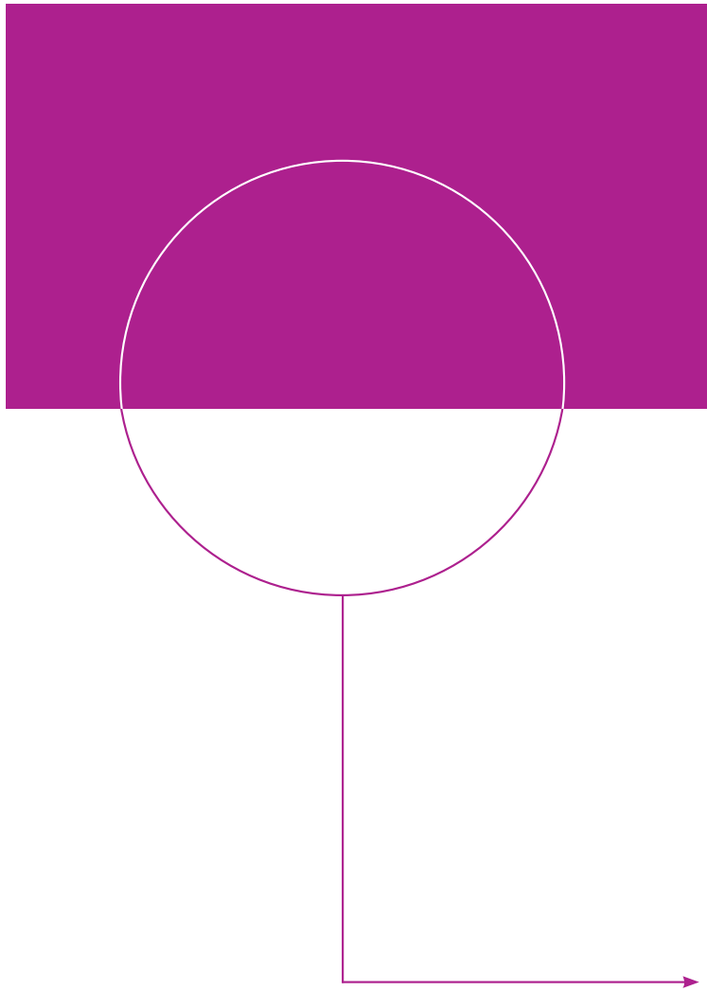


Master's thesis

2022

Vegard Haneberg, Mads Engja Rindal

NTNU
Norwegian University of
Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

Vegard Haneberg
Mads Engja Rindal

Remote Sensing of Soil Moisture Using GNSS Reflectometry

June 2022

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

Remote Sensing of Soil Moisture Using GNSS Reflectometry

Vegard Haneberg
Mads Engja Rindal

Engineering and ICT
Submission date: June 2022
Supervisor: Hossein Nahavandchi
Co-supervisor: Mostafa Hoseini

Norwegian University of Science and Technology
Department of Civil and Environmental Engineering

Abstract

The ever-growing consequences of climate change highlight the need for monitoring the water cycle over land, particularly with the increasing catastrophic events such as floods and droughts. This study shows that a novel remote sensing technology onboard micro-satellites can monitor soil moisture variations and flooding events. The remote sensing technology is based on the Global Navigation Satellite Systems (GNSS) Reflectometry (GNSS-R) concept. GNSS-R analyzes the variations of GNSS signal strength after reflection from land to retrieve information about changes in the soil's water content. Five areas with different geophysical conditions, as well as a larger area in northern India and Pakistan are investigated. The thesis also discusses some of the geophysical parameters affecting the performance of the GNSS-R measurements.

The primary dataset used in our investigations is obtained from the NASA Cyclone GNSS (CYGNSS) mission with eight micro-satellites and a revisit time of about 7 hours. The presence of moisture in the soil changes its electromagnetic properties, leading to changes in the land surface response to incoming GNSS signals. This response can be represented by a parameter called surface reflectivity, which is calculated using the CYGNSS data. In addition, Soil Moisture Active Passive (SMAP) and the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) ancillary datasets containing soil moisture measurements are used for validation purposes.

Our analysis shows a significant correlation in the spatial domain between CYGNSS surface reflectivity and SMAP soil moisture in northern India and Pakistan for January and August 2020. However, this correlation was not equally present when comparing CYGNSS to ERA5 in the same area and period. For evaluation of the CYGNSS measurements in the temporal domain, we have generated time series of surface reflectivity over the selected regions. The time series covers a period of three years from 2019 to 2021. The CYGNSS time series over India demonstrates an overall temporal correlation of 81.2% with the SMAP measurements. This case study manifests a significant degradation of the performance over mountainous sub-regions with a correlation of 28.9%. The results also demonstrate the promising potential of using GNSS-R in disaster management applications. For example, the time series of surface reflectivity measurements in Iran shows a significant increase during two flooding events.

We use machine learning to develop geophysical model functions (GMFs), converting CYGNSS observations over land to soil moisture estimations. Our analysis shows that areas with seasonal variations, low vegetation opacity, and low surface roughness are favorable for remote sensing of soil moisture, with the example of an Indian sub-region with a root mean square error (RMSE) of 0.03289 cm^3/cm^3 with respect to SMAP data.

In contrast, dry regions exhibit lower RMSEs despite demonstrating a low correlation with SMAP soil moisture. We present a novel approach to optimize data selection based on incidence angle. The proposed method shows promising results in areas where approximations of CYGNSS measurement uncertainties are producible. However, the method faces some challenges in some regions and requires further research.

We have developed an interactive web page capable of generating and visualizing soil moisture and surface reflectivity time series based on user input. The complete scientific software package developed in this thesis is available at https://github.com/vegardhaneberg/TBA4925, and the source code to the web page is available at https://github.com/Mosinor/GNSSR_Toolbox.

Sammendrag

Forsterkende konsekvenser av klimaendringer fremhever behovet for overvåkning av jordens vannsykler over land, spesielt sett i sammenheng med økningen av naturkatastrofer som flom og tørke. Denne studien presenterer en ny fjernmålingsteknikk som ved bruk av mikrosatellitter kan overvåke variasjoner i jordfuktighet samt flomdeteksjon. Fjernmålingsteknologien er basert på konseptet Global Navigation Satellite Systems (GNSS) Reflektometri (GNSS-R). GNSS-R analyserer variasjonene i GNSS signalstyrke etter refleksjon fra jordens landoverflate for å utlede informasjon relatert til endringer i bakkens vanninnhold. Fem områder med forskjellige geofysiske forhold, i tillegg til et større område i nord-India og Pakistan er brukt for å undersøke fjernmåling av jordfuktighet og flomdeteksjon. I tillegg diskuterer vi ulike geofysiske parametere som påvirker ytelsen til GNSS-R målinger.

Vi benytter data fra NASA-oppdraget Cyclone GNSS (CYGNSS), som bestående av åtte mikrosatellitter oppnår en gjenbesøkelsestid på omtrent 7 timer. Tilstedeværelsen av fuktighet i jorden endrer dets elektromagnetiske egenskaper, som fører til endringer i jordoverflatens respons til innkommende GNSS signaler. Denne responsen kan representeres ved en parameter kalt overflatereflekSjon, som kan utledes fra CYGNSS data. I tillegg brukes Soil Moisture Active Passive (SMAP) og The European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5), som blant annet inneholder målinger på jordfuktighet, til valideringsformål.

Våre analyser viser en signifikant korrelasjon i det romlige domenet mellom CYGNSS overflatereflekSjon og SMAP jordfuktighet i nord-India og Pakistan i både januar og august 2020. Korrelasjonen i området var ikke like tydelig da CYGNSS ble sammenlignet med ERA5 i samme tidsrom. Vi gjennomførte tidsserieanalyser av overflatereflekSjon for å evaluere ytelsen til CYGNSS målinger i tidsdomenet. Tidsseriene ble gjennomført fra 2019 til slutten av 2021, og et område i India viser en overordnet tidsmessig korrelasjon med SMAP på 81.2%. Dette området demonstrerer en signifikant svekkelse av ytelsen i fjellrike områder, hvor en underregion viser en korrelasjon på 28.9%. Resultatene fremhever også et lovende potensiale for utnyttelsen av GNSS-R i overvåkning og håndtering av naturkatastrofer. Eksemplifisert av tidsserier i Iran som viser en signifikant økning i overflatereflekSjon under to kjente flommer i januar 2020 og mai 2021.

Vi brukte også maskinlæring i utviklingen av geofysiske modellfunksjoner (GMF) som konverterer CYGNSS-observasjoner over land til jordfuktighetsestimater. Disse resultatene viser at områder med sesongvariasjoner, lav vegetasjonstetthet og lav overflateråffhet er foretrukket i fjernmåling av jordfuktighet, hvor et delområde i India oppnådde en kvadratisk gjennomsnittsfeil (RMSE) på 0.03289 cm^3/cm^3 sammenliknet

med SMAP data. Videre fremheves det at tørre områder oppnår lavere RMSE til tross for å ha lavere korrelasjon med SMAP jordfuktighet. Til slutt utviklet vi en ny fremgangsmåte for å optimalisere dataseleksjon basert på det transmitterte signalets innfallsvinkel ved kontakt med jordens landoverflate. Metoden viser lovende resultater i områder der tilnærminger av usikkerheten ved gjennomførte CYGNSS målinger er produserbare. Da dette kan være utfordrende i noen områder anbefales det videre utvikling av denne metoden.

Vi utviklet en interaktiv nettside som muliggjorde generering og visualisering av tidsserier for jordfuktighet og overflatereflekSjon basert på brukerinput. Kildekoden til programvaren utviklet i denne studien er tilgjengelig her https://github.com/vegardhaneberg/TBA4925, og kildekoden til nettsiden finnes her https://github.com/Mosinor/GNSSR_Toolbox.

Preface

This thesis completes our Master's degree in Engineering and ICT with a specialization in Geomatics at the Norwegian University of Science and Technology (NTNU).

Firstly, we would like to express our greatest appreciation to our supervisor Hossein Nahavandchi and co-supervisor Mostafa Hoseini for their continuous encouragement and support during the completion of this thesis.

We also want to thank our parents and grandparents for all the support we have received during our five years in Trondheim.

Finally, we would like to thank our friends and girlfriends for the moments we have shared and will share in the future.

Vegard Haneberg and Mads Engja Rindal,
June 2022.

This page is intentionally left blank

Table of Contents

This page is intentionally left blank

List of Figures

This page is intentionally left blank

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AutoML	Automated Machine Learning
BRCS	Bistatic Radar Cross Section
CNN	Convolutional Neural Network
CYGNSS	Cyclone Global Navigation Satellite System
dB	Decibel
DDM	Delay Doppler Map
DDMI	Delay Doppler Mapping Instrument
DEM	Digital Elevation Model
DRF	Distributed Random Forest
DSSS	Direct Sequence Spread Spectrum
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
GBM	Gradient Boosting Machine
GLONASS	Global Navigation Satellite System
GMF	Geophysical Model Function
GNSS	Global Navigation Satellite System
GNSS-R	GNSS Reflectometry
GPU	Graphical Processing Unit
GPS	Global Positioning System
LEO	Low Earth Orbiting
LHCP	Left-Handed Circular Polarization
LSBoost	Least-Squares Boosting
MEO	Medium Altitude Earth Orbit

MODIS	Moderate Resolution Imaging Spectrometer
NaN	Not a Number
NASA	National Aeronautics and Space Administration
NBRCS	Normalized Bistatic Radar Cross Section
NDVI	Normalized Difference Vegetation Index
NetCDF	Network Common Data Form
OPeNDAP	Open-Source Project for a Network Data Access Protocol
POBI	Previously-Observed Behavior Interpolation
PO.DAAC	Physical Oceanography Data Active Archive Center
PRN	Pseudo Random Noise
RCS	Radar Cross Section
RF	Radio Frequency
RHCP	Right-Handed Circular polarization
RMS	Root Mean Square
RMSE	Root Mean Square Error
R-V-R	Reflection-Vegetation-Roughness
SMAP	Soil Moisture Active Passive
SNR	Signal-to-Noise Ratio
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
ubRMSD	Unbiased Root Mean Square Difference
VWC	Vegetation Water Content
WAF	Woodward Ambiguity Function
WMO	World Meteorological Organization
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

The first part of this thesis will contextualize the utilization of reflected Global Navigation Satellite System (GNSS) signals for remote sensing of geophysical conditions on the surface of the Earth. In addition, we elaborate on the motivation behind this approach and define the research objectives this thesis aims to achieve. Finally, the structure of the remaining chapters is presented.

1.1 Background

The ecosystem of the Earth has experienced severe anthropogenic environmental effects in recent time, with global climate change being one of the major outcomes. Consequently, a noticeable increase in extreme weather conditions has caused human life hazards and devastating economic losses. Therefore, monitoring events related to the Earth's biosphere and climate changes has become of great interest in many areas and applications.

GNSS has been used in positioning and timing for several decades, becoming a high accuracy navigation service. In addition to the American Global Positioning System (GPS) and the Russian GLONASS, the GNSS constellation has been enriched with contributions from China and Europe, reaching a total of 136 operational satellites in 2022. The transmitted L-band frequency signals limit the impact of weather conditions, making it possible to provide global coverage independent of time.

Research has shown that GNSS signals can be utilized as signals of opportunity by focusing on the signals after reflection of the surface of the Earth, a technique known as GNSS Reflectometry (GNSS-R) [63]. GNSS-R aims to utilize the scattered GNSS signals of

the Earth's surface to e.g. derive information related to the reflective medium. Using reflected GNSS signals as a source of opportunity was first proposed by Martin-Neira in 1993, quickly followed by other studies on the subject due to the vast amount of opportunities this technology offered [48]. One of the geophysical parameters that has been studied using GNSS-R is soil moisture, defined as the amount of water stored in the soil measured in cm^3/cm^3. Zavorotny and Voronovich published the first attempt to use this remote sensing approach for soil moisture monitoring [83]. Due to the highly reflective properties of water, the initial research focused on ocean surface monitoring. However, because research has shown that the signals can be useful over land, the focus has shifted towards land [20].

In 2016 NASA launched a space mission called Cyclone GNSS(CYGNSS), with the objective of measuring wind speeds in tropical cyclones. This mission included eight low Earth orbiting (LEO) satellites installed with receivers capable of detecting both direct and reflected GNSS signals. The spread spectrum technique used in GNSS makes it possible to use low gain and low power antennas, which can be mounted on small satellites. Furthermore, by using the already available GNSS signals, CYGNSS inherits the advantage of using L-band microwave radio frequencies (RF). This leads to CYGNSS being considerably cheaper than other soil moisture missions while at the same time providing improved spatiotemporal resolution.

1.2 The Importance of Soil Moisture

Life on Earth is strongly connected to cycles where water plays the most important role. These affect humans both directly, through droughts and flooding events, and indirectly as e.g. plants require water to contribute to the production of the human vital substance Oxygen. Common for many of these happenings is that soil moisture either directly affects the events or can help monitor and survey them. Accurately measuring and administrating soil moisture, can therefore be helpful in several important applications [15].

1.2.1 Earth Environmental Cycles Including Water

The hydrological cycle is Earth's recycling system of water and is presented in Figure 1.1. It describes how water evaporates from, among others, the ocean and lakes. After evaporation, the steam rises to the atmosphere before it cools, condenses, and creates precipitation. Next, the rain falls to the surface of the Earth and eventually reemerges in lakes or oceans through aquifers. This cycle is affected equally by the ocean and land surfaces, which means that the amount of soil moisture dramatically contributes to the process through evaporation to the atmosphere and transportation to oceans [35].

Another cycle affecting life on Earth is the carbon cycle. This cycle describes how carbon is exchanged between the biosphere, pedosphere, geosphere, and atmosphere of the Earth, and its events are vital in preserving life on Earth. The changes that increase the amount of carbon in the atmosphere raise the temperature on Earth, which is one crucial climate implication humanity face at present. Both the water- and carbon cycles are connected and affect each other. For example, a larger amount of carbon in the atmosphere leads to increased temperatures, which again causes more evaporation. In some regions, this can cause more precipitation, whereas the increased temperatures can lead to droughts elsewhere.

These cycles are all linked through soil moisture, and detailed monitoring of these provides an overview of how the whole Earth system works. This includes the global climate and its future changes, which can cause substantial changes to how societies function in the future [22].

