Permanent Water and Flash Flood Detection Using Global Navigation Satellite System Reflectometry

by

© Pedram, Ghasemigoudarzi B. Sc.

A thesis submitted to the School of Graduate Studies

in partial fulfilment of the requirements for the degree of

Master of Engineering

Department of Electrical and Computer Engineering

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

May 2021

St. John's Newfoundland

Abstract

In this thesis, research for inland water extent and flash floods remote sensing using Global Navigation Satellite System Reflectometry (GNSS-R) data of the Cyclone Global Navigation Satellite System (CYGNSS) is presented.

Firstly, a high-resolution Machine Learning (ML) method for detecting inland water extent using the CYGNSS data is implemented via the Random Under-Sampling Boosted (RUSBoost) algorithm. The CYGNSS data of the year 2018 over the Congo and Amazon basins are gridded into $0.01^{\circ} \times 0.01^{\circ}$ cells. The RUSBoost-based classifier is trained and tested with the CYGNSS data over the Congo basin. The Amazon basin data that is unknown to the classifier is then used for further evaluation. Using only three observables extracted from the CYGNSS Delay-Doppler Maps (DDMs), the proposed technique is able to detect 95.4% and 93.3% of the water bodies over the Congo and Amazon basins, respectively. The performance of the RUSBoost-based classifier is also compared with an image processing based inland water detection method. For the Congo and Amazon basins, the RUSBoost-based classifier has a 3.9% and 14.2% higher water detection accuracies, respectively.

Secondly, a flash flood detection method using the CYGNSS data is investigated. Considering Hurricane Harvey and Hurricane Irma as two case studies, six different Data Preparation Approaches (DPAs) for flood detection based on the CYGNSS data and the RUSBoost classification algorithm are investigated in this thesis. Taking flood and land as two classes, flash flood detection is tackled as a binary classification problem. Eleven observables are extracted from the DDMs for feature selection. These observables, alongside two features from 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 × 500 m, and the detection accuracy for non-flooded land points is 97.20% and 71.00%, respectively.

Acknowledgements

The author wishes to thank the Faculty of Engineering and Applied Science for offering him the opportunity of conducting this work. Particularly, the supervision of Dr. Weimin Huang, Dr. Oscar De Silva, and Mr. Desmond Power and their patience and encouragement through this research are greatly appreciated. Their suggestions and insights regarding this research topic are invaluable.

The author is grateful for financial support in the form of the Natural Sciences and Engineering Research Council of Canada Discovery Grants (NSERC RGPIN-2017-04508 and RGPAS-2017-507962) to Dr. W. Huang, and the Canadian Space Agency CubeSat Grant (17CCPNFL11) to Dr. W. Huang and Mr. D. Power.

The author also would like to thank Dr. Qingyun Yun and Dr. Murilo T. Silva for their suggestions and advice at various stages of this work.

Finally, the author profoundly appreciates the patience and understanding of the author's family, especially Mrs. Robabeh Ghasemi, Dr. Fatemeh Helma Ghasemigoudarzi, and Dr. Pouria Ghasemigoudarzi. This work could not be completed without their support.

Contents

List of Tables

List of Figures

Table of Acronyms

Page number indicates the first significant reference.

GNSS-R :	Global Navigation Satellite System-Reflectometry (abstract).
CYGNSS :	Cyclone Global Navigation Satellite System (abstract).
ML :	Machine Learning (abstract).
RUSBoost :	Random Under-Sampling Boosted (abstract).
DDM :	Delay-Doppler Map (abstract).
DPA :	Data Preparation Approach (abstract).
SVM :	Support Vector Machine (abstract).
SAR :	Synthetic Aperture Radar (p. 3).
MODIS :	Terra Moderate Resolution Imaging Spectroradiometer (p. 5).
MOD44W :	MODIS Land Water Mask (p. 5).
GSW :	Global Surface Water (p. 5).
NRT :	Near Real-time (p. 5).
GFMS :	Global Flood Monitoring System (p. 5).
DFO :	Dartmouth Flood Observatory (p. 5).

GFDS :	Global Flood Detection System (p. 5).
GNSS :	Global Navigation Satellite System (p. 6).
GPS :	Global Positioning System (p. 6).
GLONASS :	Global Navigation Satellite System (p. 6).
BDS :	BeiDou Navigation Satellite System (p. 6).
SP :	Specular Point (p. 7).
GZ :	Glistening Zone (p. 7).
MEO :	Medium Earth Orbit (p. 7).
LEO :	Low Earth Orbit (p. 7).
TDS-1 :	TechDemoSat-1 (p. 7).
UPC :	Universidad Politécnica de Cataluña (p. 8).
SM :	Soil Moisture (p. 8).
SR :	Surface Reflectivity (p. 8).
GPM :	Global Precipitation Measurement (p. 9).
OW :	Open Water (p. 9).
FV :	Flooded Vegetation (p. 9).
NF :	Non-flooded Vegetation (p. 9).
MDTR :	Multiple Decision Tree Randomized Structure (p. 9).
SNR_C :	Corrected Signal to Noise Ratio (p. 10).
STD :	Standard Deviation (p. 10).
C/A :	Coarse/Acquisition (p. 12).
PRN :	Pseudo Random Noise (p. 12).
NBRCS :	Normalized Bistatic Radar Cross Section (p. 13).
WAF :	Woodward Ambiguity Function (p. 13).

BRCS :	Bistatic Radar Cross Section (p. 14).
FFZ :	First Fresnel Zone (p. 15).
NN :	Neural Networks (p. 18).
SMOTE :	Synthetic Minority Oversampling Technique (p. 19).
ADASYN :	Adaptive Synthetic Sampling Method (p. 19).
GAN :	Generative Adversarial Network (p. 19).
RUS :	Random Under-Sampling (p. 19).
GI :	Gini Impurity Factor (p. 22).
GIS :	Geographic Information System (p. 27).
$\mathrm{SNR}_{\mathrm{Peak}}$:	Peak Signal to Noise Ratio (p. 28).
RV :	Random Variable (p. 28).
INSTAAR :	Institute of Arctic and Alpine Research (p. 41).
GDACS :	Global Disaster Alert and Coordination System (p. 41).
SRTM90m DEM :	Shuttle Radar Topography Mission Digital Elevation Model (p. 41).
CIFOR :	Center for International Forestry Research (p. 42).
PALSAR :	Phased Array Type L-band Synthetic Aperture Radar (p. 42).
TES :	Trailing-edge Slope (p. 43).
LES :	Leading-edge Slope (p. 43).
DDMA :	Delay-Doppler Map Average (p. 43).
Wave-width :	Width of the Doppler Waveform (p. 43).
GLO :	Generalized Linear Observable (p. 43).
IDW :	Integrated Delay Waveform (p. 44).

PCA :	Principal Component Analysis (p. 46).
RBF :	Radial Basis Function (p. 50).
SMO :	Sequential Minimal Optimization (p. 50).
EIRP :	Effective Isotropic Radiated Power (p. 72).
SMAP :	Soil Moisture Active Passive (p. 72).
VAE :	Variational Autoencoder (p. 75).
XGBoost :	Extreme Gradient Boosting (p. 75).

Table of Symbols

Page number indicates the first significant reference. Although not all symbols are explicitly referenced below, their definitions are obvious from the context.

τ : Relative delay (p. 13).

f : Relative Doppler frequency (p. 13).

$P^{rx}_{\tau,f}$: Processed received power at each delay-Doppler bin (p. 13).

P^{tx} : GNSS transmitted power (p. 13).

G^{tx} : GNSS transmitter antenna gain (p. 13).

G^{rx} : GNSS-R receiver antenna gain (p. 13).

λ : $\lambda = 19.05cm$ is the wavelength of GPS L1-band signal (p. 13).

R^{tx} : Distance between GNSS transmitter and a point on the reflecting surface (p. 13).

R^{rx} : Distance between GNSS-R receiver and a point on the reflecting surface (p. 13).

σ^0 : Normalized bistatic radar cross section (p. 13).

χ : The Woodward Ambiguity Function (p. 13).

$\Lambda_{\tau;x,y}$: The GPS signal spreading function in delay (p. 13).

$S_{f;x,y}$: The frequency response of the GPS signal (p. 13).

$\bar{G}^{rx}_{\tau,f}$: The GNSS-R receiver antenna gain at each delay-Doppler bin (p. 13).

$\bar{A}_{\tau,f}$: Effective surface scattering area at each delay-Doppler bin (p. 13).

$\bar{R}^{tx}_{\tau,f}$: The range loss between transmitter and each delay-Doppler bin (p. 13).

$\bar{R}^{rx}_{\tau,f}$: The range loss between receiver and each delay-Doppler bin (p. 13).

$\langle\sigma^0_{\tau,f}\rangle$: The normalized bistatic radar cross section at each delay-Doppler bin (p. 13).

$\langle\sigma_{\tau,f}\rangle$: The bistatic radar cross section at each delay-Doppler bin (p. 13).

R^{rx}_{SP} : Distance between GNSS-R receiver and SP (p. 14).

R^{tx}_{SP} : Distance between transmitter and SP (p. 14).

G^{rx}_{SP} : The GNSS-R receiver antenna gain at SP (p. 14).

$P^{coh}_{\tau,f}$: Coherent power component of reflected signal (p. 14).

$P^{inc}_{\tau,f}$: Incoherent power component of reflected signal (p. 14).

$\Gamma_{\tau,f}$: Surface reflectivity at each delay-Doppler bin (p. 15).

a_{FFZ} : Semi-major axis of the FFZ (p. 15).

b_{FFZ} : Semi-minor axis of the FFZ (p. 15).

θ : Incidence angle (p. 15).

k : Wave number of GNSS signal (p. 18).

S : Imbalanced data set (p. 20).

$\mathbf{x}_i$: The feature vector of the data point i in imbalanced data set (p. 20).

$\mathbf{y}_i$: $y_i \in \{0, 1\}$ is the label of the data point i in imbalanced data set (p. 20).

S' : Balanced data set (p. 20).

$\mathbf{x'}_p$: Feature vector of the data point p in balanced data set (p. 20).

$\mathbf{y'}_p$: Label of the data point p in balanced data set (p. 20).

$D_t(i)$: Weight of data point i at iteration t in imbalanced data set (p. 20).

$D'_t(p)$: Weight of data point p at iteration t in balanced data set (p. 20).

c_t^j : Decision threshold of the feature j at iteration t (p. 22).

$\Omega_t^{r,l}$: The probability of right or left split of the decision stump t (p. 22).

$\Theta_t^{r,l}$: The Gini impurity factor of right or left split of the decision stump t (p. 23).

h_t : The weak hypothesis at iteration t (p. 23).

$\pi_{r,l}$: The label proportion of right or left split (p. 23).

$N_{r,l}(y)$: The number of $y \in \{0,1\}$ labelled points within right or left split (p. 24).

$N_{r,l}$: The total number of points within within right or left split (p. 24).

ϵ_t : The pseudo loss of the weak hypothesis at iteration t (p. 24).

α_t : The weight updating factor at iteration t (p. 24).

H : The output hypothesis (p. 24).

$\langle\sigma_m\rangle$: Maximum of BRCS DDM (p. 44).

P_{rxm} : Maximum of power DDM (p. 44).

Chapter 1

Introduction

This chapter, first, demonstrates the importance of inland permanent water and flash floods remote sensing and the significance of Global Navigation Satellite System Reflectometry (GNSS-R) signals for inland water and flash flood detection. Then, the literature about common water bodies remote sensing methods and GNSS-R applications in remote sensing, particularly inland permanent water and flash floods surveillance and monitoring, is summarized. Last, the scope of this thesis is presented.

1.1 Research Rationale

Inland permanent water bodies are key elements in their surrounding environments and various living creatures' survival depends on them [1]. Moreover, they are essential for most industrial and agricultural operations [2]. The surface water is dynamic and its extent changes due to human activities and climate variations. Thus, knowledge of high temporal water extent data is important for various disciplines.

Flash flood is a surge of water that starts and develops in a short period. The primary cause of flash floods 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 floods. Even though a flash flood dissipates quickly after its occurrence, it has consequential damages such as death and severe injuries, water contamination, financial harm, infrastructure damages, and agricultural losses [3, 4]. Hurricanes, which are a significant cause of flash floods, are tropical cyclones with high wind speed (higher than $33\,\mathrm{ms}^{-1}$ [5]) and capable of pouring massive rain over coastal regions during landfall [6]. Considering the population growth in coastal areas that are exposed to hurricanes, flood detection and monitoring can reduce these damages and increase the speed of post-disaster response [7].

Being able to continuously monitor the surface of the Earth, remote sensing technologies are more efficient compared to traditional in situ measurements. Even though remote sensing techniques require significant infrastructure and their equipment manufacturing is more complicated, they are more financially feasible in the long term since they monitor larger areas with less required labour and energy [8]. Furthermore, their operation is less impacted by the condition of the Earth's surface, which makes them a perfect solution for difficult-to-access regions [9]. The optical and microwave sensors are the two main spaceborne remote sensing instruments that are widely used for detecting surface water bodies with high-resolution.

Since the optical sensors can provide high temporal and high spatial resolutions data on a global scale, they are used as the main instrument for monitoring water bodies. Different electromagnetic wavelengths within the visible spectrum interact differently with water bodies. 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, can be detected using optical sensors [10, 11]. The water bodies detected during flash floods based on optical images are compared with water reference data sets, to estimate flood extent maps [12]. Optical remote sensing sensors are not able to detect surface water bodies when the region contains high density biomass or is covered by clouds. Furthermore, in an optical image, the cloud shadow is classified as water. Therefore, the presence of cloud and its shadow are critical challenges for water extent estimation using optical sensors [10]. The water bodies detection algorithms developed for the data of optical satellites are able to detect the surface water including floods with an accuracy higher than 97% [10, 13].

In addition to optical sensors, other sensors have been employed to detect or observe the formation of floods, including river level and rain gauge measurements [14], weather radar [15], and microwave satellite systems [16–19]. The river level and rain gauge sensors are used in designated locations such as river basins, resulting in limited coverage over a particular area. For large scale flood monitoring, a complicated network of such sensors is required, which may not be feasible due to geographic and economic reasons [20]. Weather radars are able to predict the amount and type of precipitation. In a flash flood scenario, in addition to the amount of precipitation, other factors such as topography, soil moisture, drainage of the rivers, etc. are influential [21]. Hence, in warning systems, the possibility of flooding is forecasted using weather radar data together with data sets from other sensors [22, 23].

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 flood extent maps from SAR data can be solved by using supervised and unsupervised methods [24]. In a supervised method, since the classifier is trained with labelled pixels from a region, the algorithm has local dependence. Segmentation [25], threshold determination [26], and change detection [27] 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 computation 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 [24]. In the threshold determination method, instead of a single threshold value, multiple threshold values are considered for detecting floods on a large scale [26]. Therefore, the method's accuracy is highly dependant on the accuracy of the preset threshold values. In the change detection method, prior- and post-flood SAR images are required, which is a big challenge due to the long revisit time of SAR systems [27]. Therefore, in recent works, combinations of these techniques are used [28, 29]. Depending on the region, the SAR based flood detection algorithms are able to detect floods with an accuracy ranging between 80 % to 95 % [30–32]. The SAR data require geometric correction and speckle reduction. Hence, compared to passive microwave and optical sensors, the retrieval algorithms based on SAR data are more complicated [24]. 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 [33]. Thus, due to the low temporal resolution (several days), satellite sensors might not even be able to collect data over a flooding area in time.

The optical images are used as the main resource for creating permanent water data sets. The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Land Water Mask (MOD44W Version 6) data [34] and the Global Surface Water (GSW) [10, 35], are the two main representative permanent water extent data sets created by optical sensors.

Moreover, some of the aforementioned remote sensing data and algorithms have been used by observatories to create Near Real-time (NRT) flash floods information. For example, the Global Flood Monitoring System (GFMS) uses real-time precipitation data and a hydrologic model for NRT global flood detection based on the locally defined flood thresholds [36]. In this method, the actual extent of a flash flood is not derived. The Dartmouth Flood Observatory (DFO) and NASA's Goddard Hydrology Laboratory employ the data collected by two MODIS sensors (aboard the satellites Terra and Aqua) for flood monitoring [12]. 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 [37]. Also, they employ microwave sensors data to mitigate the cloud effect to increase flood detection accuracy [12]. 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 detect riverine inundations [19].

