 4. Performance on the hand-labeled test set of flooding in Bolivia of models trained on each dataset in terms of omission and commission error for the water class.

tent with other research showing the minimal amounts of ”weak” training data can train CNN model that will still outperform traditional remote sensing methods and other common machine learning approachies like random forests or support vector machines [52]. More work can be done to explore the precise quality versus quantity tradeoffs here and compare with other machine learning methods.

5.2. Sentinel-2 provides better automatic labels for Sentinel-1 based flood detection

FCNNs trained on weak Sentinel-2 classifications performed the best compared to other models on average for 10 flood events. One explanation is that previous research has found Sentinel-2 to be superior to Sentinel-1 for near real time flood mapping over rainy season, so these labels may just be more accurate [51]. Another explanation is that since the Sentinel-2 classification labels come from data that is unavailable to the model, the model does not overfit as much. This explanation fits well with the qualitative observation that the validation curves for Sentinel-2 Weak did not show signs of overfitting while there was clear signs of overfitting on Sentinel-1 Weak and Hand-labeled. However, the Bolivia holdout test shows that the CNN trained on Sentinel-1 Weak classification has the highest IOU and OA results. This could be because there was Bolivia data in the Sentinel-1 weak training dataset and not in the Sentinel-2 Weak training data. More research is needed on a larger number of flood events to provide conclusive evidence regarding which sensor provides the best training data. Additional research could explore the value of assimilating data from various sensors (e.g. including MODIS, Lansat, Sentinel-1 and 2, and even commercial sensors) to train models, as this may prevent overfitting and improve the ability of FCNNs to adapt to a variety of flood events.

5.3. FCNNs trained on flood water perform better than those trained on permanent water alone

Permanent water prediction accuracy and flooded water prediction accuracy are not always correlated. For example, the CNN model trained on the permanent water dataset is among the best performing models for permanent water prediction but is by far the worst performing model for flooded water prediction. These results suggest that flood detection FCNNs should be trained on flood events and that training on permanent water is not necessarily transferable. This also means that testing on flooded water is important for a realistic accuracy assessment. Given that FCNNs perform well with weak labels from flood events, a logical next step is to generate much larger training datasets for flood water and flood events specifically.

5.4. FCNNs outperform thresholding algorithms to identify flooded but not permanent water

At least one of the FCNNs we trained outperformed the optimized thresholding of VH backscatter on each of the test datasets to identify flooded water. However, Otsu threshold-VH outperforms all FCNNs trained in this study to identify permanent water. This is in contrast to previous research [25] which found deep learning algorithms to substantially reduce commission errors from threshold algorithms, such as MNDWI thresholding of Landsat (commission error 0.45) to a commission error of 0.08 using a five layer CNN. However, the “conventional” radar remote sensing algorithm used in this paper for comparison is more advanced than a simple MNDWI threshold because the threshold is optimized to each flood event case. The radar remote

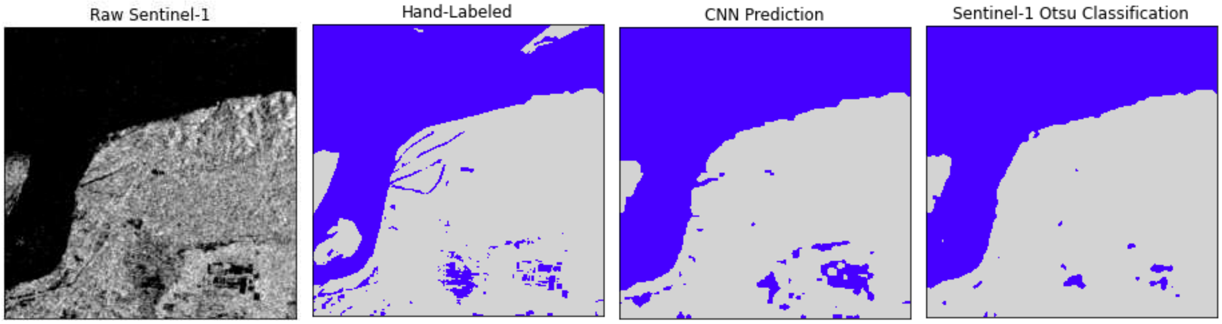


Figure 4. Example predictions on test data.