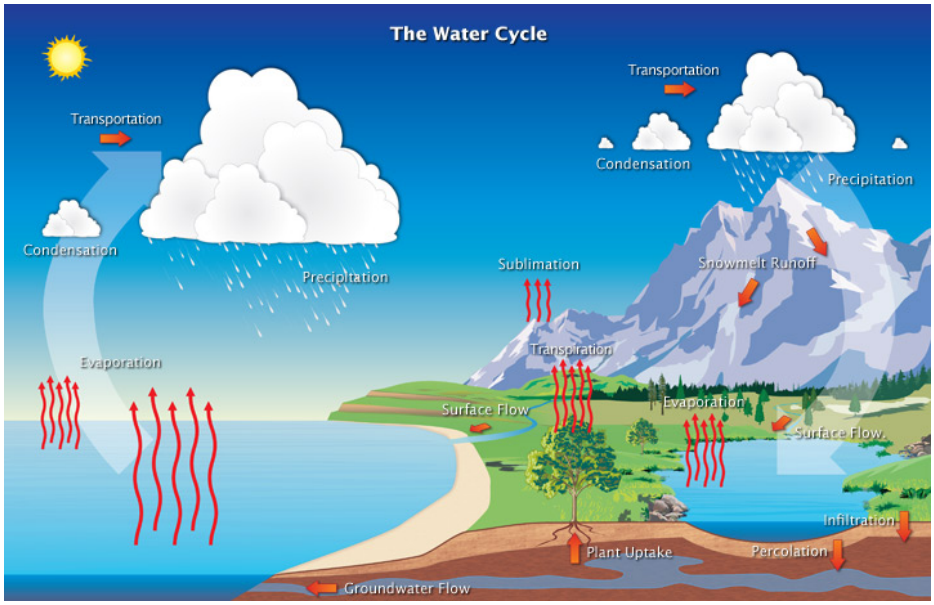


Figure 1.1: Overview of the hydrologic cycle describing the pilgrimage of water as water molecules make their way from the Earth's surface to the atmosphere and back again, in some cases below the surface.

Source: https://earthobservatory.nasa.gov

1.2.2 Monitoring of Extreme Weather and Nature Disasters

The World Meteorological Organization (WMO) published a report[1] on the 31st of August 2021, which investigated weather-related disasters over the past 50 years. This is the most comprehensive study of mortality and economic consequences from meteorological disasters to date. The WMO is an intergovernmental organization, which since 1950 has been the specialized agency for meteorology of the United Nations. The report suggests that during the past 50 years, a disaster related to a weather, climate, or water hazard occurred every single day on average. Worldwide, these disasters lead to a daily cost of about 202 million USD in addition to the loss of human lives.

WMO also investigated the trends of mortality and economic losses during the studied period. The report concluded that recent climate changes lead to extreme weather occurring more frequently by a factor of five from 1970 to 2019. This resulted in a sevenfold increase in economic losses from the 1970s to the 2010s. Mortality showed a different trend than the financial losses, where the number of deaths has decreased by a factor of almost three. One of the main reasons for this development, and the final conclusion of the report, was that early warnings save lives. This highlights the importance of early detection and monitoring the probability of future hazards.

Weather forecasts require continuous monitoring of the atmosphere, in which soil moisture is an important parameter. The amount of water in the soil directly affects the amount of water particles evaporating from the land, which again contributes to several weather cycles. The United States of America's Air Force included soil moisture data to improve their weather forecasts in 2019 [22].

Droughts and floods were two of the top three hazards leading to the most extensive human losses, according to the 2021 WMO report. Soil moisture is directly connected to the definition of drought, which is defined as a deficit in the amount of water in the soil, and is therefore the solely most important parameter to monitor in this aspect. The occurrence of wildfires is also strongly correlated with drought, thus soil moisture enables the possibility to evaluate the danger of wildfires. Furthermore, assessing how wet the soil is before rainstorms can help determine the likelihood of water overtaking its natural paths, resulting in floods. In addition, soil moisture strongly affects the amount of precipitation that runs off or infiltrates nearby streams and rivers, which is not only related to flooding but can contribute to cause e.g. landslides [22].

[1] Retreived from https://public.wmo.int/en/media/press-release/weather-related-disasters on the 27th of May 2022.

1.2.3 Famine and Agriculture

Famine is defined as a severe and prolonged hunger in a large proportion of a population and can significantly increase mortality in several regions or countries. One of the factors that can cause such a devastating situation is related to agricultural crops, which again is strongly connected to soil moisture. A study conducted in 2019 investigated the connection between famines and soil moisture in India between 1870 and 2016. Out of the six major discovered famines, the researchers concluded that five of them were linked to soil moisture agriculture drought. Monitoring the possibility of such events can lead to not only economic savings, but also spare lives [50].

Soil moisture is connected to the failure of agricultural crops and can be used as a parameter to optimize productivity in farming. Studies have shown that the number of undernourished people will increase by 5-26% worldwide by 2080. Therefore, it is of great interest to maximize the production with the same invested efforts and material. Furthermore, accurate soil moisture monitoring can help create the best conditions for healthy plant growth by improving crop yield forecasts and irrigation planning [22].

The examples mentioned above explain how soil moisture is connected to several aspects of human life. Consumption of food and water, weather conditions and extreme hazards are all related to soil moisture to some degree. The same applies to our understanding of how the Earth manages its many energy cycles.

1.3 Main Objectives and Research Questions

The main objective of this thesis is divided into two parts. The first goal is to contribute to the state of the art research in remote sensing of soil moisture using GNSS-R. This culminates in answering the following research questions:

(i) Which geophysical parameters affect the quality of remote sensing soil moisture estimates using GNSS-R?

(ii) Is it possible to reduce the amount of unwanted impact from geophysical conditions when performing remote sensing of soil moisture using GNSS-R?

(iii) Can time series analysis of surface reflectivity be used to capture flooding events?

(iv) Can machine learning be utilized to generate high performing geophysical model functions, connecting reflected GNSS signals and soil moisture?

Research questions (i) and (ii) aim to form a thorough understanding of the parameters affecting remote sensing of soil moisture using GNSS-R. This is important as future space missions can benefit from this understanding and thereby develop improved remote sensing products. In addition, time series analysis will be conducted to reveal if extreme weather events, such as flooding, leave a significant footprint which makes them possible to detect. Finally, the last research question focuses on converting the observed GNSS-R data to actual soil moisture estimates by using machine learning to create a geophysical model function.

In addition to focus on novel research within the field of remote sensing using GNSS-R, this thesis will also strive towards making the obtained results available to the public. This culminates in the following goals:

(v) Develop a web page that allows users to be presented with soil moisture related data.

(vi) Publish an article in a scientific journal to raise the awareness of this new technology of remote sensing, contributing to drive the research forward.

Goal (v) will be conducted by developing a web page that creates time series of both CYGNSS surface reflectivity and soil moisture from an ancillary data source. Users should be able to select a region, either graphically on a map or through specific coordinates, and a time frame before being presented with time series for the desired parameters. This will be performed in collaboration with the department of civil engineering

at NTNU. Goal (vi) aims at publishing significant findings to contribute to the science community.

1.4 Outline

The remainder of this thesis is organized into the following chapters: *The Theoretical Background*, *Previous Work*, *Data Products & Data Processing*, *Methodology*, *Results & Discussion*, and *Conclusion*.

In Chapter 2 - Theoretical Background, we describe the conceptual foundations necessary to understand, analyze, and use GNSS-R for remote sensing. This include both the GNSS technology itself and how the reflected signals can be utilized. The final section introduces the theory behind machine learning and how this can be used in GNSS-R and soil moisture estimation.

In Chapter 3 - Previous Approaches, we present earlier studies that have investigated different aspects of the utilization of reflected GNSS signals for soil moisture estimation. This culminates in a description of the state of the art findings within this field of research.

In Chapter 4 - Data Products and Data Processing, we present both the data sources that will be utilized and the most important data handling steps. The main focus will be directed towards CYGNSS, but SMAP and ERA5 soil moisture products are also described. Furthermore, we introduce the SRTM digital elevation model (DEM) before the choice of programming language is discussed. Finally, an interactive web page showing GNSS-R related results is presented.

In Chapter 5 - Methodology, we first present how how surface reflectivity is computed from measured SNR values available in the CYGNSS Level 1 data. Then, the analyzed areas are presented and justified before a novel framework for optimizing incidence angle interval filtering is presented. Furthermore, we present several approaches to analyze how different parameters in GNSS-R affect the quality of soil moisture estimates. Finally, a machine learning approach to soil moisture estimation is described.

In Chapter 6 - Results and Discussion, we present the results obtained from our analysis, where the main focus is on the parameters affecting the quality of soil moisture estimates using GNSS-R. In addition, we present a geophysical model function in terms of a machine learning model capable of converting raw CYGNSS observations to soil moisture estimates for local regions.

In Chapter 7 - Conclusion, we conclude on the main findings of our work and present ideas for future work.

Chapter 2

Theoretical Background

2.1 GNSS

The Global Navigation Satellite System (GNSS) consists of several satellite constellations orbiting the Earth. These constellations provide positioning and timing services to GNSS receivers located anywhere on or near the Earth's surface. The GPS is one of the fully operating GNSS constellations providing services for both civilian and military use. The constellation consists of nominally 24 operational satellites in medium altitude earth orbit (MEO) [75]. It is designed as a receive-only passive system, with the capacity to handle an unlimited number of users utilizing the system simultaneously [29]. The Russian GLObalnaya NAvigatsionnaya Sputnikovaya Sistema (GLONASS) is another constellation that has been fully operational since 2012, with 24 operational satellites [75, 26]. Other GNSS constellations include the European Union's Galileo, China's BeiDou, and India's IRNASS/- NavIC.

2.1.1 GPS Signal Structure

Because of the numerous functions a GNSS system needs to perform in order to provide the desired functionality, the structure of the transmitted GPS and other GNSS constellation signals is somewhat complex. Before deriving how the transmitted GPS signal is composed and modulated, it is worth mentioning how the navigation signals propagate through space. To circumvent the effect of changes in polarization that may occur during transmission, all current satellite navigation systems use right-handed circularly polarized (RHCP) signals [75]. This is because a linear polarized wave would turn elliptical or circular when traversing through ionized gases or the magnetic field of the Earth, res-

ulting in a change of polarization. The consistent use of RHCP signals limits possible losses caused by orientation mismatch between a receiving antenna and the incident electromagnetic field. In addition, it allows simple receivers with linear antennas to receive signals from different satellite navigation systems with minimal loss while reducing the effect of multipath [60].

The signals being transmitted from the satellites in the GPS constellation are generated using direct sequence spread spectrum (DSSS) modulation. Such a signal can be formed as the product of three components. The base component is the radio frequency (RF) carrier wave. Having RF carrier waves with frequencies within the L-band spectrum, GPS satellites utilize three different frequencies for signal transmission, all derived from the 10.23 MHz fundamental frequency f_0 [75]. The choice of L-band carrier frequencies was thoroughly studied and made based on several requirements, such as weather independency and to facilitate civilian use. Frequencies above 2 GHz would require specific receiver antennas for signal reception. Furthermore would ionospheric delays cause errors for frequency ranges below 1 GHz, making L-band frequencies advantageous for GPS signal transmission [55]. The oldest batch of the still operational satellites transmits what is known as legacy GPS signals. These satellites broadcast coarse/acquisition- (C/A-) code on the L1 carrier frequency of $154 f_0$ and precision ranging (P(Y)-) code on both L1 and the L2 carrier frequency of $120 f_0$ simultaneously. The carrier wave is a pure sinusoid modulated with a digital pseudo random noise (PRN) code and a binary data message, as presented in Figure 2.1. The message contains useful information related to satellite identification, range measurements, and parameters for data processing [75]. However, as the P code can be encrypted before it is transmitted from the GPS satellites, the broadcasted P(Y) code is only available to military equipment with a proper decryption key, and therefore inaccessible for civilian users [37].

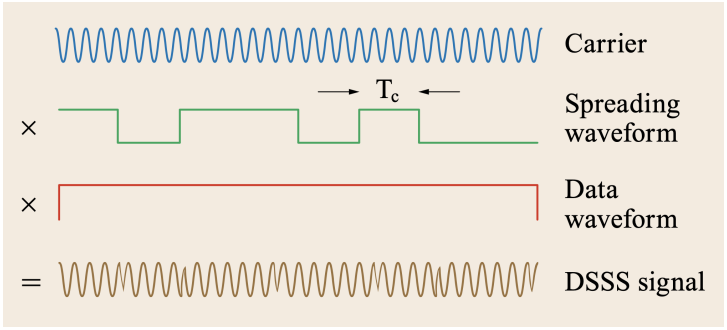


Figure 2.1: Direct sequence spread spectrum modulation.

Source: [75]

The GPS modernization initiative recently introduced a new civil GPS signal transmitted

on a new carrier frequency of $115 f_o$, referred to as L5. Modern satellites launched from 2010 and onward support this additional L5 signal, in addition to other civilian signals added to the already existing L1 and L2 frequency bands. These include the L1C and L2C, where the *C* implies civilian signals, combined with the inclusion of the military M-code signal [75].

With modulation of RF carrier waves, a general GPS signal can be expressed as in Equation 2.1,

$$Y(t) = A(t)C(t)D(t)\cos\omega t, \tag{2.1}$$

where A(t) is the amplitude modulation function representing the received signal power, C(t) is the code modulation, D(t) is the data modulation, and ω is the carrier frequency [37].

Extension of Equation 2.1 makes it possible to express L1 and L2 signals directly. The L1 signal contains both C/A- and P(Y)-code modulated onto the carrier wave in phase quadrature, resulting in a 90-degree offset between the two codes, as expressed in Equation 2.2,

$$Y(t) = A_p(t)P(t)W(t)D(t)\cos\omega_p t + A_{C/A}(t)C/A(t)D(t)\sin\omega_{C/A}t, \tag{2.2}$$

where CA(t) is the modulation function of the civil PRN codes, P(t) is the modulation function of the military PRN codes, and W(t) is an encryption modulation used for the P-code to ensure signal resistance to spoofing.

Because L2 solely contains the military P(Y)-code, the signal is defined as the first component in Equation 2.2, presented in Equation 2.3.

$$Y(t) = A(t)P(t)W(t)D(t)\cos\omega t \tag{2.3}$$

As mentioned earlier, the C/A- and P(Y)-codes provide important information related to satellite identification and ranging measurements. Each satellite transmits unique PRN sequences on L1 and L2, enabling the possibility of satellite identification. Within the data message, users obtain information such as the ephemeris and almanac used for satellite positioning, GPS week number, and a measurement of the satellite health, all used during the user position acquisition.

2.1.2 Signal Acquisition and Processing

When a GNSS receiver is turned on, the receiving antenna detects electromagnetic waves from the available satellites on the horizon and converts them to an electronic signal. This signal is then used for further processing to obtain the desired position, velocity, and timing of the user [75]. Detection of transmitted signals from satellites closer to the zenith angle will travel a shorter distance than signals coming from lower elevation angles. As a result, signals from lower elevation angles suffer from higher attenuation, introducing the importance of antenna gain patterns. The antenna gain is the ratio between the amount of power radiated in a specific direction and the power density of an isotropic source radiating the same power uniformly in all directions. The ratio is measured in isotropic decibel watts (dBWi). GPS satellites are designed in a way that makes the antenna gain along lower elevation angles higher than the antenna gain along the zenith angle. This results in the same received signal power independent of the user position as long as the satellite is above 5 degrees elevation [75]. The principle of reciprocity discussed by Neiman, stating that the transmitting and receiving properties of an antenna are identical, provides knowledge suggesting that the receiving antenna gain also plays an important role in signal retrieval. In addition, different antenna gain patterns can be constructed based on the desired application of the receiver [53].

When the signals are received, along with additional noise and interference, they are processed in the processing unit of the GNSS receiver, having dedicated channels to process signals from individual satellites and frequency bands [75]. Measurements of the SNR in the reduced receiver intermediate frequency bandwidth are important in this aspect, representing the quality of the incoming signal [29]. Since GPS signals are spread spectrum signals, meaning that the transmitted C/A code spreads the total signal power over a wide bandwidth, the received signal power usually drops as low as -130dBm. This is well below the power density of a receiver's thermal noise, resulting in the GPS signal being undetectable unless it is despread to a narrower bandwidth [30].

By generating a copy of the transmitted code, the receiver is able to utilize cross-correlation between the precisely time-aligned replica and the received signal to elevate the received signal detectability. In addition, the receiver determines from which satellite the signal originates. As the time alignment is unknown when receiving the GPS signal, searching through all possible time alignments for the generated replica is required. The channels within the receiver are capable of handling signals from different satellites. During this alignment process, keeping the receiver accurately tuned to the received carrier frequency is key to avoid poor code alignment. As GPS signal transmitters operate with frequencies at the L-band, uncertainties related to receiver position and reference oscillator, or clock, are present. Combined with uncertainties related to satellite motion, this can affect the alignment tuning. Hence, the GNSS receiver will have to also search in frequency.

