IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 4, 2020, accepted September 16, 2020, date of publication September 21, 2020,
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025302

Flash Flood Detection From CYGNSS Data Using the RUSBoost Algorithm

PEDRAM GHASEMIGOUDARZI[1], (Student Member, IEEE), WEIMIN HUANG[1], (Senior Member, IEEE), OSCAR DE SILVA[1], (Member, IEEE), QINGYUN YAN[2], (Member, IEEE), AND DESMOND T. POWER[3], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL A1B 3X7, Canada
[2]Department of Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3]Centre for Cold Ocean Resources Engineering (C-Core), St. John's, NL A1B 3X5, Canada

Corresponding author: Weimin Huang (weimin@mun.ca)

The work of Weimin Huang was supported in part by the Natural Sciences and Engineering Research Council of Canada Discovery Grant NSERC RGPIN-2017-04508 and Grant RGPAS-2017-507962, and in part by the Canadian Space Agency CubeSat Grant 17CCPNFL11. The work of Desmond T. Power was supported by the Canadian Space Agency CubeSat Grant 17CCPNFL11.

ABSTRACT Flash floods can cause massive damages because of their rapid evolution. To reduce or prevent harm caused by a flash flood, it is vital to have information about its formation and spread. Hence, providing real-time surveillance flood is essential. Considering Hurricane Harvey and Hurricane Irma as two case studies, six different data preparation approaches (DPAs) for flood detection based on the Cyclone Global Navigation Satellite System (CYGNSS) data and the Random Under-Sampling Boosted (RUSBoost) classification algorithm are investigated in this article. Taking flood and land as two classes, flash flood detection is tackled as a binary classification problem. Eleven observables are extracted from the delay-Doppler maps (DDMs) for feature selection. These observables, alongside two features from an ancillary data, are considered in feature selection. All the combinations of these observables with and without ancillary data are fed into the classifier with 5-fold cross-validation one by one. Based on the test results, five observables with the ancillary data are selected as a suitable feature vector for flood detection here. Using the selected feature vector, six different DPAs are investigated and compared to find the best one for flash flood detection. Then, the performance of the proposed method is compared with that of a support vector machine (SVM) based classifier. For Hurricane Harvey and Hurricane Irma, the selected method is able to detect 89.00% and 85.00% of flooded points, respectively, with a resolution of 500 m $\times$ 500 m, and the detection accuracy for non-flooded land points is 97.20% and 71.00%, respectively.

INDEX TERMS Flood detection, CYGNSS, global navigation satellite system reflectometry (GNSS-R), random under-sampling boosted (RUSBoost), support vector machine (SVM).

I. INTRODUCTION

A flash flood is a surge of water that starts and develops in a short period. The primary cause of flash flood is heavy rain. Additionally, dam breakage, ice and snow meltdown, and events in which a large amount of water is released to dry areas can also cause flash flood. Even though a flash flood dissipates quickly after the occurrence, it has consequential damages such as death and severe injuries, water contamination, financial harms, infrastructure damages, and agricultural losses [1], [2]. Hurricanes, which are a significant cause of flash floods, are tropical cyclones with high wind speed (higher than 33 ms^{-1} [3]) and capable of pouring massive rain over coastal regions during landfall [4]. Considering the population growth in coastal areas that are exposed to hurricanes, flood detection and monitoring its extent are important to reduce these damages and increase the speed of post-disaster response [5].

Being able to monitor the surface of Earth continuously, remote sensing technologies have been used as reliable resources for flood detection. In order to observe the extent of floods, different methods based on active and passive remote sensing satellite systems operating at various frequencies have been applied [6]–[9]. Different electromagnetic

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

waves within the visible spectrum interact differently with water bodies, depending on their wavelengths. For instance, blue bands penetrate the water, while red bands are partially absorbed and near-infrared bands are fully absorbed. Therefore, by defining certain thresholds, the water bodies, including flash floods, can be detected using optical sensors [10]. The detected water extent are then compared with water reference data sets to estimate floods extent maps [8]. The water bodies detection algorithms are able to detect the surface water including floods with an accuracy higher than 97% [11], [12]. However, in an optical image, the cloud shadows are classified as flood. Therefore, the cloud shadows are the main challenges for flood extent estimation using optical sensors [12]. Active microwave satellites such as synthetic aperture radar (SAR) systems work in day or night providing high spatial resolution data. They are able to see through obstacles such as clouds and certain biomass. Similar to any other classification problem, creating floods extent maps from SAR data can be solved by using supervised and unsupervised methods [13]. In a supervised method, since, the classifier is trained with labeled pixels from a region, the algorithm has local dependence. Segmentation [14], threshold determination [15], and change detection [16] are three main methods for unsupervised classification. Even though these methods are able to detect floods effectively, they have drawbacks. The segmentation method requires heavy computations compared with the other two methods. Moreover, since it ignores small flooded clusters surrounded by large non flooded ones and vice versa, it is less precise [13]. In the threshold determination method, instead of a single threshold value, multiple threshold values are considered for detecting floods on a large scale [15]. Therefore, its accuracy is highly dependant on how accurate the preset threshold values are. In the change detection method, prior- and post-flood SAR images are required, which is a big challenge due to the revisit time of SAR systems [16]. Therefore, in recent works, combinations of these techniques are used [17], [18]. Depending on the region, the SAR based flood detection algorithms are able to detect floods with an accuracy ranging between 80 % to 95 % [19]–[21]. The SAR data requires geometric correction and speckle reduction. Hence, compared to passive microwave and optical sensors, the retrieval algorithms based on SAR data are more complicated [13]. Moreover, since in active SAR systems such as RADARSAT-1/2, TerraSAR-X, and Sentinel-1 the transmitter and receiver are placed on the same platform, obtaining a large constellation is costly and their constellations are usually small [22]. Thus, due to the low temporal resolution (several days), satellite might not even be able to collect data over a flash flooded area in time. Some of these remote sensing data and algorithms have been used by observatories to create near real-time (NRT) flash floods information. For example, the Dartmouth Flood Observatory (DFO) and the NASA Goddard's Hydrology Laboratory employ the data collected by two MODIS sensors (aboard the satellites Terra and Aqua) for flood monitoring [8]. By computing the MODIS reflectance ratio of Band 1 (red) and Band 2 (near-infrared) as well as a threshold on Band 7 (shortwave infrared) to estimate water extent and comparing with reference data, they determine the flash flooded areas [23]. Also, they employ microwave sensors data to mitigate the cloud effect to increase flood detection accuracy [8]. The Global Flood Detection System (GFDS) uses AMSR-E passive microwave remote sensing data to detect riverine flooding globally. In this system, the value of calibrated surface brightness is compared with a threshold to detected riverine inundations [9].

The Global Navigation Satellite System Reflectometry (GNSS-R) is a well-established technique for remote sensing [24]. The GNSS-R receivers collect the Global Navigation Satellite System (GNSS) signals reflected from the surface of the Earth in a bistatic radar configuration. Since it takes advantage of existing signals of opportunity, a GNSS-R system does not require any onboard transmitter. Hence, they are cost-efficient, which makes it possible to achieve a larger constellation and, consequently, high temporal resolution (hours) [25]. Since the GNSS-R receivers operate at L-Band, they can see through clouds what exist when most flash floods happen. Therefore, with a large operational constellation, the GNSS-R data is ideal for flash flood remote sensing. The GNSS-R has shown a great capacity for various applications such as altimetry [26], sea surface wind [27]–[31], soil moisture (SM) [32]–[34], target detection [35], tsunami [36], [37], sea ice [38]–[43], inland water detection [44], and seasonal flood classification [45]. However, its application for flash floods detection is yet to be investigated.

Among GNSS-R data sets collected by different GNSS-R receivers, only the Cyclone Global Navigation Satellite System (CYGNSS) GNSS-R data, which is collected by the constellation of eight satellites [46], was shown to have the potential for capturing the variations of surface reflectivity caused by flash floods over affected regions [47], which are consistent with the changes in precipitation data and brightness temperature data [48]. Unlike seasonal floods, which are developed in specific periods, and they last over more extended time [49], the flash floods are difficult to monitor. Therefore, it is vital to develop a method that can detect and monitor flash floods. To the best of the authors' knowledge, the capacity of the GNSS-R technique for flash flood detection has not been quantitatively analyzed and this article is the first one that investigates such a topic by using the CYGNSS GNSS-R data. Therefore, the objective of this work is to quantitatively investigate the ability of the GNSS-R technique to detect flash floods. The flash flood detection problem is a binary classification problem with two classes (flood and land). Various ML algorithms can be implemented for solving a binary supervised classification problem, such as the Neural Networks (NN), Super Vector Machines (SVMs), and Decision Trees, which are among the most commonly used classifiers in remote sensing [50]. By combining decision trees as basic classifiers, a classifier that outperforms

the constituent classifiers is created, which is called an ensemble classifier. Stacking, blending, bagging, and boosting are four main approaches for creating an ensemble classifier [51].

Since floods are usually localized, when a large area is considered, the number of points collected over flooded regions is smaller than those obtained from the non-flooded one, which creates an imbalanced data set. For instance, the flood labeled data points are only 4.38 % and 8.37 % of the Harvey and Irma data sets, respectively. These two imbalanced factors are calculated as the ratios of the number of flood labeled data points (m_f) to the total number of data points (m) in each data set. In an imbalanced data set, information provided by the minor class is considered less important due to the unequal ratio between major and minor classes. However, the minor class results could be more vital at higher costs, despite its smaller size. Various strategies for tackling imbalanced data sets have been developed [52]. At the data level, the leading solutions for handling imbalanced data include cost-sensitive learning and data sampling. In cost-sensitive learning, each class is assigned with a misclassification cost and the goal is to minimize the overall misclassification cost instead of maximizing the accuracy of the model [52]. In data sampling, by creating new instances in the minor class (oversampling methods) or eliminating instances from the major class (under-sampling methods), the imbalanced data becomes balanced [52]. The Synthetic Minority Oversampling Technique (SMOTE) and the Adaptive Synthetic Sampling Method (ADASYN) are two renowned oversampling methods, in which synthetic instances are generated from existing instances in the minor class [53], [54]. As a powerful tool, the Generative Adversarial Network (GAN) is another method for creating artificial instances in the minor class [55]. In such a method, two neural networks compete to optimize their objective functions that are contradictory to each other [56]. The Random Under-Sampling (RUS) is an under-sampling method that balances the data via random elimination of instances from the major class [52]. The balancing techniques are applied to different classifiers, such as ensembles methods, leading to various developed algorithms for classifying imbalanced data [52]. Among different methods developed for classifying imbalanced data, in this study, the Random Under-Sampling Boosted (RUSBoost) algorithm is selected for classification due to its efficient computational time, accuracy, and widely available resources [57]–[59]. Moreover, the support vector machines (SVM) algorithm, which is a representative ML method, is implemented for comparison purpose.

In this article, six different data preparation approaches (DPAs) for flood detection with high-resolution (500 m × 500 m) are investigated using the RUSBoost based algorithm. After comparison, the best technique is selected for flash flood detection using CYGNSS data. The performance of the proposed method is compared with that of a SVM based classifier. The contributions of this study are as follows

1) The first method for detecting flash floods using the GNSS-R technique is proposed.
2) Based on the eleven different CYGNSS observables and an ancillary data set, a suitable feature vector for flash flood detection is determined.
3) Six different DPAs for detecting flash floods using the CYGNSS data are investigated, showing Approach 3 as the best one.

This study is a well-detailed follow-up of what has been done in [60]. This work is outlined as follows: Section II introduces employed data sets. In Section III, eleven different CYGNSS observables, and the RUSBoost-based and SVM-based algorithms are described. Section IV provides the results of feature selection, flood detection, and comparison with the SVM classifier. Conclusions are provided in Section V.

II. DATA SETS

A. CYGNSS

In this work, we employed level 1 V2.1 of the CYGNSS data [61] that are available for the public through [62].