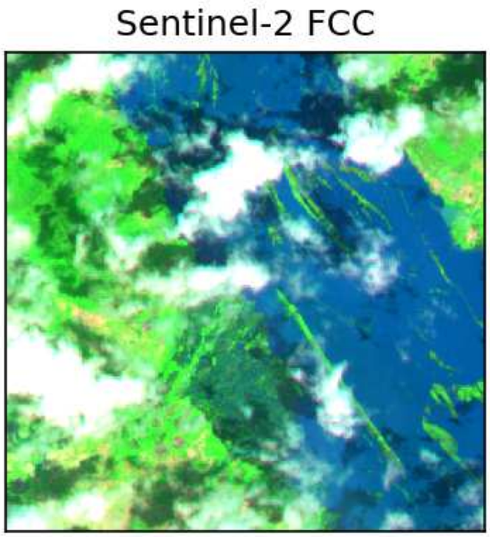
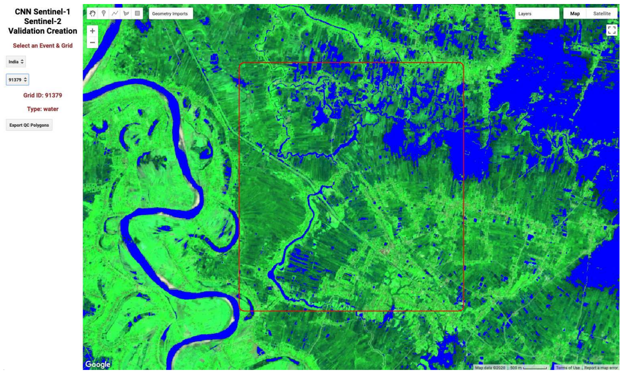


Figure 5. Sentinel-2 Imagery from a flood in Nigeria. Cloud shadows, like in this image, are sometimes confused for water by water detection algorithms.



India

91379

Grid ID: 91379

Type: water

Figure 6. GUI for labeling data.

sensing algorithm in this paper reduces speckle with focal mean filters which may remove small water features, but using refined Lee filter could have mitigated this problem. The initial results in this paper suggest that deep learning could provide substantial gains in accuracy to detect floods in particular by reducing both errors of commission and omission. Further accuracy gains could be achieved via simple improvements such as different data augmentation schemes and optimized hyperparameters. FCNNs could be compared to other machine learning methods, or improved thresholding methods and algorithms using, for example, refined Lee filters. We encourage others to leverage and build upon the training and validation data provided in this paper to improve models further. Our validation dataset does not include any urban flood events, in part because Sentinel-1 algorithms often struggle to map floods in urban areas. However, recent research shows CNNs have promising results in mapping urban floods especially using interferometric information (beyond just backscatter thresholding used here) [31]. Increasing the existing training dataset to include urban floods would be an important next step.

5.5. Conclusion

In conclusion, while computer vision is already contributing substantially to improve remote sensing of land and cloud cover, there are no datasets available to train algorithms for flood detection using publicly available satellite imagery. Here we focus on surface water detection for floods specifically, to contribute to efforts to operationalize monitoring for humanitarian and relief efforts. We provide Sen1Floods11 for other researchers to train deep learning algorithms for flood detection without the overhead of generating training and validation datasets. Expanding this dataset to include urban flood events, including additional sensors, and training models with additional radar information (e.g. interferometry and change in coherence) are important future avenues for research to build on this current effort.