$\pm 6kHz$ is considered a standard search space, as satellite motion alone is responsible for approximate changes in frequency of 4.2 kHz utilizing the L1 signal [75]. The cross-correlation function for a satellite p can be formulated as in Equation 2.4,

$$C_p(\Delta t) = \frac{1}{\tau}\int_{t_0}^{t_0+\tau}\left[\sum_{i=1}^{n} Y_i(t)D_i(t) + noise\right]Y_p(t+\Delta t)dt, \tag{2.4}$$

where n is the number of satellites whose signals are being received. This cross-correlation must be carried out for each satellite p, with i running from 1 to n.

Because of the unique PRN codes transmitted from each satellite, integrating the product of the received signal and various PRN code replicas based on different delays and frequencies will result in a correlation peak. This peak indicates the desired parameter values for both code delay and Doppler frequency shift. An example of such a peak in correlation is presented in Figure 2.2.

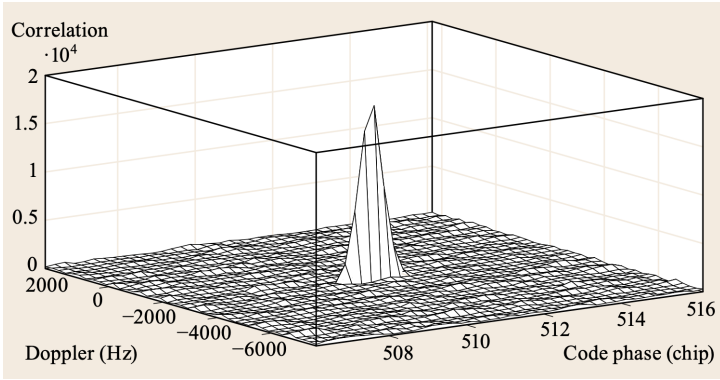


Figure 2.2: GPS C/A correlation function computed for PRN1 during signal acquisition.

Source: [75]

A typical GNSS receiver use code and carrier tracking loops to generate error signals that keep the receiver tuned to the correct frequency as changes in Doppler occurs. If the signal power drops below a certain threshold, the channel can then declare the signal as lost, before restarting the acquisition process. However, as long as the receiver monitors the signal, the tracking channel synchronizes to the broadcaster navigation data message and decodes it. The parameters obtained with the data message are then used to generate the main measurements of GNSS, ultimately leading to the receiver obtaining its position, velocity, and timing.

2.1.3 Utilization of Reflected GNSS Signals

In addition to being considered a service for user positioning and navigation, transmitted GPS and other GNSS constellation signals can be useful in a number of other applications. GNSS signals transmitted through space are prone to several potential error sources such as multipath and interference [75]. However, as opposed to solely focusing on reducing the effects of error sources, Martin-Neira suggested a low-cost and efficient system to perform Earth observations through a set of simultaneous remote sensing observations [48]. Further development of the remote sensing technique has made what is commonly known as GNSS reflectometery (GNSS-R) emerge as a popular field of research, providing information about the reflecting surface covering broad geographical zones [75].

2.2 GNSS Reflectometry for Remote Sensing of Soil Moisture

GNSS-R aims to utilize the scattered GNSS signals off the Earth's surface to derive information of the reflected medium. The transmitting satellite and receiver are separated by a considerable distance and form a bistatic radar system, as shown in Figure 2.5 [3]. On the contrary, a mono-static radar is constructed such that the transmitter and receiver are colocated [71]. The concept of bistatic radar systems allow remote sensing, i.e. obtaining information without being physically present [47].

2.2.1 Scattering

In bistatic remote sensing systems, the scattering distribution depends on the type and roughness of the reflecting medium, vegetation canopy, and the soil dielectric properties [13, 21]. Coherent scattering occurs when the electromagnetic wave is reflected in a single direction. On the other hand, if the outgoing distribution of the wave is spread, the phenomena of incoherent scattering occurs [3]. The surface point in which the incident and reflected angles are equal is known as the specular point. This is the point that usually contributes to the peak value of the reflected signal [72]. Surrounding this point is an area called the glistening zone. The size of this zone is, among others, determined by the roughness of the surface. If a completely smooth surface is illuminated by an electromagnetic wave, all scattering will occur in the coherent direction, as shown in Figure 2.3 (a). As the surface roughness increases, an incoherent component of the scattered signal appears, which is shown in Figure 2.3 (b). This incoherent component increases with the surface roughness, as demonstrated in Figure 2.3 (c).

A practical demonstration of this phenomenon is shown in Figure 2.4. In the absence

of wind, the water surface is smooth and coherent scattering of the visible light creates a clear image of the reflected scene. A slight presence of wind introduces roughness to the water surface, which increases the incoherent scattering and prevents the creation of clear reflection images.

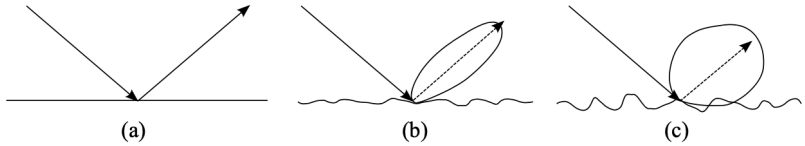


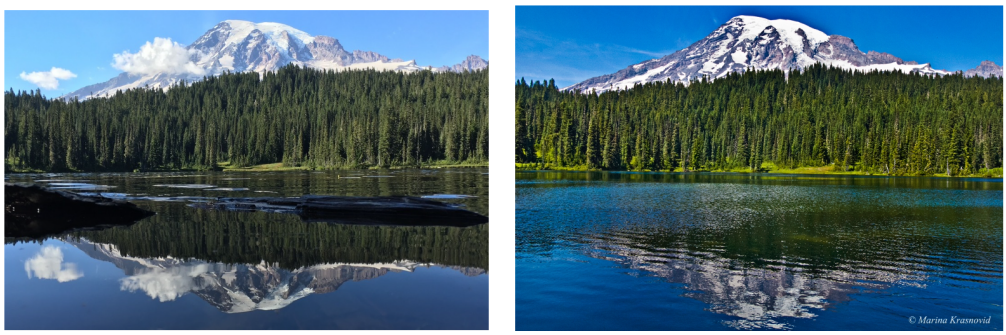
Figure 2.3: The scattering distribution of coherent scattering of a smooth surface (a), and incoherent scattering of a slightly rough (b) and a more rough surface (c).

Source: [21]

Several criteria exist to classify a surface as rough or smooth. One of the most commonly used criteria is the Rayleigh parameter and is calculated using Equation 2.5,

$$R_a = \frac{2\pi}{\lambda} h \cos\theta_{inc}, \tag{2.5}$$

where R_a represents the Rayleigh parameter, λ is the wavelength of the electromagnetic wave, h is the root mean square (RMS) of the surface heights, and θ_{inc} is the incident angle. The surface is classified as rough if the Rayleigh parameter is larger than one and smooth if it is smaller than one [3].



(a) **(b)**

Figure 2.4: Coherent scattering of visible light on a smooth water surface (a) and incoherent scattering on a rougher water surface (b).

Source: [3]

2.2.2 Delay Doppler Maps

Delay Doppler Maps (DDMs) are scattering images where the Radar Cross Section (RCS) is varied with different time delays and Doppler shifts. Time delay and Doppler shifts form the coordinate axes in DDMs, as shown in Figure 2.6. The differences in time of arrival between the direct and the scattered GPS signal constitutes the delay, and the difference in frequency between the two signals makes the Doppler shift [62].

Both delays and Doppler shifts are varied in ranges that include the specular point to create the DDM. The geometry of the delay and Doppler shift in a GNSS-R bistatic radar system is shown in Figure 2.5. Variations in delay shorter than the one for the specular point correspond to locations above the surface of the Earth. Approximately no scattering occurs in these regions [62]. On the other hand, longer delays can be mapped to iso-delay contours, illustrated with a blue color in Figure 2.5. This also applies to Doppler shift, where different frequency shifts are mapped into iso-Doppler lines, drawn in black in Figure 2.5. The DDM thus illustrates the incoherent scattering in adjacency to the specular point.

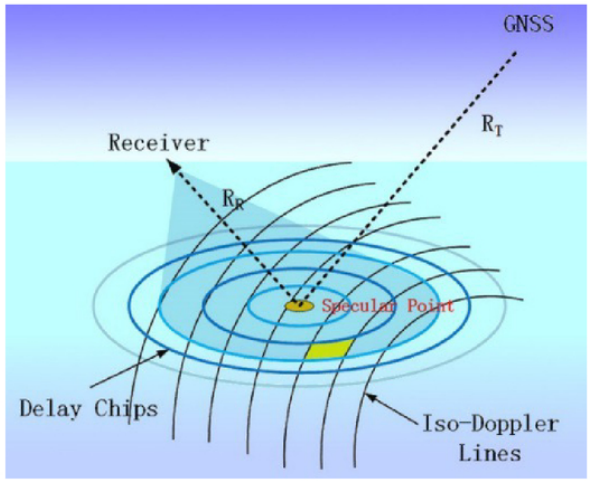


Figure 2.5: The geometry of a GNSS-R bistatic radar system with delay and Doppler shift.

Source: [38]

A noticeable property of the DDMs is the ambiguity of some DDM cells. Each individual cell in the DDM domain does not necessarily correspond to a single cell in the spatial domain. This is illustrated in Figure 2.6, where two cells in the spatial domain map to a single cell in the DDM domain. However, the specular point is located in the center of the spatial domain and is therefore unambiguous [62]. The line of equal distance between

the transmitter and the receiver forms an unambiguous line in the spatial domain. The receiver measures raw uncalibrated units called counts. However, these values do not only depend on the signal itself but include noise from thermal emissions by the Earth and the receiver. Therefore, the raw measured counts are processed to produce values in the desired unit of watts. Firstly, the measured signal is the product of all input signals and the receiver gain. This is shown in Equation 2.6,

$$C = G(P_a + P_r + P_g), \tag{2.6}$$

where C is the DDM values in counts at each bin, P_a is the thermal noise power generated by the antenna, P_r is the thermal noise power generated by the receiving instrument, P_g is the scattered signal power at the instrument, and G is the instrument gain. P_a, P_r and P_g are measured in watts and G has the unit counts per watt. The DDMs are then corrected so that the remaining DMM is only influenced by the signal. The corrected DDM counts can then be expressed as in Equation 2.7,

$$C_g = C - C_N = GP_g, \tag{2.7}$$

where C_N is the thermal noise power of the antenna and the instrument multiplied with the instrument gain. C_N is assumed independent of delay and Doppler shifts, meaning that the pure noise contribution to the raw counts can be estimated [62]. C_g is thus the raw counts produced by only the scattered GPS signal.

The instrument gain can be estimated periodically when the receiver passes over open oceans, using a global open ocean mask [62]. This results in the following equation:

$$C_B = G(P_B + P_r) \tag{2.8}$$

In Equation 2.8, C_B is the blackbody target measurement in raw counts, G is the instrument gain, P_B is the blackbody target power and P_r is the receiver noise power.

Solving Equation 2.7 and Equation 2.8 for the instrument gain, G, and also combining the two, an expression for the scattered signal power, P_G, is obtained and shown in Equation 2.9.

$$P_g = \frac{(C - C_N)(P_B - P_r)}{C_B} \tag{2.9}$$

Due to the fact that C is dependent on both delay and Doppler shift, P_g has the same property. In conclusion, these steps convert the raw counts acquired by the receiver to

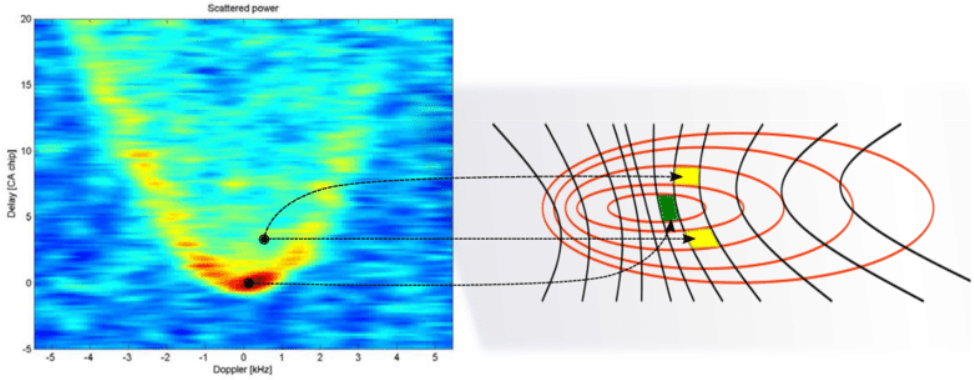


Figure 2.6: Delay Doppler Map in relation to the spatial domain.

Source: [21]

power in watts, while removing unwanted noise occurring from thermal emissions from the Earth and the instrument [62].

2.2.3 Normalized Bistatic Radar Cross Section

The Radar Cross Section (RCS) is a measure of an object's ability to scatter incident energy into a certain direction [40]. In a bistatic radar system the value is referred to as Bistatic RCS (BRCS) and is calculated using Equation 2.10,

$$\bar{\sigma}_{\hat{\tau},\hat{f}} = \frac{P_{g,\hat{\tau},\hat{f}}(4\pi)^3 L_{a1} L_{a2} I_{\hat{\tau},\hat{f}}}{P^T \lambda^2 G_{SP}^T G_{SP}^R R_{SP}^{Total}}, \tag{2.10}$$

where the individual parameters are described in Table 2.1.

The BRCS accounts for surface-related terms and must be normalized in order for geophysical parameters to be estimated [62]. Calculating the Normalized Bistatic Radar Cross Section (NBRCS) is done using Equation 2.11,

$$\bar{\sigma}_0 = \frac{\bar{\sigma}_{total}}{\bar{A}_{total}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} \bar{\sigma}_{\tau_i,f_j}}{\sum_{i=1}^{N}\sum_{j=1}^{M} \bar{A}_{\tau_i,f_j}}, \tag{2.11}$$

where N and M represent the delay and Doppler bins and $\bar{A}_{\tau_i,f_j}$ is the effective scattered

Table 2.1: The parameters used in calculation of Bistatic Radar Cross Section.

Parameter	Description
$\bar{\sigma}_{\hat{\tau},\hat{f}}$	The BRCS for a given delay ($\hat{\tau}$) and Doppler shift ($\hat{f}$)
$P_{g,\hat{\tau},\hat{f}}$	The signal power at a given delay ($\hat{\tau}$) and Doppler shift ($\hat{f}$). This value is presented in Equation 2.9.
L_{a1}	The estimated atmospheric loss corrections from the transmitter to the surface.
L_{a2}	The estimated atmospheric loss corrections from the surface to the receiver.
$I_{\hat{\tau},\hat{f}}$	The correction for losses introduced by the Delay Doppler Mapping Instrument (DDMI).
P^T	The GPs satellite transmit power.
λ	The signal wave length.
G^T_{SP}	The GPS antenna gain at the specular point.
G^R_{SP}	The receiver antenna gain at the specular point.
R^{Total}_{SP}	The total range loss from the transmitter to the surface and the surface to the receiver at the specular point.

area of the respective DDM bin. Standard practice distinguishes between the physical and effective area of a DDM bin. The actual physical area can be calculated by integrating over the area, as shown in Equation 2.12.

$$A_{\hat{\tau},\hat{f}} = \iint_A dxdy \tag{2.12}$$

A DDM using only the physical scattering area contains no scattering for delays shorter than the specular point delay. However, power in these cells occurs due to the spreading of the power by the GPS ambiguity function from physical bins adjacent to the specular point. The effective scattering area is given in Equation 2.13,

$$\bar{A}_{\hat{\tau},\hat{f}} = \iint_A \Lambda^2_{\hat{\tau};x,y} S^2_{\hat{f};x,y} dxdy, \tag{2.13}$$

where $\Lambda^2_{\hat{\tau};x,y}$ is the delay spreading function and $S^2_{\hat{f};x,y}$ is the Doppler spreading function. These two functions are integrated over the physical area to get the effective scattering area.

2.2.4 Fresnel Coefficients

In addition to offer information about a medium's ability to reflect signals, the Normalized BRCS (NBRCS) may also be further calibrated to provide surface characteristics such as the mean square slope (MSS). The MSS is an important quantity, especially of the ocean surface, and is crucial for understanding the physical process at the air-sea interface and interpreting altimeter and scatterometer measurements [82].