The GNSS-R is a well-established technique for remote sensing. Since it takes advantage of existing signals of opportunity, receivers are cost-efficient, which makes it possible to achieve a large constellation and, consequently, high temporal resolution (hours) [38]. The Cyclone Global Navigation Satellite System (CYGNSS) is a GNSS-

R constellation that operates between 38° S and 38° N latitudes [39]. This area is impacted by various flash floods and contains massive permanent water bodies. Considering the benefits of the GNSS-R method and availability of the CYGNSS data, in this work, we focus on the application of the CYGNSS GNSS-R data in detecting permanent water bodies and flash floods. In this thesis, instead of aforementioned methods, the Machine Learning (ML) algorithms are used.

1.2 Literature Review

Global Navigation Satellite System (GNSS) is a term describing spaceborne systems that provide geospatial positioning information. Based on the distances between the target and four GNSS transmitting satellites, positioning, navigation and timing are provided for users. The United States' Global Positioning System (GPS) and Russia's Global Navigation Satellite System (GLONASS), with China's BeiDou Navigation Satellite System (BDS) and the European Union's Galileo are four operational GNSSs [40].

In the late 1980s, it was proposed that the multipath signals could be used as a source for remote sensing in a bistatic (multistatic) radar technique called GNSS-R since the multipath signals carry information about the surfaces from which they were reflected [41]. In the GNSS-R technique, the reflected signal is correlated with a direct signal for different values of delays and Doppler frequencies, and the result is plotted in a Delay-Doppler Map (DDM) as the output of the system. The geospatial scheme of this method is illustrated in Figure 1.1. The GNSS-R receiver collects the GNSS signals that are reflected off the surface of the Earth and transfers them into DDMs.

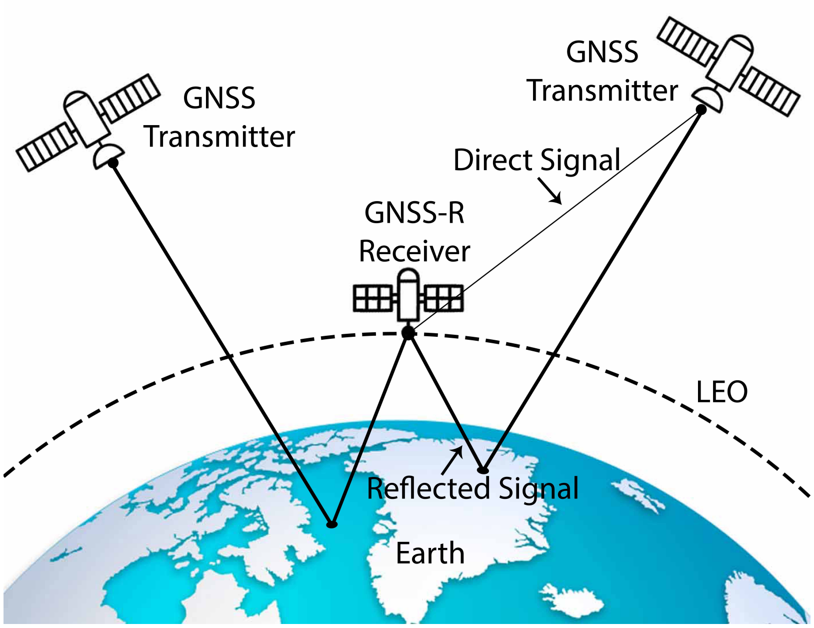


Figure 1.1: A schematic of the GNSS-R technique, with two GNSS transmitters in MEO and a GNSS-R receiver in LEO.

Each DDM represents an area around the Specular Point (SP) called the Glistening Zone (GZ). The GNSS constellations positioned in the Medium Earth Orbit (MEO) are used as transmitters. Hence, only passive receiving satellites located in the Low Earth Orbit (LEO) are demanded, which makes the GNSS-R technique less expensive [38].

The first spaceborne observation of a GNSS reflected signal was found by the SIR-C radar aboard the Shuttle [42]. As a part of the Disaster Monitoring Constellation, the UK-DMC 1 was the first operational GNSS-R satellite, which was launched to an altitude of 686 km in September 2003. Following this mission, the UK TechDemoSat-1 (TDS-1) was launched in 2014 and is still operating with global coverage [43, 44]. The TDS-1 daily data are provided via an internet link [45] and the first sea ice product based on TDS-1 data has been officially published since 2019. In 2016, a 6-U CubeSat called ^{3}Cat2 that was designed by the Universidad Politécnica de Cataluña (UPC) was launched into LEO [46]. In the same year, NASA launched the CYGNSS constellation to 510 km for tracking cyclones and estimating winds speed over the ocean. The CYGNSS consists of eight satellites equipped with GNSS-R payloads. Each satellite is able to scan up to four reflection swaths simultaneously [39]. Therefore, it has a higher temporal resolution compared to other GNSS-R based instruments. The CYGNSS daily data are posted two days after their collection and are available to the public through [47]. By having the CYGNSS operating over land with a high temporal resolution, the course of GNSS-R land remote sensing has been improved since its launch.

The GNSS-R technique has shown a great capacity for various ocean remote sensing applications such as altimetry [48], sea surface wind [49–53], target detection [54], tsunami [55, 56], and sea ice [57–62]. Moreover, it also has been used for land remote sensing applications, especially for the Soil Moisture (SM) [63–65].

There have been some attempts at modelling the received GNSS signal reflected from the regions containing water bodies [66, 67]. The primary results show that these models are able to estimate the value of Surface Reflectivity (SR) by knowing the characteristics of the scattering surface. However, further studies are required in order to extract the characteristics of the scattering surface from the SR value.

The CYGNSS GNSS-R SR changes when the water bodies are changed due to a flash flood. The difference between SR values of areas with and without water is $\sim 12\,\mathrm{dB}$ [68]. Even though considering a threshold can illustrate the presence of water, it is not the optimum approach to obtain precise extent maps for water bodies.

Moreover, there is a high correlation between the CYGNSS SR and Global Precipitation Measurement (GPM) data. Comparing the precipitation, brightness temperature, and the CYGNSS SR over the surfaces impacted by typhoons (tropical cyclones located in the Northwest Pacific Ocean) shows that the CYGNSS data is as reliable as validated reference data sets [69]. Therefore, it is possible to monitor dynamic phenomena using the CYGNSS GNSS-R data.

It has been shown in [70] using the CYGNSS data alongside two other ancillary data sets, that seasonal inundation over the Pacaya-Samiria Natural Reserve, a tropical wetland complex located in the Peruvian Amazon, can be classified into three categories, which are Open Water (OW), Flooded Vegetation (FV), and Non-flooded Vegetation (NF). For classification, an ML method called Multiple Decision Tree Randomized Structure (MDTR) based on the random forest algorithm was implemented. In such a study, various GNSS-R observables were extracted from the CYGNSS data. The results show that the classification accuracies for OW, FV, and NF labels were 65.4%, 60.26%, and 94.75, respectively. Unlike seasonal floods and permanent water which are developed in specific periods and last over more extended times [71], flash floods are difficult to monitor. Therefore, it is vital to develop a method that can detect and monitor flash floods. Hence, one objective of this work is to quantitatively investigate the ability of the GNSS-R technique to detect flash floods.

Furthermore, for inland water detection, an image processing based technique, which in this thesis is referred to as the watermask detection method, is able to produce water extent maps with a resolution of $0.01^{\circ} \times 0.01^{\circ}$ [72]. The watermask detection method is a multi-step procedure that processes the map of the corrected Signal to Noise Ratio ($\mathrm{SNR_C}$). The first step is to remove small clusters with out-of-range $\mathrm{SNR_C}$ values. Then, each removed value is replaced with an estimated value computed based on the nearest-neighbour interpolation. After that, a Standard Deviation (STD) map is created by comparing the value of each grid cell with the cells around it within a window. At the last step, random walker segmentation is applied to the STD map to create a water extent map. The parameters of this technique are optimized by comparing the detected water extent map with a hand-drawn watermask map. The full description of this method can be found in [72]. Since the CYGNSS data collection is random for having a suitable $\mathrm{SNR_C}$ map as the input for this method, at least three months' worth of the CYGNSS collected data is demanded. Moreover, the whole procedure must be applied step by step for creating each water extent map, making the process heavily computational. Hence, as another objective, a method for inland water detection using the CYGNSS data is developed.

1.3 The Scope of the Thesis

This thesis contains an investigation of methods for detecting surface water extents using CYGNSS data and ML algorithms with a focus on permanent water and flash floods.

The thesis is organized as follows: Chapter 2 provides the GNSS-R theoretical background and a description of the classification algorithms employed. In Chapter 3, a method for detecting permanent water using CYGNSS data and the RUSBoost ML algorithm is investigated. The CYGNSS data over the Congo and Amazon basins are employed as two case studies for permanent water detection. The results of the proposed method are compared to the watermask detection method proposed in

[72]. In Chapter 4, after feature selection, six different Data Preparation Approaches (DPAs) for flash flood detection based on the CYGNSS data are investigated using the RUSBoost ML algorithm. The DPA with the best performance is recommended for flash flood detection. Using selected DPA, the performance of the RUSBoost-based classifier is compared with the SVM-based classifier. The method with better performance is selected as the proposed flash flood detection method. For flash flood detection, the areas impacted by Hurricane Harvey and Hurricane Irma are considered as two case studies. Chapter 5 presents a summary of the main conclusions from the previous four chapters. A few suggestions for future work are also provided.

The research presented in this thesis has been accepted for publication in two refereed journals as listed below:

1. P. Ghasemigoudarzi, W. Huang, O. De Silva, Q. Yan and D. Power, "A Machine Learning Method for Inland Water Detection Using CYGNSS Data," in *IEEE Geosci. Remote Sens. Lett.*, (in press, DOI: 10.1109/LGRS.2020.3020223).

   This paper presents the RUSBoost-based method for permanent surface water detection based on the CYGNSS GNSS-R data (Chapter 3).

2. P. Ghasemigoudarzi, W. Huang, O. DeSilva, Q. Yan, and D. Power, "Flash Flood Detection from CYGNSS Data Using the RUSBoost Algorithm," in *IEEE Access*, vol. 8, pp. 171864–171881, 2020.

   This paper presents the RUSBoost-based method for flash flood detection from the CYGNSS GNSS-R data (Chapter 4).

Chapter 2

GNSS-R Theoretical Background and Classification Algorithms

In this chapter, first, a brief GNSS-R theoretical background is presented in Section 2.1. Then, a description of two machine learning algorithms that are used to develop the methods for permanent water and flash flood detection in this study are introduced in Section 2.2.

2.1 GNSS-R Theoretical Background

The CYGNSS creates power DDMs using the reflected and direct GPS L1-band Coarse/Acquisition (C/A) Pseudo Random Noise (PRN) codes. Each CYGNSS DDM consists of 17 delay bins and 11 Doppler bins with each delay bin equalling 249.4 ns (0.25 of a chip) and each Doppler bin equalling 500 Hz. In a bistatic configuration, the processed power DDM is described as [73, 74]

$$P^{rx}_{\tau,f} = \frac{\lambda^2 P^{tx}}{(4\pi)^3} \iint_{A_s} \frac{G^{tx}_{x,y} G^{rx}_{x,y} \sigma^0_{x,y}}{(R^{rx}_{x,y})^2 (R^{tx}_{x,y})^2} |\chi_{\tau,f;x,y}|^2 \; dxdy \tag{2.1}$$

where τ is the relative delay, f is the relative Doppler frequency, P^{tx} is the transmitted power, $G^{rx}_{x,y}$ and $G^{tx}_{x,y}$ are the transmitter and receiver gains, λ is the GPS wavelength, which is 19.05 cm, $R^{tx}_{x,y}$, and $R^{rx}_{x,y}$ are the distances between a point on the surface and the transmitter and receiver, respectively, A_s is the GZ, and $\sigma^0_{x,y}$ is the Normalized Bistatic Radar Cross Section (NBRCS) of the scattering surface, and $\chi_{\tau,f;x,y}$ is the Woodward Ambiguity Function (WAF) , which is obtained by [75, 76]

$$|\chi_{\tau,f;x,y}|^2 = (\Lambda_{\tau;x,y})^2 |S_{f;x,y}|^2 \tag{2.2}$$

where $\Lambda_{\tau;x,y}$ is the GPS signal spreading function in delay determining equi-delay zones and $S_{f;x,y}$ is the frequency response of the GPS signal determining equi-Doppler-frequency zones. Using $\chi_{\tau,f;x,y}$, the surface around an SP is gridded with delay-Doppler bins.

By considering the effective values of variables under the integral of (2.1) (except σ^0), the processed power of each delay-Doppler bin is simplified to [74]

$$P^{rx}_{\tau,f} = \frac{P^{tx} \lambda^2 G^{tx}_{\tau,f} \bar{G}^{rx}_{\tau,f} \langle \sigma^0_{\tau,f} \rangle \bar{A}_{\tau,f}}{(4\pi)^3 (\bar{R}^{tx}_{\tau,f})^2 (\bar{R}^{rx}_{\tau,f})^2} \tag{2.3}$$

where $\bar{G}^{rx}_{\tau,f}$ is the receiver antenna gain at each delay-Doppler bin, $\bar{R}^{tx}_{\tau,f}$ and $\bar{R}^{rx}_{\tau,f}$ are the range losses at each delay-Doppler bin, and $\bar{A}_{\tau,f}$ is the effective surface scattering area at each delay-Doppler bin. For each delay-Doppler bin we have

$$\langle \sigma_{\tau,f} \rangle = \langle \sigma^0_{\tau,f} \rangle \bar{A}_{\tau,f} \tag{2.4}$$

where $\langle\sigma_{\tau,f}\rangle$ is the Bistatic Radar Cross Section (BRCS) of each delay-Doppler bin. By substituting (2.4) into (2.3) and solving for $\langle\sigma_{\tau,f}\rangle$, we have

$$\langle\sigma_{\tau,f}\rangle = \frac{P^{rx}_{\tau,f}(4\pi)^3(R^{tx}_{SP})^2(R^{rx}_{SP})^2}{P^{tx}G^{tx}G^{rx}_{SP}\lambda^2}. \tag{2.5}$$

where R^{tx}_{SP} and R^{rx}_{SP} are the distances between the SP and the transmitter and receiver, and G^{rx}_{SP} is the receiver antenna gain at the SP. An example of BRCS DDM is shown in Figure 2.1(a). Unlike normalized BRCS that is only valid for ocean surfaces, $\langle\sigma\rangle$ is valid for both land and ocean since it is computed based on geometrical and instrumental corrections [70].

Depending on the surface roughness, the GNSS signal can be reflected coherently or incoherently. Therefore, the computed power for each delay-Doppler bin described by (2.3) is a summation of coherent and incoherent power components. [75, 77]

$$P^{rx}_{\tau,f} = P^{coh}_{\tau,f} + P^{inc}_{\tau,f}. \tag{2.6}$$

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 have low roughness and the reflected signals from them are predominantly coherent. Thus, the reflection from the water bodies surrounded by dense biomass can be considered coherent, which is the case for our permanent water case studies (the Amazon and Congo basins). On the other hand, during severe conditions such as a hurricane, the high-speed winds can increase the roughness of inland surface water bodies. However, the presence of high-speed winds of a hurricane over land is shorter than the flood caused by its landfall. In addition, as a hurricane reaches land, the winds speed decreases gradually due to the increased surface roughness [78, 79]. Moreover, the coherent components of received power during severe typhoons are consistent with the changes caused by floods, as investigated in [69]. Therefore, following similar assumptions made in the literature [65, 69, 70], in this work, it is assumed that the reflected signals for both permanent water and flash floods are coherent. Based on the Friis radar equation, the coherent power component is given as [77]

$$P_{\tau,f}^{coh} = \frac{P^{tx} G^{tx} G_{SP}^{rx} \lambda^2}{(4\pi)^2 ((R_{SP}^{tx}) + (R_{SP}^{rx}))^2} \Gamma_{\tau,f}, \tag{2.7}$$

where $\Gamma_{\tau,f}$ is the SR DDM.