References

[1] Lorenzo Alfieri, Sagy Cohen, John Galantowicz, Guy JP Schumann, Mark A Trigg, Ervin Zsoter, Christel Prudhomme, Andrew Kruczkiewicz, Erin Coughlan de Perez, Zachary Flamig, et al. A global network for operational flood risk reduction. *Environmental science & policy*, 84:149–158, 2018. 1

[2] Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, Apr. 2016. 2

[3] Mirco Boschetti, Francesco Nutini, Giacinto Manfron, Pietro Alessandro Brivio, and Andrew Nelson. Comparative analysis of normalised difference spectral indices derived from modis for detecting surface water in flooded rice cropping systems. *PloS one*, 9(2), 2014. 2

[4] G.R. Brakenridge. Global active archive of large flood events. http://floodobservatory.colorado.edu/Archives/index.html. 3

[5] Robert Brakenridge and Elaine Anderson. Modis-based flood detection, mapping and measurement: the potential for operational hydrological applications. In *Transboundary floods: reducing risks through flood management*, pages 1–12. Springer, 2006. 2

[6] Serena Ceola, Francesco Laio, and Alberto Montanari. Satellite nighttime lights reveal increasing human exposure to floods worldwide. *Geophysical Research Letters*, 41(20):7184–7190, 2014. 1

[7] Stephen Chignell, Ryan Anderson, Paul Evangelista, Melinda Laituri, and David Merritt. Multi-Temporal Independent Component Analysis and Landsat 8 for Delineating Maximum Extent of the 2013 Colorado Front Range Flood. *Remote Sensing*, 7(8):9822–9843, July 2015. 2

[8] Marco Chini, Renaud Hostache, Laura Giustarini, and Patrick Matgen. A Hierarchical Split-Based Approach for Parametric Thresholding of SAR Images: Flood Inundation as a Test Case. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):6975–6988, Dec. 2017. 2

[9] Marco Chini, Ramona Pelich, Luca Pulvirenti, Nazzareno Pierdicca, Renaud Hostache, and Patrick Matgen. Sentinel-1 InSAR Coherence to Detect Floodwater in Urban Areas: Houston and Hurricane Harvey as A Test Case. *Remote Sensing*, 11(2):107, Jan. 2019. 2

[10] Fabio Cian, Mattia Marconcini, and Pietro Ceccato. Normalized Difference Flood Index for rapid flood mapping: Taking advantage of EO big data. *Remote Sensing of Environment*, 209:712–730, May 2018. 2

[11] EMDAT CRED. Unisdr. 2017. the human cost of weather-related disasters 1995-2016. Technical report, Technical report CRED, EM-DAT, and UNISDR. 1

[12] Ben DeVries, Chengquan Huang, John Armston, Wenli Huang, John W Jones, and Megan W Lang. Rapid and robust monitoring of flood events using sentinel-1 and landsat data on the google earth engine. *Remote Sensing of Environment*, 240:111664, 2020. 2

[13] Ben DeVries, Chengquan Huang, Megan Lang, John Jones, Wenli Huang, Irena Creed, and Mark Carroll. Automated Quantification of Surface Water Inundation in Wetlands Using Optical Satellite Imagery. *Remote Sensing*, 9(8):807, Aug. 2017. 2

[14] Gennadii Donchyts, Jaap Schellekens, Hessel Winsemius, Elmar Eisemann, and Nick van de Giesen. A 30 m Resolution Surface Water Mask Including Estimation of Positional and Thematic Differences Using Landsat 8, SRTM and OpenStreetMap: A Case Study in the Murray-Darling Basin, Australia. *Remote Sensing*, 8(5):386, May 2016. 2, 4

[15] G. Doncyths and I. Hausman. Geeviz: A repository of gee python code modules for general data processing, analysis, and visualization. https://github.com/gee-community/geeViz. 4