As an alternative to the representation of the NBRCS presented in Equation 2.11, the geometric optics limit of the Kirchhoff approximation represents the term as shown in Equation 2.14,

$$\sigma_0 = \pi |\,\Re|^2 \,(q/q_z)^4 P(-q_\perp/q_z), \tag{2.14}$$

where $\vec{q}$ is the scattering vector as a function of the coordinate $\vec{\rho}$ in the mean surface plane, and $\Re$ is the complex Fresnel coefficient. The Fresnel coefficient is an important parameter in GNSS-R as it describes the reflection and transmission of electromagnetic radiation when incident on an interface between different mediums occurs. Accordingly, the coefficient is dependent on the signal polarization, the local incidence angle, and the complex dielectric constant of the reflecting surface, ε.

GNSS-R mainly operates with reflected signals, resulting in the polarization state of the signals mostly being left-handed circular polarized (LHCP). The final factor P($\vec{s}$) is the probability density function of large-scale smoothed surface slopes. Combining the approximation of P($\vec{s}$) by the use of the anisotropic bivariate Gaussian distribution given in the wind direction along the x- or y-axis, Equation 2.14 can be extended to the one presented in Equation 2.15, connecting the MSS components to the NBRCS measurements.

$$\sigma_0(\vec{q}) = \frac{|\,\Re|^2\,(q/q_z)^4}{2\sqrt{mss_x mss_y(1-b^2_{x,y})}} \exp\left[-\frac{1}{2q_z^2(1-b^2_{x,y})}\left(\frac{q_x^2}{mss_x} - \frac{2b_{x,y}q_xq_y}{\sqrt{mss_x mss_y}} + \frac{q_y^2}{mss_y}\right)\right] \tag{2.15}$$

In this equation, mss_x and mss_y are mean square slopes of the sea surface for the denoted axis along and across the wind direction, and $b_{x,y}$ is the correlation coefficient between

the two slope components. Equation 2.16 presents the formula for computing the MSS components, solely determined by the wave-number integral from the ocean elevation spectrum, $\Psi(\vec{\kappa})$ times $\kappa_{x,y}^2$, also known as the slope spectral density.

$$mss_{x,y} = \iint_{\kappa<\kappa_*} \kappa_{x,y}^2 \Psi(\vec{\kappa}) d^2\kappa \tag{2.16}$$

Further simplification of Equation 2.15 can be done by assuming that $mss_x = mss_y = mss/2$, $b_{x,y}^2 = 0$ and the use of $\sigma_0(\theta)$ in the specular point where $\theta_1 = \theta_2 = \theta$ and $\phi = 0$. The assumptions results in Equation 2.17 with the Fresnel coefficient described by the incidence angle θ, presenting the important relation between the NBRCS and the MSS [82].

$$\sigma_0(\theta) = \frac{|\Re(\theta)|^2}{mss} \tag{2.17}$$

Separating co- and cross-polar Fresnel coefficients, Semmling et al. defined the coefficients as in Equation 2.18 and Equation 2.19, where $\Re_\parallel$ and $\Re_\perp$ are the components parallel and perpendicular to the incidence plane [64].

$$\Re^{co} = \frac{1}{2}(\Re_\parallel + \Re_\perp) \tag{2.18}$$

$$\Re^{cross} = \frac{1}{2}(\Re_\parallel - \Re_\perp) \tag{2.19}$$

Because GNSS-R mostly operates with reflected, and hence LHCP signals, the complex cross-polar Fresnel reflection coefficient presented in Equation 2.19, at the interface between air and another medium with a complex dielectric permittivity ε_{med}, can be described as in Equation 2.20. This equation can then be utilized to highlight the importance of parameters such as incidence angle and the electric permittivity of a medium affecting its ability to reflect electromagnetic signals.

$$\Re(\theta) = \frac{1}{2}\left[\frac{\varepsilon_{med}\sin e - \sqrt{\varepsilon_{air}\varepsilon_{med} - (\varepsilon_{air}\cos e)^2}}{\varepsilon_{med}\sin e + \sqrt{\varepsilon_{air}\varepsilon_{med} - (\varepsilon_{air}\cos e)^2}} - \frac{\varepsilon_{air}\sin e - \sqrt{\varepsilon_{air}\varepsilon_{med} - (\varepsilon_{air}\cos e)^2}}{\varepsilon_{air}\sin e + \sqrt{\varepsilon_{air}\varepsilon_{med} - (\varepsilon_{air}\cos e)^2}}\right] \tag{2.20}$$

2.3 Machine Learning

Machine learning is a subset of artificial intelligence (AI) and can be defined as computational methods and algorithms programmed to utilize past experience to improve performance or predict outcomes [51]. Another definition was formed by the American professor and computer scientist Tom Mitchell. He defined it as a computer program, learning from experience, with respect to some class of tasks and performance measure [28]. Common for both definitions is that some computer model or algorithm attempts to discover patterns in available data, referred to as experience. This experience can be structured in different ways, which again affects the machine learning task. Firstly, if all data points are stored with their corresponding ground truth value, the machine learning task is defined as supervised learning. The goal is to train based on the labeled data to be able to predict the correct value for unlabelled and unseen data. On the contrary, data points in unsupervised learning tasks do not contain ground truth values. Within the field of supervised learning, there exist several subclasses of tasks. Supervised regression refers to the set of tasks where the target value, i.e. the ground truth, is a continuous number. Furthermore, supervised classification appears when the target value belongs to one of a finite set of classes [28].

Since the 1950s, AI and machine learning have experienced several periods of hype and skepticism. However, due to the increased computational power of modern computers, as well as the emergence of e.g. graphical processing units (GPUs), several new applications of machine learning have seen their light in the 21st century. These applications belong among others within the public sector, healthcare, finance, and logistics, where two concrete examples are cancer diagnostics and improved sustainability in smart production [41, 14].

Machine learning has also been applied to the field of soil moisture estimation [61, 65, 36], which will be further elaborated in chapter 3. This can be approached as a supervised regression task if sufficient ground truth data is available. In-situ measurements or other GNSS-R data sources can be colocated both in the spatial and the temporal domain, thus serving as ground truth for the model. Due to the fact that soil moisture values range continuously from zero to one, the problem is defined as a regression task.

2.3.1 Supervised Regression

Despite different supervised machine learning algorithms using distinct procedures when solving tasks, there exist several common traits which will be elaborated in this section. Common for all models is their goal of inferring a mathematical function to map input to output before utilizing this function on unseen data. Furthermore, the models need

to quantify the correctness of their prediction function in order to evaluate their performance. This is done by using a problem-specific cost function. This cost function is then minimized with respect to the parameters in the prediction function and the training data. The minimization process is referred to as training because the model adapts to the provided data, thus attempting to learn from past experience [19].

One important goal of the machine learning model is the ability to generalize based on the provided data. A lot of models can perfectly learn every detail of their training data, thus providing a training cost of zero, while still being useless on unseen data. This phenomenon is called overfitting and occurs when the training loss is significantly lower than the test loss. On the other hand, underfitting occurs when the model is too simple to capture the trends in the dataset. In addition to the cost function, variance and bias are also used to evaluate the performance of the model. Variance describes the change in the model when using different parts of the training set, and bias is the systematic error of a model, which tends to draw the results in a certain direction. Models with high variance are prone to overfitting. Several measures can be taken to avoid this unfavorable trait. Firstly, the dataset should be separated into one training set and one test set, with a low bias distinguishing them. Furthermore, it is possible to impose a constraint of regularity on the cost function, which is called regularization. This can be interpreted as a way of limiting the complexity of the model to avoid overfitting [19].

Every machine learning model has its own set of hyperparameters, which can be defined as parameters affecting the learning process. This includes parameters regarding regularization, optimization, and the architecture of the model, like the number of neurons and hidden layers in an artificial neural network. The domain of one single hyperparameter can be continuous numeric, integer, or binary based on what it describes. Selecting the optimal set of hyperparameters is one important and often time-consuming challenge. Early approaches based the selection on human engineering to manually select the best parameters. This procedure is still used at present, but more recent approaches include grid search, black box optimization, and Bayesian Optimization [24].

2.3.2 Regression Models

There are many different approaches on how to perform and solve regression problems. Linear regression is the most simple form of regression, finding the best linear fit to the provided data points. Solely reducing the distance between a linear fit and provided data makes this method prone to outliers, possibly causing the linear fit to deviate from the overall trend. The development of more robust forms of regression has therefore been thoroughly studied within the field of statistics, with both linear and non-linear regression methods having been presented in recent time [34].

Distributed Random Forest (DRF) is one of the recently presented methods for classification and regression. Developed as an extension to the easy-to-use and high-performing Random Forest algorithm introduced by Breiman in 2001, DRF generates a forest of regression trees as opposed to the single tree produced using random forest. Each tree is then built on a subset of the data before the average of all trees is computed and used to perform the desired prediction [6, 49]. Another regression method is the Gradient Boosting Machines (GBM), which is an ensemble method sequentially building fully distributed regression trees on all relevant features in parallel. This means that a GBM model is obtained by training different regression trees based on modified versions of the original dataset before returning the best combination of weights to different variables found during training. The Extreme Gradient boosting (XGBoost) algorithm is an example of a highly rated GBM model. Designed to be flexible and efficient, XGBoost has become a popular machine learning algorithm due to its tendency to yield high accuracy results in a variety of applications [8].

2.3.3 Automated Machine Learning

Automated machine learning (AutoML) addresses a difficult decision when applying machine learning to solve real life problems, namely model selection. As of today, there exists no single model that outperforms every other model in every task, nor any algorithm that is guaranteed to select the best model for a given task. AutoML approaches this problem by providing an interface to train a variety of different machine learning models without making compromises on the performance. A single function, rather than potentially long and demanding implementation processes, is utilized to easily be able to run and compare several models [45]. This approach can be used both as a replacement for manually selecting models but also as an indication of which model types that outperforms others for the given task. In addition to the performance assessment of different machine learning models, some AutoML libraries also include a stacked ensemble of the tested base models. This is a similar concept to that already being performed in DRF and GBM, where the model finds the optimal combination of different ensembles. However, stacking differs from the mentioned methods as it utilizes multiple regression algorithms, possibly resulting in improved performance compared to the usage of a single algorithm [57]. Another important feature of AutoML is its aim of providing automatic hyperparameter selection for the machine learning models, using different search algorithms [24].

Chapter 3

Previous Approaches

Soil moisture estimation has been performed for several decades. The earliest approaches utilized handheld measurement units that were placed in the soil before measuring the direct contact of water as opposed to soil. While being useful in e.g. farming, handheld measurements have severe limitations in terms of spatial coverage and timeliness. Measuring the soil moisture over large fields or monitoring the changes over time is unfavorable using this method.

The utilization of scattered GNSS signals originates from the late 1980s. Hall and Cordey was the first to propose the usage of reflected GNSS signals for scatterometry purposes in 1988 [31]. A few years later, in 1993, Martin-Neira presented the idea of using reflected GNSS signals as signals of opportunity in an ocean altimetry application [48]. Using these signals for surface reflectivity computations was first mentioned in 1998 by Kavak et al.. This study used the Interferometric Pattern Technique (IPT) by measuring both the direct and the reflected GPS signals. The two signals create either constructive or disruptive interference patterns that cause an oscillator in the apparatus to vibrate, whereas the difference between the maximum and the minimum interference pattern can be related to the surface reflectivity [39].

Several studies and experiments have later been performed to develop the theory and equipment necessary to obtain high enough SNR to be able to perform remote sensing of soil moisture over land. Among the used apparatus in these studies is the digital mobile radio. It has been mounted on top of tall stationary towers and airplanes to calibrate and develop the understanding of parameters affecting the remote sensing during the beginning of the 2000s [81, 76]. The research has continuously improved with, among others, NASA taking part. NASA launched a space mission called Soil Moisture Active Passive (SMAP) in 2015 with the aim of utilizing both passive and active remote sensing

techniques to estimate soil moisture on a global scale. While being developed to estimate wind speeds in cyclones, the space mission CYGNSS has also formed the possibility of improving the results in this field.

In 2020, the available soil moisture products were limited to passive or active microwave data whose temporal revisit time are 2-3 days or more than a week, respectively. An example of such a product is SMAP. In an attempt to reduce this revisit time when using remote sensing in soil moisture investigation, Chew and Small developed a product providing soil moisture estimates gridded to 36 km on both daily and sparsely sampled 6-hour intervals starting from 2017. The authors called this product the UCAR/CU product [10]. This soil moisture product was developed by calibrating CYGNSS reflectivity observations to soil moisture retrievals from SMAP and validating the retrievals against observations from 171 in-situ soil moisture probes. Based on the assumption that the signal power measured from CYGNSS observations is linearly related to SMAP soil moisture, the algorithm calculates the soil moisture based on linear regression in $3km$ x $3km$ sub-cells between SMAP and CYGNSS measured on the same calendar day. All sub-cell soil moisture values were then averaged within each 36km SMAP pixel cell, producing the resulting CYGNSS soil moisture value. Both the SMAP and UCAR/CU datasets were validated against in-situ measurements, where SMAP produced an ubRMSE of $0.045cm^3/cm^3$ and UCAR/CU $0.049cm^3/cm^3$. Furthermore, the need to upscale CYGNSS data to match the SMAP 36km resolution leads to the available CYGNSS data not being utilized to its full potential. For these reasons, the dataset should only be considered complementary to SMAP and be used in applications where the importance of improved revisit time outweighs the slightly higher noise. The temporal resolution of UCAR/CU is 63.4% better than the likes of SMAP. Finally, the authors used time series analysis to show that the UCAR/CU product might has the ability to capture events occurring too rapidly for SMAP, thus highlighting the difference in temporal resolution [10].

With the historical background in context and available soil moisture products described, the remaining of this section will take a closer look at some of the most recent contributions within this field. First, four studies focusing on data quality assessment and data processing will be presented. The next part focuses on comparisons between CYGNSS and other data sources, where time series is a widely used technique. Then, several machine learning approaches to soil moisture estimation will be presented before a conclusion on the state of the art follows.

3.1 Data Pre Processing

With the discovered possibility to utilize CYGNSS measurements also over land, Gleason et al. proposed a new processing algorithm used to geolocate CYGNSS L1 measurements from the WGS84 reference ellipsoid to the actual surface topography in 2020. This would allow for correction of topographical effects during measurements of the reflected GNSS signals. Ocean surface RMS heights are usually large compared to the 19 cm electromagnetic wavelength within the L-band, which leads to incoherent scattering. Hence, CYGNSS L1 ocean calibration can focus on incoherent scattering alone, unlike CYGNSS land observations exhibiting both coherent and incoherent scattering due to varieties in topography and surface cover. Besides differences in scattering mechanisms, Gleason et al. highlighted the increased complexity of geolocating land reflections compared to those measured over the ocean due to local terrain variations. In an attempt to assess this issue, the authors proposed a land geolocation algorithm using three criteria for identifying the dominant scattering location on the surface of the Earth, rather than on the WGS84 ellipsoid. These criteria were focused on delay iso-surface range agreement, Doppler iso-surface frequency agreement, and forward reflection Snell angle geometry. As a result, Gleason et al. found it possible to accurately estimate the geolocation point of the land reflected signal over more than 77.2% of the Earth's land surface [27].

Two other recent studies have investigated the effect of interpolation in the extraction of soil moisture estimates. In 2021, Senyurek et al. applied three different interpolation techniques to the CYGNSS L1 data from March 2017 to December 2019 [67]. The methods were linear interpolation, cubic interpolation, and natural interpolation. Measurements with less than -30 decibel (dB) and with an incidence angle greater than 65° were excluded, as they were considered significantly affected by noise. SMAP Enhanced L3 Radiometer Global Daily 9-km EASE-Grid SM was used to analyze the performance of each method. The interpolation was applied both spatially, i.e. in two dimensions, and spatiotemporally in three dimensions. For the 2D interpolation, one interpolation function was produced for each day, which means that nearest neighbour interpolation of one day was applied in the temporal domain. This means that the overall resolution in this study was one day temporal and $9km$ x $9km$ spatial. Two main findings occurred when evaluating the performance of the interpolation techniques. Firstly, performing the interpolation only in two dimensions performed slightly better than in 3D, with a RMSE of 0.0309 m^3/m^3. Secondly, natural interpolation performed best, being slightly better than the linear interpolation, both in 2D and 3D [67].

In 2021, Chew proposed a method to interpolate CYGNSS observations. Spatial gaps are normally handled in two ways in the literature, either by aggregating the CYGNSS data over a longer period or by increasing the size of the spatial grid boxes. Neither of these options is optimal as they lower the spatiotemporal resolution, thus limiting the de-

sired potential of CYGNSS. Furthermore, traditional interpolation techniques like linear, cubic, and spline interpolation assume the property of smoothly changing observables, which may be inadequate for CYGNSS measurements. For these reasons, the author proposed a novel interpolation technique, namely the Previously-Observed Behavior Interpolation (POBI) [12].