Considering $P_{\tau,f}^{rx} = P_{\tau,f}^{coh}$ and substituting (2.7) into (2.5), $\Gamma_{\tau,f}$ can be found as

$$\Gamma_{\tau,f} = \frac{(R_{SP}^{tx} + R_{SP}^{rx})^2}{4\pi (R_{SP}^{tx})^2 (R_{SP}^{rx})^2} \langle \sigma_{\tau,f} \rangle. \tag{2.8}$$

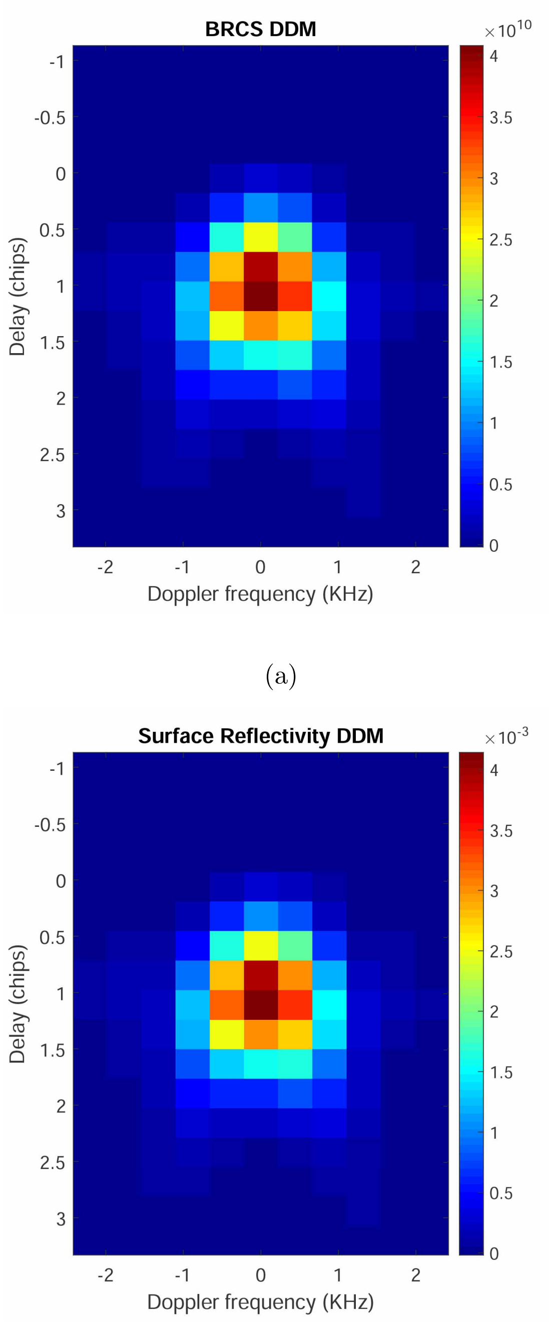
Similar to [65, 70, 80], the BRCS DDM is used for computing the SR DDM described by (2.8). The corresponding SR DDM calculated from BRCS DDM shown in Figure 2.1(a) is depicted in Figure 2.1(b).

In terms of spatial resolution, the maximum area that a CYGNSS DDM can cover is 25 km × 25 km around the SP [39]. This area consists of both coherent and incoherent reflections. However, the majority of coherent reflections happen within the First Fresnel Zone (FFZ). The semi-major and semi-minor axes of the FFZ around each SP are defined as [81, 82]

$$a_{FFZ} = \frac{1}{\cos(\theta)} \left(\frac{R_{SP}^{tx} R_{SP}^{rx} \lambda}{R_{SP}^{tx} + R_{SP}^{rx}} \right)^{0.5}, \tag{2.9}$$

$$b_{FFZ} = \left(\frac{R_{SP}^{tx} R_{SP}^{rx} \lambda}{R_{SP}^{tx} + R_{SP}^{rx}} \right)^{0.5}. \tag{2.10}$$

where θ is the incidence angle at SP. Using the CYGNSS data, the semi-major and semi-minor axes for SPs with incidence angles between 0° and 70° are shown in



(a)

(b)

Figure 2.1: (a): an example of BRCS DDM, and (b): its SR DDM calculated from (2.8).

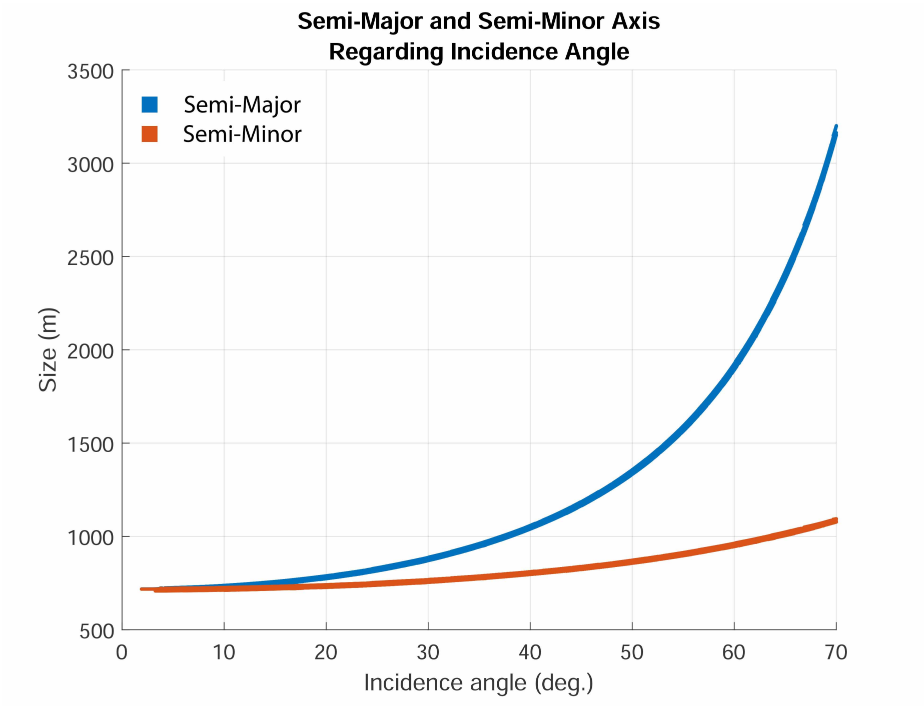


Figure 2.2: The dimensions of semi-major and semi-minor axes of the FFZ with respect to incidence angle using the CYGNSS data.

Figure 2.2. Both axes start at around 700 m near $\theta = 0°$. As θ increases, the semi-major and semi-minor axes increase till at $\theta = 70°$ they reach 3190 m and 1090 m, respectively. Considering the case where the semi-major axis aligns with the satellite's along-track direction, the along-track footprint varies between 6.6 km and 8.8 km [70]. Hence, the final size of the FFZ ellipse varies from 700 m × 6.6 km to 1090 m × 8.8 km for incidence angles less than 70°.

Even though the first Fresnel zone is the area where the majority of reflections are coherent, it cannot determine the spatial resolution of a DDM. Since the reflection is highly dependent on surface roughness and geometry, the resolution is dynamic. In this work, a predetermined area in the range of FFZ around each SP or for each grid cell is considered as corresponding region for coherent reflections and spatial resolution.

Before proceeding to the next section, it is worth mentioning that in addition to (2.7), another approximation for coherent reflected power is suggested in [67] that describes the reflected signal from heterogeneous smooth surfaces with the surface diffraction integral given as

$$P_{\tau,f}^{coh} = \Gamma_{\tau,f} \frac{P^{tx} \lambda^2 G^{tx} G_{SP}^{rx}}{(4\pi)} \left| \iint_S \frac{jk \cos(\theta)}{2\pi (R_{x,y}^{tx} R_{x,y}^{rx})^2} e^{jk(R_{x,y}^{tx} + R_{x,y}^{rx})} dxdy \right|^2 , \tag{2.11}$$

where S is an area around SP and k is the signal wave number. The surface diffraction integral is calculated over an area larger than the FFZ. In contrast, we assume that the received signal is reflected from an area within the range of FFZ. Moreover, our case studies consist of both rough and smooth surfaces. Therefore, here, (2.7) is considered for deriving $\Gamma_{\tau,f}$.

More information about the GNSS-R theory can be found in [38, 39, 74, 83].

2.2 Classification Algorithms

The water detection problem is a binary classification problem with two classes (water/flood and land). Various ML algorithms can be implemented for solving a binary supervised classification problem, such as the Neural Networks (NN), SVMs, and Decision Trees, which are among the most commonly used classifiers in remote sensing [84]. 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 [85].

Since approximately 5.8% of the land is covered by surface water bodies [86] and flash floods are localized, when a large area is considered, the number of points collected over water bodies is much smaller than those obtained from land. This creates an imbalanced 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 [87]. 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 [87]. 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 [87]. 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 [88, 89]. As a powerful tool, the Generative Adversarial Network (GAN) is another method for creating artificial instances in the minor class [90]. In such a method, two neural networks compete to optimize their objective functions that are contradictory to each other [91]. The Random Under-Sampling (RUS) is an under-sampling method that balances the data via random elimination of instances from the major class [87]. The balancing techniques are applied to different classifiers, such as ensembles methods, leading to various developed algorithms for classifying imbalanced data [87]. Among different methods, in this thesis, the

RUSBoost algorithm is selected for classification due to its efficient computational time, accuracy, and widely available resources [92–94]. Moreover, in Chapter 4, the performance of the proposed method for flash flood detection is compared with that of an SVM-based classifier for better perception.

2.2.1 RUSBoost

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 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 [95] is depicted in Figure 2.3. The training data set, is the imbalanced set $S = \{(\mathbf{x}_i, y_i) \mid i = 1, ..., m\}$, in which $\mathbf{x}_i = [x_{i,1}, ..., x_{i,J}]$ 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 features and y_i can be either land (0) or water/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 $S'_t = \{(\mathbf{x'}_p, y'_p)| \; p = 1, ..., 2n\} \subset S$ 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

```
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 $c_t^k(q_k)$ that minimizes the Gini impurity factor.
      ii Return $N_{r,l}(y)$ and $N_{r,l}$ regarding $S'$ and $c_t^k(q_k)$
    c Calculate the label proportion
      $\pi_{r,l} = N_{r,l}(y)/N_{r,l}$.
    d Create a weak hypothesis
      $h_t(\mathbf{x}_i, y) = \begin{cases} \pi_r(y) & \text{if } x_{i,k} > c_t^k(q_k), \\ \pi_l(y) & \text{otherwise.} \end{cases}$
    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 $D_{t+1}$
      $D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum\limits_{i=1}^{m} D_{t+1}(i)}$.
  3 Output the final hypothesis
    $H(\mathbf{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \sum_{t=1}^{T} h_t(\mathbf{x}, y) \log \frac{1}{\alpha_t}$.
```

Figure 2.3: The Pseudo-code of the RUSBoost algorithm recreated from [95].

employed for training weak learners based on the idea of reducing the classification error iteratively (step 2a) [95].

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, $(j = 1, ..., J)$, 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. Considering the features with 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 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 $c^j_t(q)$, $(q = 1, ..., 2n+1)$, the weighted Gini Impurity Factor (GI_t) of the decision stump t is obtained as:

$$\mathrm{GI}_t(q) = \Omega^r_t(q)\Theta^r_t(q) + \Omega^l_t(q)\Theta^l_t(q) \tag{2.12}$$

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:

$$\Omega^r_t(q) = \sum_{p=1}^{2n} D'_t(p)[[x'_{p,j} > c^j_t(q)]] \tag{2.13}$$

$$\Theta_t^r(q) = 1 - \sum_y \theta_t^r(y), \tag{2.14}$$

where

$$\theta_t^r(y) = \left(\frac{\sum\limits_{p=1}^{2n} D_t'(p)[[y_p' = y]][[x_{p,j}' > c_t^j(q)]]}{\sum\limits_{p=1}^{2n} D_t'(p)[[x_{p,j}' > c_t^j(q)]]} \right)^2 \tag{2.15}$$

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 (2.14) and (2.13) to $[[x_{p,j}' \leq c_t^j(q)]]$ [96]. It is worth mentioning that $\theta_t^r(y)$ is a weighted probability that determines how likely a y labelled point is located in the right split.

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) [97]. 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)

$$h_t(\mathbf{x}_i, y) = \begin{cases} \pi_r(y) & \text{if } x_{i,k} > c_t^k(q_k), \\ \pi_l(y) & \text{otherwise} \end{cases} \tag{2.16}$$

where $\pi_{r,l} = N_{r,l}(y)/N_{r,l}$ is the label proportion, which is the ratio between the number of $y \in \{0, 1\}$ labelled 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)

$$\epsilon_t = \sum_{i=1}^{m} D_t(i)(1 - h_t(\mathbf{x}_i, y_i) + h_t(\mathbf{x}_i, 1 - y_i)), \tag{2.17}$$

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

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}. \tag{2.18}$$

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

$$D_{t+1}(i) = D_t(i)\alpha_t^{\frac{1}{2}(1+h_t(\mathbf{x}_i,y_i)-h_t(\mathbf{x}_i,1-y_i))}, \tag{2.19}$$

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum\limits_{i=1}^{m} D_{t+1}(i)}. \tag{2.20}$$

When the hypothesis of the weak learner is correct for the training data set, which means that the weak learner is 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):

$$H(\mathbf{x}) = \underset{y\in\{0,1\}}{\operatorname{argmax}} \sum_{t=1}^{T} h_t(\mathbf{x}, y) \log \frac{1}{\alpha_t}, \tag{2.21}$$

where $\mathbf{x}$ is a feature vector of the test data. The criteria for the trained RUSBoost classifier are to find the label that maximizes the summation of the hypothesis of weak learners with respect to α_t [95, 98]. 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.

2.2.2 SVM

In Chapter 4, 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 [57, 99–101]. SVMs classify data by determining the optimal hyperplane for maximizing the margin between classes [102, 103]. 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 [102, 103].

Chapter 3

Permanent Water Detection

In this chapter, a permanent water detection method is presented based on the RUSBoost algorithm described in Section 2.2.1 and the CYGNSS data. The results of the proposed method are compared with those of the watermask detection method considering two different regions as two case studies. This chapter is outlined as follows: Section 3.1 introduces employed data sets. In Section 3.2, four different CYGNSS observables and the RUSBoost-based algorithm are described. Section 3.3 provides the results of the classification and the comparison between RUSBoost-based classifier and watermask detection methods. General chapter summary is provided in Section 3.4.

3.1 Data Sets

3.1.1 CYGNSS

The CYGNSS daily data are posted two days after their collection and are available for the public through [47]. In this chapter, the level 1 V2.1 data of the year 2018 over the Congo and Amazon basins are employed. These two regions are located in geographic coordinates of $[2^\circ\,\mathrm{S}, 3^\circ\,\mathrm{N}][17^\circ\,\mathrm{E}, 27^\circ\,\mathrm{E}]$ and $[10^\circ\,\mathrm{S}, 0^\circ\,\mathrm{N}][50^\circ\,\mathrm{W}, 75^\circ\,\mathrm{W}]$, respectively.

3.1.2 Global Surface Water

In the current chapter, the GSW seasonality data of the year 2018, which has a high-resolution ($30\,\mathrm{m} \times 30\,\mathrm{m}$) and has been employed in various studies as reliable sources, is used as the reference for labelling and evaluation. The GSW seasonality data describes the inter-annual behaviour of surface water and such a data set was created based on the optical images taken by Landsat during 2018 [10]. The permanent water Geographic Information System (GIS) data maps are publicly available through [35]. We used the QGIS and MATLAB to convert the GIS data to TIFF maps and assign each $30\,\mathrm{m} \times 30\,\mathrm{m}$ grid cell with a value of 0 or 1 where 0 means land without water content and 1 is water. Since the CYGNSS resolution is worse than that of GSW seasonality data, values of $30\,\mathrm{m}\times 30\,\mathrm{m}$ grid cells located within a $0.01^\circ \times 0.01^\circ$ cell are averaged and compared with a threshold. If the average value of a cell is higher/lower than the threshold, such a cell is labeled as water/land. Empirically, 0.6 was found to be the optimal threshold.

3.2 Water Detection Methodology

3.2.1 CYGNSS Observables

In this chapter, four different observables, including SNR_C, kurtosis, maximum, and variance are computed for each SP. All the observables except SNR_C are computed using the SR DDM described by equation (2.8).

- The SNR_Peak, which is the ratio of the maximum value in a DDM to average noise per bin ($10\log(S_{max}/N_{avg})$), is provided in the CYGNSS data set. This value is then corrected to SNR_C via [68]

$$\text{SNR}_\text{C} = 10\log\left(\frac{(R_{tx}+R_{rx})^2\lambda^2}{P_{tx}G_{tx}G_{rx}(4\pi)^2}10^{\text{SNR}_\text{Peak}/10}\right). \tag{3.1}$$

As in [68, 72], in order to make the values of SNR_C intuitively meaningful, the mean of the lowest 5% of SNR_C values, which is equal to 139.7 dB, was subtracted from them.

- The maximum is the maximum value of the SR DDM.