In the CYGNSS constellation, each satellite is an along-track scanner which collects the GNSS reflected signal in the direction of the satellite passing over a region with an onboard GNSS-R payload. Hence, when a disaster occurs in a few days (5 to 10 days), the CYGNSS receivers are only able to cover a portion of the flooded area and for some areas, there is no data. Considering this limitation, among all the floods that have happened since 2016 (the year CYGNSS was launched) to 2019, we considered two significant events, Hurricane Harvey and Hurricane Irma. These two hurricane events are among the harshest and costliest ones that have affected the United States significantly [63].

Hurricane Harvey reached the coast of the USA on Aug. 25th, 2017, and according to media, the inundation lasted till Sep. 8th, 2017. Hurricane Irma landed the coast of the USA on Sep. 10th, 2017 and caused a 6-day flood. The affected areas of Hurricane Harvey and Hurricane Irma are located in geographic coordinates of [26.7° N, 32.29° N][91° W, 100° W] and [24.5° N, 29.2° N] [79.2° W, 93° W], respectively. Since Hurricane Harvey affected a larger area compared to Hurricane Irma, it has more data points. In other words, the data of Irma might not be enough to fully train a classifier, which may lead to an underfitted model. Therefore, in Section IV-A, for feature selection, both data sets are combined and used for classification by a 5-fold cross-validation evaluation. Furthermore, in Section IV-B, we decided to use 50% of the Harvey data for training and then validate the trained classifier with the remaining 50%. This trained classifier is then tested with the data of Irma, which is unknown to the machine.

The CYGNSS constellation collects the reflected GPS L1-band signals and converts them into DDMs. A DDM is a projection of the scattered signal from the surface around a specular point (SP). In the CYGNSS data set, each SP has a

17 delay bins in 11 Doppler bins DDM with a delay bin equals 249.4 ns and a Doppler bin equals to 500 Hz.

B. ANCILLARY DATA

In current work, the flood maps created by the DFO are used as reference data for training and validation. The DFO is a remote sensing research lab of Institute of Arctic and Alpine Research (INSTAAR), at the University of Colorado Boulder. As a part of the Global Disaster Alert and Coordination System (GDACS) project, they create and provide flood maps using data from multiple sources, including NASA MODIS, ESA Sentinel 1, ASI Cosmo SkyMed, Copernicus Sentinel 1, and Radarsat 2 [8]. In this work, the regions impacted by Hurricane Harvey and Hurricane Irma are considered as two case studies, one flood map for each event is obtained from the DFO geographic information system (GIS) data. The GIS data of Hurricane Harvey and Hurricane Irma and more details on them are available through [68], and [64], respectively.

Since the water tends to move to places at low altitudes, the elevation data can impact the accuracy of classification. Therefore, altitude data of the Shuttle Radar Topography Mission Digital Elevation Model (SRTM90m DEM) is employed as an ancillary data [65]. This data set alongside the extracted GNSS-R observables are used as the input for training and testing of the classifier.

The flood reference map is created based on the changes of the surface during the flood. However, areas with water bodies such as permanent waters and some regions of wetlands might have similar characteristics to flood, but SPs over such regions could be detected as either flood and land. One solution is to exclude the points that are located over such areas. Therefore, for excluding such data points, the Global Wetland V3 data provided by the Center for International Forestry Research (CIFOR) [66] and Global Surface Water (GSW) Occurrence data [12], [67], are used. The CIFOR Global Wetland data set indicates the distribution of wetland, peatland and peat depth that covers the tropics and subtropics. This data set is created using products from the MODIS sensors, the phased array type L-band synthetic aperture radar (PALSAR) data, and other ancillary data sets [69]. Even though this data set is not validated due to the unavailability of ground truth, it agrees well with other commonly used data sets [69]. The GSW data set is generated based on optical images collected by Landsat [12]. The GSW Occurrence data shows the extent of permanent water from 1984 to 2019. Hereafter, we refer to the Global Wetland CIFOR and GSW Occurrence data sets as CIFOR and GSW, respectively. The key parameters of the employed datasets are listed in Table 1. Although their accuracies were not available, the CIFOR and CYGNSS data sets are benchmark data sets that have been widely adopted for analysis in literature.

In this work, various georeferenced data sets with different spatial resolutions are considered. Therefore, a comprehensive approach for matching the flood reference and ancillary data to each GNSS-R data point must be taken. Similar to other GNSS-R systems, the coherent footprint of CYGNSS is dynamic. In [45], it has been shown that DDMs can be gridded into cells of size 500 m × 500 m. Following the literature, in this study, we assume that the DDM of each SP represents a 500 m × 500 m region around it. Therefore, for assigning a flood/land label to an SP, the number of flood pixels of the reference flood map within an area of 500 m × 500 m around the SP is counted. When the percentage of flood pixels around the SP is higher than 75%, it is labeled as flood; otherwise, it is labeled as land. We investigated all the possible values for flood threshold and 75% is the optimum value. For SRTM90m DEM, assigned value to each SP is the average of the reference data within the area of 500 m × 500 m around each SP. Moreover, whether an SP is located within wetlands or permanent water bodies is determined by the value of a grid cell in CIFOR or GSW data sets that is closed to the SP.

III. METHODOLOGY

In this section, first, eleven different observables for CYGNSS data are discussed. Then, the RUSBoost algorithm and a brief description of the SVM algorithm are presented. The RUSBoost classification in this study follows the flow chart depicted in Figure 1. Lastly, the six DPAs that are investigated for flash flood detection are described.

A. CYGNSS OBSERVABLES

In GNSS-R technique, the received power is a combination of coherent and incoherent scattered components. When the surface is smooth, the reflection is mostly coherent. As the surface roughness increases, the reflected signal becomes more incoherent. Under stable and calm weather conditions, inland surface water bodies are smooth. Thus, the reflected signal from them is predominantly coherent. On the other hand, during a severe condition such as a hurricane, the high speed winds can increase the roughness of inland water bodies. However, the presence of high speed winds of a hurricane over land is shorter than the flash flood caused by its landfall. In addition, as a hurricane reaches the land its wind speed decreases gradually due to the higher roughness of land [70], [71]. Moreover, as investigated in [48], during severe typhoons (tropical storms in the Northwest Pacific Ocean) the coherent components of reflected signal are still consistent with flash floods. Therefore, following similar assumption in the literature [34], [45], [48], in this study we regard the reflected signals as coherent. The surface reflectivity (SR) DDM can be calculated by [45]

$$\Gamma = \frac{(R_{tx} + R_{rx})^2}{4\pi R_{tx}^2 R_{rx}^2} \langle \sigma \rangle. \tag{1}$$

where R_{tx} and R_{rx} are the distances between SP and transmitter and receiver, and $\langle \sigma \rangle$ is the calibrated incoherent bistatic radar cross section (BRCS) that is reported as DDM [72]. It is worth mentioning that another approximation for coherent reflected power from heterogeneous smooth surfaces is suggested by [73], in which the reflected signal is described by

TABLE 1. Summary of employed data sets.

Data Set	Resolution	Spatial Coverage	Temporal Coverage	Accuracy
CYGNSS [61]	incoherent: 25 km coherent: dynamic	38° S to 38° N	Daily from March 13, 2017	NA
DFO [8], [64]	< 250 m	50° S to 70° N	Flood Events 2000 - Present	Geolocation accuracy: ± 50 m
SRTM [65]	90 m	56° S to 60° N	11-22 February 2000	Vertical accuracy 6 m
CIFOR [66]	236 m	60° S to 40° N	2009-2017	NA
GSW [12], [67]	30 m	50° S to 80° N	1984-2019	98%

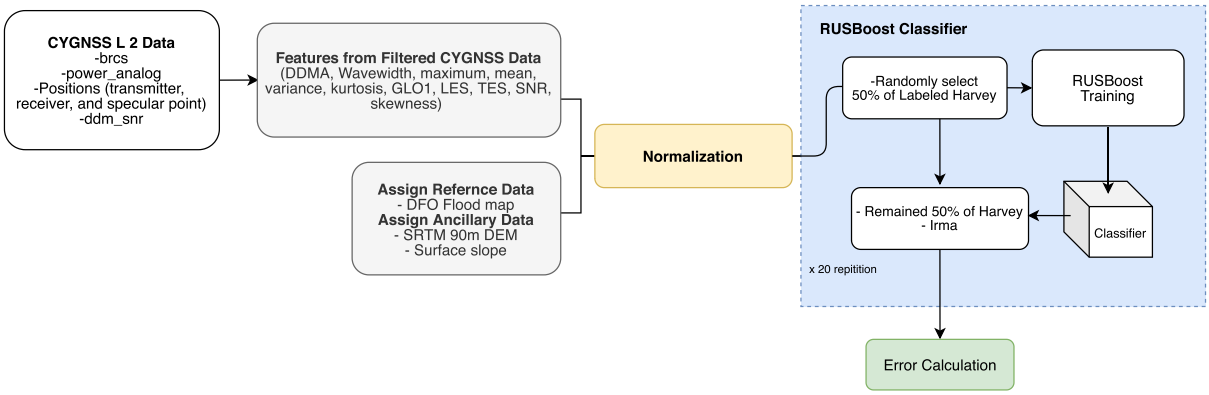


FIGURE 1. Block diagram of the classification.

surface diffraction integral. The surface diffraction integral is calculated over an area larger than the first Fresnel zone (FFZ) whose radius varies from 300 m to 800 m depending on the incidence angle [45]. However, our case studies consist of both rough and smooth surfaces. In addition, in this study, the coherent reflection comes from an area of 500 m $\times$ 500 m around each SP, which is within the range of FFZ. Therefore, here, (1) is considered.

In this study, instead of working with the whole DDM, eleven different observables including corrected signal to noise ratio ($\mathrm{SNR_C}$), trailing edge slope (TES), leading-edge slope (LES), delay-Doppler map average (DDMA), the width of the waveform (Wave-width), the first generalized linear observable ($\mathrm{GLO_1}$), kurtosis, maximum, mean, skewness, and variance are extracted for each SP. All the observables except $\mathrm{SNR_C}$ are computed using the SR DDM, which is calculated as described in [45]. The first seven observables ($\mathrm{SNR_C}$, LES, TES, DDMA, Wave-width, $\mathrm{GLO_1}$, and maximum) are obtained as follows [45], [74]

- Provided in the CYGNSS data set, $\mathrm{SNR_{Peack}}$ is the ratio between the maximum value in a DDM to its average noise per bin ($10\log(S_{max}/N_{avg})$). This value is then corrected to $\mathrm{SNR_C}$ [45]:

$$\mathrm{SNR_C} = \frac{(R_{tx}+R_{rx})^2\lambda^2\langle\sigma_m\rangle}{P_{rxm}R_{tx}^2R_{rx}^2(4\pi)^3}\mathrm{SNR_{Peack}} \tag{2}$$

where λ is the GPS wavelength that is 19.05 cm and $\langle\sigma_m\rangle$ is the maximum value of BRCS DDM and P_{rxm} is the maximum value in power DDM [45]. The maximum of BRCS and maximum of DDM are computed using the BRCS DDM and power DDM of each SP and vary with different DDMs.