The POBI uses a combination of historical data and linear regression to compute the correlation between the surface reflectivity in different grid cells. The general idea is to use the correlations to discover cells that follow similar patterns. If a cell lacks a measurement at the desired point in time, the historically computed correlation with nearby cells can therefore be utilized compute a weighted average by using Equation 3.1,

$$\Gamma_{interpolated} = \frac{\sum_{i=1}^{n} w_i(a_i\Gamma_i + b_i)}{\sum_{i=1}^{n} w_i}, \tag{3.1}$$

where i represents the neighboring grid cells, Γ_i is the surface reflectivity of grid cell i, and w_i is the weight for cell i, which is equal to the correlation coefficient squared. Finally, a_i and b_i are the slope and the intercept of the best fitting line between cell i and the cell of interest [12].

The researchers implemented the interpolation algorithm, where neighboring cells were defined as being located within eight cells from each other, and concurrent measurements were defined as occurring within one week. The spatial resolution was $3km$ x $3km$. CYGNSS data from January 2018 until December 2020 was used as calibration data to calculate the historical correlations and then validated on data from 2017 in a region covering the northern parts of India and Pakistan. This area was chosen due to its significant variety in soil moisture before and after the monsoon season. Furthermore, the interpolated values were compared to 30 day averaged CYGNSS soil moisture values to retrieve information regarding the compliance of the two methods. The mean error between the 30 days aggregation and POBI was 0.17 dB with a standard deviation of 1.96 dB, suggesting that POBI can be applicable in applications where keeping a high spatial-temporal resolution is important. In other words, if the necessity of having high spatial-temporal resolution outweighs the cost of uncertainty and expensive computational power, POBI can be a suitable interpolation technique [12].

Another study focusing on processing the data before applying other models was performed in 2022. Tang and Yan highlighted the importance of data quality in data driven methods, such as machine learning, for soil moisture prediction. The authors suggest that increasing the quality of the input data will lead to improved performance of the soil moisture estimates. Based on this assumption, a quality control method was developed for assessing and eliminating poor CYGNSS data. After implementing this method, the data was further processed using a support vector machine (SVM) to produce soil mois-

ture estimates. Ancillary data, such as normalized difference vegetation index (NDVI), elevation, and slope of the surface specular point was also used as input to the SVM. These results were compared with in-situ soil moisture in Baoxie, which is located in Wuhan in China [74].

The quality control process consisted of four steps. CYGNSS data comes with different quality flags that describe the conditions at the CYGNSS receivers and the quality of the data. The first step in the quality control evaluated different flags and discarded the CYGNSS data with poor quality flags. The next step excluded measurements where the incidence angle was above 50°. The third step evaluated the distribution of the power in the produced DDMs. A DDM over land should only have one peak, hence DDMs with more than one peak were removed. This was evaluated by examining the ratio between average power around the specular point and the average power elsewhere in the DDM. If this ratio was larger or equal to 0.25, the DDM was removed. Finally, impervious surfaces were excluded by creating a spatial filter using ancillary data from the Gaofen-1 satellite (GF-1) in Baoxie. This final step removed measurements falling on buildings and inland water bodies [74].

After the quality control method was applied, 46.2% of the dataset was removed. The researchers tested different parameter combinations as input to the SVM, and the best performing model achieved a correlation coefficient of 0.87 with the in-situ soil moisture data. This value increased from 0.67 without using quality control, highlighting the importance of data quality. However, the RMSE of 0.23 cm^3/cm^3 is higher than other studies have shown. These will be presented in section 3.3. Finally, the researchers conclude that this approach should be applied to different soil types in the future to assess a global usability [74].

3.2 CYGNSS Performance Investigation

Several studies have compared the performance of CYGNSS surface reflectivity with ancillary soil moisture data products. This includes both remote sensing applications, such as SMAP, and other in-situ stations.

An early comparison between CYGNSS and SMAP was conducted in 2018 by Chew and Small [11]. The study aimed to investigate the relationship between the power of the forward scattered GNSS signals and colocated soil moisture. The power of the reflected GNSS signals, i.e. the surface reflectivity, was calculated using data produced by the CYGNSS constellation, and SMAP soil moisture was used as ground truth in the performed analysis. Initially, surface reflectivity was calculated for an area covering north India and Pakistan for May 2017 and August 2017 and compared with SMAP soil

moisture. The surface reflectivity was calculated using Equation 3.2,

$$P_{r,eff} = 10log\Gamma_{rl} \propto 10logP_{rl}^{c} - 10logN - 10logG^{r} - 10logG^{t} - 10logP_{r}^{t} + 20log(R_{ts} + R_{sr}), \tag{3.2}$$

where P_{rl}^{c} is the coherent contribution of the surface scattered signal power, P_{r}^{t} is the transmitted RHCP power, G^{t} and G^{r} is the gain of the transmitting and receiving antenna, respectively, R_{ts} is the distance between the transmitter and the specular reflection point, R_{sr} is the distance between the specular reflection point and the receiver, and N is a correction for background noise. The contribution of incoherent scattering was omitted in this study [11].

The correlation between the direct magnitude of the two variables showed a correlation of 0.45 in May and 0.65 in August. However, the temporal differences between the two months showed a stronger correlation of 0.84. This result formed the foundation of further investigation of temporal differences in CYGNSS surface reflectivity in the form of time series analysis. The time series started on the 18th of March 2017 and lasted until the 17th of February 2018. The area of interest was divided into a grid of 36 km by 36 km. Daily averaged surface reflectivity was computed for each cell before the difference from the mean value of that specific time period was extracted. This value was compared to SMAP soil moisture calculated the same way and Cosmic-ray Soil Moisture Observing System in-situ soil moisture data. The resulting correlation varied with the type of climate on the observing site. Climates with more significant fluctuations in soil moisture showed the strongest correlation of 0.74. However, despite drier climates illustrating a weaker correlation, these areas have the lowest ubRMSE of 0.017 cm^3/cm^3. The comparison with in-situ soil moisture requires that the surface reflectivity is converted to a standard soil moisture unit. This conversion was achieved using linear regression with SMAP values as ground truth, and the correlation was between 0.04 cm^3/cm^3 and 0.06 cm^3/cm^3. Finally, based on these analyses the researchers conclude that CYGNSS is sensitive to variations both in the spatial and temporal domains. This was true both for the comparison with SMAP and in-situ soil moisture. However, the researchers conclude that a soil moisture extraction method independent of SMAP is preferable and should be investigated further [11].

Clarizia et al. attempted to utilise ancillary data to describe surface related terms in their attempt to produce soil moisture estimates [17]. The approach is named the Reflection-Vegetation-Roughness (R-V-R) algorithm and it utilizes a trilinear regression to to compute the soil moisture estimates. The function will therefore depend on CYGNSS surface reflectivity, SMAP vegetation opacity, and SMAP roughness coefficient, calculated using Equation 3.3,

$$u_{CYGNSS} = a\Gamma'_{dB} + b\tau' + c\sigma' + d, \tag{3.3}$$

where Γ'_{dB} is the standardized reflectivity in the unit of dB, τ' is the standardized vegetation opacity, and σ' is the standardized roughness coefficient. The standardization is performed to directly compare the coefficients after the regression, allowing a better understanding of the effect of different surface conditions. CYGNSS level 3 data from April 2017 to August 2017 was used to implement the regression, where half of the dataset was used for training and the other half for testing by randomly splitting the dataset. This was done to minimize possible biases. The whole CYNGSS footprint was included, meaning that the R-V-R algorithm will operate semi globally after the training and testing steps are performed [17].

Table 3.1: Globally computed R-V-R coefficients, where a is the coefficient for surface reflectivity, b is for vegetation opacity, c is for surface roughness, and d is the constant term.

a	**b**	**c**	**d**
0.88	0.54	-0.18	-0.43

The coefficients for the globally computed R-V-R algorithm are presented in Table 3.1 [17]. The coefficient for the surface reflectivity is positive and the largest. This means that an increase in surface reflectivity implies an increase in soil moisture, which is according to expectations. The vegetation opacity also obtained a positive coefficient, as the algorithm has to compensate for the reduction in coherent scattering that increased vegetation opacity causes. Finally, it is noted that the surface roughness has a smaller coefficient. The semi globally RMSE was 0.07 cm^3/cm^3 [17].

Another study performed in 2019 used the NBRCS computed from CYGNSS measurements to estimate soil moisture in time series. The results were evaluated against the SMAP soil moisture product. The NBRCS property was computed using Equation 3.4,

$$\sigma_0 = \frac{\Gamma_{LR}(\varepsilon_s, \theta_i)}{MSS} e^{-\tau_v sec\theta_i}, \tag{3.4}$$

where σ_0 is the NBRCS measured at incidence angle θ_i and $\Gamma_{LR}(\varepsilon_s, \theta_i)$ is the Fresnel reflectivity of the surface. The LR subscript represents the left-handed circular polarization of the scattered signal, and ε_s is the surface relative complex permittivity. Finally, MSS is the mean square slope of surface roughness. The authors perform two filtering steps on the CYGNSS data before computing the time series. Firstly, measurements with low SNR are excluded as they are considered affected by considerable noise. Furthermore, inland

water bodies contribute to unwanted coherent scattering, and are filtered out. This is performed by computing the correlation with the CYGNSS Woodward ambiguity function (WAF). Coherent scattering leads to a strong correlation, and measurements with high correlation are excluded. The thresholds for the SNR and correlation were calculated location-specific, meaning that they were calibrated to minimize the error. Despite vegetation attenuation influencing the measured SNR of CYGNSS, the researchers did not correct this effect. They concluded that the error caused by this phenomenon was limited to below 1%. Finally, the NBRCS was scaled for incidence angle by using a simple curve fit [2].

The time series were computed for the whole CYGNSS footprint on a spatial resolution of 0.2° by 0.2° with a temporal resolution of one day between December 2017 and May 2018. Because SMAP has a temporal resolution of three days, the computed time series values were averaged over three days to make them comparable with SMAP. This led to a correlation of 0.82 and a RMSE of 0.040 cm^3/cm^3, yielding promising results for this type of analysis given the increased temporal resolution of CYGNSS. However, the method is dependent on SMAP data to solve the underdetermined system of equations used to compute the time series. This dependency is a weakness because SMAP is the same dataset used to evaluate the time series. It is therefore concluded that further work should focus on reducing the dependence of SMAP, as well as compensating for the effect of vegetation attenuation [2].

Three years later, in 2022, Al-Khaldi and Johnson published a study aiming at improving the work from 2019. The study replaced the location specific thresholds for SNR and CYGNSS WAF correlation with an improved Level-1 coherence detector as an attempt to reduce the computational complexity without significantly increasing the errors in soil moisture estimation. Furthermore, Al-Khaldi and Johnson investigated the possibility of limiting the method's dependency on SMAP. The importance of selecting optimal thresholds from SMAP to solve the underdetermined system of equations was analyzed. It was shown that this factor plays a significant role in the performance of the method. Al-Khaldi and Johnson then suggests that fixing the boundaries year-to-year can remove the dependency of contemporary SMAP data while also maintaining high performance. The time series was computed for 27 months from January 2018 to May 2020 on a 36km grid, leading to a RMSE of 0.06 cm^3/cm^3. Compared to the results presented in [2], this is slightly higher, but the computational improvements to the algorithm and the reduced dependency of SMAP can compensate for these limitations [1].

3.3 Machine Learning

In order to convert the CYGNSS produced surface reflectivity or SNR to actual soil moisture values, a geophysical model function is needed. This task can be approached in different ways, where e.g. linear regression has been presented [17]. Another approach involves training machine learning models for this purpose. As will be elaborated in this section, using machine learning allows for non-linear transformations between surface reflectivity and soil moisture to be developed.

A common approach when using machine learning to estimate soil moisture has been to utilize SMAP soil moisture data as the ground truth in supervised learning. As an attempt to avoid the dependency of SMAP, Senyurek et al. presented a study were in-situ soil moisture was used as ground truth instead of SMAP, in 2020 [65]. The resulting models where then evaluated against SMAP soil moisture. The approach built on findings presented in a previous publication by the same authors, suggesting that random forest trained using a least-squares boosting (LSBoost) ensemble strategy provided enhanced the compared to other tested ML techniques. This model was therefore used in further studies [66]. Instead of utilizing SMAP values for ground truth soil moisture values, Senyurek et al. used daily averaged in-situ SM data of 170 sites selected from the International Soil Moisture Network (ISMN), with the measured value assumed representative for a $9km$ x $9km$ area around a given ISMN site.

SMAP values with a standard spatial resolution of $36km$ x $36km$ and the enhanced SMAP soil moisture product with a resolution of $9km$ x $9km$ measured over nearly three years (18 March 2017–31 December 2019) were used for evaluation of the results. The mean unbiased RMS difference (ubRMSD) between all 3-day averaged CYGNSS and SMAP soil moisture 9km-cells globally was found to be $0.049 cm^3/cm^3$. In addition, regions that were flagged by SMAP as being poor quality, such as the Amazon rainforest and central Africa, showed a higher ubRMSD. Because these areas are less represented by the ISMN site networks used in the model training, the results coincided with the initial expectations. Excluding the grid values measured in areas that did not meet the recommended SMAP quality control factors reduced the ubRMSD to $0.041 cm^3/cm^3$, despite the model being trained independently of SMAP data. The results presented by Senyurek et al. indicate that accurate CYGNSS soil moisture estimation is possible in regions with little to light vegetation cover. However, CYGNSS soil moisture estimation in dense vegetation areas requires further research to produce results with higher accuracy [65].

Random forest was also used by Lei et al. in 2021 to estimate soil moisture based on a combination of CYGNSS data and extensive use of ancillary data. The researchers combined ancillary data from several remote sensing sources to complement the CYGNSS data. The goal was to allow the random forest to use the information about the surface

conditions to better isolate the soil moisture estimates and thereby increase performance. The soil moisture estimates were performed quasi-globally, meaning for the entire CYGNSS footprint between March 2017 and December 2019 in 3 km by 3 km grid cells [46].

A large part of the preprocessing in this approach consisted of gathering, merging, and colocating the data sources. The data included elevation from the DEM GTOPO30, soil texture data from the 250 m Global Gridded Soil Information (SoilGrids), NDVI from the Moderate Resolution Imaging Spectrometer (MODIS), vegetation water content (VWC), and dominant land cover for each grid cell from the MODIS 2018 yearly land cover type product (MCD12Q1). Common for all datasets is that they were upscaled or averaged to $3km$ x $3km$ grid cells. After the data collection, clustering was used to classify the soil type of each grid cell, where one random forest model was trained for each soil type. Furthermore, the random forest models, using 60 decision trees each, were trained with 80% of the data randomly selected and tested on the remaining 20%. The validation function was 5-fold cross-validation, and SMAP soil moisture data was used for testing. The global median correlation coefficient was 0.3502, and the median ubRMSD was 0.0426 m^3/m^3. A discovery made in this study is that regions with high vegetation cover lead to higher ubRMSD, up to 0.08 m^3/m^3. The impact of VWC is also significant, where the ubRMSD was reduced to 0.034 m^3/m^3 in regions containing a VWC of less than 4 kg/m^2. For this reason, the authors suggest that further work should pivot towards reducing the error in these regions [46].

Another study attempting to obtain global CYGNSS soil moisture estimates was done in 2021 by Jia et al.. This work wanted to investigate the possibility of using minimal ancillary data, to limit their approach's dependency on other products. Their approach was based on machine learning regression and preclassification. Furthermore, the study included testing and comparisons between the overall performance of CYGNSS soil moisture estimation utilizing different machine learning algorithms with and without preclassification. The obtained results were also compared to those presented in previous studies. In this way, Jia et al. were able to map the influence of the proven performance-enhancing ancillary data, known to be both heavy-loaded and technically challenging to obtain [36].

The preclassification strategy employs resampling and submodeling procedures based on a machine learning regression approach to minimize the influence of different land types. First, all samples of varying land types were grouped according to the International Geosphere-Biosphere Programme land type classification provided by SMAP in the resampling step. Then, distinct submodels were trained to estimate soil moisture for each land type. The estimations were based on CYGNSS measurements and SMAP soil moisture as ground truth from June 2018 to June 2020, combined with SMAP surface roughness and vegetation opacity as the only ancillary data. In addition, both SMAP and

CYGNSS soil moisture estimations were compared to in-situ soil moisture measurements from 301 different stations [36].