The surface roughness changes the pattern of scattered signals. Eventually, it will change the statistical characteristics and histogram of a DDM. The statistical moments including mean, variance, kurtosis, and skewness, are general parameters that can describe the shape of a histogram of Random Variables (RV). By considering the magnitude of the SR DDM as an RV, the statistical moments can indicate a comprehensive explanation about the impact of water on its histogram [104]. The maps of the four statistical moments were created and by comparing them to the water reference map, we found that among them, the variance and kurtosis are more correlated with the watersheds and permanent water bodies. The variance and kurtosis are defined as follows

- Variance, which is a measure of the deviation of an RV from its mean and shows how the distribution of an RV is centralized around the mean, is given by

  $$variance = \mathrm{E}[(X - \mathrm{E}[X])^2] = \frac{1}{187}\sum_{i=1}^{17}\sum_{j=1}^{11}(\Gamma_{\tau_i,f_j} - mean)^2, \qquad (3.2)$$

  where

  $$mean = \mathrm{E}[X] = \frac{1}{187}\sum_{i=1}^{17}\sum_{j=1}^{11}\Gamma_{\tau_i,f_j}. \qquad (3.3)$$

- Kurtosis is a measure of the tailedness of the distribution of an RV. In other words, kurtosis assesses the data points that are outside of the standardized data region (outliers) and is defined as

  $$kurtosis = \mathrm{E}\left[\left(\frac{X - mean}{\sqrt{variance}}\right)^4\right] = \frac{\sum\limits_{i=1}^{17}\sum\limits_{j=1}^{11}(\Gamma_{\tau_i,f_j} - mean)^4/187}{(variance)^2}. \qquad (3.4)$$

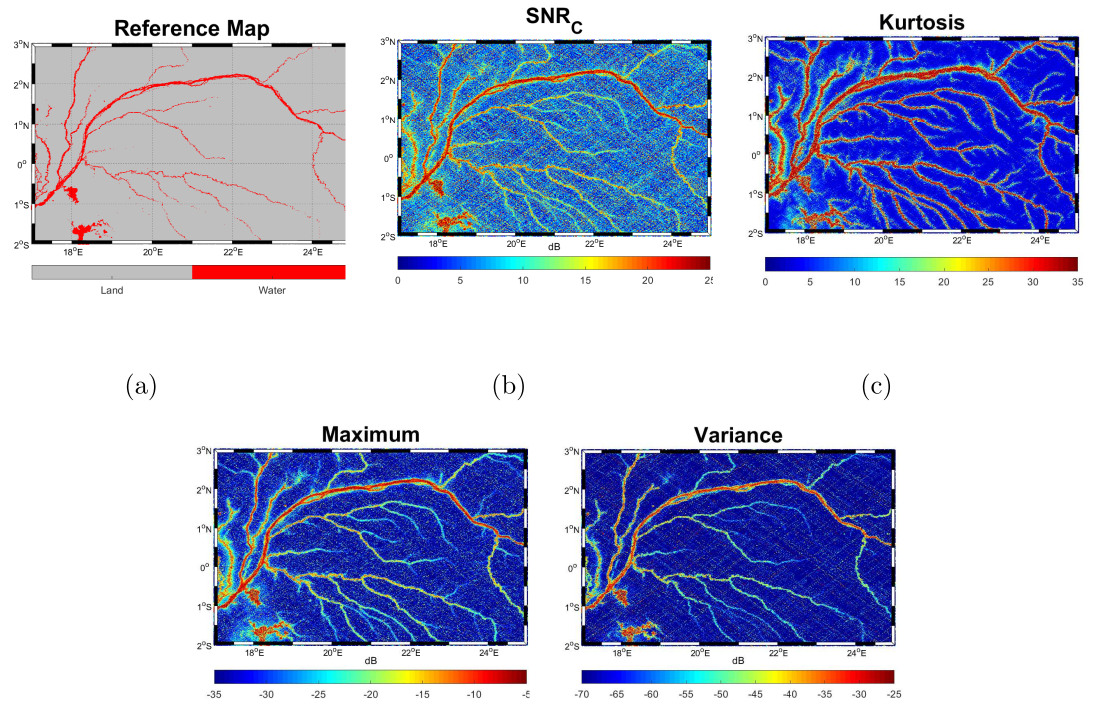
Since the ranges of features are different, as a part of the data cleaning step, the values of maximum, variance, and kurtosis are normalized. The minimum to maximum normalization ranges of maximum, variance, and kurtosis are obtained based on self-observations and are equal to $[-35\,, -5]$, $[-70\,, -25]$, and $[0\,, 35]$, respectively. The $\mathrm{SNR_C}$ is not normalized since it is not used with other features and is only utilized in watermask detection method.

In order to conduct a fair comparison between the proposed method and that in [72], the extracted CYGNSS observables are gridded into a $0.01^\circ \times 0.01^\circ$ cell as done in [72]. For each observable, values of SPs that are located in the same grid cell are averaged.

The surface water reference and calculated observables maps of the Congo basin are depicted in Figure 3.1. Comparing the water reference map with the maps of the extracted observables indicates that the $\mathrm{SNR_C}$ map is noisier, meaning that there are land grid cells whose $\mathrm{SNR_C}$ value is in the same range as those of water grid cells. On the other hand, the kurtosis values of most land grid cells are less than 5. Also, the grid cells with high kurtosis values correspond to the network of land areas with low altitudes that might contain and channel the surface water bodies or in other words the watersheds. However, there is a similar pattern in the $\mathrm{SNR_C}$, maximum, variance maps that correspond to the permanent water whose extent map is depicted in Figure 3.1(a). The watermask detection method attempts to detect this pattern that grid cells with high value $\mathrm{SNR_C}$ create. On the other hand, the RUSBoost-based classifier determines the label of each grid cell only by its corresponding observable values.

3.2.2 Classifier

Classification of each grid cell as water or land is determined by applying the RUSBoost-based classifier to GNSS-R data. While selecting the training and testing data sets, a random shuffling and a random selection are applied to the data from the considered region. Therefore, the training data set consists of the data points from all over the considered region. Thus, the trained classifier is not over-fitted to any specific areas within the considered region. Using RUSBoost method, at the first step each



(a) (b) (c)

(d) (e)

Figure 3.1: The Congo basin (a): permanent water reference map obtained from the GSW seasonality data, (b): SNR$_C$ map, (C): kurtisos map, (d): maximum map, and (e): variance map.

data point in the training data set is assigned to a weight initialized with an equal value. Each iteration is allocated with a separate weak learner. At each iteration, an imbalanced training data set is randomly undersampled creating a balanced data set [95]. The balanced data set is then passed through the weak learner of that iteration, which creates a weak hypothesis [98]. 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 the training data points correctly, the weights of observations remain unchanged. Otherwise, the weights of the misclassified points will be higher than those of correctly classified points. Therefore, in the next iteration, the weak learner will be biased to classify the misclassifications of the previous weak learner, which translates to increasing the variance step by step. This process that consists of random undersampling, creating a weak hypothesis, and updating observations weights is repeated for each iteration. At the last iteration, when all the weak learners are trained, the output hypothesis is created as a weighted vote of weak hypotheses.

In this chapter, RUSBoost-based classification is conducted using MATLAB R2018 software. The decision stump is chosen as the weak learner, and 150 of decision stumps are trained with a learning rate of 0.1 in 150 iterations. The output hypothesis is then used for testing and validation with other separate data sets. The block diagram of the RUSBoost-based classifier for permanent water detection is shown in Figure 3.2.

3.3 Results and Discussion

The CYGNSS data include high altitude measurements and noisy DDMs that must be discarded in the data cleaning stage. In this thesis, we refer to all discarded

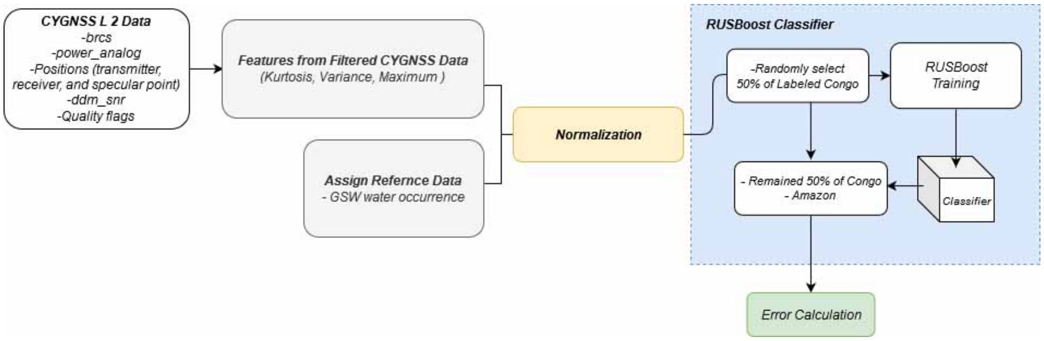


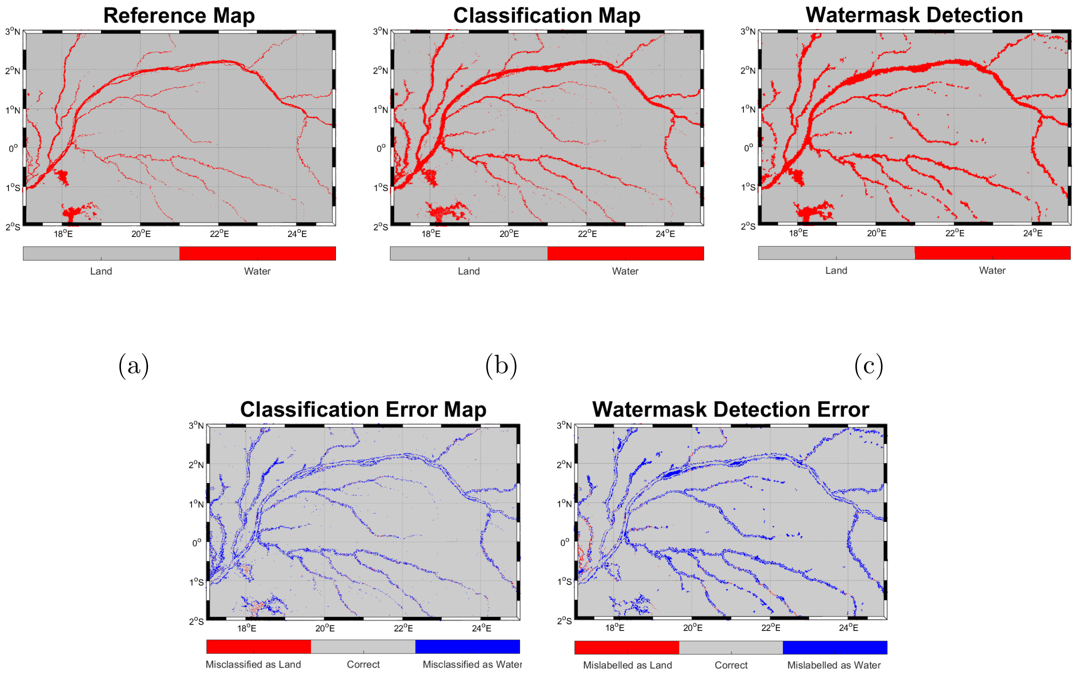
Figure 3.2: Block diagram of the RUSBoost classification. After training the classifier with 50% of the Congo data, it is applied to the remaining 50% and the Amazon data.

DDMs as noisy DDMs. In order to eliminate noisy DDMs, the CYGNSS quality flags, mentioned in Table 3.1, are applied to the data in the preprocessing step [80]. Moreover, it has been shown in [105] that when the incidence angle of an SP is between 15° and 60°, the received signal is more sensitive to the surface water content. Hence, SPs with incidence angles out of this range are removed.

Among the four observables extracted from the CYGNSS data, $\mathrm{SNR_C}$ is used for implementing the comparison watermask detection method described in [72]. The kurtosis, maximum, and variance are used as inputs of the RUSBoost-based classifier. Each grid cell is labeled using the GSW data. As a supervised classification, the RUSBoost-based classifier is trained with 50% of the CYGNSS data (144791 data points) from the Congo basin. The remaining 50% is used for testing. Since the parameters of the watermask detection method are optimized for the Congo basin, in order to compare the two methods, we also used the data from the same area for training and testing the classifier [72]. To further investigate the capability of these two methods, they are applied to the data from the Amazon basin that it is not included in the training data and is unknown to the two methods.

Table 3.1: CYGNSS data quality flags (suggested in [80]) considered in this thesis.

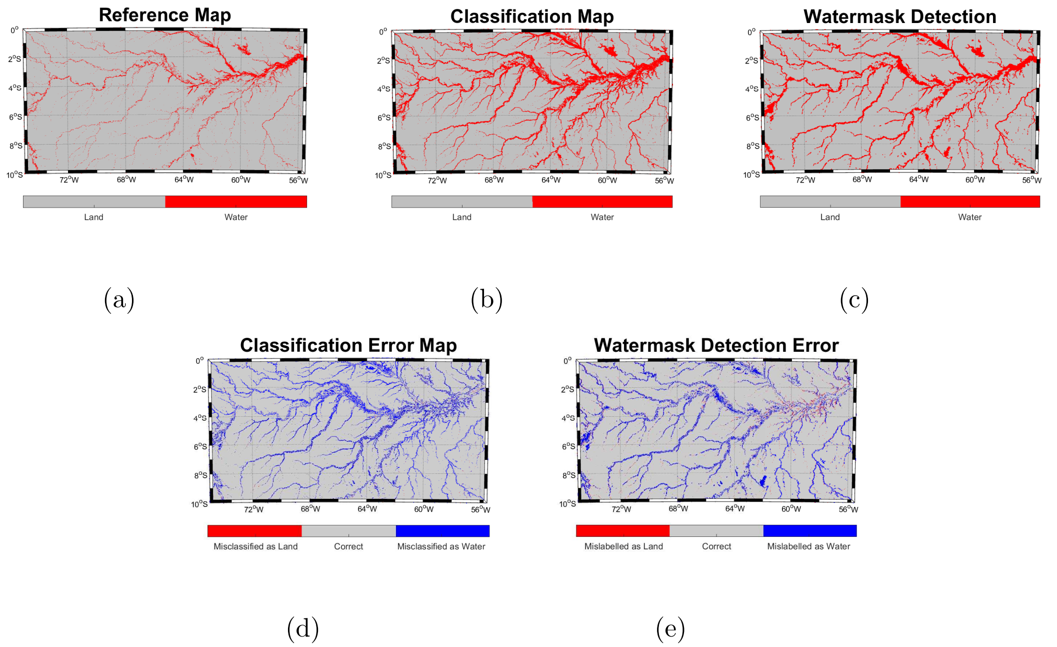
Quality Flag	Flagged in Analysis
Poor Overall Quality	No
S Band Powered Up	Yes
Small Spacecraft Attitude Error	No
Large Spacecraft Attitude Error	Yes
Blackbody DDM	Yes
DDMI Reconfigured	Yes
Spacewire CRC Invalid	Yes
DDM is Test Patten	Yes
Channel Idle	Yes
Low Confidence DDM Noise Floor	No
SP Over Land	No
SP Very Near Land	No
SP Near Land	No
Large Step Noise Floor	No
Direct Signal in DDM	Yes
Low Confidence GPS EIRP Estimate	Yes
Radio-frequency Interference Detected	Yes
BRCS DDM SP Bin Delay Error	No
BRCS DDM SP Bin Doppler Error	No
Negative BRCS Value Used for NRBCS	No
GPS Position, Velocity and Time SP3 Error	No
SP Non Existent Error	Yes
BRCS Look-up Table Range Error	No
Antenna Data Look-up Table Range Error	No
Blackbody Framing Error	Yes



(a) (b) (c)

(d) (e)

Figure 3.3: The Congo basin (a): the permanent water reference map using the GSW data, (b): the RUSBoost classification map, (C): the result map of the watermask detection method, (d): the RUSBoost error map classification, and (e): the error map of the watermask detection method. The areas within the two dashed rectangles include small land regions that are surrounded by water bodies.



(a)

(b)

(c)

(d)

(e)

Figure 3.4: The Amazon basin (a): the permanent water reference map using the GSW data, (b): the RUSBoost classification map, (C): the result map of the watermask detection method, (d): the RUSBoost error map classification, and (e): the error map of the watermask detection method.

Table 3.2: Accuracies of Inland Water Detection Methods.

#	Case	Method	Class	Accuracy
1	Congo Basin	RUSBoost Classification	Land	94.63 %
			Water	**95.4** %
2	Congo Basin	Watermask Detection	Land	94.8 %
			Water	91.5 %
3	Amazon Basin	RUSBoost Classification	Land	91.9 %
			Water	**93.3** %
4	Amazon Basin	Watermask Detection	Land	92.2 %
			Water	79.1 %

The GSW seasonality reference map, labeling and classification results, and error maps of the two methods over the Congo and Amazon basins are depicted in Figures 3.3 and 3.4, respectively. The accuracies of the watermask detection method and the RUSBoost-based classifier are shown in Table 3.2. The comparison between the RUSBoost-based method and the watermask detection method indicates that the former has a better performance regarding water detection. When both methods are applied to the Amazon basin, the difference is more significant and the proposed method outperforms by 14.2% in water detection accuracy. In terms of land detection, the watermask detection method has a slightly higher accuracy. The differences of land detection accuracy for the Congo and Amazon basins is 0.17% and 0.3%, respectively.