[16] Yun Du, Yihang Zhang, Feng Ling, Qunming Wang, Wenbo Li, and Xiaodong Li. Water Bodies' Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the SWIR Band. *Remote Sensing*, 8(4):354, Apr. 2016. 2, 4

[17] Markus Enenkel, Daniel Osgood, Martha Anderson, Bristol Powell, Jessica McCarty, Christopher Neigh, Mark Carroll, Margaret Wooten, Greg Husak, Christopher Hain, et al. Exploiting the convergence of evidence in satellite data for advanced weather index insurance design. *Weather, Climate, and Society*, 11(1):65–93, 2019. 1

[18] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 3

[19] Lian Feng, Chuanmin Hu, Xiaoling Chen, Xiaobin Cai, Liqiao Tian, and Wenxia Gan. Assessment of inundation changes of poyang lake using modis observations between 2000 and 2010. *Remote Sensing of Environment*, 121:80–92, 2012. 2

[20] Gudina L. Feyisa, Henrik Meilby, Rasmus Fensholt, and Simon R. Proud. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment*, 140:23–35, Jan. 2014. 2

[21] Laura Giustarini, Renaud Hostache, Patrick Matgen, Guy J.-P. Schumann, Paul D. Bates, and David C. Mason. A Change Detection Approach to Flood Mapping in Urban Areas Using TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2417–2430, Apr. 2013. 2

[22] Stephane Hallegatte, Adrien Vogt-Schilb, Mook Bangalore, and Julie Rozenberg. *Unbreakable: building the resilience of the poor in the face of natural disasters*. World Bank Publications, 2016. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[25] Furkan Isikdogan, Alan C Bovik, and Paola Passalacqua. Surface water mapping by deep learning. *IEEE journal of selected topics in applied earth observations and remote sensing*, 10(11):4909–4918, 2017. 2, 7

[26] Akm Saiful Islam, Sujit Kumar Bala, and Anisul Haque. Flood inundation map of bangladesh using modis surface reflectance data. In *International conference on water and flood management (ICWFM) Dhaka, Bangladesh*, volume 2, pages 739–748. Citeseer, 2009. 2

[27] Jian Li and David Roy. A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring. *Remote Sensing*, 9(9):902, Aug. 2017. 2

[28] Igor Klein, Andreas Dietz, Ursula Gessner, Stefan Dech, and Claudia Kuenzer. Results of the Global WaterPack: a novel product to assess inland water body dynamics on a daily basis. *Remote Sensing Letters*, 6(1):78–87, Jan. 2015. 2

[29] Scott A Kulp and Benjamin H Strauss. New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. *Nature communications*, 10(1):1–12, 2019. 1

[30] Jong-Sen Lee. Refined filtering of image noise using local statistics. 15(4):380–389. Publisher: Elsevier. 2

[31] Yu Li, Sandro Martinis, and Marc Wieland. Urban flood mapping with an active self-learning convolutional neural network based on terrasar-x intensity and interferometric coherence. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:178–191, 2019. 3, 8

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 5

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[36] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, June 2019. 2

[37] Sandro Martinis, André Twele, and Stefan Voigt. Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-X data. *Natural Hazards and Earth System Sciences*, 9(2):303–314, 2009. 2

[38] D.C. Mason, L. Giustarini, J. Garcia-Pintado, and H.L. Cloke. Detection of flooded urban areas in high resolution Synthetic Aperture Radar images using double scattering. *International Journal of Applied Earth Observation and Geoinformation*, 28:150–159, May 2014. 2

[39] 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. *Physics and Chemistry of the Earth, Parts A/B/C*, 36(7-8):241–252, Jan. 2011. 2

[40] Perry C Oddo and John D Bolten. The value of near real-time earth observations for improved flood disaster response. *Frontiers in Environmental Science*, 7:127, 2019. 1

[41] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. 4

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 5

[43] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 2016. 2, 3

[44] Maryam Rahnemoonfar, Robin Murphy, Marina Vicens Miquel, Dugan Dobbs, and Ashton Adams. Flooded area detection from uav images based on densely connected recurrent neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1788–1791. IEEE, 2018. 2