The results presented in the study show strong indications that the soil moisture estimations using different machine learning algorithms all show substantial improvement using the preclassification strategy. XGBoost turned out to be the model with the highest performance, obtaining a RMSE of 0.052 cm^3/cm^3. Comparing the CYGNSS and SMAP soil moisture estimates to the in-situ networks resulted in median ubRMSEs of 0.049 cm^3/cm^3 and 0.046 cm^3/cm^3 respectively. The study also presents results further indicating that reflectivity plays the primary role in soil moisture estimation. It also highlighted the effect of vegetation, as the RMSE increased in areas with dense vegetation, while surface roughness could be of more importance in dry areas [36].

A novelty to the field of research related to soil moisture estimation utilizing GNSS-R was proposed by Shi et al. in 2021. The authors referred to the absence of real-time meteorological data considered used in previous work, introducing the inclusion of precipitation and soil moisture temperature as relevant variables in soil moisture inversion. Using the so called Genetic Algorithm-Back Propagation neural network, Shi et al. combined surface reflectivity obtained from CYGNSS, surface environmental elements, meteorological data, and soil moisture data to analyse the effect the two real-time meteorological variables have on soil moisture inversion [69].

Another difference between most of the previous work being presented and the work performed by Shi et al. is the surface reflectivity value used in the analysis. In this case, the surface reflectivity was calculated from the BRCS, as expressed in Equation 3.5.

$$\tau(\theta) = \frac{\sigma(r_{st} + r_{sr})^2}{(4\pi) r_{st}^2 r_{sr}^2}, \tag{3.5}$$

where σ is the peak value of the BRCS, and r_{st} and r_{sr} are the distances between the specular point and the GNSS transmitter and receiver, respectively. The authors highlight the increased correlation with soil moisture as the reasoning behind the choice of surface reflectivity value. In addition, the researchers utilized ancillary data to estimate soil moisture. This included incidence angle, NDVI, land cover type, rainfall, temperature, and four types of topographic data - slope, slope direction, shading, and elevation. Using CYGNSS L1 and ancillary data collected over the western continental United States for January and February 2019, the finalized result showed indications of meteorological data improving the soil moisture estimations. A RMS error of 0.0344 cm^3/cm^3 was obtained when including rainfall and temperature, compared to 0.0395 cm^3/cm^3 when no meteorological data was considered [69].

Another novelty within this research field was presented by Roberts et al. in 2021. In

contrast to existing machine learning analyses relying on single-valued feature extraction from CYGNSS DDMs, this approach has the potential to extract additional information from the entire two-dimensional DDM using convolutional neural networks (CNNs). The introduction of CNNs made it possible to achieve improved soil moisture measurements, especially in regions where simple linear relations yield high uncertainty [61].

The complete neural network used in the work by Roberts et al. consisted of three parts. As the main input to the network was a 2D DDM, the first part consisted of a stand-alone CNN processing the DDM input and returning one of two different land covers as provided by the GlobCover 2009 data product. The information returned from the CNN was then combined with other ancillary data related to spacecraft and different surfaces. These include range and antenna configurations and geometry of a given measurement, incidence angle, latitude and longitude, land type, and surface topography. In addition, surface reflectivity calculated from CYGNSS measurements, as displayed in Equation 3.6, was provided as input, where P is power, G is gain, R is separation, and the t, r, and s subscripts represent the transmitter, receiver, and specular reflection point, respectively. The data was then concatenated and used to estimate soil moisture utilizing findings of modern CNN architectures like ResNet, VGG, and AlexNet, adding dropout layers, image augmentation, and the usage of smaller kernels in order to improve model performance [61].

$$P_r = \frac{P^t G^t G^r}{(R_{ts} + R_{sr})^2} \left(\frac{\lambda}{4\pi} \right)^2 \Gamma \tag{3.6}$$

The use of a CNN compared to an artificial neural network (ANN) is justified by comparing model performance when classifying bare or forested land cover areas. Using an ANN with the single-valued peak power as input yielded an F1-score of 76%, while a flattened DDM input resulted in an F1-score of 87% using a similar network. On the contrary, a CNN taking the original DDM as an input obtained an F1-score of 89% and was therefore used throughout the presented analysis, with a goal of making direct comparisons between the proposed method and mature CYGNSS-based soil moisture products like SMAP and UCAR.

Global comparisons were made between the three products on the lowest possible common resolution, being a 36km averaged daily soil moisture distribution. Overall global trends showed similarities. However, the CNN estimated less soil moisture variation in areas with exceedingly high soil moisture content measured by SMAP. This is a result of SMAP and UCAR values in the respective regions being considered of low quality and were removed from the training process in the network of Roberts et al.. Another advantage of the proposed approach is the obtainable resolution of 3km, being a significant improvement from the 36 km used in the other products. Even though the predictions are

inherently limited to the accuracy of SMAP measurements, having an unbiased RMSE of 0.04 m^3/m^3, the network showed trends of better soil moisture estimations in areas where SMAP and UCAR are prone to low quality values, such as over the Amazon and central Africa [61].

3.4 State of the Art in Remote Sensing of Soil Moisture

The research area of soil moisture estimation using remote sensing has developed over the last decades. The spatial and temporal resolution of the data products continues to improve with the inclusion of new research projects. However, recent studies tend to focus on specific areas of the problem, resulting in isolated improvements that have not been investigated combined.

One major challenge in global soil moisture estimation is the lack of global ground truth measurements. Previous studies have used a variety of datasets derived from ground based methods or other remote sensing missions like SMAP. The main issue when comparing ground based measuring methods to CYGNSS soil moisture estimates is the difference in spatial coverage. One of the main benefits of using CYGNSS is the semi global coverage, which is difficult to assess when compared to in-situ measurements. This problem can be dealt with by using SMAP as the basis for performance evaluation. However, SMAP has its sources of errors and configurations that will lead to differences compared to CYGNSS. Therefore, it is challenging to determine which of the sources are wrong when they differ from each other. Products like UCAR/CU are mainly based on CYGNSS measurements and should therefore not be used to evaluate CYGNSS.

Gleason et al. discussed the challenges regarding varying topography. As CYGNSS satellites initially were designed to measure reflections over the ocean, land topography has not been taken into account. Therefore, the study proposed an algorithm for geolocating the measurements over land to the actual surface topography, which can lead to improved performance in remote sensing applications [27].

The work by Tang and Yan showed that performing data quality assessment before applying data driven methods in soil moisture estimation can improve the resulting errors. Their models did not produce state of the art RMSEs but demonstrated an improvement when filtering the CYGNSS data based on different criteria. The effect of interpolation when colocating data sources or filling data gaps was also discussed. POBI was proposed as state of the art, but the algorithm requires immense computational power. Natural interpolation, or even linear interpolation, can therefore be considered for practical reasons.

Several studies comparing CYGNSS with other sources were also presented. The studies differed in their approaches to converting the CYGNSS measurements to soil moisture,

varying from linear regression to more complex methods. However, the trends in these results indicate a correlation between CYGNSS and other soil moisture products. Some studies also focused on the parameters affecting the CYGNSS measurements, where vegetation opacity was one of the more important. Areas like the Amazon rainforest proved challenging due to their dense vegetation cover. In drier regions, the surface roughness proved to be of greater importance. In conclusion, analyzing the optimal CYGNSS conditions proved to be challenging due to the large variety of dependencies that exist. Therefore, analyzing this area can be helpful not only for CYGNSS but for improving future space missions.

Machine learning has been the most adopted method to create a geophysical model function capable of converting CYGNSS measurements to soil moisture estimates. These techniques are also considered to be state of the art. Different approaches, such as random forest, ANNs, CNNs, and gradient boosting algorithms have all been tested and produced RMSE of similar magnitude. The choice of the model itself seems to be of less importance than which parameters and ancillary data are selected. Furthermore, optimal machine learning approaches should be applied in combination with data quality assessment to feed the models with the most accurate data.

Chapter 4

Data Products and Data Processing

The datasets used to investigate GNSS-R for remote sensing of soil moisture are presented below. CYGNSS data is the primary source, which will be evaluated against SMAP and ERA5. In addition, a DEM is presented and used as ancillary data. All data sources use the WGS84 reference ellipsoid to define the surface of the Earth.

4.1 Cyclone Global Navigation Satellite System

The Cyclone Global Navigation System (CYGNSS) is a space mission conducted by NASA, initially with the primary goal collecting space-based measurements of wind speed in tropical cyclones to support and improve hurricane forecasting. CYGNSS space segments are designed to utilize signals transmitted from operational satellites in the American GPS constellation, in addition to the reflection of the same satellite's signals reflected of the surface of the Earth to gather information about the Earth's surface. On the 15th of December 2016 NASA launched eight LEO micro-satellites, which in combination with already transmitting GPS satellites, form bistatic radar systems as previously explained in section 2.2. Prior to CYGNSS, a typical technique when conducting spaceborn wind speed measurements over oceans was scatterometry. It consisted of a radar instrument aboard a satellite sending a signal to the ground and then measuring the strength of the reflected signal. However, building both sending and receiving capabilities into a single instrument is considerably more expensive than the CYGNSS approach. Signal degradation of the transmitted microwave pulses when passing through intense rainfall is also a well-known limitation of traditional scatterometry, causing difficulties in retrieving observations of high wind speeds in critical regions of tropical storms. On the other hand, reflected GPS signals operate at lower microwave frequencies, enabling

the possibility to penetrate thick clouds and precipitation. Thus, utilization of CYGNSS satellites will result in both a more robust measuring technique and significant cost savings [63].

In addition to providing ocean wind speed measurements at a reduced cost, with a total mission cost of 150 million USD, the CYGNSS constellation contributes to an improved spatiotemporal resolution compared to traditional single, polar-orbiting satellites. Using an orbital inclination of 35 degrees from the equator, CYGNSS measures ocean surface winds between approximately $\pm$ 38 degrees latitude and $\pm$ 180 degrees longitude, representing the critical latitude band for tropical cyclone formation and movement. The predefined inclination results in CYGNSS providing nearly gap-free Earth coverage of the $\pm$ 38 $^\circ$ latitude band, with a mean revisit time of 7.2 hours, a median revisit time of 2.8 hours, and a spatial resolution of 25x25 km over rough surfaces such as the ocean [62, 16].

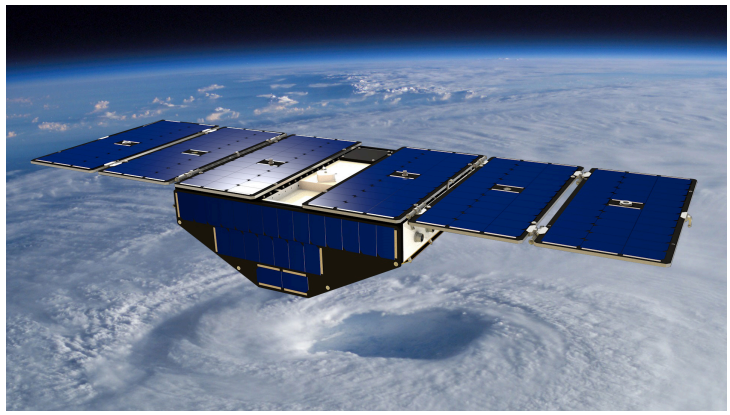


Figure 4.1: Visualization of a CYGNSS micro-satellite deployed in space above a hurricane.

Source: https://www.nasa.gov/

4.1.1 CYGNSS Satellites

Figure 4.1 presents a visualization of a CYGNSS micro-satellite deployed in space. The CYGNSS satellite is a three-axis-stabilized, nadir pointed vehicle using a star tracker for primary altitude knowledge and a reaction wheel triad for control. The spacecraft consists of a micro-satellite platform hosting a DDMI used to generate relevant data from the measurements made by the satellite. In addition to the DDMI, CYGNSS satellites are equipped with one RHCP antenna on the zenith side of the spacecraft and one LHCP antenna on the nadir side of the spacecraft. The RHCP antenna receives signals directly from the transmitting GPS satellites, and the LHCP receives the scattered signal. While

the direct signal is used for positioning of the DDMI, the reflected signal can be used to derive properties of the reflecting medium [7, 11, 17]. Each DDMI measure signals at 2 Hz and can receive four signals simultaneously. This means that the eight LEO DDMIs of CYGNSS conduct 64 measurements per second in total [62]. Even though the CYGNSS micro-satellite was designed for ease of manufacture and integration to provide a cost-effective solution across the constellation, the satellite structure and shape are specifically configured to allow a clear nadir and zenith field of view for DDMI antennas on board. Besides the LHCP and RHCP L-band antennas used for scientific measurements, CYGNSS satellites are also equipped with an S-band (2GHz) transceiver providing a low-cost, radiation-tolerant communication system developed not to interfere with ongoing science operations [42].

The electrical power subsystem is designed to perform battery charging without interrupting science data acquisition and is based on a 28±4 V dc primary power bus with electrical power generated by fixed solar arrays. Using triple junction solar cells, with most panels facing outwards, CYGNSS satellites can generate 70 W of power, resulting in a 30% margin of the power necessary to operate. A simplified overview of a CYGNSS micro-satellite and where the different components are located on the spacecraft is presented in Figure 4.2. Because the figure is viewed from underneath, some presented components such as the S-band and RHCP L-band antennas are not visible. The DDMI is also hidden inside the lightweight 25kg spacecraft [63].

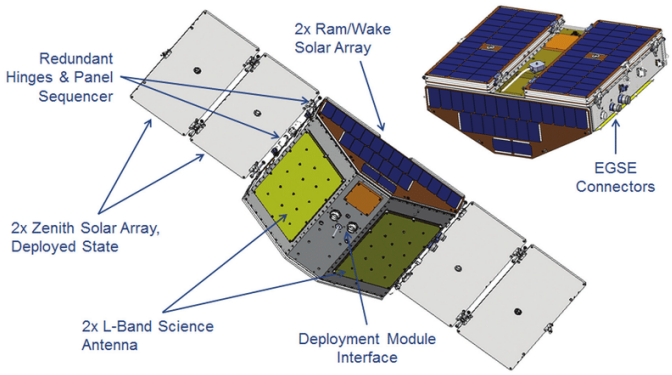


Figure 4.2: A simplified overview of a CYGNSS micro-satellite.

Source: [62]

It is also important to highlight how Figure 4.2 is only a simplified overview of a CYGNSS micro-satellite and its components. A more technical description of the satellites and the science behind the mission operations can be found in official mission documents [62] or related work [42].

4.1.2 CYGNSS Data Products

The main objective of the CYGNSS mission is to obtain more accurate information related to tropical cyclones and ocean wind speed. Hence, returned data from CYGNSS satellites are expected to expand previous knowledge of the rapidly changing environment in the core of these tropical cyclones under development. CYGNSS produces four levels of data products, of which all are publicly available, and the CYGNSS science operations center is responsible for the processing, development, and dissemination of these. The different levels are referred to as level 0 to level 3 data products. Levels 1, 2, and 3 are available in the format Network Common Data Form (NetCDF) and are published through the NASA Physical Oceanography Data Active Archive Center (PO.DAAC) at maximum data latency from spacecraft downlink to availability of six days [62]. NetCDF is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data and has proceeded to be a community standard in the latter. Developed and maintained by the Unidata Program Center, the netCDF format has become popular for its many beneficial attributes. These include the format being self-describing, easily accessible, and scalable [59]. A more descriptive presentation of the advantageous attributes of the netCDF format is found in Table 4.1.

The level 1 data product includes calibration DDMs of received power and calibrated DDMs of BRCS. The values have been converted from the measured raw counts (level 0) via received power in watts to unnormalized BRCS, as explained in subsection 2.2.2 and subsection 2.2.3. In order to calculate the NBRCS using Equation 2.11, the data product also includes a DDM of effective scattering areas. The spatial resolution of both DDMs is 17 delay x 11 Doppler, corresponding to a physical surface area of approximately 50 km^2. In addition to the DDMs, the data product also contains metadata used to convert the level 0 data to level 1 data.

Data available in the level 2 data product contains spatially averaged wind speed and MSS over a $25km$ x $25km$ region, centered at the specular point geolocated in spacecraft time and space coordinates. This data product will not be used in this study. However, the fully calibrated data product of this level is commonly used by the CYGNSS science team for their own research, as MSS and wind speed are important variables in tropical cyclone investigation [62, 18].

Level 3 includes surface wind speed averaged in space and time on a 0.2° x 0.2° latitude and longitude grid and will not be used in this study. An overview of the different data products and latency is found in Table 4.2.

Table 4.1: Important attributes of the netCDF format.