- LES and TES are computed as the slopes between the maximum point and the points at two delay bins before and after the maximum point in the SR delay waveform (SR DDM integrated over Doppler axis) [45].
- The DDMA is the arithmetic mean of SR DDM within a window around the maximum value. In this study the size of window is chosen as 3 delay bins $\times$ 5 Doppler bins [75].
- The width of the waveform is the number of Doppler bins whose intensity is higher than $1/e$ of the maximum of the SR Doppler waveform (SR DDM integrated over the delay axis).
- The Nth generalized linear observable (GLO) is defined as [76]:

$$\mathrm{GLO}_N = \sum_{i=i_{max}-3}^{i_{max}+3} a_N(i)\Gamma_{del}(i), \tag{3}$$

where Γ_{del} is SR delay waveform, $a_N(i)$ is the Nth weight of SR in the delay bin i and it is computed by

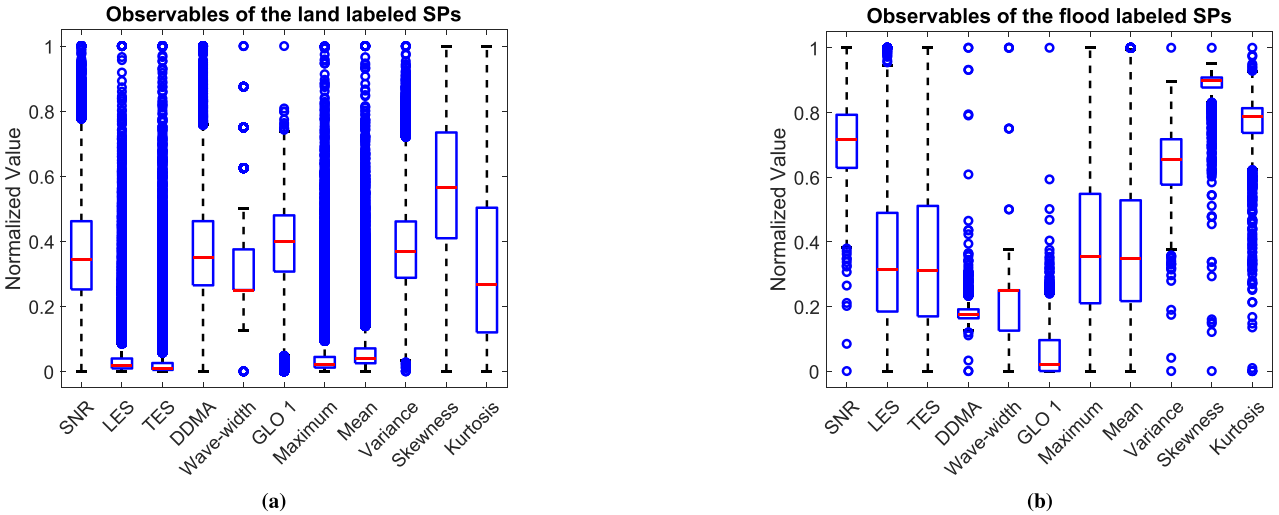


FIGURE 2. The box plots of the eleven observables (a): SPs labeled as land, and (b): SPs labeled as flood.

applying principal component analysis (PCA) to the SR delay waveform. The summation is calculated considering ± 3 delay bins around the delay bin of the maximum of SR delay waveform (i_{max}). We only consider the first GLO (GLO_1), since it has been proven that it is more correspondent to the inundation over land [45].

- The maximum (Maximum) that is the maximum value of the SR delay waveform is also considered as another observable.

A DDM represents the pattern of the scattered power from a surface. When the surface roughness changes, the scattered power and its pattern changes as well. These variations impact the statistical characteristics and histogram of the DDM. Moreover, a histogram can be described by statistical moments such as mean, variance, kurtosis, and skewness. Therefore, by considering the SR delay waveform as a random variables (RV) and analyzing its statistical moments, the impact of flood on a DDM can be studied [77], [78]. The statistical moments considered are described as

- Mean shows the position of the central mass of an RV;
- Variance is the squared differences of an RV from its mean. In other words, it measures how far values of an RV are located from the mean in a histogram;
- Skewness is an indicator of the asymmetry of the probability distribution of an RV. When the distribution is symmetrical, skewness equals to zero. However, when the distribution is skewed to the values higher or lower than mean the skewness is a negative or positive value, respectively;
- Kurtosis is a value that estimates the tailedness of the shape of a histogram by taking into account the outliers values.

More explanation about statistical moments and their mathematical formulas are given in [79]. It is worth mentioning that the number of observables is not confined. Other observables can be defined and computed based on different aspects of the GNSS-R data.

TABLE 2. Ranges of observables in normalization step.

Observable	Min	Max	Observable	Min	Max
DDMA	3	12	LES	0	0.35
TES	0	0.4	Wave-width	1	9
SNR_C (dB)	105	130	GLO_1	-35	-5
Kurtosis	1.5	4.5	Skewness	-1	1.8
Mean	0	0.1	Maximum	0	0.5
Variance(dB)	-70	-10			

Since the ranges of observables are different, as a part of the data cleaning step, they are normalized based on the normalization ranges mentioned in Table 2. The value of each parameter is projected to the interval of [0, 1] using its Min to Max. These values are obtained based on self-observations.

Depending on the labels of the SPs, their observables show different characteristics as depicted by the box plots in Figure 2, for which the SPs located over permanent water bodies and wetlands are excluded using the GSW and CIFOR data sets. Comparing Figure 2(a) with Figure 2(b) indicates that the values of SNR, LES, TES, mean, maximum, variance, skewness, and kurtosis of the flood labeled SPs are higher than those labeled as land. On the other hand, the flood labeled SPs have lower values in DDMA, Wave-width, and GLO_1.

The DDMs whose maximum is not between delay bins 4 and 14 are discarded as noise. The discarded DDMs include high altitude measurement and noisy DDMs. This range is determined by observing DDMs and comparing the delay bins of their maximum values. Moreover, when the incidence angle is between 15° and 60°, the reflected signal is more correlated with the water extent around an SP, as

shown in [80]. Since we intend to detect flash flood that is a type of surface water body, the SPs with incidence angles out of this range are removed. In addition to these conditions, quality flags, mentioned in [81], are also considered in the preprocessing step. It is worth mentioning that the speckle noise impact is negligible since each DDM is obtained from 1 s incoherent integration of 1000 DDMs [82].

B. CLASSIFICATION ALGORITHM

1) RUSBoost

Flood detection is a classification problem with two classes (land/flood). Since the DFO reference maps include flooded areas and only some/incomplete permanent water bodies but no wetland, both permanent water and wetland are not included in the training and testing. The class of each SP is determined by using the trained RUSBoost based classifier and its GNSS-R extracted observables and ancillary features. After selecting the features, all the observations in the Harvey data set, which is allocated for training and testing the classifier, are shuffled together. Then by a random selection, two separate equal sets for training and testing are generated. This unit that contains random shuffling and random selection is added to the RUSBoost classifier. For better perception, the pseudo-code of the RUSBoost classifier recreated from [83] is depicted in Figure 3. The training data set, is the imbalanced set $S = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, m\}$, in which $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,J}]$ is a vector in the J dimensional feature space and $y_i \in \{0, 1\}$ is its respective class label. In our case, $\mathbf{x}_i$ is a vector containing selected observables and y_i can be either land (0) or flood (1). At the first step, each point in S is assigned with an initial weight of $D_1(i) = 1/m$ prior to the first iteration (step 1).

Using the RUSBoost method, at iteration t, balanced temporary subset $S'_t = \{(\mathbf{x'}_p, y'_p) | \ p = 1, \ldots, 2n\} \subset S$ is created containing all the n points of minor class and n randomly selected points from major class. Knowing the indices of the selected data points from S that are members of S'_t, another temporary subset containing their corresponding weights $D'_t \subset D_t$ is obtained. These two temporary sets are then employed for training weak learners based on the idea of reducing the classification error iteratively (step 2a) [83].

When the data points in S'_t are passed to the tth decision stump, it divides them into two splits which in this work are referred to as right and left. Having a decision threshold for the feature j, $(j = 1, \ldots, J)$, the observation p in S'_t is positioned into the right or left split based on whether $x'_{p,j}$ is higher or lower than the value of decision threshold, respectively. Hence, the performance of decision stump depends on the feature and its decision threshold. Since all the employed features except Wave-width have continuous values, the decision threshold can take an infinite number of values. However, these thresholds do not necessarily result in different results. When $2n$ points of S' are sorted regarding their values of the same feature, between every two adjacent points, infinite thresholds can be considered. However, since

Algorithm RUSBoost

Given: Set $S = \{(\mathbf{x}_i, y_i) \mid i = 1, ..., m\}$ with feature vector $\mathbf{x}_i = [x_{i,1}, ..., x_{i,J}]$ and minority class $y = 1$, $y \in \{0, 1\}$

Weak learner, decision stump

Number of iterations, T

Desired percentage of total instances to be represented by the minority class, 50

1 Initialize $D_1(i) = \frac{1}{m}$.

2 Do for t = 1,2,...,T

a Create temporary training dataset S'_t with distribution D'_t using random undersampling.

b Call decision stump, providing it with examples S'_t and their weights D'_t.

i Select $x_{i,k} \in \mathbf{x}_i$ with decision threshold $c^k_t(q_k)$ that minimizes the Gini impurity factor.

ii Return $N_{r,l}(y)$ and $N_{r,l}$ regarding S' and $c^k_t(q_k)$

c Calculate the label proportion

$\pi_{r,l} = N_{r,l}(y)/N_{r,l}$.

d Create a weak hypothesis

$$h_t(\mathbf{x}_i, y) = \begin{cases} \pi_r(y) & \text{if } x_{i,k} > c^k_t(q_k), \\ \pi_l(y) & \text{otherwise.} \end{cases}$$

e Calculate the pseudo-loss (for S and D_t)

$\epsilon_t = \sum_{i=1}^{m} D_t(i)(1 - h_t(\mathbf{x}_i, y_i) + h_t(\mathbf{x}_i, 1 - y_i))$.

f Calculate the weight update parameter

$\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$.

g Update D_t

$D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(\mathbf{x}_i,y_i)-h_t(\mathbf{x}_i,1-y_i))}$.

h Normalize D_{t+1}

$D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_{i=1}^{m} D_{t+1}(i)}$.

3 Output the final hypothesis

$H(\mathbf{x}) = \underset{y\in\{0,1\}}{\text{argmax}} \sum_{t=1}^{T} h_t(\mathbf{x}, y) \log \frac{1}{\alpha_t}$.

FIGURE 3. The Pseudo-code of the RUSBoost algorithm recreated from [83].

they all have a similar result, only one of them should be considered. Therefore, instead of trying infinite numbers of thresholds, $2n - 1$ values between sorted points plus 0 and 1 are enough to be considered as the values of the decision threshold. Hence, $(2n + 1)J$ combinations of thresholds and features can be used for examining all the possible outputs. Considering the combination of jth feature and its qth decision threshold $c^j_t(q)$, $(q = 1, \ldots, 2n + 1)$, the weighted Gini impurity factor (GI_t) of the decision stump t is obtained as:

$$\text{GI}_t(q) = \Omega^r_t(q)\Theta^r_t(q) + \Omega^l_t(q)\Theta^l_t(q) \tag{4}$$

where $\Omega^{r,l}_t(q)$ is the probability of right or left split and $\Theta^{r,l}_t(q)$ is the Gini impurity factor of right or left split. For the right split $\Omega^r_t(q)$ and $\Theta^r_t(q)$ are defined as:

$$\Omega^r_t(q) = \sum_{p=1}^{2n} D'_t(p)[[x'_{p,j} > c^j_t(q)]] \tag{5}$$

$$\Theta^r_t(q) = 1 - \sum_{y} \theta^r_t(y), \tag{6}$$

where

$$\theta_t^r(y) = \left(\frac{\sum_{p=1}^{2n} D_t'(p)[[y_p' = y]][[x_{p,j}' > c_t^j(q)]]}{\sum_{p=1}^{2n} D_t'(p)[[x_{p,j}' > c_t^j(q)]]} \right)^2 \tag{7}$$

and $[[\cdot]]$ is a Boolean-valued function, with $[[\text{true}]] = 1$ and $[[\text{false}]] = 0$. Similarly, for the left split, $\Omega_t^l(q)$ and $\Theta_t^l(q)$ are computed by changing the condition $[[x_{p,j}' > c_t^j(q)]]$ in Equation (6) and (5) to $[[x_{p,j}' \leq c_t^j(q)]]$ [84].