The two dashed rectangles in Figures 3.3(d) and 3.3(e) indicate small land areas that are surrounded by water. By comparing these two maps, when a small land area is localized within water bodies, the watermask detection method mislabels it as water, while the RUSBoost-based method classifies it correctly. Therefore, even though the presence of water is overestimated, the resulting map of the RUSBoost-based classifier is more precise.

Since the proposed method provides more precise water maps and its water detection accuracy is higher, it has a better overall performance than the watermask detection method described in [72].

3.4 General Chapter Summary

In this chapter, a high-resolution water detection technique based on CYGNSS data and the RUSBoost algorithm is presented. By considering three observables, the classifier was trained with half of the CYGNSS data from the Congo basin. Then the classifier was tested with the remaining half of the data. To further evaluate the performance of the classifier, it was applied to the data of the Amazon basin that is unknown to the machine. Moreover, the results of the two cases are compared with the watermask detection method proposed in [72].

Chapter 4

Flash Flood Detection

In this chapter, six different DPAs for detecting flash floods using CYGNSS data and RUSBoost ML algorithm are investigated. Eleven different observables are extracted for each SP. By considering five features, the optimum flood threshold is determined. Based on the optimum flood threshold, all different features combinations are investigated to find a suitable one for flash flood detection. Using the selected features, the performance of the classifier for each approach is computed and by comparing the results, the best one is proposed as the recommended approach for flash flood detection. Then, the performance of the RUSBoost classifier is compared with an SVM classifier using the recommended approach. This chapter is outlined as follows: first, employed data and reference assignment are discussed in Section 4.1; then, the methodology for flash flood detection is described in Section 4.2. Section 4.3 presents the feature selection, flood detection results, and a comparison between two classifiers. Lastly, a general summary of the chapter is provided in Section 4.4.

4.1 Data Sets

4.1.1 CYGNSS

Similar to Chapter 3, in this chapter, we employed level 1 V2.1 of the CYGNSS data [106] that are available for the public through [47].

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 as mentioned in Section 1.1. 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 [107].

Hurricane Harvey reached the coast of the USA on Aug. 25th, 2017, and according to media, the inundation lasted until Sept. 8th, 2017. Hurricane Irma landed on 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 within geographic coordinates of $[26.7^\circ\,\mathrm{N}, 32.29^\circ\,\mathrm{N}][91^\circ\,\mathrm{W}, 100^\circ\,\mathrm{W}]$ and $[24.5^\circ\,\mathrm{N}, 29.2^\circ\,\mathrm{N}][79.2^\circ\,\mathrm{W}, 93^\circ\,\mathrm{W}]$, 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 4.3.1, for feature selection, both data sets are combined and used for classification by a 5-fold cross-validation evaluation. Furthermore, in Section 4.3.2, 50% of the Harvey data is used 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.

4.1.2 Ancillary Data

In this chapter, 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 [12]. In this chapter, 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 GIS data. The GIS data of Hurricane Harvey and Hurricane Irma and more details on them are available through [108], and [109], 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 ancillary data [110]. 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, which leads to overestimation.

Table 4.1: Summary of employed data sets.

Data Set	Resolution	Spatial Coverage	Temporal Coverage	Accuracy
CYGNSS [106]	incoherent: 25 km coherent: dynamic	38° S to 38° N	Daily from March 13, 2017	NA
DFO [12, 109]	< 250 m	50° S to 70° N	Flood Events 2000 - Present	Geolocation accuracy: ±50 m
SRTM [110]	90 m	56° S to 60° N	11-22 February 2000	Vertical accuracy 6 m
CIFOR [111]	236 m	60° S to 40° N	2009-2017	NA
GSW [10, 35]	30 m	50° S to 80° N	1984-2019	98%

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) [111] and GSW Occurrence data [10, 35], 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 [112]. Even though this data set is not validated due to the unavailability of ground truth, it agrees well with other commonly used data sets [112]. The GSW data set is generated based on optical images collected by Landsat [10]. 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 4.1. 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 chapter, 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 [70], 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 a threshold, it is labeled as flood; otherwise, it is labeled as land. For SRTM90m DEM, the 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.

4.2 Flash Flood Detection Methodology

4.2.1 CYGNSS Observables

In this chapter, instead of working with the whole DDM, eleven different observables including SNR_C, Trailing-edge Slope (TES), Leading-edge Slope (LES), Delay-Doppler Map Average (DDMA), the Width of the Waveform (Wave-width) defined in the fourth bullet point in the following, the first Generalized Linear Observable (GLO_1), kurtosis, maximum, mean, skewness, and variance are extracted for each SP. All the observables except SNR_C are computed using the SR DDM, which is calculated as described in [70]. The first seven observables (SNR_C, LES, TES, DDMA, Wave-width, GLO_1, and maximum) are obtained as follows [70, 113]:

- In this chapter, another equivalent correcting expression is used to calculate SNR_C, as suggested in [70]

$$\text{SNR}_\text{C} = \frac{(R_{tx} + R_{rx})^2 \lambda^2 \langle \sigma_m \rangle}{P_{rxm} R_{tx}^2 R_{rx}^2 (4\pi)^3} \text{SNR}_\text{Peack} \tag{4.1}$$

  where $\langle \sigma_m \rangle$ is the maximum value of BRCS DDM, and P_{rxm} is the maximum value in power DDM [70]. 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.

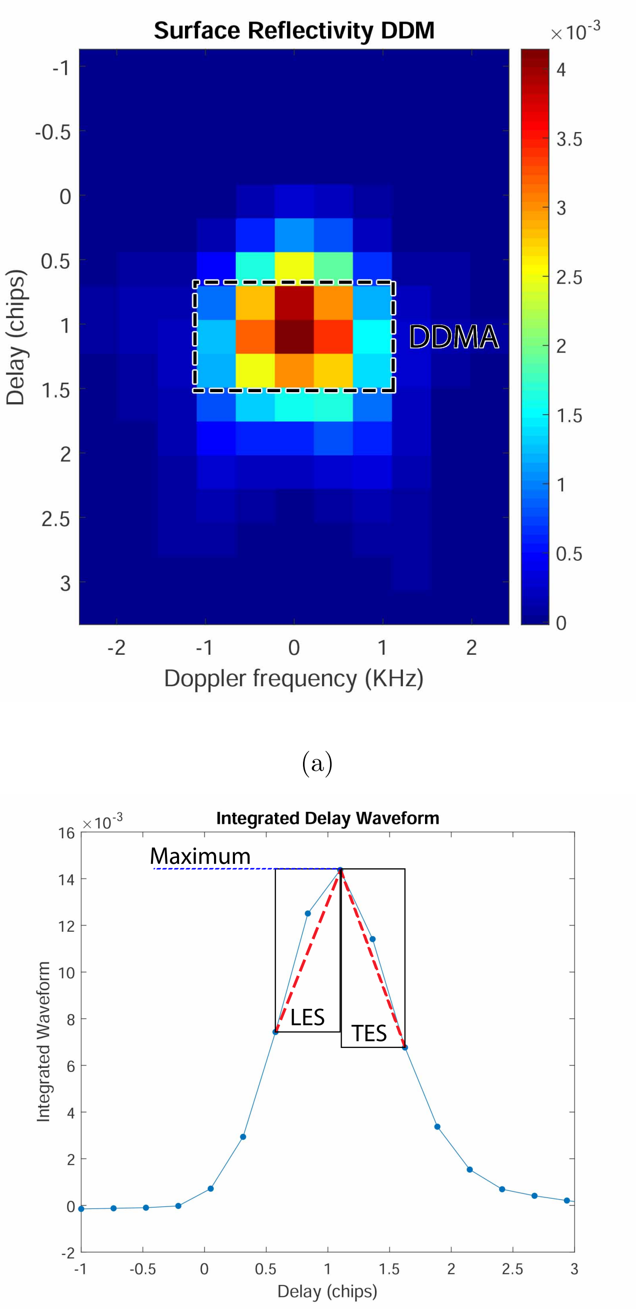
- As depicted in Figure 4.1(b), LES and TES are computed as the slopes between the maximum point and the points at two delay bins before and after the maximum point in the SR Integrated Delay Wavefrom (IDW) [70].

- The DDMA is the arithmetic mean of SR DDM within a window around the maximum value. In this thesis the size of the window is chosen as 3 delay bins × 5 Doppler bins as shown in Figure 4.1(a). The DDMA is described as [114]

$$\text{DDMA} = \frac{1}{15} \sum_{i=i_{max}-1}^{i_{max}+1} \sum_{j=j_{max}-2}^{j_{max}+2} \Gamma_{\tau_i, f_j}, \tag{4.2}$$

  where i_{max} and j_{max} are the delay and Doppler bins of the maximum SR value in the SR DDM, respectively.

- 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) [70].



(a)

(b)

Figure 4.1: (a): An example of SR DDM described by (2.8), and (b): its SR IDW. A 3 × 5 window is considered for calculating DDMA.

- The Nth Generalized Linear Observable (GLO) is defined as [115]:

$$\mathrm{GLO}_N = \sum_{i=i_{max}-3}^{i_{max}+3} a_i^N \Gamma_{\tau_i}^{del}, \tag{4.3}$$

  where Γ^{del} is SR IDW, a_i^N is the Nth weight of SR in the delay bin i and it is computed by applying Principal Component Analysis (PCA) to the SR IDW. The summation is calculated considering ±3 delay bins around the delay bin of the maximum of SR IDW (i_{max}). We only consider the first GLO (GLO_1), since it has been proven that it is more correspondent to the inundations over land [70].

- The Maximum that is the maximum value of the SR IDW is also considered another feature.

As mentioned in Chapter 3, the impact of flood on a DDM can be studied by considering the SR IDW (or SR DDM) as an RV and analyzing its statistical moments. In this chapter, all statistical moments including mean, variance, kurtosis, and skewness are considered as observables, which are obtained as

- Mean shows the position of the central mass of an RV;

$$mean = \mathrm{E}[X] = \frac{1}{17} \sum_{i=1}^{17} \Gamma_{\tau_i}^{del}, \tag{4.4}$$

- Variance is the squared differences of an RV from its mean;

$$variance = \mathrm{E}[(X - \mathrm{E}[X])^2] = \frac{1}{17} \sum_{i=1}^{17} (\Gamma_{\tau_i}^{del} - mean)^2, \tag{4.5}$$

- Skewness is an indicator of the asymmetry of the probability distribution of an RV. When the distribution is symmetrical, skewness equals zero. In asymmetrical distributions, when the skewness is negative, the tail of the distribution is on the left side of the mean, but when the skewness of the distribution is positive, the tail is on the right side of the mean;

$$skewness = \mathrm{E}\left[\left(\frac{X - mean}{\sqrt{variance}}\right)^3\right] = \frac{\sum\limits_{i=1}^{17}(\Gamma_{\tau_i}^{del} - mean)^3/17}{(variance)^{3/2}} \tag{4.6}$$

- Kurtosis that estimates the tailedness of the shape of a histogram;

$$kurtosis = \mathrm{E}\left[\left(\frac{X - mean}{\sqrt{variance}}\right)^4\right] = \frac{\sum\limits_{i=1}^{17}(\Gamma_{\tau_i}^{del} - mean)^4/17}{(variance)^{2}}. \tag{4.7}$$

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.

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 4.2. 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.

Table 4.2: Ranges of observables in normalization step.

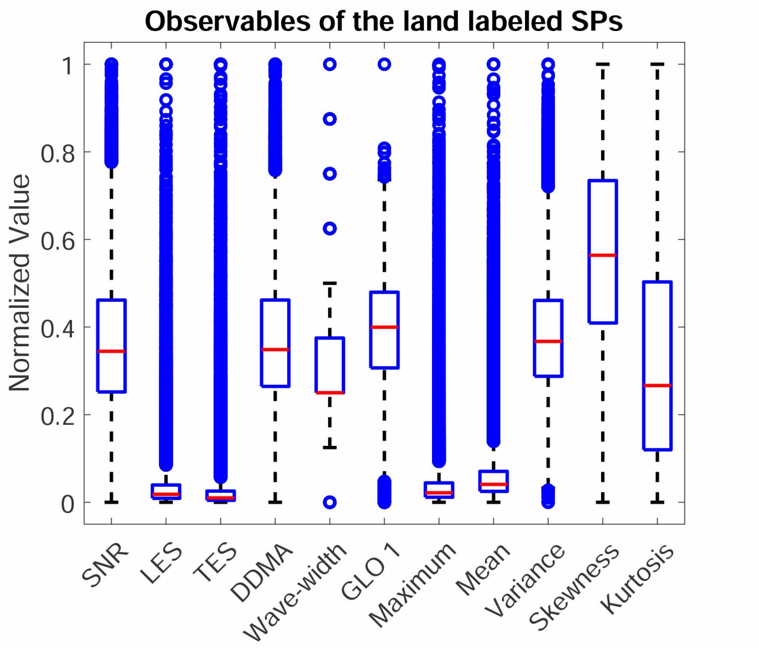
Parameter	Min	Max	Parameter	Min	Max
DDMA	3	12	LES	0	0.35
TES	0	0.4	Wave-width	1	9
$\mathrm{SNR_C}$ (dB)	105	130	$\mathrm{GLO_1}$	-35	-5
Kurtosis	1.5	4.5	Skewness	-1	1.8
Mean	0	0.1	Maximum	0	0.5
Variance(dB)	-70	-10			

Depending on the labels of the SPs, their observables show different characteristics as depicted by the box plots in Figure 4.2, for which the SPs located over permanent water bodies and wetlands are excluded using the GSW and CIFOR data sets. Comparing Figure 4.2(a) with Figure 4.2(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.

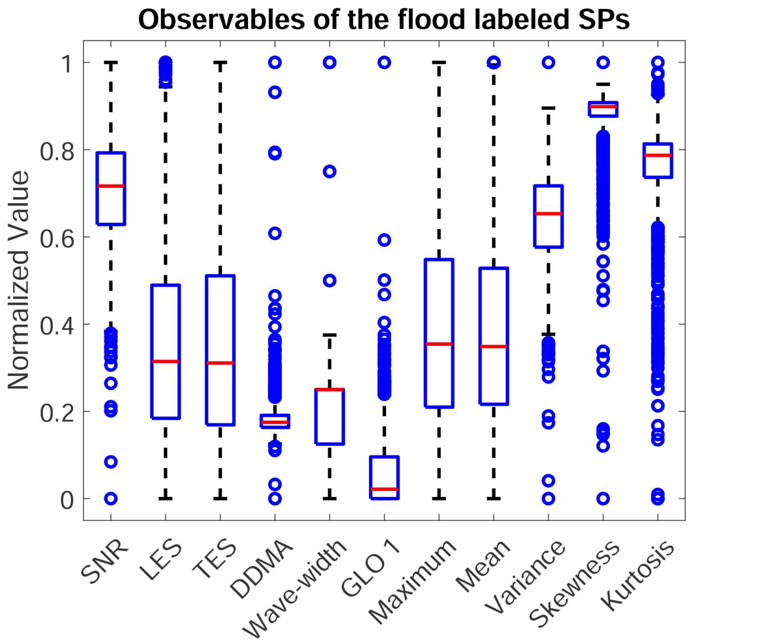
The DDMs whose maxima are not between delay bins 4 and 14 are discarded as noise. The discarded DDMs include high altitude measurement and noisy DDMs. This range is determined by observing DDMs and comparing the delay bins of their maximum values. Moreover, when the incidence angle is between 15° and 60°, the reflected signal is more correlated with the water extent around an SP since within this range the coherent scattering is dominant over incoherent one, as shown in [105]. Since the intention here is to detect a flash flood, which 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 Table 3.1, are also considered in the preprocessing step [80]. It is worth mentioning that the speckle noise impact is negligible since each DDM is obtained from 1 s incoherent integration of 1000 DDMs [116].

4.2.2 Classifiers

In this chapter, the RUSBoost-based classification is implemented in MATLAB R2018 using the Statistics and Machine Learning Toolbox. A total number of 150 weak



(a)



(b)

Figure 4.2: The box plots of the eleven observables (a): SPs labeled as land, and (b): SPs labeled as flood.

learners are trained with a learning rate of 0.1. Different combinations of the number of weak learners and learning rate values were investigated 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.

Moreover, an SVM based classifier is implemented using the Statistics and Machine Learning Toolbox of MATLAB R2018. Similar to RUSBoost classifier, 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) [92, 93]. 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 [117].

Since the selected training data from Harvey is random, to provide a better perception of the performance, the classification was repeated 20 times for both SVM and RUSBoost classifiers. At each repetition, half of the data points in Harvey are randomly selected for training the classifiers.