Attribute	Description
Self-Describing	A netCDF file includes information about the data it contains.
Portable	A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
Scalable	Small subsets of large datasets in various formats may be accessed efficiently through netCDF interfaces, even from remote servers.
Appendable	Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
Sharable	One writer and multiple readers may simultaneously access the same netCDF file.
Archivable	Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

Source: [59]

4.1.3 CYGNSS Data Download and Processing

After the launch at the end of 2016, CYGNSS satellites have produced vast amounts of data available for public users since early 2017, and the data products are continuously updated. This study will use the level 1 data product with the intention of utilizing calibrated DDMs to obtain better knowledge related to soil moisture at a higher spatiotemporal resolution than existing methods. The data product is structured in at most eight netCDF files each day, corresponding to a unique DDMI in the CYGNSS constellation, with six to eight CYGNSS micro-satellites retrieving data each day under normal conditions. In addition to the DDMs calibrated into received power and BRCS, each file from the level 1 data product contains a large number of other engineering and science measurement parameters, such as sets of quality flags and indicators, error estimates, and a variety of orbital, spacecraft/sensor health, timekeeping, and geolocation parameters.

CYGNSS satellites measure different parameters every half a second, resulting in a huge amount of data being accumulated and available in each file. Downloading and processing this data over several years on a global scale require immense time- and storage

Table 4.2: CYGNSS data products and publication latency.

CYGNSS Product	Description	Latency
Level 0	Raw data of total system power (received signal + instrument noise).	6 days
Level 1	Calibrated DDMs of receiver power and bi-static radar cross section.	6 days
Level 2	Spatially averaged mean square slopes and wind speed over a 25 km x 25 km region centered at the specular point, geolocated in spacecraft time and space coordinates.	6 days
Level 3a	Wind speed, gridded in space and time (1/4° latitude and longitude, 3 hours).	6 days
Level 3b	Wind speed, gridded and optimized for observing system experiment data assimilation (optimized spatial and temporal resolution).	6 days

Source: [42]

capacities, which in most cases is adverse for individual users. However, PO.DAAC has made all data available through the Open-source Project for Network Data Access Protocol (OPeNDAP) data access framework. This makes it possible to remotely request specific parts of the available data, depending on the application. An approach to retrieve data from the PO.DAAC storage drives using OPeNDAP, and avoid having to download all available data, is to generate specific retrieval-URLs based on a specific year, month, day, satellite number, and the desired parameters. In combination with the Pydap client, being a Python library implementing OPeNDAP used to access scientific datasets in an efficient way, the generated URLs are used to access and download the desired data. Code 4.1 presents how the URLs are generated in this study.

```
def generate_url(year: int, month: int, day: int, satellite_nr: int):

    day_of_year = datetime(year, month, day).timetuple().tm_yday
    date_string = str(year) + str(month).zfill(2) +
                  str(day).zfill(2)

    base_url = 'https://podaac-opendap.jpl.nasa.gov/opendap/'
               'hyrax/allData/cygnss/L1/v3.0/'
```

```
specific_url = str(year) + '/' + str(day_of_year).zfill(3)+
               '/cyg0' + str(satellite_nr) + '.ddmi.s' +
               date_string  + '-000000-e' + date_string +
               '-235959.l1.power-brcs.a30.d31.nc'

data_url = base_url + specific_url

return data_url + '?sp_lat,sp_lon,ddm_timestamp_utc,ddm_snr,'
      'gps_tx_power_db_w,gps_ant_gain_db_i,rx_to_sp_range,'
      'tx_to_sp_range,sp_rx_gain,spacecraft_num,'
      'prn_code,track_id,quality_flags,quality_flags_2,'
      'sp_inc_angle', day_of_year
```

Code 4.1: Code for generating specific URLs for optimized data retrieval. The Pydap client is then used to retreive data based on the generated URL.

With code for generating URLs based on different dates and satellites, data for the years 2019, 2020, and 2021 is collected for use in this study. Due to the introduced normality of not receiving data from all eight satellites, a complete overview of missing satellite measurements and at what time these were measured is found in Table A.1 in Appendix A. As displayed in the *base_url* variable in Code 4.1 version 3.0 of the data product is collected. The reasoning behind this is that version 3.0 is the most recent version of level 1 data available at PO.DAAC. It supersedes version 2.1 with improvements related to changes in calibration and processing, including adjustments to level 1 calibration parameters and allowance for BRCS calibration to be corrected for variations in GPS transmit power.

Specific parameters were chosen based on the desire to investigate surface reflectivity and soil moisture. As to be presented in section 5.1, Equation 5.5 highlights six important parameters in the computation of surface reflectivity from CYGNSS level 1 data. These include the peak received signal power, the power of the transmitted GPS signal, and distances from the reflecting surface to both the transmitting GPS satellite and the orbiting CYGNSS satellite receiving the reflected signal. In addition to the parameters used in surface reflectivity calculations, the features selected and downloaded for processing purposes are presented in Table 4.3.

To facilitate data filtration and manipulation based on desired locations, time intervals, and data quality, all collected data was stored and processed in Pandas Data Frames. Data Frames are two-dimensional data structures containing tabular data. Using this data structure simplified the introduction of additional data columns constructed based on the already available data to be used in further analysis. This includes features such as *ddm_channel*, day_of_year, $time\_of\_day$, and $hours\_after\_jan\_1st\_2019$ representing one of four DDM measurement channels located on each spacecraft, the overall day number to when the measurement is made, the time of day of the measurement, and the

Table 4.3: The additional CYGNSS parameters downloaded for processing purposes.

Parameter	Description
sp_lat	The latitude coordinate of the specular point.
sp_lon	The longitude coordinate of the specular point.
ddm_timestamp_utc	The number of seconds since the start of the day.
spacecraft_num	The CYGNSS spacecraft number of the received signal.
prn_code	The PRN code of the satellite transmitting the received reflected signal by CYGNSS.
track_id	The id of the track, or line of continuous measurements being made each half second, the reflected signal is a part of.
quality_flags	A bit string indicating the presence of different conditions having affected the measured values.
quality_flags_2	A bit string indicating the presence of a number of additional conditions having affected the measured values.
sp_inc_angle	The specular point incidence angle.

total number of hours since January 1st, 2019 at the time of measurement respectively.

As some of the collected features were used to calculate surface reflectivity values, all rows containing pre-determined fill values were removed. These are added if a CYGNSS spacecraft fails to measure a specific value and could preferably, as a result, be neglected in order to prevent results affected by wrong input.

Furthermore, measurements containing bad quality values were removed. The CYGNSS L1 data product includes quality control flags designed to indicate potential problems with the measurements [62]. These quality flags made it possible to filter out measurements with specific quality notations to produce better results in terms of quality and reliability. The most intuitive quality flag to take into account would be flag 1, indicating bad overall quality. However, as the CYGNSS project initially was designed for data collection and use over the ocean, most land data would be removed if filtered by quality flag 1. As a consequence, the measurements indicating bad overall quality remained, while the specific quality flags used to filter available data were 2, 4, 5, 8, 16, and 17. These flags are related to the S-band transmitter being powered up, spacecraft attitude error, black body DDM, DDM being a test pattern channel, the direct signal in DDM and

low confidence in the GPS EIRP estimate, respectively, and measurements consisting of any of these flags were removed as suggested by the work done by Rajabi et al. [58]. The quality flags are stored as binary bit strings. For example, a measurement with the quality flags 2 and 4 will have the number 1 on positions two and four of its corresponding quality flag bit string and zeros elsewhere. By applying the bitwise AND operator between the quality flag and a mask containing ones on locations 2, 4, 5, 8, 16, and 17, the unwanted observations can be filtered out.

The provided track ids are only unique for each satellite within one day only. Therefore, overall unique track ids were computed using the provided track id, PRN code, satellite number, and the day of the year to distinguish between different tracks. The PRN codes were also used to categorize the satellites into different blocks based on satellite configuration and type. This made it possible to e.g. filter out data obtained from aging legacy satellites, analyze potential differences in transmitting power due to satellite blocks, or inspect modernized satellites only. A complete list of which PRN codes belong to each satellite block type is presented in Table 4.4, retrieved from the U.S. Coast Guard Navigation Center[1].

Table 4.4: A complete overview over the transmitting GPS satellites and what kind of satellite block, representing the satellite generation and configuration, they belong to. As of 20th January 2022 satellites with PRN 11, 22 and 28 are unavailable as 28 is retired, 22 is unhealthy and 11 is testing in preparation for operation.

Satellite Block	Satellites PRNs
IIR	2, 13, 16, 19, 20, 21, 22, 28
IIR-M	5, 7, 12, 15, 17, 29, 31
IIF	1, 3, 6, 8, 9, 10, 25, 26, 27, 30, 32
III	4, 11, 14, 18, 23, 24

Lastly, surface reflectivity values were scaled down by subtracting the minimum surface reflectivity value from all other values, as done in previous work by Rajabi et al. and Chew et al. [58, 13]. The processed data was then filtered based on desired location and time interval to facilitate comparison and in-depth analysis in combination with other datasets.

[1] https://www.navcen.uscg.gov/?Do=constellationStatus

4.2 Soil Moisture Active Passive

Soil Moisture Active Passive (SMAP) is a NASA conducted space mission consisting of one spacecraft, launched on 31st of January 2015. Initially, the mission lifetime was three years, thus SMAP has been in extended operation since 2018 and is still operational at present. NASA's budget was slightly below one billion USD, which includes design, development, launch, and operations [22].

One of the main goals of SMAP is to develop enhanced flood prediction and drought monitoring capabilities, which includes spaceborne remote sensing of soil moisture over land. Furthermore, classifying soil in colder areas of the Earth as frozen or thawed is another output of SMAP.

The orbital design is near-polar, Sun-synchronous at 685 kilometers altitude, and uses 98.5 minutes for one complete Earth orbit. As a consequence of the orbital design, the temporal resolution differs near the poles and at the equator. While two days at the poles, the temporal resolution is three days at the equator.

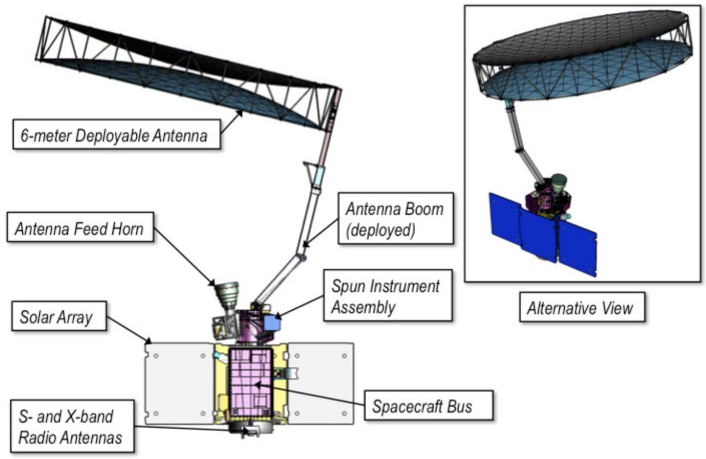


Figure 4.3: Overview of the SMAP satellite.

Source: [22]

4.2.1 SMAP Satellite

The SMAP satellite, shown in Figure 4.3, consists of two instruments, a radar and a radiometer, which makes the satellite capable of performing both active and passive remote sensing of the surface of the Earth. Active measurements possess the attribute of

high spatial resolution but lower soil moisture sensibility. On the other hand, passive measurements are sensitive to soil moisture but are limited to a coarse spatial resolution. By joint processing of these two techniques, the goal was to develop a product sensitive to changes in soil moisture with an improved spatiotemporal resolution compared to existing sources.

The radar emits RF signals at L-band (1.26 GHz) to limit the impact of weather conditions and vegetation cover, thus being responsible for the active measurements. After the signals backscatter off the Earth's surface, the radar measures the amount of returned signal. The radiometer is a sensitive receiver that measures the naturally emitted RF energy of the Earth at L-band. It works similarly to an infrared camera, where warmer objects appear brighter than colder ones. At the specific operating frequency, water appears cold and dry soil appears warm, distinguished by a significant distance of approximately 200K. This makes it possible to measure soil moil moisture with higher accuracy than the radio. Both the radio and the radiometer share one common antenna, shown in Figure 4.3. The antenna is six meters and is tilted 40 degrees such that the actual measurement distance to the Earth is 500km instead of the 685km orbital altitude. The antenna feed horn illuminates the large golden plated mesh at the reflector, creating a beam that reaches the Earth. The whole antenna system, including the feed horn and the boom, rotates around its own axis, thus illuminating a larger area on the Earth. In addition to the above mentioned components, the satellite carries a solar array that transforms sunlight into electrical power and S- and X-band radio antennas used to communicate with the Earth.

4.2.2 SMAP Data products

SMAP produces data on five different levels, made publicly available through Alaska Satellite Facility and National Snow and Ice Data Center. The data products are described in Table 4.5, where the latency is the time from observation acquisition of instrument data unit it is publicly available.

In this thesis, the Level 3 data will be used as ancillary data to evaluate the performance of CYGNSS. This level contains soil moisture measurements of the top 5cm of the soil, re-sampled to a spatial resolution of 36km by 36km and a temporal resolution of 1 day. The dataset covers $\pm 85°$ latitude and the whole range of longitudes [22].

In addition to soil moisture, SMAP also produces ancillary data products, which contribute to describing geophysical conditions on the surface of reflection. Two such features are vegetation opacity and surface roughness. Vegetation opacity is a global daily composite description of the density of vegetation gridded on the same spatiotemporal resolution as the soil moisture product [22]. SMAP's roughness coefficient represents a

Table 4.5: SMAP data products and publication latency.

SMAP Product	Description	Latency
Level 0	Reconstructed, unprocessed instrument data.	n/a
Level 1	Time referenced, geolocated, calibrated, and corrected level 0 data.	12h
Level 2	Geophysical parameters, such as soil moisture, derived from the Level 1 data.	24h
Level 3	Geophysical parameters, such as soil moisture, derived from the Level 1 or Level 2 data, that have been spatiotemporally resampled to a global grid.	50h
Level 4	Geophysical parameters derived by assimilating the lower level data and land surface models.	7 days

daily global composite description of the surface roughness coefficient as a scaled value between 0 and 1. The coefficient is identical to the *h* parameter in the *tau-omega* model, commonly used in soil moisture retrieval algorithms, representing irregularities in the soil surface [68, 23].

4.2.3 SMAP Data Processing

Similar to the CYGNSS data, SMAP soil moisture was downloaded as NET-CDF files. However, each day of data was separated into different files. This meant that traversing file structures were necessary before acquiring the data, as shown in Code 4.2. First, all subfolders are explored to retrieve the path of all SMAP netCDF files. Then, the data in all these files are gathered in one Pandas DataFrame and returned. Each filename is time referenced, and this is used to validate that the correct files are processed. After the data was read, the processing continued similar to the one of CYGNSS, as explained in subsection 4.1.3.

```
def get_smap(root_path: str, years: list, months: list, days: list):
    first = True
    sub_dirs = []
    filenames = []
```

```
    for dir_name, sub_dir_list, file_list in os.walk(root_path):
        if first:
            sub_dirs = sub_dir_list
            first = False
        else:
            filenames.append(file_list[0])

    smap_df = pd.DataFrame()

    for i in progressbar(range(len(sub_dirs))):
        current_day = int(filenames[i].split('_')[5][6:8])
        current_month = int(filenames[i].split('_')[5][4:6])
        current_year = int(filenames[i].split('_')[5][:4])

        if (current_day in days) and (current_year in years)
                                 and (current_month in months):
            current_path = root_path + '/' + sub_dirs[i] + '/'
                         + filenames[i]
            current_df = get_smap_df(current_path)
            smap_df = smap_df.append(current_df)

    return smap_df
```

Code 4.2: Code for reading and processing SMAP data.

4.3 ERA5

The European Centre for Medium-Range Weather Forecasts (ECMWF) uses retrospective analysis (reanalysis) to produce a dataset of the global atmosphere, land surface, and ocean waves from 1950 until the present. This means that the method differs from CYGNSS and SMAP by not being a method solely dependent on spaceborne remote sensing. The final product is available with a latency of two to three months, but a preliminary product with higher uncertainty is published after five days. This dataset contains global coverage soil moisture estimates on a regular latitude/longitude grid with a spatial resolution of 0.25 by 0.25 degrees and a temporal resolution of one hour [32].

The reanalysis in the ERA5 soil moisture product is re-sampled over areas covering lakes. These areas were removed using another dataset from ECMWF called Land Sea Mask. This dataset contains values for the proportion of land as opposed to the ocean or inland water bodies in grid boxes with the resolution of 0.1° x 0.1°. Because lakes remain constant in the spatial domain, the dataset is considered invariant in time. Each value ranges from zero to one and is dimensionless.