Moreover, in boosting methods, the performance of a weak learner needs to be slightly better than the random guess (Gini impurity factor of 0.5) [85]. Hence, by randomly selecting a limited number of pairs of thresholds and features, the one that minimizes the Gini impurity factor is selected for creating the weak hypothesis. It should be mentioned that since S_t' is balanced, minimizing the Gini impurity factor translates to maximizing the Gini gain. We assume that among all features, $x_{i,k} \in \mathbf{x}_i$, with decision threshold $c_t^k(q_k)$, meets the requirements (step 2bi). From the feature, its decision threshold, and the number of points in each split regarding S' (step 2bii), the weak hypothesis is constructed as (step 2 d)

$$h_t(\mathbf{x}_i, y) = \begin{cases} \pi_r(y) & \text{if } x_{i,k} > c_t^k(q_k), \\ \pi_l(y) & \text{otherwise} \end{cases} \tag{8}$$

where $\pi_{r,l} = N_{r,l}(y)/N_{r,l}$ is the label proportion, which is the ratio between number of $y \in \{0, 1\}$ labeled points within a split $N_{r,l}(y)$, and its total number of points $N_{r,l}$ (step 2c). The pseudo loss of the weak hypothesis for all the points in S is calculated as (step 2e)

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)(1 - h_t(\mathbf{x}_i, y_i) + h_t(\mathbf{x}_i, 1 - y_i)), \tag{9}$$

where $1 - y_i$ is the incorrect label of observation i. A weights updating factor, α_t, is calculated as (step 2f)

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}. \tag{10}$$

Then, a new set of weights are computed and normalized as (step 2g-2h)

$$D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(\mathbf{x}_i,y_i)-h_t(\mathbf{x}_i,1-y_i))}, \tag{11}$$

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_{i=1}^{m} D_{t+1}(i)}. \tag{12}$$

When the hypothesis of the weak learner is correct for the training data set, which means that the weak learner was able to classify all of the training data points correctly, ϵ_t will be equal to zero, and the new weights will be equal to the previous ones. Otherwise, the weights of the misclassified points will be higher than the ones of correctly classified points. Therefore, in the next iteration, the weak learner will be biased to classify the misclassifications of the previous decision tree, which translates to increasing the variance step by step.

The procedure of random undersampling, creating a weak hypothesis, and updating observations weights is repeated for T iterations. At the last iteration, when the training of all of T weak learners is finished, the output hypothesis is created as a weighted vote of weak hypotheses (step 3):

$$H(\mathbf{x}) = \underset{y \in \{0,1\}}{\text{argmax}} \sum_{t=1}^{T} h_t(\mathbf{x}, y) \log \frac{1}{\alpha_t}, \tag{13}$$

where $\mathbf{x}$ is a feature vector of the test data. The criteria for the trained RUSBoost classifier is to find the label that maximizes the summation of the hypothesis of weak learners with respect to α_t [83], [86]. Since the number of weak learners affects the structure of the trained classifier and its performance, the number of weak learners (T) is the hyperparameter of our model. Also, the learning rate, which determines the step size at each iteration, is another important parameter of our model.

In this article, the RUSBoost-based classification is implemented in MATLAB R2018 using the Statistics and Machine Learning Toolbox. A total number of 150 of weak learners are trained with a learning rate of 0.1. We investigated different combinations of the number of weak learners and learning rate values in terms of classification error and the selected combination gives the minimum error. Each weak learner is a decision stump. At each iteration, among 150 random combinations of different features and decision thresholds, one of them is chosen. The trained classifier is used for testing and evaluation.

2) SVM

In this article, the SVM classifier is considered for comparison with the proposed method. As a well-known supervised ML algorithm, SVM has been used in various remote sensing applications [38], [87]–[89]. SVMs classify data by determining the optimal hyperplane for maximizing the margin between classes [90], [91]. For nonlinear data, computing the hyperplane is achieved by using the kernel trick, which maps the data in a higher dimensional space. More details on the SVM ML algorithm can be found in [90], [91].

In this article, an SVM based classifier is implemented using the Statistics and Machine Learning Toolbox of MATLAB R2018. As in Section III-B1, selecting training data points from Harvey consists of random shuffling and random selection. For balancing the imbalanced data sets, RUS is applied to the training data set since it requires a much lower computational load compared to oversampling methods (e.g., SMOTE) [57], [58]. The radial basis function (RBF) kernel is selected as the kernel function. The values of hyperparameters are optimized using the sequential minimal optimization (SMO) algorithm proposed in [92].

Since the selected training data from Harvey is random, for having a better perception of the performance, the classification was repeated 20 times for both SVM and RUSBoost

classifiers as shown in the block diagram of the RUSBoost classifier depicted in Figure 1.

C. DATA PREPARING APPROACHES

In this section, six different DPAs for flash flood detection are described. As mentioned in Section II-B, water bodies that are not caused by flash floods, e.g., permanent water bodies and some regions of wetlands, can be mislabeled as flood. Two main DPAs could be taken for solving this issue. One solution is to use reference data sets and exclude SPs that are located over water bodies. Another one is to use the variation between the CYGNSS data during flood and the CYGNSS data collected during a period that flood did not happen. Therefore, six different DPAs are investigated in this study that are described as

- In Approach 1, all inland SPs collected during floods are used. Even though some SPs are located over wetlands or permanent waters, in order to investigate the errors caused by water bodies other than flood, the non-flooded SPs are labeled as land.
- In Approach 2, based on the GSW and CIFOR data sets, SPs located over wetlands and water bodies are excluded. This method was previously used in Section IV-A for feature selection.
- In Approach 3 GSW data set is used for excluding the SPs located over permanent waters.
- Approach 4 consists of three steps: detecting water bodies, excluding the SPs associated with water detection results, and flood detection. Using the 2018 CYGNSS data and inland water detection method described in [78], water bodies over Harvey and Irma are detected. The detected water extent is then used as a reference for corresponding excluding SPs.
- In Approach 5, the impact of flood is investigated by considering the changes caused by flood with respect to the CYGNSS data collected one month prior to flood.
- Similar to Approach 5, in Approach 6, the variations caused by flood are considered. In this DPA, the CYGNSS data of three months dry season of the year 2018 are considered as background data.

In Approach 5 and Approach 6, for calculating the changes of selected observables, each SP in the CYGNSS flood data set is matched with the closest data point from the background data set. The distance between SP in the flood data set and its match from the background data set is to be less than 1.5km. When the distance between two points is higher than 1.5km, the SP is excluded. We investigated different values for determining this distance and 1.5km was the optimum value with respect to data exclusion amount and classification error. Since in Approach 1 all data points collected during floods are used for classification, its result includes possible misclassifications. Comparing the classification results of other DPAs with Approach 1 can indicate the advantages and disadvantages of them. The coverage of the CYGNSS is low and in some DPAs, a portion of data is not even considered due to the data exclusion. Hence, the percentage of excluded data points alongside the accuracy of the classifier are two factors that are used in Section IV-B to evaluate the overall performance of each DPA. Since the amount of excluded data for each DPA is different, it is not possible to evaluate them using exactly same validation data.

TABLE 3. Accuracies for five best combinations with and without ancillary data obtained from 5-fold cross-validated classification with 150 weak learners.

#*	Inputs	Class	Accuracy
1	All features	Land	95.69 %
		Flood	79.20%
2	SNR_C, TES, Wave-width, DDMA	Land	94.14 %
	GLO_1, Kurtosis, SRTM90m DEM	Flood	81.19%
3	SNR_C, LES, Wave-width,	Land	93.99 %
	GLO_1, SRTM90m DEM	Flood	80.85%
4	Kurtosis, Maximum, Variance	Land	**96.08** %
	Mean, Skewness, SRTM90m DEM	Flood	79.41%
5	Kurtosis, Maximum, Variance	Land	95.73%
	DDMA, Wave-width, SRTM90m DEM	Flood	**83.32**%
6	All observables	Land	92.91 %
	(All features except SRTM90m DEM)	Flood	69.54%
7	SNR_C, TES, Wave-width,	Land	92.63 %
	DDMA,GLO_1, Kurtosis	Flood	71.04%
8	SNR_C, LES, Wave-width,	Land	92.71 %
	GLO_1	Flood	70.07 %
9	kurtosis, Maximum, Variance,	Land	93.63 %
	Mean, Skewness	Flood	69.48 %
10	Kurtosis, Maximum, Variance	Land	93.57 %
	DDMA, Wave-width	Flood	72.16%

* Combinations from 1 to 5 are with ancillary data (SRTM90m DEM), and combinations from 6 to 10 are without it.

IV. RESULTS AND DISCUSSION

In Section IV-A, all the combinations of eleven observables and two features from SRTM90m DEM are separately used as inputs of a RUSBoost classifier with 5-fold cross-validation to select a suitable combination of features. With the selected features, the detecting flood performances associated with the six DPAs are evaluated in Section IV-B. By comparing their results, the best method for flood detection is selected. The performance of the recommended RUSBoost classifier is then compared with that of an SVM based classifier.

A. FEATURES SELECTION

In this section, we want to select the features that are proper for flood detection. Therefore, by using GSW and CIFOR data sets, SPs located over wetlands and permanent water are excluded. This ensures that the remaining SPs used in feature selection are either flood or bare land.

The eleven observables described in Section III-A and the surface elevation and terrain from the SRTM90 DEM data set are considered as thirteen features. In this section, both

TABLE 4. Accuracies of Approach 1 to Approach 6.

#	Event	Class	Accuracy	Excluded Data	Event	Class	Accuracy	Excluded Data
Approach 1	Harvey	Land	94.00 %	0.0 %	Irma	Land	63.00 %	0.0 %
		Flood	90.90 %	0.0 %		Flood	87.30%	0.0 %
Approach 2	Harvey	Land	97.68 %	22.0 %	Irma	Land	79.20 %	48.0%
		Flood	81.60 %	38.0 %		Flood	87.50%	55.0 %
Approach 3	Harvey	Land	97.20%	5.0 %	Irma	Land	71.00 %	14.0 %
		Flood	89.00 %	0.0 %		Flood	85.00%	0.0 %
Approach 4	Harvey	Land	98.30 %	29.50 %	Irma	Land	79.50 %	57.1 %
		Flood	76.00 %	80.1 %		Flood	83.00%	70.5 %
Approach 5	Harvey	Land	95.50%	48.0 %	Irma	Land	86.00%	44.0 %
		Flood	86.80%	45.0 %		Flood	45.00%	47.0 %
Approach 6	Harvey	Land	91.50 %	0.0 %	Irma	Land	61.00%	0.0 %
		Flood	86.90%	0.0 %		Flood	78.20%	0.0 %

Harvey and Irma data sets are combined and the idea of recursive feature elimination is implemented to determine a suitable feature combination out of all the 8191 different combinations. For each combination, the accuracy of the classifier with 150 weak learners is evaluated by 5-fold cross-validation in which same subsets of data are used for all combinations. The accuracies of the best five combinations with and without the two features from SRTM90 DEM are shown in Table 3. From Table 3 it is clear that using the elevation and terrain from SRTM90 DEM (combinations 1 to 5 in Table 3) has improved the accuracy (around 10% for flood and 2.5% for land). Due to the gravity, water accumulates in lower altitudes. Hence, it is unlikely that flood would occur in an area located on the slope. Overall, knowing the elevation and terrain of an area gives a better insight into the regions with a higher possibility of flooding.

In terms of land detection, features combination 4 in Table 3 has the best accuracy. However, it has a lower flood detection accuracy compared to features combination 5, which has the best performance for flood detection. Therefore, features combination 5 in Table 3 has a better performance overall.

Here, the hyperparameter (i.e., the number of weak learners (T)) of the RUSBoost classifier is set to 150 after evaluating the classification error for various values of T by considering all the features as the input feature vector. Next, the top five combinations in terms of classification error are found based on the selected hyperparameter. Since the value of T can impact the classification results, we further investigated the variation of the overall classification error with T for the other four feature combinations listed in Table 3. As shown in Figure 4, as the number of weak learners increases, the classification errors decrease. After a certain value of T (140 here), the accuracy will not change significantly. The optimal values of T are 150 for Combinations 1, 3, and 5 and 149 for Combinations 2 and 4. Although there is a small difference between the optimal and selected values for Combinations 2 and 4, the results obtained from $T = 150$ are still appropriate since the difference between the accuracies with the selected and optimal hyperparameters is less than 0.1%.