The block diagrams of the classifiers implemented in Section 4.3.1 and Section 4.3.2 are depicted in Figure 4.3 and Figure 4.4, respectively.

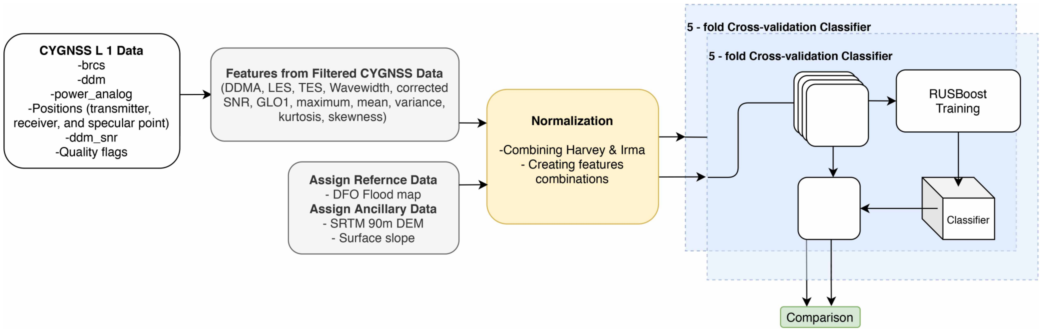

CYGNSS L 1 Data
-brcs
-ddm
-power_analog
-Positions (transmitter, receiver, and specular point)
-ddm_snr
-Quality flags
Features from Filtered CYGNSS Data
(DDMA, LES, TES, Wavewidth, corrected SNR, GLO1, maximum, mean, variance, kurtosis, skewness)
Assign Refernce Data
- DFO Flood map
Assign Ancillary Data
- SRTM 90m DEM
- Surface slope
Normalization
-Combining Harvey & Irma
- Creating features combinations
5 - fold Cross-validation Classifier
5 - fold Cross-validation Classifier
RUSBoost Training
Classifier
Comparison

Figure 4.3: Block diagram of 5-fold cross-validation classifier used for feature selection.

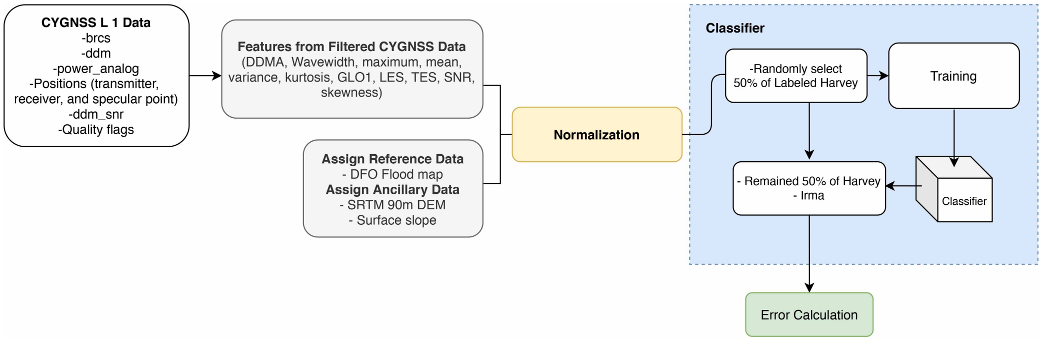


Figure 4.4: Block diagram of the classification.

4.2.3 Data Preparation Approaches

In this section, six different DPAs for flash flood detection are described. As mentioned in Section 4.1.2, 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 4.3.1 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 Chapter 3, water bodies over Harvey and Irma are detected. The detected water extent is then used as a reference for correspondingly 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 1.5 km. When the distance between two points is higher than 1.5 km, the SP is excluded. We investigated different values for determining this distance and 1.5 km 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 their advantages and disadvantages. 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 4.3.2 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.

4.3 Results and Discussion

As mentioned in Section 4.2.1, eleven different observables are extracted from the CYGNSS data.

First, the optimum flood threshold is determined by considering five observables and two ancillary features. Based on the the optimum flood threshold, all different combinations of eleven observables and two features from SRTM90m DEM are used as inputs of a RUSBoost classifier with 5-fold cross-validation to select a suitable combination of features in Section 4.3.1. With the selected features, six different approaches for detecting flood are studied in Section 4.3.2. By comparing their results, the best approach for flood detection is selected. The performance of the recommended RUSBoost classifier is then compared with that of an SVM based classifier in Section 4.3.3.

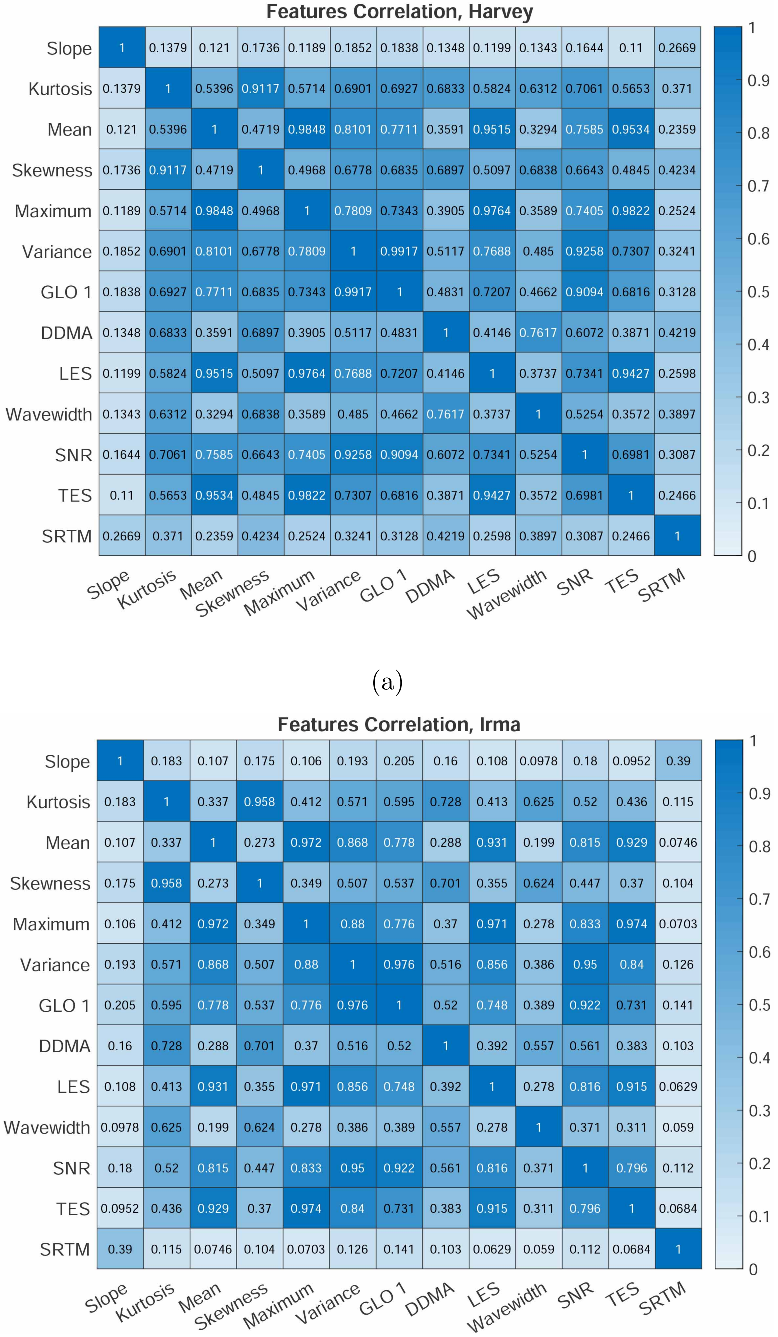
4.3.1 Features Selection

In this section, we want to select the features that are proper for flood detection. Therefore, by using GSW and CIFOR GW 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 4.2.1 and the surface elevation and terrain from the SRTM90 DEM data set are considered as thirteen features. Even though the extracted features are different at the first look, some of them may carry the same information and employing them all may lead to more computations without any improvement. One way for clarifying the relation between variables is to crosscorrelate them. When two variables are highly correlated (cross-correlation between 0.9 and 1), they provide similar information, but they are not identical. As shown in Figures 4.5(a) and 4.5(b), there are high correlations (more than 0.9) between the maximum and TES, LES, and mean, between variance and $\mathrm{GLO_1}$ and $\mathrm{SNR_C}$ and between kurtosis and skewness. Therefore, in this section all combinations are considered for determining a suitable features combination.

Before proceeding to the feature selection, the optimum value of flood threshold is determined by considering all the possible values. The combination of Kurtosis, Maximum, Variance, DDMA, Wave-width, and two ancillary features (surface elevation and terrain from the SRTM90 DEM) are used as input for 5-fold classifier whose classification errors are shown in Figure 4.6. The average flood classification error starts at $\sim$ 20% and gradually declines to $\sim$ 11%. The average land classification error starts at $\sim$ 14%, and it reaches its minimum, $\sim$ 5%, when the flood threshold is 25%. After this point, as the flood threshold increases, the error increases. The flood error at a flood threshold of 25% is relatively lower than other values within 20% to 30% interval. Therefore, 25% is selected as the optimum value of the flood threshold by considering both land and flood classifications errors.

By knowing the optimum flood threshold, both 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 the 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 4.3. From Table 4.3 it is clear that using the



(a)

(b)

Figure 4.5: Correlations between features, (a): Harvey, and (b): Irma.

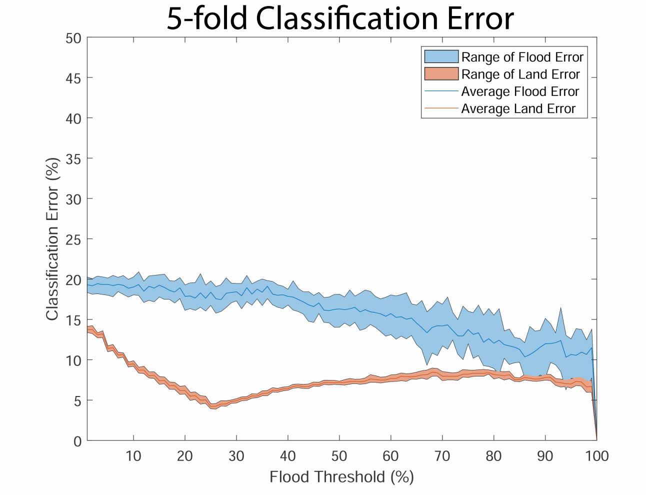


Figure 4.6: Error values of 5-fold cross validated classifier for different flood thresholds by using Kurtosis, Maximum, Variance, DDMA, Wave-width, and SRTM90m DEM as the input feature vector.

elevation and terrain from SRTM90 DEM (combinations 1 to 5 in Table 4.3) has improved the accuracy (around 10% for flood and 2.5% for land). Due to 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 4.3 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 4.3 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 combina-

Table 4.3: 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.

tions 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 4.3. As shown in Figure 4.7, as the number of weak learners increases, the classification errors decrease since the classifier becomes more adapted to the data. 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

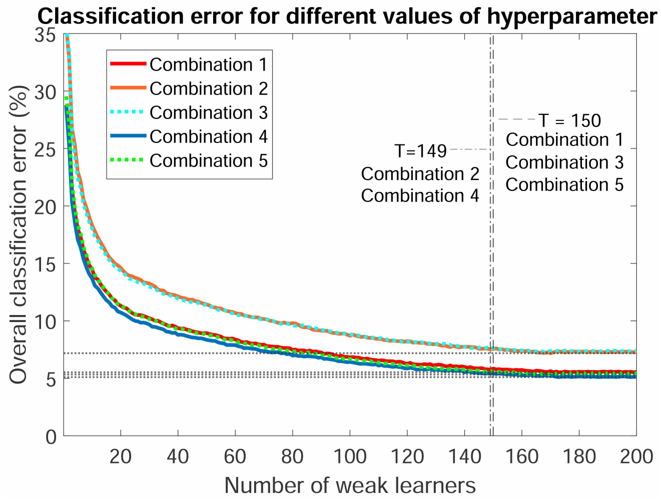


Figure 4.7: The classification error of the classifier with 5-fold cross-validation with respect to the number of weak learners (i.e. the value of the hyperparameter, T).

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%.

4.3.2 Flood Detection

By knowing the best feature vector from Section 4.3.1, in this section the intention is to find the best DPA for flood detection using the CYGNSS data through evaluating the six DPAs mentioned in Section 4.2.

Unlike Section 4.3.1 where a classifier with 5-fold cross-validation was used, here, the RUSBoost based classifier depicted in Figure 4.4 is trained and tested by using the features combination 5 in Table 4.3 for each DPA.

The results shown in Table 4.4 are the accuracies and the percentage of excluded data for both Harvey and Irma. For choosing the best method, first, the percentage of the discarded data of each DPA is compared with other DPAs. Then, a 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, Approach 4, and Approach 5 have an acceptable accuracy in some instances, 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 4.3.1, where the intention is to find a suitable feature vector for detecting flood, here, we are investigating different DPAs for flash flood detection. Moreover, as shown in Table 4.4, 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 4.8 to 4.11, respectively. The GSW and CIFOR-GW references shown in Figures 4.8(a) and 4.10(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 4.8(a) and 4.10(a) in blue colour. Based on the high-resolution DFO flood reference maps depicted in Figures 4.8(b) and 4.10(b), in each DPA, considered SPs are labelled as flood/land. Due to data exclusion, the flood reference map of each DPA could be different from others. As shown in Figures 4.8(b) and 4.10(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, the 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-GW, 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 [118]. Using these observables, the RUSBoost classifier from Chapter 3, which is trained with data from the Congo basin, is applied for detecting water bodies of Harvey and Irma. 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. The employed data sets for water detection are collected throughout a year in different

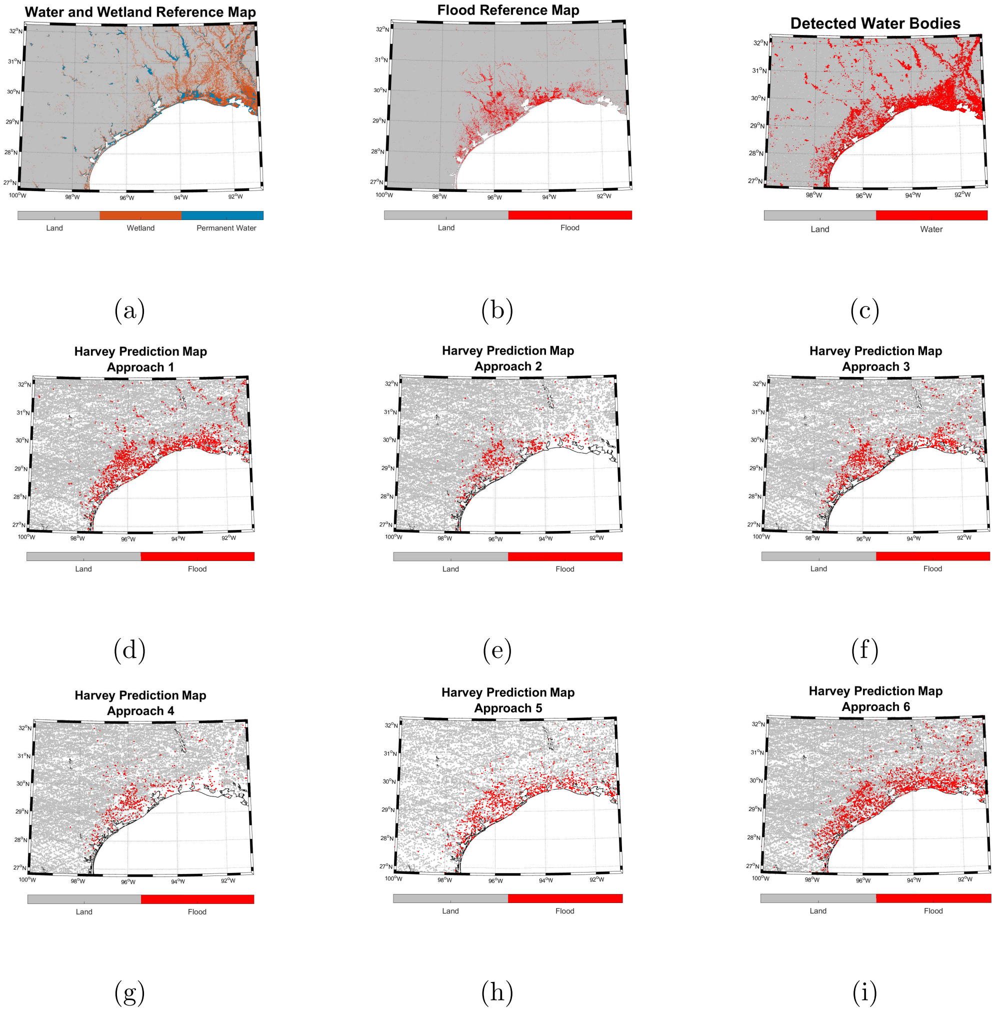
Table 4.4: Accuracies of Approach 1 to Approach 6.