[45] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019. 2

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[47] Stefan Schlaffer, Patrick Matgen, Markus Hollaus, and Wolfgang Wagner. Flood detection from multi-temporal SAR data using harmonic analysis and change detection. *International Journal of Applied Earth Observation and Geoinformation*, 38:15–24, June 2015. 2

[48] 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. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:153–160, Sept. 2019. 3

[49] Xinyi Shen, Emmanouil N Anagnostou, George H Allen, G Robert Brakenridge, and Albert J Kettner. Near-real-time non-obstructed flood inundation mapping using synthetic aperture radar. *Remote Sensing of Environment*, 221:302–315, 2019. 2

[50] Beth Tellman, Jonathan Sullivan, and Colin Doyle. Global Flood Observation with Multiple Satellites: Applications in Rio Salado, Argentina, and the Eastern Nile Basin. In *Global Drought and Flood: Monitoring, Prediction, and Adaptation*, AGU Books. 3

[51] Beth Tellman, Sam Weber, Lisa Landuyt, Tyler Anderson, JIayong Liang, Emmalina Glinskis, Jeff C. Ho, Colin Doyle, and Jonathan Sullivan. From publishable to operational: new metrics to more honestly measure the ability of remote sensing algorithms to consistently monitor flooded assets and

populations in near real time. In *Global Floods: Forecasting, Monitoring, Risk Assessment, and Socioeconomic Response I.*, San Francisco CA, Dec. 2019. 4, 7

[52] Sherrie Wang, William Chen, Sang Michael Xie, George Azzari, and David B. Lobell. Weakly supervised deep learning for segmentation of remote sensing imagery. 12(2):207. 2, 7

[53] Marc Wieland, Yu Li, and Sandro Martinis. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sensing of Environment*, 230:111203, Sept. 2019. 2

[54] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 5

[55] Michael A. Wulder, Thomas R. Loveland, David P. Roy, Christopher J. Crawford, Jeffrey G. Masek, Curtis E. Woodcock, Richard G. Allen, Martha C. Anderson, Alan S. Belward, Warren B. Cohen, John Dwyer, Angela Erb, Feng Gao, Patrick Griffiths, Dennis Helder, Txomin Hermosilla, James D. Hipple, Patrick Hostert, M. Joseph Hughes, Justin Huntington, David M. Johnson, Robert Kennedy, Ayse Kilic, Zhan Li, Leo Lymburner, Joel McCorkel, Nima Pahlevan, Theodore A. Scambos, Crystal Schaaf, John R. Schott, Yongwei Sheng, James Storey, Eric Vermote, James Vogelmann, Joanne C. White, Randolph H. Wynne, and Zhe Zhu. Current status of Landsat program, science, and applications. *Remote Sensing of Environment*, 225:127–147, May 2019. 2

[56] Xiangming Xiao, Stephen Boles, Steve Frolking, Changsheng Li, Jagadeesh Y Babu, William Salas, and Berrien Moore III. Mapping paddy rice agriculture in south and southeast asia using multi-temporal modis images. *Remote Sensing of Environment*, 100(1):95–113, 2006. 2

[57] Hanqiu Xu. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033, July 2006. 4

[58] Kang Yang, Manchun Li, Yongxue Liu, Liang Cheng, Yuewei Duan, and Minxi Zhou. River Delineation from Remotely Sensed Imagery Using a Multi-Scale Classification Approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(12):4726–4737, Dec. 2014. 2

[59] Brittany Zajic. How flood mapping from space protects the vulnerable and can save lives, Jun 2019. 1

[60] Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt J Kusner, and Duncan Watson-Parris. Cumulo: A Dataset for Learning Cloud Classes. page 11, 2019. 2

[61] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, Dec. 2017. 2