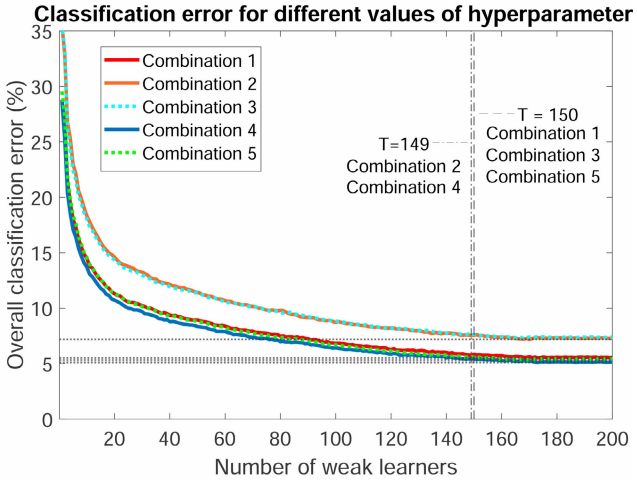


FIGURE 4. The classification error of the classifier with 5-fold cross-validation with respect to the number of weak learners (i.e. the value of the hyperparameter, T).

B. FLOOD DETECTION

By knowing the best feature vector from Section IV-A, in this section we intend to find the best DPA for flood detection using the CYGNSS data through evaluating the six DPAs mentioned in Section III-C.

Unlike Section IV-A where a classifier with 5-fold cross-validation was used, here, the RUSBoost based classifier depicted in Figure 1 is trained and tested by using the features combination 5 in Table 3 for each DPA.

The results shown in Table 4 are the accuracies and the percentage of excluded data for both Harvey and Irma. For choosing the best method, first, the percentage of the

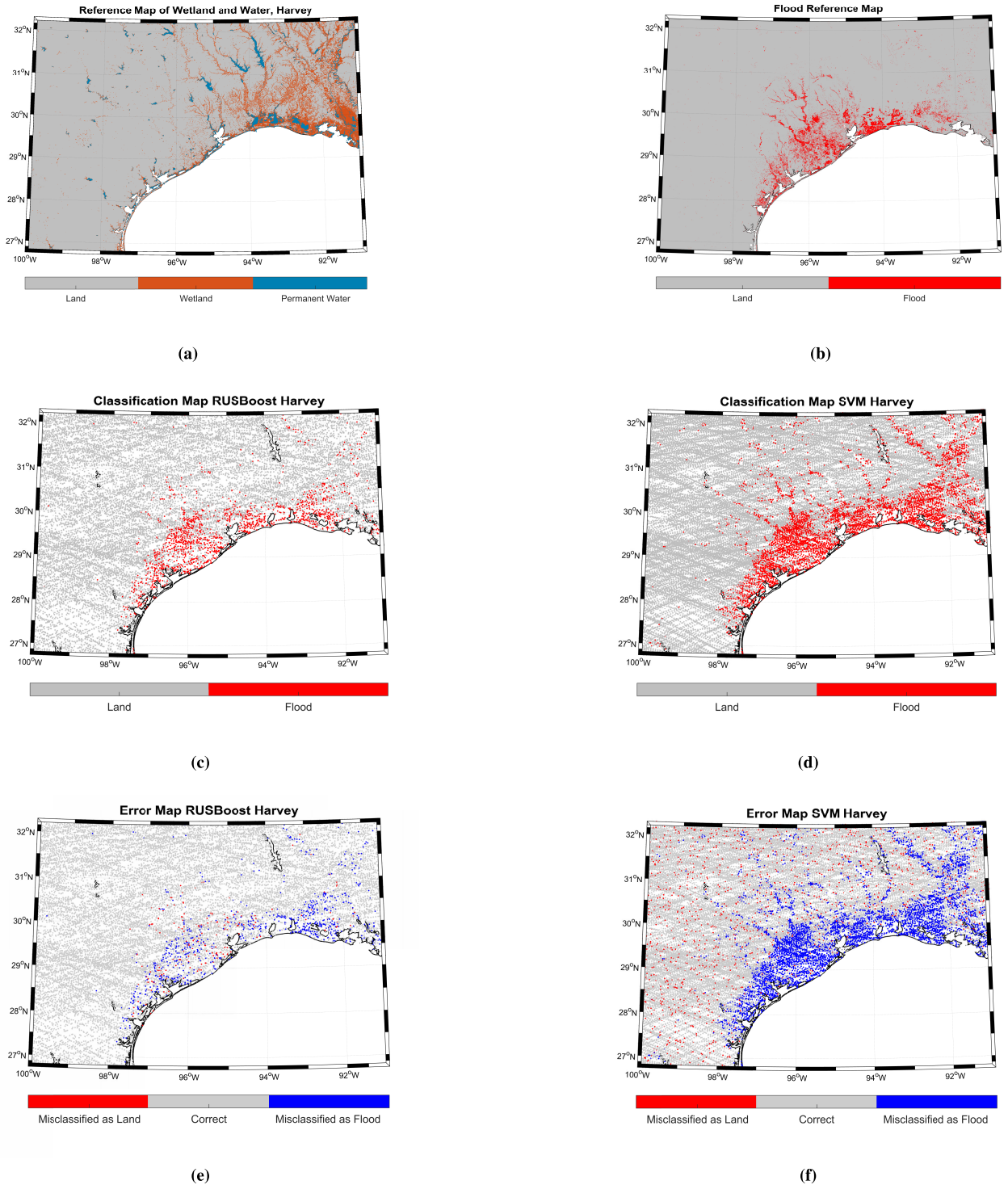


FIGURE 5. Maps of Harvey (a): GSW water, and CIFOR reference maps, (b): DFO flood reference map, (c): RUSBoost-based classification result map, (d): SVM-based classification result map, (e): RUSBoost-based classification error map, and (f): SVM-based classification error map. In (c)-(f) the results are obtained via data exclusion of Approach 3.

discarded data of each DPA is compared with other DPAs. Then, suitable method for flood detection is selected based on the accuracy. Among all the DPAs, Approach 1 and Approach 6 have no data exclusion. The excluded data points in Approach 3 are located over permanent water. This is reasonable since permanent water area does not need to be determined to be flooded or not, i.e., there is no overlap between permanent water and the reference flooding regions. Due to the overestimation of detected water extent, Approach 4 has the highest data exclusion. Even though Approach 2,

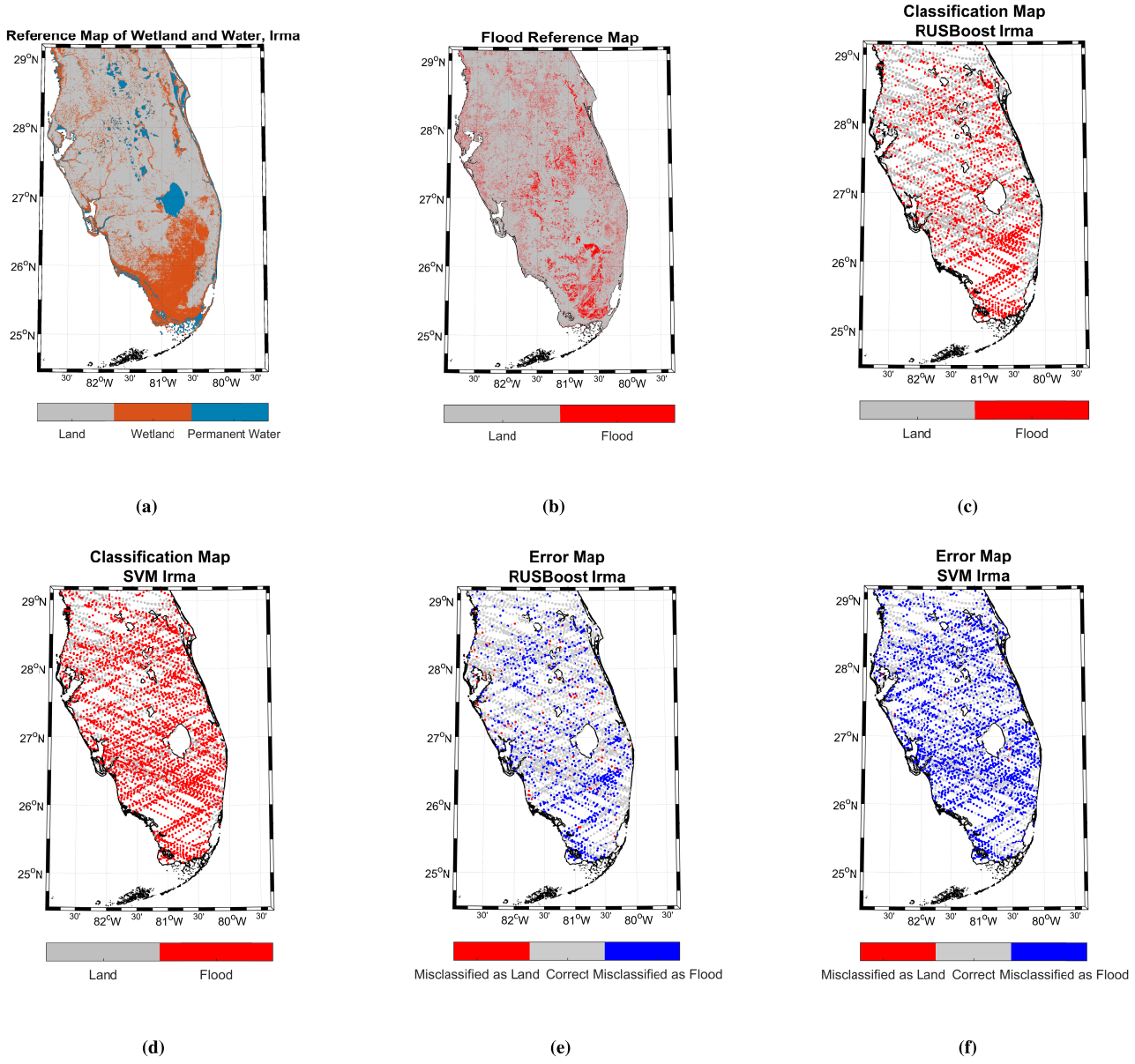


FIGURE 6. Maps of Irma (a): GSW water, and CIFOR reference maps, (b): DFO flood reference map, (c): RUSBoost-based classification result map, (d): SVM-based classification result map, (e): RUSBoost-based classification error map, and (f): SVM-based classification error map. In (c)-(f) the results are obtained via data exclusion of Approach 3.

Approach 4, and Approach 5 have an acceptable accuracy in some occasions, they do not seem to be proper options for flash flood detection due to their high percentage of data exclusion. It should be pointed out that unlike Section IV-A, where we intended to find a suitable feature vector for detecting flood, here, we are investigating different DPAs for flash flood detection. Moreover, as shown in Table 4, the accuracies of Approach 2 and Approach 3 are comparable, but Approach 2 is not suggested since more data points are excluded. Approach 5 has the lowest flood detection accuracy for Irma. Among Approach 1, Approach 3, and Approach 6, Approach 3 has the highest land detection accuracy and Approach 1 has the highest flood detection accuracy for both Harvey and Irma. The flood and land detection accuracies of Approach 6 are less than those of Approach 1 and Approach 3. Therefore, Approach 1 and Approach 3 are the final candidates. In terms of flood detection, Approach 1 outperforms Approach 3 by 1.9% in Harvey and 2.3% in Irma. However, Approach 3 is able to detect land with a higher accuracy (3.2% in Harvey and 8% in Irma). The intention in this study is to detect flash flood. However, since the land points outnumber the flood points, the overestimation of flood is also crucial. Hence, Approach 3 seems like the proper method for flash flood detection.