#	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 %

weather conditions. As shown in Figures 4.8(e), 4.8(g), 4.10(e) and 4.10(g), when a large portion of SPs is discarded, detected flood does not correspond well to the actual event shown in Figures 4.8(b) and 4.10(b). The error maps depicted in Figures 4.9(a) to 4.9(f) and 4.11(a) to 4.11(f) indicate that most error points are located close to the flooded regions shown in Figures 4.8(b) and 4.10(b). Comparing the classification results of Approach 3 and flood reference maps depicted in Figures 4.8(f) and 4.10(f) and Figures 4.8(b) and 4.10(b), respectively, shows that even with small coverage, Approach 3 is capable of identifying the flash flood extent.

4.3.3 Comparison to SVM Classifier

For comparison, an SVM-based classifier is trained using the selected features in Section 4.3.1 and same data that is produced by Approach 3 and employed for building the RUSBoost classifier. The classification results and error maps of two classifiers

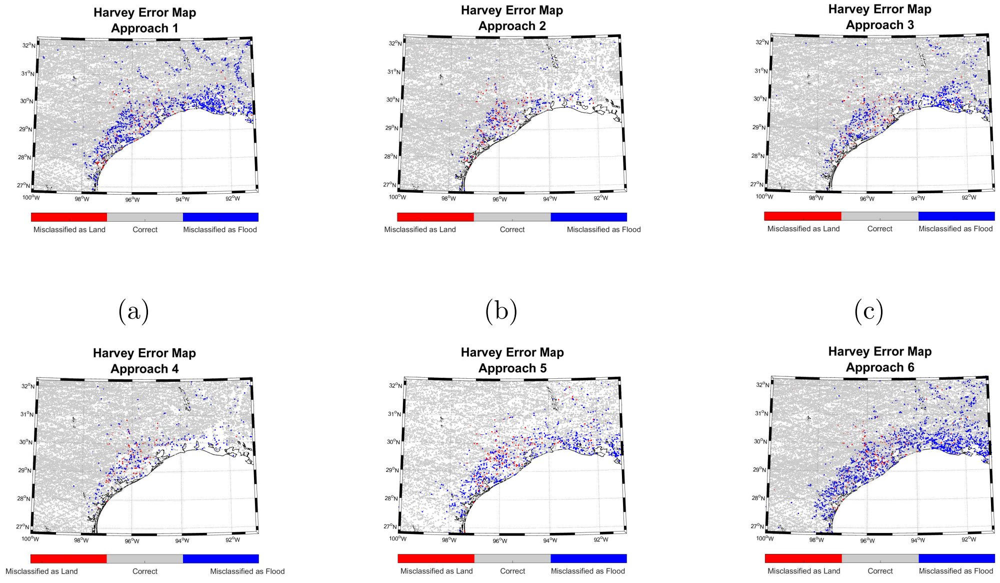


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.8: Maps of Harvey (a): GSW water and CIFOR-GW, (b): DFO flood reference map, (c): detected water bodies, classification result map of (d): Approach 1, (e): Approach 2, (f): Approach 3, (g): Approach 4, (h): Approach 5, (i): Approach 6



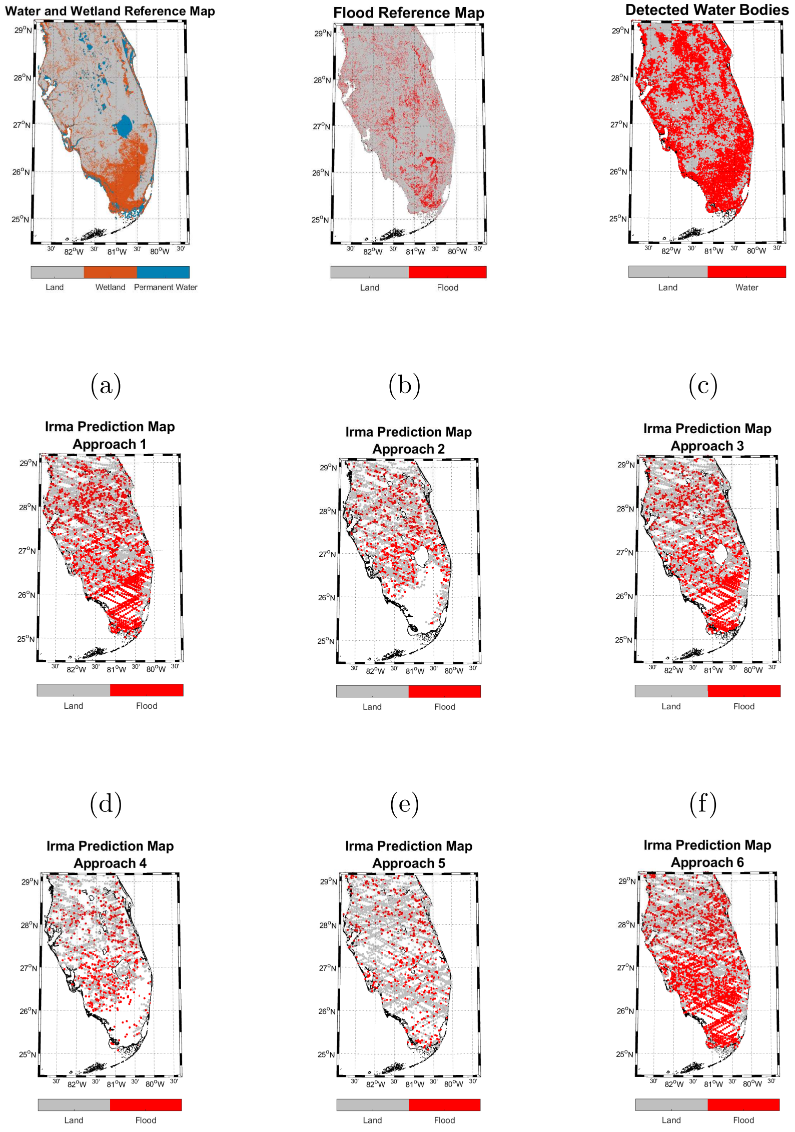
(a) (b) (c)

(d) (e) (f)

Figure 4.9: Harvey classification error map (a): Approach 1, (b): Approach 2, (c): Approach 3, (d): Approach 4, (e): Approach 5, and (f): Approach 6.

are depicted in Figures 4.12 and 4.13. As mentioned in Section 2.2.2, the parameters of the SVM-based classifier are optimized using the SMO algorithm.

As shown in Table 4.5, 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 4.12(a) to 4.12(d) and 4.13(a) to 4.13(d). 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.

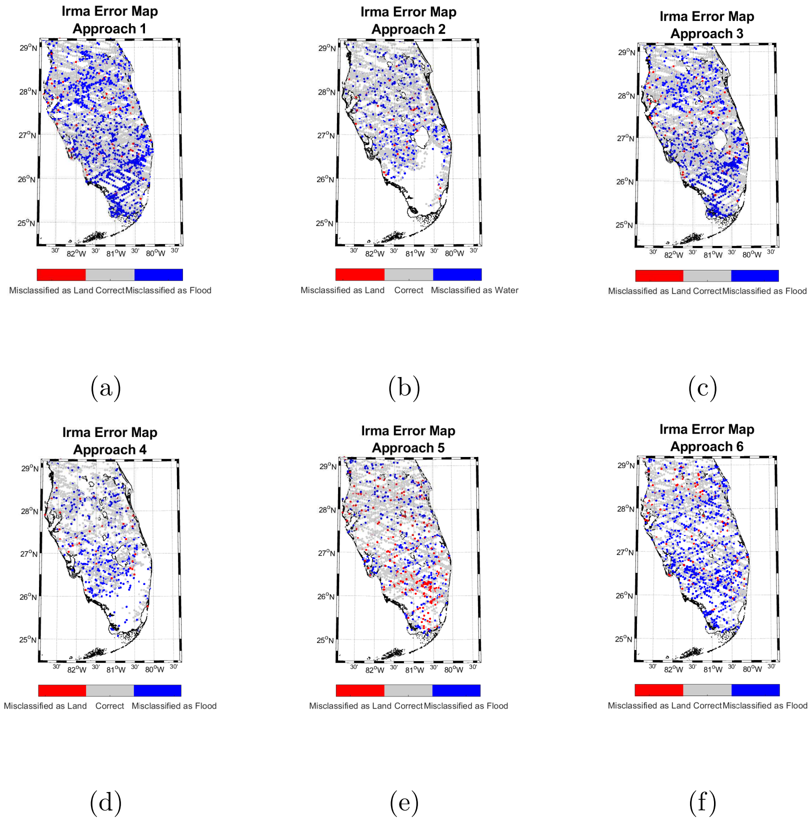

Water and Wetland Reference Map
Land
Wetland
Permanent Water
Flood Reference Map
Land
Flood
Detected Water Bodies
Land
Water
Irma Prediction Map Approach 1
Irma Prediction Map Approach 2
Irma Prediction Map Approach 3
Irma Prediction Map Approach 4
Irma Prediction Map Approach 5
Irma Prediction Map Approach 6
Land
Flood
29°N
28°N
27°N
26°N
25°N
30'
82°W
30'
81°W
30'
80°W
30'

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.10: Maps of Irma (a): GSW water and CIFOR-GW, (b): DFO flood reference map, (c): detected water bodies, classification result map of (d): Approach 1, (e): Approach 2, (f): Approach 3, (g): Approach 4, (h): Approach 5, (i): Approach 6



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.11: Irma classification error map (a): Approach 1, (b): Approach 2, (c): Approach 3, (d): Approach 4, (e): Approach 5, and (f): Approach 6.

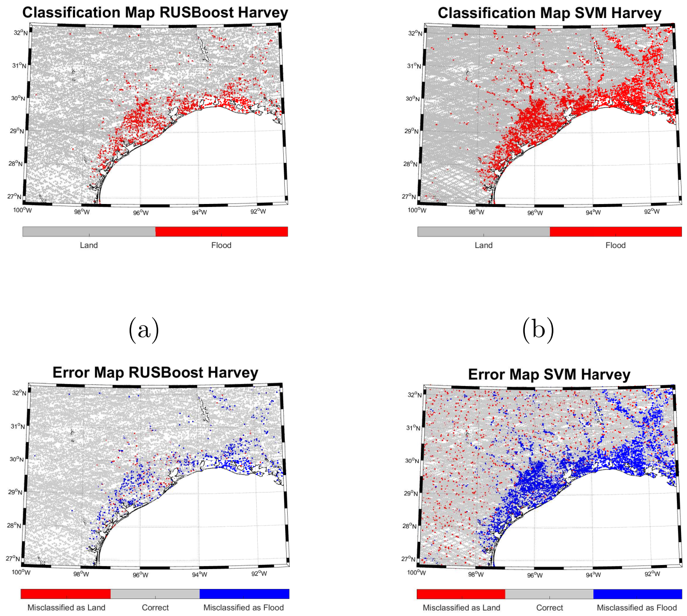
Table 4.5: Results of SVM and RUSBoost classifiers.

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

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.

4.4 General Chapter Summary

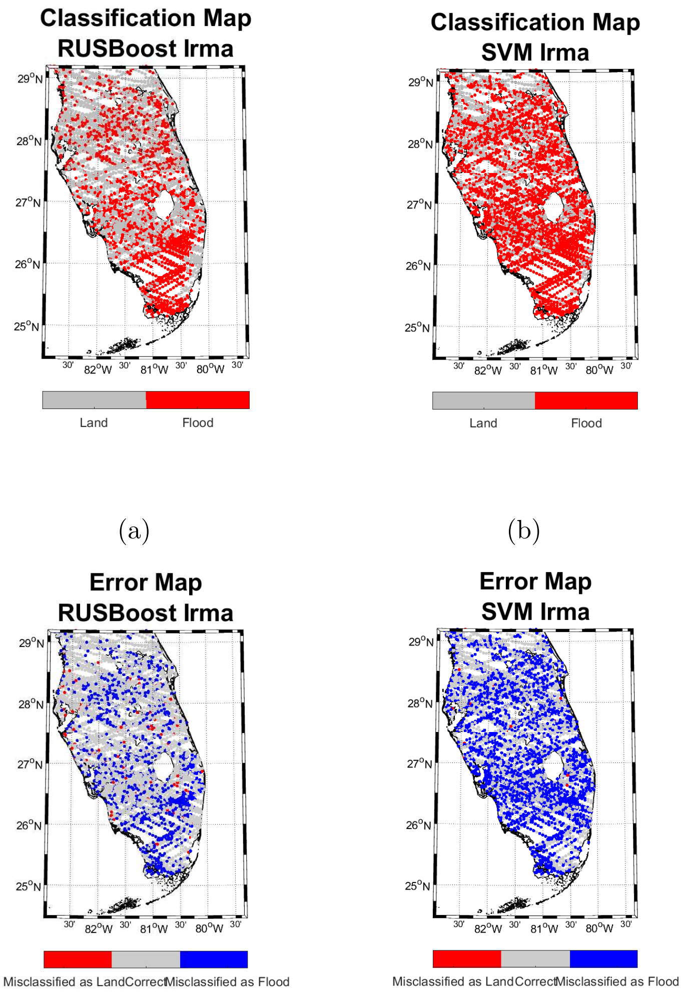
In this chapter, a flood detection method based on CYGNSS data has been conducted using the RUSBoost based classification. Eleven different features have been extracted from CYGNSS data. Considering all the possible flood thresholds, the optimum value regarding the classification error is determined. 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, a suitable features combination was selected. By using the selected features combination, six different DPAs for detecting flash flood were investigated. Subsequently, the performance of the RUSBosst-based classifier was compared with an SVM-based classifier using data exclusion in the selected DPA.



(a) (b)

(c) (d)

Figure 4.12: Comparison between RUSBoost and SVM classifiers for Harvey (a): classification result map RUSBoost classifier, (b): classification result map SVM classifier, (c): error map RUSBoost classifier, and (d): classification error map SVM classifier.



(a)

(b)

(c)

(d)

Figure 4.13: Comparison between RUSBoost and SVM classifiers for Irma (a): classification result map RUSBoost classifier, (b): classification result map SVM classifier, (c): error map RUSBoost classifier, and (d): error map SVM classifier.

Chapter 5

Conclusion

5.1 General Synopsis and Significant Results

In this thesis, research for permanent water and flash floods remote sensing using GNSS-R data of the CYGNSS is presented. First, a permanent water detection method was investigated using three CYGNSS observables and RUSBoost-based classifier [118]. Its results over two case studies involving data from the Congo and Amazon basins are compared with the watermask detection method proposed in [72]. Comparison results show that the RUSBoost-based classifier can detect the water bodies more accurately with more details. Though the watermask detection method is slightly more accurate in terms of land detection, it mislabels the small land areas that are surrounded by water. Furthermore, the evaluation results from the Amazon basin indicate that the proposed technique is more general compared to the watermask detection technique.

Due to the satellite movement during incoherent integration, the along-track coherent resolution of the CYGNSS receivers is approximately 7 km [70]. Therefore, at a resolution of $0.01^{\circ} \times 0.01^{\circ}$ both techniques overestimate the presence of water. Moreover, the GSW data is created based on the data collected by optical sensors. Hence, the water reference data does not include the data of cloudy days or over regions with dense biomass. In other words, the water content in the GSW data is underestimated. Thus, CYGNSS, which is not sensitive to clouds and certain biomass, detects water content that is not included in reference data, leading to water overestimation.

The proposed method is able to classify a grid cell by using its features vector and without any dependency on neighbouring grid cells. The classifier is not designed for any specific region or time period and once it is trained, it can classify the data with a low computational load. Hence, it does not require a large CYGNSS data set. In addition, the preprocessing of the CYGNSS data and the extraction of observables do not require any heavy calculations. Moreover, it could detect small water bodies with a size comparable to the minor axis of the FFZ (~ 600 m).

As the second objective, a method based on the CYGNSS data and RUSBoost-based classifier was proposed for flash flood detection [119, 120]. Eleven different observables were extracted from the CYGNSS data. A suitable combination of features and the optimum flood threshold were determined. Considering the selected features and optimum flood threshold, six approaches for detecting flash floods were investigated. By comparing their classification results and the percentage of excluded data, the best approach was selected for flash flood detection. The results of the proposed method are also compared with those of an SVM-based classifier as a representative ML method.