The maps of employed reference data sets, classification result and error maps of Harvey and Irma are depicted

in Figures 5 and 6, respectively. The GSW and CIFOR references shown in Figures 5(a) and 6(a) are used for data exclusion in Approach 2. For Approach 3, discarded SPs are selected by using the GSW reference data that is depicted in Figures 5(a) and 6(a) in blue colour. Based on the high-resolution DFO flood reference maps depicted in Figures 5(b) and 6(b), in each DPA, considered SPs are labeled as flood/land. Due to data exclusion, the flood reference map of each DPA could be different from others. As shown in Figures 5(b) and 6(b) the area flooded by Hurricane Harvey is concentrated over the coastline, and most of the inland areas were not impacted. While in Hurricane Irma, affected areas are scattered over the land. Furthermore, in order to make the flash flood detection method independent of other data sets such as the GSW and CIFOR, in Approach 4, we attempt to detect the water extent over Harvey and Irma and use the results as a water extent reference for excluding data. Considering the CYGNSS data of the year 2018, three observables, including Kurtosis, Maximum, and Variance, are extracted. Using these observables, a RUSBoost classifier trained with data from the Congo basin is used for detecting water bodies of Harvey and Irma [78]. The water detection method overestimates the presence of water bodies, which leads to excluding a large portion of data. The two investigated regions consist of various dynamic water bodies, to which the CYGNSS is sensitive, including wetlands, permanent waters, and farmlands. Therefore, water overestimation is inevitable. Comparing the classification results of Approach 3 and flood reference maps depicted in Figures 5(c) and 6(c) and Figures 5(b) and 6(b), respectively, shows that even with small coverage, Approach 3 is capable of identifying the flash flood extent.

C. COMPARISON TO SVM CLASSIFIER

For comparison, an SVM-based classifier is trained using the selected features in Section IV-A and same data that is produced by Approach 3 and employed for building the RUSBoost classifier. As mentioned in Section III-B2, the parameters of the SVM-based classifier are optimized using the SMO algorithm.

As shown in Table 5, compared to the RUSBoost classifier with Approach 3, the SVM classifier can detect flash floods with an accuracy of 6.1% and 11.3% higher for Harvey and Irma, respectively. However, in terms of land detection, the RUSBoost-based classifier is 8.98% and 32.2% more accurate for Harvey and Irma, respectively. The SVM classifier overestimates flash floods as depicted in Figures 5(d), 5(f), 6(d) and 6(f). For the SVM based classifier, due to the disproportion of imbalanced data sets, the number of misclassified land points is much higher than correctly detected flood points. Therefore, the RUSBoost classifier with Approach 3 is better than the SVM classifier.

It is worth mentioning that the overall run-time of the RUSBoost and the SVM classifiers are 12.10 s and 0.48 s, respectively.

TABLE 5. Results of SVM and RUSBoost classifiers.

Event	Classifier	Class	Accuracy
Harvey	RUSBoost	Land	97.20 %
		Flood	89.00 %
Harvey	SVM	Land	88.22%
		Flood	95.10%
Irma	RUSBoost	Land	71.00 %
		Flood	85.00%
Irma	SVM	Land	38.80 %
		Flood	96.30%

V. CONCLUSION

In this article, a flood detection method based on CYGNSS data has been conducted using the RUSBoost based classification. Eleven different features have been extracted from the CYGNSS data, and each point is labeled as land/flood, as discussed in Section II-B. For feature selection, by excluding wetland and permanent water, the CYGNSS flood data set only includes SPs that are either flood or land. Using this data, after investigating the accuracies of all the combinations of thirteen features via a classifier with 5-fold cross-validation, the most effective features were selected. Even though the accuracy of various combinations is roughly in the same range, the combination of Kurtosis, DDMA, Maximum, Variance, Wave-width, and SRTM90m DEM has the best performance overall. The selected feature combination might not be the best one, but the classification results indicate that it is a suitable option for flash flood detection.

By using the selected features combination, six different DPAs for detecting flash flood were investigated, among which Approach 3 gives the best performance in terms of accuracy and data exclusion. Moreover, the comparison between the RUSBosst-based and SVM-based classifiers, which were trained using data exclusion in Approach 3, indicates that the RUSBosst-based classifier has a better overall performance. Therefore, it is recommended as the method for flood detection using CYGNSS data.

The GSW and SRTM90m DEM data sets are the only ancillary data sets employed in the recommended DPA, i.e., Approach 3. Both of them are available for the public with global coverage. In addition, unlike GLO_1, the selected observables do not require any heavy preprocessing, and for each SP, they are computed based on provided parameters in the CYGNSS data for that SP without any dependency on a region or time period. As mentioned in Section II-A, all classifiers in Section IV-B were trained with half of the Harvey data, which was randomly selected. Then the trained classifiers were tested with the remaining half of the Harvey data and all of the data of Irma.

The CYGNSS data set involves non-geophysical uncertainties. The effective isotropic radiated power (EIRP) of GPS transmitters that is used in the CYGNSS data process

is not subtle [93]. Due to the different designs of space vehicles and the transmitting antenna panel, the EIRP of GPS transmitters fluctuates that leads to inaccuracy of CYGNSS measurements and impacts the results of this study. Since August 2018, by monitoring the transmitted power of GPS satellites, the fluctuations are compensated [94], [95]. However, due to the limitation of available CYGNSS data that is associated with significant flash flood, we selected the two representative events that happened in 2017.

The main drawback of the proposed method is flood overestimation with respect to the DFO reference data. This problem was also reported in [96], where data from Soil Moisture Active Passive (SMAP) was employed for flood detection. This may be because both CYGNSS and SMAP use L-band signals which are sensitive to SM. Flash flood is a complicated matter, and it depends on various conditions. In addition to the massive surge of water, various factors such as soil moisture, soil type, vegetation, subsurface flows, elevation, etc. can impact the development of flood [49], [97], [98]. Moreover, the scattering from the surface at L-band primarily depends on two factors: roughness and soil moisture, as investigated in [99]. Due to heavy precipitation during a flash flood, the SM increases till the soil becomes saturated. This increase in SM can be an explanation for this problem since the reflected signal from an SP with high SM can be as coherent as the reflection from a flooded SP, which causes flood overestimation. The low accuracy of flood over Irma in Approach 5 shows the impact of SM on flood detection. Both Hurricane Harvey and Hurricane Irma occurred during the high season (July to November) [100], [101] and during the month prior to flood, several precipitations happened in those areas especially for Irma [102]. Since having high SM does not necessarily indicate that an SP is flooded [103], SM is not the only source of error. Another parameter that also has a major role in the reflected signal is the roughness [104]. The coherent reflection from a smooth surface can lead to flood overestimation. Therefore, the regions that are relatively flat with high SM, such as Irma, are overestimated by the proposed method. Moreover, as mentioned before, the high-speed wind of a hurricane would increase the surface roughness of water in flooded regions. As the surface becomes rougher, its root-mean-square-height increases. Consequently, surface reflectivity decreases exponentially [105]. In other words, the incoherent components become more dominant. Therefore, in the early stages of our case studies where a high-speed wind is present, there are flood points whose scattered power is predominantly incoherent, which leads to flood underestimations. Furthermore, some flooded areas are heterogeneous, meaning that there is a diversity of land and flood in them. The heterogeneity can impact the scattering pattern. which results in overestimation or underestimation of flood. Despite the overestimation and underestimation, based on the obtained results, the proposed method is able to detect a flash flood with high accuracy. It is worth mentioning that similar to other microwave systems [106], [107], the turbidity of the water does not significantly impact the scattering of the GNSS signals from the water bodies since the signals cannot penetrate into the water too much. Thus, the turbidity of water may not be a major source of error in our work.

Compared to optical satellites, similar to other spaceborne microwave systems, CYGNSS is not affected by clouds, which makes it a reliable source for monitoring flash floods. Compared to other remote sensing satellites such as SAR and optical, the GNSS-R technique has a lower quality in terms of spatial resolution and accuracy [108]. On the other hand, the revisit time of the CYGNSS satellites is shorter than SAR systems, hence it is able to detect flash floods. Moreover, due to the less expensive receiver of GNSS-R, larger constellations can be obtained that leads to better coverage.

The proposed method has two main limitations. Firstly, it cannot detect urban flash floods since the impacted regions include various human-made obstacles causing incoherent reflection. Moreover, due to the gap between CYGNSS constellation tracks, it is unlikely to have enough collected data for flash flood monitoring when a small flash flood happens. Hence, the proposed method is more suitable for observing extensive flash floods. However, this limitation can be solved by having more GNSS-R receivers.

The parameters, such as high-speed winds during hurricanes and soil type, were not included in this study. These parameters should be considered in future studies. Moreover, in this study, the proposed method was only evaluated over two case studies and it was compared with an SVM-based classifier. In addition, feature selection was based on the optimal hyperparameter found for one combination rather than the corresponding optimal value of each combination. In the future, more advanced feature selection methods could be investigated. Also, other oversampling (e.g. GAN, Variational Autoencoder (VAE), and SMOTE) and undersampling methods developed for tackling imbalanced data along with other ML algorithms such as the random forest, Extreme Gradient Boosting (XGBoost) may be investigated for flash flood detection from the CYGNSS GNSS-R data.

REFERENCES

[1] K. Sene, *Flash Floods—Forecasting Warning*. Amsterdam, The Netherlands: Springer, 2013.

[2] S. L. Cutter, C. T. Emrich, M. Gall, and R. Reeves, "Flash flood risk and the paradox of urban development," *Natural Hazards Rev.*, vol. 19, no. 1, pp. 05017005-1–05017005-12, 2018.

[3] P. J. Fitzpatrick, *Hurricanes: A Reference Handbook*, 2nd ed. Santa Barbara, CA, USA: ABC-CLIO, 2006.

[4] J. Weinkle, R. Maue, and R. Pielke, Jr., "Historical global tropical cyclone landfalls," *J. Climate*, vol. 25, no. 13, pp. 4729–4735, 2012.

[5] E. Fussell, S. R. Curran, M. D. Dunbar, M. A. Babb, L. Thompson, and J. Meijer-Irons, "Weather-related hazards and population change: A study of hurricanes and tropical storms in the united states, 1980–2012," *Ann. Amer. Acad. Political Social Sci.*, vol. 669, no. 1, pp. 146–167, Jan. 2017.

[6] V. Tsyganskaya, S. Martinis, P. Marzahn, and R. Ludwig, "SAR-based detection of flooded vegetation—A review of characteristics and approaches," *Int. J. Remote Sens.*, vol. 39, no. 8, pp. 2255–2293, Apr. 2018.

[7] D. C. Mason, I. J. Davenport, J. C. Neal, G. J.-P. Schumann, and P. D. Bates, "Near real-time flood detection in urban and rural areas using high-resolution synthetic aperture radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3041–3052, Aug. 2012.

[8] J. Nigro, D. Slayback, F. Policelli, and G. R. Brakenridge. (2014). *NASA/DFO MODIS Near Real-Time (NRT) Global Flood Mapping Product Evaluation of Flood and Permanent Water Detection*. Accessed: Jun. 25, 2020. [Online]. Available: https://floodmap.modaps.eosdis.nasa.gov//documents/NASAGlobalNRTEvaluat%ionSummary_v4.pdf

[9] Z. Kugler and T. De Groeve, *The Global Flood Detection System*, vol. 45. Luxembourg City, Luxembourg: Office for Official Publications of the European Communities, 2007.

[10] Z. Musa, I. Popescu, and A. Mynett, "A review of applications of satellite SAR, optical, altimetry and DEM data for surface water modelling, mapping and parameter estimation," *Hydrol. Earth Syst. Sci.*, vol. 19, no. 9, p. 3755, 2015.

[11] X. Tong, X. Luo, S. Liu, H. Xie, W. Chao, S. Liu, S. Liu, A. N. Makhinov, A. F. Makhinova, and Y. Jiang, "An approach for flood monitoring by the combined use of landsat 8 optical imagery and COSMO-SkyMed radar imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 136, pp. 144–153, Feb. 2018.

[12] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, Dec. 2016.

[13] X. Shen, D. Wang, K. Mao, E. Anagnostou, and Y. Hong, "Inundation extent mapping by synthetic aperture radar: A review," *Remote Sens.*, vol. 11, no. 7, p. 879, Apr. 2019.

[14] L. Pulvirenti, M. Chini, N. Pierdicca, L. Guerriero, and P. Ferrazzoli, "Flood monitoring using multi-temporal COSMO-SkyMed data: Image segmentation and signature interpretation," *Remote Sens. Environ.*, vol. 115, no. 4, pp. 990–1002, Apr. 2011.

[15] S. Martinis and C. Rieke, "Backscatter analysis using multi-temporal and multi-frequency SAR data in the context of flood mapping at river saale, germany," *Remote Sens.*, vol. 7, no. 6, pp. 7732–7752, Jun. 2015.

[16] L. Landuyt, A. Van Wesemael, G. J.-P. Schumann, R. Hostache, N. E. C. Verhoest, and F. M. B. Van Coillie, "Flood mapping based on synthetic aperture radar: An assessment of established approaches," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 722–739, Feb. 2019.

[17] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A hierarchical split-based approach for parametric thresholding of SAR images: Flood inundation as a test case," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6975–6988, Dec. 2017.

[18] M. Chini, R. Pelich, L. Pulvirenti, N. Pierdicca, R. Hostache, and P. Matgen, "Sentinel-1 InSAR coherence to detect floodwater in urban areas: Houston and hurricane harvey as a test case," *Remote Sens.*, vol. 11, no. 2, p. 107, Jan. 2019.

[19] K. Uddin, M. A. Matin, and F. J. Meyer, "Operational flood mapping using multi-temporal sentinel-1 SAR images: A case study from bangladesh," *Remote Sens.*, vol. 11, no. 13, p. 1581, Jul. 2019.

[20] J. Cohen, H. Riihimäki, J. Pulliainen, J. Lemmetyinen, and J. Heilimo, "Implications of boreal forest stand characteristics for X-band SAR flood mapping accuracy," *Remote Sens. Environ.*, vol. 186, pp. 47–63, Dec. 2016.

[21] S. Grimaldi, J. Xu, Y. Li, V. R. N. Pauwels, and J. P. Walker, "Flood mapping under vegetation using single SAR acquisitions," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111582.

[22] G. P. Petropoulos and T. Islam, *Remote Sensing of Hydrometeorological Hazards*. Boca Raton, FL, USA: CRC Press, 2018.

[23] G. R. Brakenridge, "Flood risk mapping from orbital remote sensing," in *Global Flood Hazard: Applications in Modeling, Mapping, and Forecasting*, G. J. Schumann, P. D. Bates, H. Apel, and G. T. Aronica, Eds. Hoboken, NJ, USA: Wiley, 2018, ch. 3, pp. 43–54.

[24] S. Jin, E. Cardellach, and F. Xie, *GNSS Remote Sensing*. Hoboken, NJ, USA: Springer, 2014.

[25] V. U. Zavorotny, S. Gleason, E. Cardellach, and A. Camps, "Tutorial on remote sensing using GNSS bistatic radar of opportunity," *IEEE Geosci. Remote Sens. Mag.*, vol. 2, no. 4, pp. 8–45, Dec. 2014.

[26] W. Li, E. Cardellach, F. Fabra, A. Rius, S. Ribó, and M. Martín-Neira, "First spaceborne phase altimetry over sea ice using TechDemoSat-1 GNSS-R signals," *Geophys. Res. Lett.*, vol. 44, no. 16, pp. 8369–8376, Aug. 2017.

[27] M. P. Clarizia and C. S. Ruf, "Wind speed retrieval algorithm for the cyclone global navigation satellite system (CYGNSS) mission," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4419–4432, Aug. 2016.

[28] N. Rodriguez-Alvarez, D. M. Akos, V. U. Zavorotny, J. A. Smith, A. Camps, and C. W. Fairall, "Airborne GNSS-R wind retrievals using delay-Doppler maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 626–641, Jan. 2013.

[29] E. Valencia, V. U. Zavorotny, D. M. Akos, and A. Camps, "Using DDM asymmetry metrics for wind direction retrieval from GPS ocean-scattered signals in airborne experiments," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3924–3936, Jul. 2014.

[30] M. P. Clarizia, C. P. Gommenginger, S. T. Gleason, M. A. Srokosz, C. Galdi, and M. Di Bisceglie, "Analysis of GNSS-R delay-Doppler maps from the UK-DMC satellite over the ocean," *Geophys. Res. Lett.*, vol. 36, no. 2, pp. 3924–3936, 2009.

[31] C. Li and W. Huang, "An algorithm for sea-surface wind field retrieval from GNSS-R delay-Doppler map," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2110–2114, Dec. 2014.

[32] A. Camps, H. Park, M. Pablos, G. Foti, C. Gommenginger, P.-W. Liu, and J. Judge, "Sensitivity of GNSS-R spaceborne observations to soil moisture and vegetation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4730–4742, Jul. 2016.

[33] Y. Jia, P. Savi, D. Canone, and R. Notarpietro, "Estimation of surface characteristics using GNSS LH-reflected signals: Land versus water," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4752–4758, Jul. 2016.

[34] O. Eroglu, M. Kurum, D. Boyd, and A. C. Gurbuz, "High spatio-temporal resolution CYGNSS soil moisture estimates using artificial neural networks," *Remote Sens.*, vol. 11, no. 19, p. 2272, Sep. 2019.

[35] J. W. Cheong, B. J. Southwell, and A. G. Dempster, "Blind sea clutter suppression for spaceborne GNSS-R target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5373–5378, Dec. 2019.

[36] Q. Yan and W. Huang, "GNSS-R delay-Doppler map simulation based on the 2004 Sumatra-Andaman tsunami event," *J. Sens.*, vol. 2016, Dec. 2016, Art. no. 2750862.

[37] Q. Yan and W. Huang, "Tsunami detection and parameter estimation from GNSS-R delay-Doppler map," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4650–4659, Oct. 2016.

[38] Q. Yan and W. Huang, "Detecting sea ice from TechDemoSat-1 data using support vector machines with feature selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1409–1416, May 2019.

[39] D. Schiavulli, F. Frappart, G. Ramillien, J. Darrozes, F. Nunziata, and M. Migliaccio, "Observing sea/ice transition using radar images generatedfrom TechDemoSat-1 delay Doppler maps," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 734–738, Mar. 2017.

[40] Q. Yan and W. Huang, "Spaceborne GNSS-R sea ice detection using delay-Doppler maps: First results from the UK TechDemoSat-1 mission," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4795–4801, Oct. 2016.

[41] N. Rodriguez-Alvarez, B. Holt, S. Jaruwatanadilok, E. Podest, and K. C. Cavanaugh, "An arctic sea ice multi-step classification based on GNSS-R data from the TDS-1 mission," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111202.

[42] B. J. Southwell and A. G. Dempster, "Sea ice transition detection using incoherent integration and deconvolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 14–20, 2020.

[43] Y. Zhu, J. Wickert, T. Tao, K. Yu, Z. Li, X. Qu, Z. Ye, J. Geng, J. Zou, and M. Semmling, "Sensing sea ice based on Doppler spread analysis of spaceborne GNSS-R data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 217–226, 2020.

[44] C. Gerlein-Safdi and C. S. Ruf, "A CYGNSS-based algorithm for the detection of inland waterbodies," *Geophys. Res. Lett.*, vol. 46, no. 21, pp. 12065–12072, 2019.

[45] N. Rodriguez-Alvarez, E. Podest, K. Jensen, and K. C. McDonald, "Classifying inundation in a tropical wetlands complex with GNSS-R," *Remote Sens.*, vol. 11, no. 9, p. 1053, May 2019.

[46] C. Ruf, P. S. Chang, M.-P. Clarizia, S. Gleason, Z. Jelenak, S. Majumdar, M. Morris, J. Murray, S. Musko, D. Posselt, D. Provos, D. Starkenburg, and V. Zavorotny, *CYGNSS Handbook*. Ann Arbor, MI, USA: Michigan Publishing, 2016.

[47] C. Chew, J. T. Reager, and E. Small, "CYGNSS data map flood inundation during the 2017 atlantic hurricane season," *Sci. Rep.*, vol. 8, no. 1, p. 9336, Dec. 2018.

[48] W. Wei, B. Liu, Z. Zeng, and X. Chen, "Using CYGNSS data to monitor China's flood inundation during Typhoon and extreme precipitation events in 2017," *Remote Sens.*, vol. 11, no. 7, p. 854, 2019.

[49] K. Smith and R. Ward, *Floods: Physical Processes and Human Impacts*. Hoboken, NJ, USA: Wiley, 1998.

[50] G. Bonaccorso, *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning*. Birmingham, U.K.: Packt Publishing, 2018.

[51] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, O. Maimon and L. Rokach, Eds. Boston, MA, USA: Springer, 2005.

[52] A. Fernández, *Data Mining for Imbalanced Datasets: An Overview*. New York, NY, USA: Springer, 2018.

[53] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[54] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE WCCI*, Jun. 2008, pp. 1322–1328.

[55] K. T. Chui, R. W. Liu, M. Zhao, and P. O. De Pablos, "Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020.

[56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[57] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[58] P. Lim, C. K. Goh, and K. C. Tan, "Evolutionary cluster-based synthetic oversampling ensemble (ECO-ensemble) for imbalance learning," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2850–2861, Sep. 2017.

[59] Y. Zhang, X. Li, L. Gao, L. Wang, and L. Wen, "Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning," *J. Manuf. Syst.*, vol. 48, pp. 34–50, Jul. 2018.

[60] P. Ghasemigoudarzi, W. Huang, and O. DeSilva, "Detecting floods caused by tropical cyclone using CYGNSS data," in *Proc. IEEE MFI*, Karlsruhe, Germany, 2020.

[61] C. Ruf, S. Asharaf, R. Balasubramaniam, S. Gleason, T. Lang, D. McKague, D. Twigg, and D. Waliser, "In-orbit performance of the constellation of CYGNSS hurricane satellites," *Bull. Amer. Meteorological Soc.*, vol. 100, no. 10, pp. 2009–2023, Oct. 2019.

[62] CYGNSS. (2018). *CYGNSS Level 1 Science Data Record Version 2.1. Ver. 2.1. PO.DAAC, CA, USA*. Accessed: Feb. 7, 2020. [Online]. Available: https://doi.org/10.5067/CYGNS-L1X21

[63] E. S. Blake, "The 2017 atlantic hurricane season: Catastrophic losses and costs," *Weatherwise*, vol. 71, no. 3, pp. 28–37, May 2018.

[64] G. Brakenridge and A. J. Kettner. DFO Flood Event # 4516. University of Colorado, Boulder, CO, USA. [Online]. Available: http://floodobservatory.colorado.edu/Events/2017USA4516/GISData/

[65] A. Jarvis, H. I. Reuter, A. Nelson, and E. Guevara. (2008). *Hole-Filled Seamless SRTM Data Version 4*. [Online]. Available: http://srtm.csi.cgiar.org

[66] CIFOR. *Global Wetlands*. Accessed: Sep. 1, 2020. [Online]. Available: https://www.cifor.org/global-wetlands/

[67] *Global Surface Water Data Access*. Accessed: Jun. 22, 2020. [Online]. Available: https://global-surface-water.appspot.com/download

[68] G. Brakenridge and A. J. Kettner. *DFO Flood Event # 4510*. Univ. Colorado, Boulder, CO, USA. Accessed: Sep. 1, 2020. [Online]. Available: http://floodobservatory.colorado.edu/Events/2017USA4510/GISData/

[69] T. Gumbricht, R. M. Roman-Cuesta, L. Verchot, M. Herold, F. Wittmann, E. Householder, N. Herold, and D. Murdiyarso, "An expert system model for mapping tropical wetlands and peatlands reveals south america as the largest contributor," *Global Change Biol.*, vol. 23, no. 9, pp. 3581–3599, Sep. 2017.

[70] P. Zhu, "Impact of land-surface roughness on surface winds during hurricane landfall," *Quart. J. Roy. Meteorological Soc.*, vol. 134, no. 633, pp. 1051–1057, 2008. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.265

[71] H. Jiang, J. B. Halverson, J. Simpson, and E. J. Zipser, "Hurricane 'rainfall potential' derived from satellite observations aids overland rainfall prediction," *J. Appl. Meteor. Climatol.*, vol. 47, no. 4, pp. 944–959, 2008, doi: 10.1175/2007JAMC1619.1.