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 parameters provided in the CYGNSS data for that SP without any dependency on a region or time period. 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.

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 fluctuates [121]. Due to the different designs of space vehicles and the transmitting antenna panel, the EIRP of GPS transmitters fluctuates and 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 [122, 123]. However, due to the limitation of available CYGNSS data that is associated with significant flash flood, the two representative events that happened in 2017 were selected.

The main drawback of the proposed method is flood overestimation with respect to the DFO reference data. This problem was also reported in [124], 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 flood development [71, 125, 126]. Moreover, the scattering from the surface at L-band primarily depends on two factors:

roughness and soil moisture, as investigated in [127]. Due to heavy precipitation during a flash flood, the SM increases until 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) [128, 129] and during the month prior to flood, several precipitations happened in those areas especially for Irma [130]. Since having high SM does not necessarily indicate that an SP is flooded [131], SM is not the only source of error. Another parameter that also has a major role in the reflected signal is the roughness [132]. The coherent reflection from a smooth surface can lead to flood overestimation. Therefore, floods in regions that are relatively flat with high SM, such as those impacted by 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 [133]. 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 and can result in an 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 [134, 135], 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. Thus, the turbidity of water may not be a major source of error in our work.

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.

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 [136]. On the other hand, the revisit time of the CYGNSS satellites is shorter than SAR systems. Hence, it is able to detect permanent water and flash floods. Moreover, due to the relatively low cost of GNSS-R receivers, larger constellations can be formed, leading to better and more economical coverage.

5.2 Suggestions for Future Work

The proposed methods are able to detect permanent water and flash floods as long as the reflecting signal is predominantly coherent over the water bodies. Wind could increase the roughness of surface water bodies. The high-speed winds during hurricanes or winds over large inland water bodies (e.g. large lakes) are two examples of this issue. When the roughness of a surface increases, its reflection becomes more incoherent. Distinguishing between water and land under such a situation is more challenging and requires further study.

Furthermore, the proposed methods were only applied to some case studies and might not be general. In order to achieve more comprehensive methods, further studies should be conducted.

Moreover, the feature selection in Chapter 4 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 such as the recursive feature elimination and the minimum redundancy maximum relevance could be investigated.

In this thesis, only the RUSBoost classifier among other methods that are developed for tackling imbalanced data was used. In the future, 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.

When various types of water bodies such as permanent water and wetlands are present in the same region, it is challenging to differentiate between them using the GNSS-R technique. Moreover, the impact of various factors, such as SM, soil type, vegetation, and subsurface flow, were not included in this work. Further work is required to increase the performance of this technique for water extent monitoring, especially for highly dynamic regions.

Moreover, the relationship between the extracted features and the surface scattering mechanism, which was not investigated in this thesis, should be investigated in the future.

Bibliography

[1] M. Leira and M. Cantonati, "Effects of water-level fluctuations on lakes: an annotated bibliography," in *Ecological effects of water-level fluctuations in lakes*, vol. 613, pp. 171–184, 3300 AA Dordrecht, Netherland: Springer, 2008.

[2] A. Karimi and R. Ardakanian, "Development of a dynamic long-term water allocation model for agriculture and industry water demands," *Water Resour. Manag.*, vol. 24, no. 9, pp. 1717–1746, 2010.

[3] K. Sene, *Flash Floods – forecasting and warning.* Springer Netherlands, 2013.

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

[5] P. J. Fitzpatrick, *Hurricanes : a reference handbook.* ABC-CLIO, second ed., 2006.

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

[7] 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. Am. Acad. Pol. Soc. Sci.*, vol. 669, no. 1, pp. 146 – 167, 2017.

[8] G. A. Schultz and E. T. Engman, *Remote sensing in hydrology and water management.* SSBM, 2012.

[9] S. Khorram, F. H. Koch, C. F. van der Wiele, and S. A. Nelson, *Remote sensing.* SSBM, 2012.

[10] 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, pp. 418–422, 2016.

[11] 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.

[12] J. Nigro, D. Slayback, F. Policelli, and G. R. Brakenridge, “NASA/DFO MODIS near real-time (NRT) global flood mapping product evaluation of flood and permanent water detection.” `https://floodmap.modaps.eosdis.nasa.gov//documents/NASAGlobalNRTEvaluationSummary_v4.pdf`, 2014. Accessed on 25 June 2020.

[13] X. Tong, X. Luo, S. Liu, H. Xie, W. Chao, S. Liu, S. Liu, A. Makhinov, A. Makhinova, and Y. Jiang, “An approach for flood monitoring by the com-

bined use of Landsat 8 optical imagery and COSMO-SkyMed radar imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 136, pp. 144–153, 2018.

[14] J. P.Looper and B. E.Vieux, "An assessment of distributed flash flood forecasting accuracy using radar and rain gauge input for a physics-based distributed hydrologic model," *J. Hydrol.*, vol. 412–413, pp. 114 – 132, 2012.

[15] V. Montesarchio, M. Napolitano, F.and Rianna, E. Ridolfi, F. Russo, and S. Sebastianelli, "Comparison of methodologies for flood rainfall thresholds estimation," *Nat. Hazards*, vol. 75, no. 1, pp. 909 – 934, 2015.

[16] A. D'Addabbo, A. Refice, D. Capolongo, G. Pasquariello, and S. Manfreda, "Data fusion through bayesian methods for flood monitoring from remotely sensed data," in *Flood Monitoring through Remote Sensing*, pp. 181–208, Springer, 2018.

[17] 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, 2018.

[18] 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, 2012.

[19] Z. Kugler and T. De Groeve, "The global flood detection system," *Office for Official Publications of the European Communities: Luxembourg*, vol. 45, 2007.

[20] B. Vieux and J. Vieux, "Rainfall accuracy considerations using radar and rain gauge networks for rainfall-runoff monitoring," *J. Water Manag. Model.*, 2005.

[21] K. Sene, *Flood Warning, Forecasting and Emergency Response.* Springer Berlin Heidelberg, 2008.

[22] C. Corral, M. Berenguer, D. Sempere-Torres, L. Poletti, F. Silvestro, and N. Rebora, "Comparison of two early warning systems for regional flash flood hazard forecasting," *J. Hydrol.*, vol. 572, pp. 603–619, 2019.

[23] M. Acosta-Coll, F. Ballester-Merelo, M. Martinez-Peiró, D. la Hoz-Franco, *et al.*, "Real-time early warning system design for pluvial flash floods—a review," *Sensors*, vol. 18, no. 7, p. 2255, 2018.

[24] 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, 2019.

[25] 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, 2011.

[26] 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, 2015.

[27] L. Landuyt, A. Van Wesemael, G. J.-P. Schumann, R. Hostache, N. E. Verhoest,

and F. M. 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, 2018.

[28] 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, 2017.

[29] 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, 2019.

[30] 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, 2019.

[31] 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, 2016.

[32] S. Grimaldi, J. Xu, Y. Li, V. R. Pauwels, and J. P. Walker, “Flood mapping under vegetation using single SAR acquisitions,” *Remote Sens. Environ.*, vol. 237, p. 111582, 2020.

[33] G. P. Petropoulos and T. Islam, *Remote sens. of hydrometeorological hazards*. Boca Raton, FL : CRC Press, Taylor & Francis Group, 2018.

[34] "MOD44W data access." https://lpdaac.usgs.gov/products/mod44wv006/.

[35] "Global Surface Water data access." https://global-surface-water.appspot.com/download. accessed on 22 June 2020.

[36] H. Wu, R. F. Adler, Y. Hong, Y. Tian, and F. Policelli, "Evaluation of global flood detection using satellite-based rainfall and a hydrologic model," *J. Hydrometeorol.*, vol. 13, no. 4, pp. 1268–1284, 2012.

[37] 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.), ch. 3, pp. 43–54, John Wiley & Sons, 2018.

[38] 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, 2014.

[39] C. Ruf *et al.*, *CYGNSS Handbook.* Ann Arbor, MI, USA: Michigan Publishing, 2016.

[40] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle, *GNSS–global navigation satellite systems: GPS, GLONASS, Galileo, and more.* SSBM, 2007.

[41] A. Gurvich, "Navigation satellites for radio sensing of the Earth's atmosphere," *Sov. J. Remote Sens.*, vol. 6, pp. 89–93, 1990.

[42] S. T. Lowe, J. L. LaBrecque, C. Zuffada, L. J. Romans, L. E. Young, and G. A. Hajj, “First spaceborne observation of an Earth-reflected GPS signal,” *Radio Sci.*, vol. 37, no. 1, pp. 7–1, 2002.

[43] M. Unwin, P. Jales, J. Tye, C. Gommenginger, G. Foti, and J. Rosello, “Spaceborne GNSS-reflectometry on TechDemoSat-1: Early mission operations and exploitation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4525–4539, 2016.

[44] P. Teunissen and O. Montenbruck, *Springer handbook of global navigation satellite systems*. Springer, 2017.

[45] “TDS-1 data access.” `http://merrbys.co.uk/data-access`. Accessed on 5 June 2020.

[46] H. Carreno-Luengo, A. Camps, P. Via, J. F. Munoz, A. Cortiella, D. Vidal, J. Jané, N. Catarino, M. Hagenfeldt, P. Palomo, *et al.*, “3Cat-2—an experimental nanosatellite for GNSS-R earth observation: Mission concept and analysis,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4540–4551, 2016.

[47] “CYGNSS. 2018. CYGNSS Level 1 Science Data Record Version 2.1. Ver. 2.1. PO.DAAC, CA, USA..” `https://doi.org/10.5067/CYGNS-L1X21`. Dataset accessed [2020-07-02].

[48] 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,” *Geophy. Res. Lett.*, vol. 44, no. 16, pp. 8369–8376, 2017.

[49] 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, 2016.

[50] 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, 2013.

[51] 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, 2014.

[52] 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.

[53] 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, 2014.

[54] 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, 2019.

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

[56] 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, 2016.

[57] 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, 2019.

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

[59] 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, 2016.

[60] 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, p. 111202, 2019.

[61] 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, 2019.

[62] Y. Zhu, T. Tao, K. Yu, Z. Li, X. Qu, Z. Ye, J. Geng, J. Zou, M. Semmling, and J. Wickert, “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.

[63] 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, 2016.

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

[65] 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, 2019.

[66] C. Chew and E. Small, "Estimating inundation extent using CYGNSS data: A conceptual modeling study," *Remote Sens. Environ.*, vol. 246, p. 111869, 2020.

[67] E. Loria, A. O'Brien, V. Zavorotny, B. Downs, and C. Zuffada, "Analysis of scattering characteristics from inland bodies of water observed by CYGNSS," *Remote Sens. Environ.*, vol. 245, p. 111825, 2020.

[68] C. Chew, J. T. Reager, and E. Small, "CYGNSS data map flood inundation during the 2017 Atlantic hurricane season," *Sci. Rep.*, vol. 8, p. 9336, 2018.

[69] 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.

[70] 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, 2019.

[71] K. Smith and R. Ward, *Floods: Physical Processes and Human Impacts*. Wiley, 1998.

[72] 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.

[73] V. U. Zavorotny and A. G. Voronovich, "Scattering of GPS signals from the ocean with wind remote sensing application," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 951–964, 2000.

[74] C. Ruf, J. Scherrer, R. Rose, and D. Provost, "Algorithm theoretical basis document level 1B DDM calibration." `https://clasp-research.engin.umich.edu/missions/cygnss/reference/ATBD%20L1B%20DDM%20Calibration%20R1.pdf`. Accessed on 2 July 2020.

[75] A. G. Voronovich and V. U. Zavorotny, "Bistatic radar equation for signals of opportunity revisited," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, pp. 1959 –1968, April 2018.

[76] J. C. Curlander and R. N. McDonough, *Synthetic aperture radar*, vol. 11. Wiley, New York, 1991.

[77] R. D. De Roo and F. T. Ulaby, "Bistatic specular scattering from rough dielectric surfaces," *IEEE Trans. Antennas Propag.*, vol. 42, pp. 220–231, Feb 1994.

[78] P. Zhu, "Impact of land-surface roughness on surface winds during hurricane landfall," *Q. J. Roy. Meteor. Soc.*, vol. 134, no. 633, pp. 1051–1057, 2008.

[79] 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.

[80] 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, 2018.

[81] A. Ghasemi, A. Abedi, and F. Ghasemi, *Propagation engineering in wireless communications*. Springer, 2012.

[82] P. Beckmann and A. Spizzichino, *The scattering of electromagnetic waves from rough surfaces*. Oxford, UK: Pergamon Press, 1987.

[83] N. J. Willis, *Bistatic Radar*. SciTech, 2005.

[84] G. Bonaccorso, *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd, 2018.

[85] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview.* Boston, MA: Springer US, 2005.

[86] P. H. Gleick, “Water in crisis,” *Oxford Univ. Press.*, vol. 9, p. 473, 1993.

[87] A. Fernández, *Learning from Imbalanced Data Sets.* Springer, 2018.

[88] 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, 2002.

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

[90] 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.

[91] 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*, pp. 2672–2680, 2014.

[92] 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*, vol. 42, no. 4, pp. 463–484, 2012.

[93] 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, 2016.

[94] 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, 2018.

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

[96] L. Breiman, *Classification and regression trees.* Wadsworth statistics/probability series, Belmont, Calif.: Wadsworth International Group, 1984.

[97] R. E. Schapire, *Boosting : foundations and algorithms.* Adaptive computation and machine learning, Cambridge, MA: MIT Press, 2012.

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

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

[100] 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, 2018.

[101] 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, 2020.

[102] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Adv. Neural Inf. Process Syst.*, pp. 831–838, 1992.

[103] V. Vapnik, *The nature of statistical learning theory.* New York, NY, USA: Springer, 2013.

[104] 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, p. 111944, 2020.

[105] 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, 2018.

[106] 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. Am. Meteorol. Soc.*, vol. 100, no. 10, pp. 2009–2023, 2019.

[107] E. S. Blake, "The 2017 Atlantic hurricane season: catastrophic losses and costs," *Weatherwise*, vol. 71, no. 3, pp. 28–37, 2018.

[108] G. Brakenridge and A. J. Kettner, "DFO flood event # 4510." `http://floodobservatory.colorado.edu/Events/2017USA4510/GISData/`.

[109] G. Brakenridge and A. J. Kettner, "DFO flood event # 4516." `http://floodobservatory.colorado.edu/Events/2017USA4516/GISData/`.

[110] A. Jarvis, H. I. Reuter, A. Nelson, and E. Guevara, "Hole-filled seamless SRTM data version 4." `http://srtm.csi.cgiar.org`, 2008.

[111] "CIFOR Global Wetlands." `https://www.cifor.org/global-wetlands/`.

[112] 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," *Glob. Chang. Biol.*, vol. 23, no. 9, pp. 3581–3599, 2017.

[113] 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.

[114] 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, 2019.

[115] 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.

[116] 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 Aerospace Conference*, pp. 1–13, 2013.

[117] 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.

[118] P. Ghasemigoudarzi, W. Huang, O. DeSilva, Q. Yan, and D. Power, "A machine learning method for inland water detection using CYGNSS data," *IEEE Geosci. Remote Sens. Lett.*, (in press, DOI: 10.1109/LGRS.2020.3020223).

[119] P. Ghasemigoudarzi, W. Huang, O. DeSilva, Q. Yan, and D. Power, "Flash flood detection from CYGNSS data using the RUSBoost algorithm," *IEEE Access*, vol. 8, pp. 171864–171881, 2020.

[120] P. Ghasemigoudarzi, W. Huang, and O. DeSilva, "Detecting floods caused by tropical cyclone using CYGNSS data," in *Proc. IEEE MFI*, (Karlsruhe, Germany), pp. 212–215, 2020.

[121] 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, 2019.

[122] 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, 2019.

[123] 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*, pp. 8725–8728, 2019.

[124] M. S. 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, 2019.

[125] 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, 2016.

[126] 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. Hydrometeor.*, vol. 18, no. 7, pp. 1997–2009, 2017.

[127] 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, 2016.

[128] E. Linacre, *Climate data and resources: a reference and guide.* Psychology Press, 1992.

[129] 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.

[130] "Ventusky precipitation history." https://www.ventusky.com/?p=27.26;-81.42;6&l=rain-3h&t=20170901/2100.

[131] 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.

[132] 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, 2020.

[133] 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.

[134] 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, 2003.

[135] G. Wu, J. de Leeuw, A. K. Skidmore, Y. Liu, and H. H. Prins, "Performance of Landsat TM in ship detection in turbid waters," *Int. J. Appl. Earth Obs.*, vol. 11, no. 1, pp. 54–61, 2009.

[136] 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, 2015.