[72] C. Ruf, J. Scherrer, R. Rose, and D. Provost. *Algorithm Theoretical Basis Document Level 1B DDM Calibration*. Accessed: Jul. 2, 2020. [Online]. Available: https://clasp-research.engin.umich.edu/missions/cygnss/reference/ATBD%2%0L1B%20DDM%20Calibration%20R1.pdf

[73] E. Loria, A. O'Brien, V. Zavorotny, M. Lavalle, C. Chew, R. Shah, and C. Zuffada, "Analysis of wetland extent retrieval accuracy using cygnss," in *Proc. IEEE IGARSS*, Jul. 2019, pp. 8684–8687.

[74] J. Cartwright, C. J. Banks, and M. Srokosz, "Sea ice detection using GNSS-R data from TechDemoSat-1," *J. Geophys. Res. Oceans*, vol. 124, no. 8, pp. 5801–5810, 2019.

[75] S. Gleason, C. S. Ruf, A. J. O'Brien, and D. S. McKague, "The CYGNSS level 1 calibration algorithm and error analysis based on on-orbit measurements," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 37–49, Jan. 2019.

[76] N. Rodriguez-Alvarez and J. L. Garrison, "Generalized linear observables for ocean wind retrieval from calibrated GNSS-R delay-Doppler maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 1142–1155, 2016.

[77] Q. Yan, W. Huang, S. Jin, and Y. Jia, "Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111944.

[78] P. Ghasemigoudarzi, W. Huang, O. De Silva, Q. Yan, and D. Power, "A machine learning method for inland water detection using CYGNSS data," *IEEE Geosci. Remote Sens. Lett.*, early access, Sep. 7, 2020, doi: 10.1109/LGRS.2020.3020223.

[79] M. N. Das, *Statistical Methods and Concepts*. Hoboken, NJ, USA: Wiley, 1989.

[80] H. Carreno-Luengo, G. Luzi, and M. Crosetto, "Impact of the elevation angle on CYGNSS GNSS-R bistatic reflectivity as a function of effective surface roughness over land surfaces," *Remote Sens.*, vol. 10, no. 11, p. 1749, Nov. 2018.

[81] K. Jensen, K. McDonald, E. Podest, N. Rodriguez-Alvarez, V. Horna, and N. Steiner, "Assessing L-band GNSS-reflectometry and imaging radar for detecting sub-canopy inundation dynamics in a tropical wetlands complex," *Remote Sens.*, vol. 10, no. 9, p. 1431, Sep. 2018.

[82] R. Rose, C. Ruf, D. Rose, M. Brummitt, and A. Ridley, "The CYGNSS flight segment; a major NASA science mission enabled by micro-satellite technology," in *Proc. IEEE Aerosp. Conf.*, Mar. 2013, pp. 1–13.

[83] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

[84] L. Breiman, *Classification and Regression Trees* (Wadsworth Statistics/Probability Series). Belmont, CA, USA: Wadsworth International Group, 1984.

[85] R. E. Schapire, *Boosting: Foundations and Algorithms* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2012.

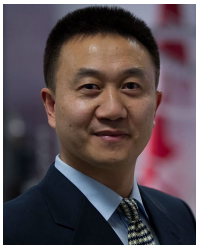
[86] F. Yoav and S. Robert, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 148–156.

[87] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 1007–1011, Mar. 2005.

[88] L. Liu, W. Huang, and C. Wang, "Hyperspectral image classification with kernel-based least-squares support vector machines in sum space," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1144–1157, Apr. 2018.

[89] X. Chen, W. Huang, C. Zhao, and Y. Tian, "Rain detection from X-band marine radar images: A support vector machine-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2115–2123, Mar. 2020.
[90] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Adv. Neural Inf. Process Syst.*, 1992, pp. 831–838.
[91] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[92] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, no. Dec, pp. 1889–1918, 2005.
[93] C. S. Ruf, S. Gleason, and D. S. McKague, "Assessment of CYGNSS wind speed retrieval uncertainty," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 87–97, Jan. 2019.
[94] T. Wang, C. S. Ruf, B. Block, D. S. McKague, and S. Gleason, "Design and performance of a GPS constellation power monitor system for improved CYGNSS L1B calibration," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 26–36, Jan. 2019.
[95] T. Wang, C. Ruf, S. Gleason, B. Block, D. McKague, and A. O'Brien, "A real-time EIRP level 1 calibration algorithm for the CYGNSS mission using the zenith measurements," in *Proc. IEEE IGARSS*, Jul. 2019, pp. 8725–8728.
[96] M. Rahman, L. Di, E. Yu, L. Lin, C. Zhang, and J. Tang, "Rapid flood progress monitoring in cropland with NASA SMAP," *Remote Sens.*, vol. 11, no. 2, p. 191, Jan. 2019.
[97] W. R. Berghuijs, R. A. Woods, C. J. Hutton, and M. Sivapalan, "Dominant flood generating mechanisms across the united states," *Geophys. Res. Lett.*, vol. 43, no. 9, pp. 4382–4390, May 2016.
[98] S. Ye, H.-Y. Li, L. R. Leung, J. Guo, Q. Ran, Y. Demissie, and M. Sivapalan, "Understanding flood seasonality and its temporal shifts within the contiguous united states," *J. Hydrometeorol.*, vol. 18, no. 7, pp. 1997–2009, Jul. 2017.
[99] M. Parrens, J.-P. Wigneron, P. Richaume, A. Mialon, A. Al Bitar, R. Fernandez-Moran, A. Al-Yaari, and Y. H. Kerr, "Global-scale surface roughness effects at L-band as estimated from SMOS observations," *Remote Sens. Environ.*, vol. 181, pp. 122–136, Aug. 2016.
[100] E. Linacre, *Climate Data Resources: A Reference Guide*. East Sussex, U.K.: Psychology Press, 1992.
[101] A. Arguez, I. Durre, S. Applequist, R. S. Vose, M. F. Squires, X. Yin, R. R. Heim, Jr., and T. W. Owen, "NOAA's 1981–2010 US climate normals: An overview," *Bull. Amer. Meteor. Soc.*, vol. 93, no. 11, pp. 1687–1697, 2012.
[102] *Ventusky Precipitation History*. Accessed: Sep. 1, 2020. [Online]. Available: https://www.ventusky.com/?p=27.26;-81.42;6&l=rain-3h&t=20170901/2100
[103] H. Laachrate, A. Fadil, and A. Ghafiri, "Soil moisture mapping using SMOS applied to flood monitoring in the Moroccan context," *ISPRS*, vol. 42, no. 4/W12, pp. 105–111, 2019.
[104] D. Stilla, M. Zribi, N. Pierdicca, N. Baghdadi, and M. Huc, "Desert roughness retrieval using CYGNSS GNSS-R data," *Remote Sens.*, vol. 12, no. 4, p. 743, Feb. 2020.
[105] B. J. Choudhury, T. J. Schmugge, A. Chang, and R. W. Newton, "Effect of surface roughness on the microwave emission from soils," *J. Geophys. Res. Oceans*, vol. 84, no. C9, pp. 5699–5706, 1979. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC084iC09p05699
[106] Y. Zhang, J. T. Pulliainen, S. S. Koponen, and M. T. Hallikainen, "Water quality retrievals from combined landsat TM data and ERS-2 SAR data in the gulf of finland," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 3, pp. 622–629, Mar. 2003.
[107] G. Wu, J. de Leeuw, A. K. Skidmore, Y. Liu, and H. H. T. Prins, "Performance of landsat TM in ship detection in turbid waters," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 11, no. 1, pp. 54–61, Feb. 2009.
[108] S. Martinis, C. Kuenzer, A. Wendleder, J. Huth, A. Twele, A. Roth, and S. Dech, "Comparing four operational SAR-based water and flood detection approaches," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3519–3543, Jul. 2015, doi: 10.1080/01431161.2015.1060647.

PEDRAM GHASEMIGOUDARZI (Student Member, IEEE) was born in Aligudarz, Iran. He received the B.Sc. degree in telecommunications engineering from the Iran University of Science and Technology, Tehran, Iran, in 2017. He is currently pursuing the M.Eng. degree in electrical engineering with the Memorial University of Newfoundland, St. John's, NL, Canada. His research interests include flood detection and inland water remote sensing using global navigation satellite system-reflectometry.



WEIMIN HUANG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in radio physics from Wuhan University, Wuhan, China, in 1995, 1997, and 2001, respectively, and the M.Eng. degree in electrical engineering from the Memorial University of Newfoundland, St. John's, NL, Canada, in 2004.

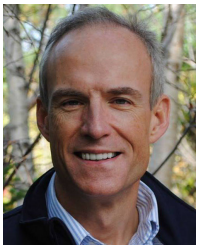
From 2008 to 2010, he was a Design Engineer with Rutter Technologies, St. John's. Since 2010, he has been with the Faculty of Engineering and Applied Science, Memorial University of Newfoundland, where he is currently a Professor. He has authored over 250 research articles. His research interests include the mapping of oceanic surface parameters via high-frequency ground wave radar, X-band marine radar, and global navigation satellite systems.

Dr. Huang has been a Technical Program Committee Member. He serves as a Regular Reviewer over 50 international journals and a Reviewer for many IEEE international conferences, such as RadarCon, ICC, GLOBECOM, IGARSS, and Oceans. He received the Postdoctoral Fellowship from the Memorial University of Newfoundland. In 2017, he received the Discovery Accelerator Supplements Award from the Natural Sciences and Engineering Research Council of Canada. He was a recipient of the IEEE Geoscience and Remote Sensing Society 2019 Letters Prize Paper Award. He has served as the Technical Program Co-Chair of the IEEE Newfoundland Electrical and Computer Engineering Conference in 2012 and 2013. He is also an Associate Editor of IEEE Access, the IEEE Canadian Journal of Electrical and Computer Engineering and an Editorial Board Member of *Remote Sensing*.

OSCAR DE SILVA (Member, IEEE) received the B.Sc. degree in engineering from the University of Moratuwa, Sri Lanka, in 2009, and the Ph.D. degree from the Memorial University of Newfoundland (MUN), St. John's, NL, Canada, in 2015. Following postdoctoral work with the ABS Harsh Environment Technology Center, St. John's, he joined MUN as a Faculty Member, in 2016. He is currently an Assistant Professor with the Faculty of Engineering and Applied Science, MUN. His main research interests include autonomous robotics, navigation system design, and machine learning.

QINGYUN YAN (Member, IEEE) was born in Haimen, China. He received the B.Eng. degree in electronic science and engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014, and the M.Eng. and Ph.D. degrees in electrical engineering from the Memorial University of Newfoundland, St. John's, NL, Canada, in 2015 and 2020, respectively.

He is currently with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology. His research interests include tsunami, sea ice, and land remote sensing using global navigation satellite system-reflectometry. He was a recipient of the 2019 IEEE GRSS Letters Prize Paper Award from the IEEE Geoscience and Remote Sensing Society.



DESMOND T. POWER (Member, IEEE) was born in North River, NL, Canada, in 1967. He received the Bachelor and Master of Engineering degrees from the Memorial University of Newfoundland.

He is currently the Vice President of Remote Sensing at C-CORE and manages a group of 40 individuals with expertise on geomatics, radar systems, earth observation, and geomatics. He has 29 years of professional experience. In the first seven years of his career, he primarily worked on high frequency over the horizon radar. In 1998, he started work on microwave radar systems and satellite synthetic aperture radar. He has since branched out into work on earth observation (electro-optical and SAR) and geomatics, with a strong focus on research and applications development. The research performed by Desmond's group include iceberg and vessel detection, SAR-based winds, EO water quality and river ice monitoring, ice charting and sea ice thickness measurement, satellite and ground-based interferometric SAR (for ground deformation monitoring), vehicle detection and monitoring, wetlands monitoring, and oil slick detection.

Mr. Power is a member at large of the executive of Canadian Remote Sensing Society and a P.Eng registered with Professional Engineers and Geoscientists of Newfoundland and Labrador.