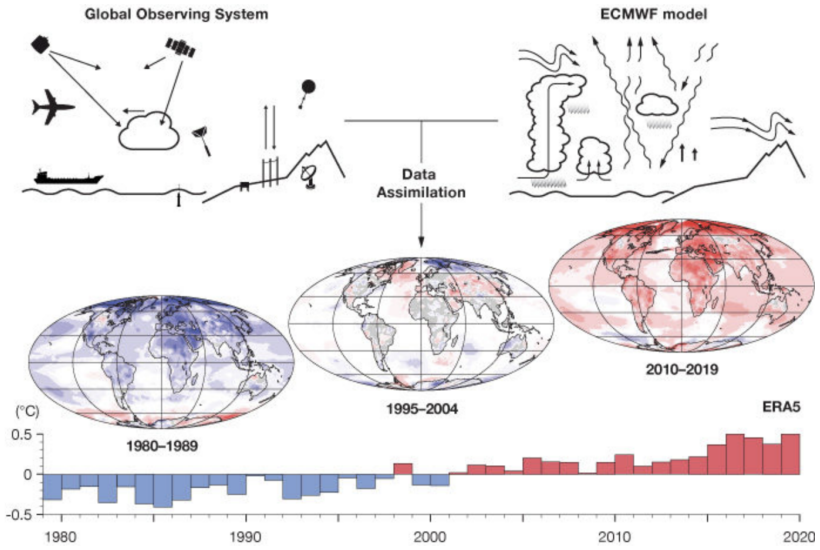


Figure 4.4: Reanalysis overview.

4.3.1 Reanalysis

Reanalysis combines unevenly distributed measurements, both in the temporal and spatial domain, and physical models to produce e.g. climate models. A schematic overview of the approach is illustrated in Figure 4.4, where the assimilation of numerous observation techniques and models makes it possible to not only produce climate data for recent years but also to have a valid representation of the situation several decades in the past. This property allows e.g. monitoring long-term consequences of natural catastrophes such as earthquakes and volcanic eruptions. However, the data assimilation often includes heavy numerical computations, which require intense computing power.

4.3.2 ERA5 Data Processing

ERA5 soil moisture data was collected as NET-CDF files[2], similar to the likes of CYGNSS and SMAP. All data was downloaded as one single file, making file traversing unnecessary. The main variable is volumetric soil water level 1 (vswl1), which gives the amount of water in the top seven centimeters of the soil in the unit m^3/m^3. Furthermore, soil moisture values over lakes were filtered out by applying the Land Sea Mask dataset. All areas with a Land Sea Mask value larger than 0.5, meaning that the grid cell contains

[2]Downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form

more water than land, were removed, resulting in a dataset to evaluate the performance of CYGNSS.

4.4 SRTM's Digital Elevation Model

The Shuttle Radar Topography Mission (SRTM) is a joint project between the National Geospatial-Intelligence Agency and NASA. Launched aboard the Endeavour space shuttle in February 2000, the 11 days lasting mission aimed to collect digital elevation data for all Earth land areas between 60° north and 56° south latitude. SRTM used a technique called radar interferometry for data collection, meaning that two radar images are taken from slightly different locations before the differences between these allow calculation of surface elevation. Orbiting the Earth 16 times each day, SRTM collected 12 terabytes of raw radar echo data during the up-time period. This data was later used to provide a gridded DEM with a spatial resolution of 1-arc-second, equivalent to approximately 30 meters, and a vertical accuracy of 16 meters with 90% accuracy [80].

4.5 Data Demonstration

4.5.1 Spatiotemporal Resolution Comparison between CYGNSS and SMAP

One of the essential differences between the CYGNSS and SMAP space missions is their spatiotemporal resolutions, as described in section 4.1 and section 4.2. To demonstrate this difference, CYGNSS measurements from the first 90 minutes, seven hours, and 24 hours of 2020 have been plotted against SMAP measurements from the first day, the first two days, and the first three days of 2020 in Figure 4.5. The spatial coverage is slightly worse around the equator than nearer the poles for both CYGNSS and SMAP. This is a result of the orbital inclination of 35° and 90° for CYGNSS and SMAP, respectively. Furthermore, it is clear that the spatial coverage of CYGNSS is considerably better than SMAP between $\pm 38°$ latitude.

4.5.2 CYGNSS Surface Reflectivity Demonstration

To provide an intuition of how CYGNSS surface reflectivity responds to different surface types, this value was plotted over Kenya for the first three days of 2020. The surface reflectivity computation will be elaborated in section 5.1. This specific region, which is shown in Figure 4.6 (a), was selected because it contains a large lake that makes it possible to observe the behavior of surface reflectivity moving from land to water bodies.

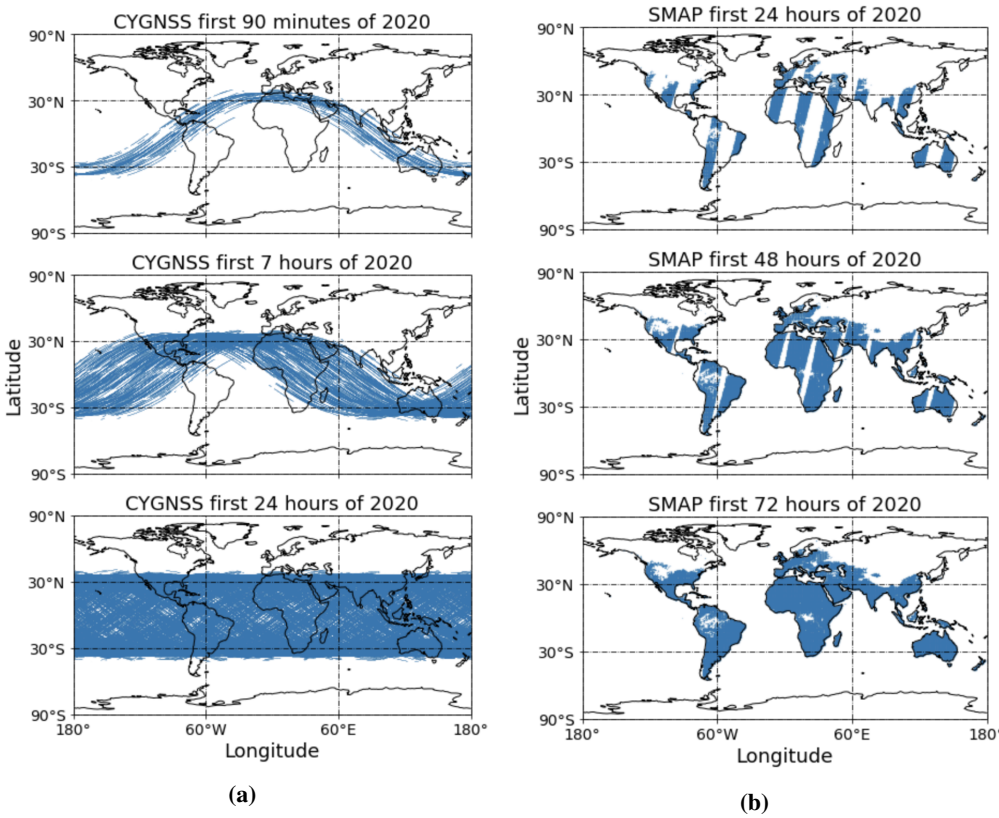


Figure 4.5: CYGNSS spatial coverage during the first day of 2020 (a) and SMAP spatial coverage during the first three days of 2020 (b).

Figure 4.6 (b) shows the plotted CYGNSS measurements. Areas covering lakes and rivers experience an increase in surface reflectivity, indicated by the yellow color in the figure. On the other hand, the northwest corner suggests dryer land.

4.6 Programming Language

One of the necessities of the selected programming language is its performance on extensive datasets. During the analysis in this thesis, data exceeding 1 TeraByte will be downloaded and processed, which will be difficult without a programming language capable of performing well on such tasks. Furthermore, the language must allow rapid code development to make it possible to achieve the desired results within this thesis' time frame.

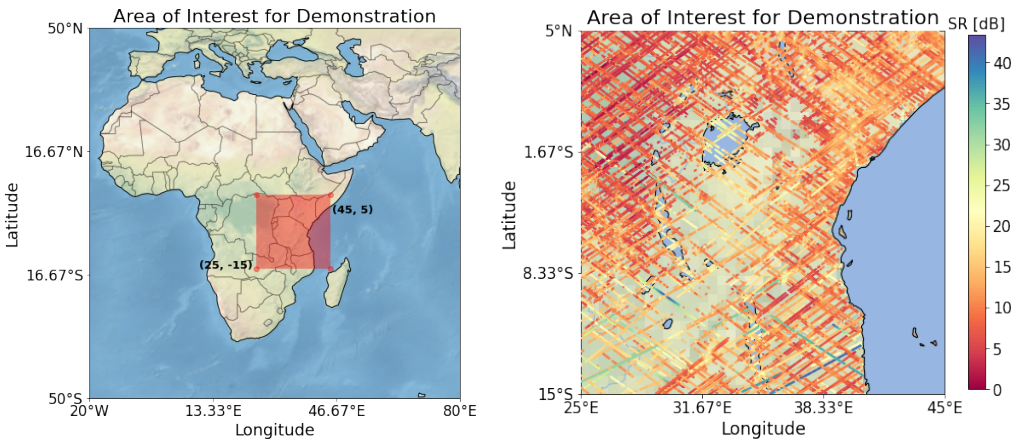


Figure 4.6: The selected area for the demonstration of the behaviour of CYGNSS (a) and the measured surface reflectivity in the unit of decibels (b), using measurements that were collected from January 1st to January 3rd 2020.

The dynamic and interpreted programming language Python was selected to suit the needs mentioned above. Dynamic programming languages perform numerous operations at run time that static languages perform at compile time [56, 44]. Consequently, dynamic languages lack the possibility of performing several performance optimization steps relied on by static languages. However, numerous Python packages, such as Numba, Pandas, and Numpy, have been developed to overcome this issue, which again has increased Python's popularity in scientific programming. These packages allow high-performance programming in Python. A study conducted in 2019 compared more than 20 Python libraries to evaluate the performance of Python as a tool for data mining and big data analysis and concluded that Pandas is a well suited option for data preparation [73]. In addition, Python is well documented and, web pages like Stack Overflow provide support if needed. These reasons made Python a fitting choice for the analysis.

4.7 Interactive Web Page

In collaboration with the department of civil and environmental engineering at NTNU an interactive web page with the aim of publishing GNSS-R related results was developed. This web page is a toolbox providing the opportunity to download NASA's CYGNSS data for a selected area and time frame, visualise data related to each individual satellite track, and a selection of ground based GNSS-R data visualizations. As a part of this thesis, the toolbox was extended with an option for visualizing CYGNSS- and SMAP

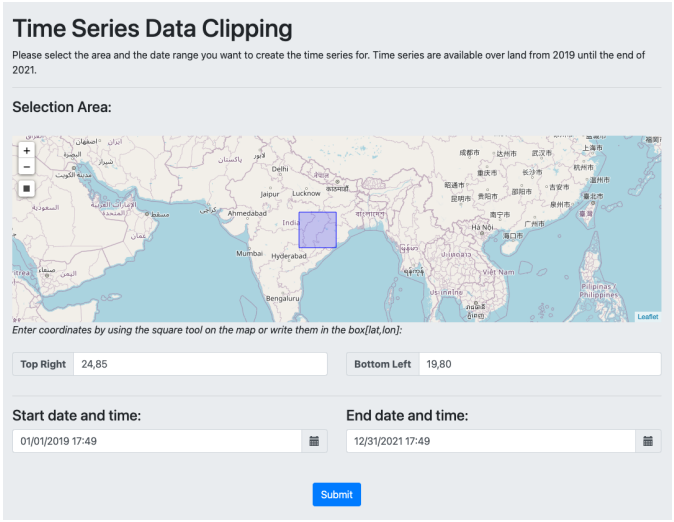


Figure 4.7: A screenshot of the web page showing the time series data filtering where the 5° x 5° area in India is selected.

time series.

The web page is written using Python Django, which is a high-level framework that allows rapid development of web applications. One advantage of using Python for web development is the uncomplicated process of adding packages to support additional functionality. Basemap and Leaflet were used to plot maps, where e.g. the interactive map was possible due to Leaflet. Leaflet provides the opportunity for users to draw a rectangle over the desired area instead of manually typing the coordinates. GDAL was used to translate between raster and vector geospatial data. Furthermore, Chart.js is a JavaScript based library enabling the creation of interactive charts.

The home page presents the users with the different GNSS-R tools. The additional time series tool allows the user to select a spatial and temporal filter before being presented with both CYGNSS surface reflectivity and SMAP soil moisture time series for the selected region. A screenshot of the data selection page is shown in Figure 4.7, where time series over 5° x 5° area in India from 2019 until 2021 is selected. After the user submits the form, the time series graph containing both CYGNSS and SMAP appears, as shown in Figure 4.8.

Due to performance challenges regarding the download of CYGNSS data, the server has been extended with a processed CYGNSS dataset from 2019 to 2021 on a 0.5° x 0.5° spatial resolution. The same applies for SMAP, which means that it is only possible to produce time series in this period.

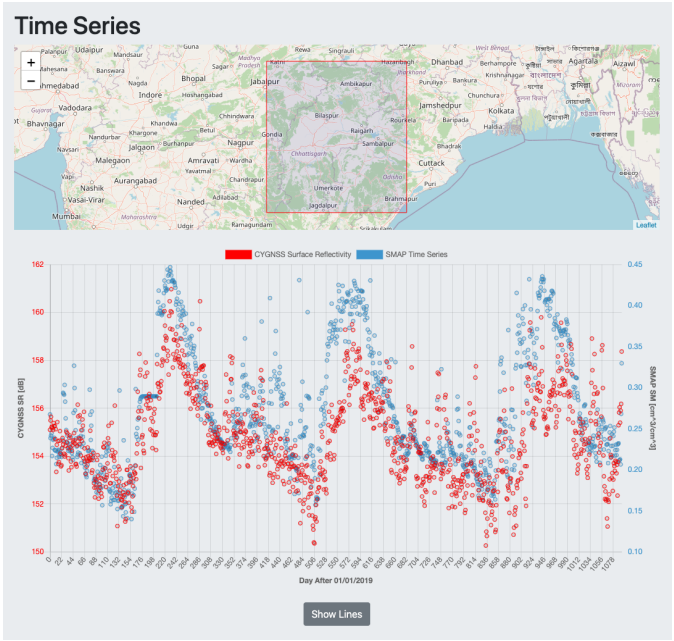


Figure 4.8: A screenshot of the web page showing time series over the 5° x 5° area in India from 2019 until 2021.

The version control software Git was used to enable cooperation with the previous developer of the web site. The project had a dependency to the Python package Basemap that was installed locally on the previous contributor's computer, which makes it difficult for others to continuously add functionality to the web page. Significant effort was therefore put into making it easier for future developers to start working with the source code. This included removing the local Basemap dependency and replacing it with the possibility of downloading it from the internet. Because the Python package installer, PIP, has issues with downloading Basemap, an alternative process of downloading it through Anaconda is proposed in the projects ReadMe. Another package dependency that can be troublesome to download, both through PIP and Anaconda is GDAL. The project ReadMe was therefor also extended with a paragraph explaining how to use Homebrew for this specific package. These changes to the project reduce the necessary effort needed to start contributing to the web site.

This page is intentionally left blank

Chapter 5

Methodology

Figure 5.1 presents the general overview of the performed analysis in this thesis. Data was collected from four independent sources and stored in an external hard drive. The total amount of downloaded data exceeded 1 TeraByte in total. After the data was downloaded, it was processed and prepared for the upcoming analysis. One step of the processing includes a novel approach for incidence angle optimization, which will be presented in section 5.3.

The above-mentioned flowchart formed the basis for the analysis performed in this thesis. The initial parts of the analysis will focus on developing a thorough understanding of the parameters affecting the CYGNSS measurements. This includes an investigation of the effect of the differences between CYGNSS and SMAP in spatiotemporal resolution by applying Gaussian smoothing, explained in section 5.4. Furthermore, the consequences of different mission configurations and noise impacts between CYGNSS and SMAP are discussed before a time series analysis is performed and elaborated in section 5.5. Then, the resulting insight from these studies will be gathered in an attempt to utilize machine learning for soil moisture prediction, as further elaborated in section 5.6.

5.1 Surface Reflectivity Computations

An approach to obtain information related to soil moisture is to analyze the surface reflected signal power using bistatic radar equations [58]. Assuming coherent surface reflections, Rajabi et al. presented how the coherent component of scattered power can be expressed from observations derived from DDMs as in Equation 5.1,

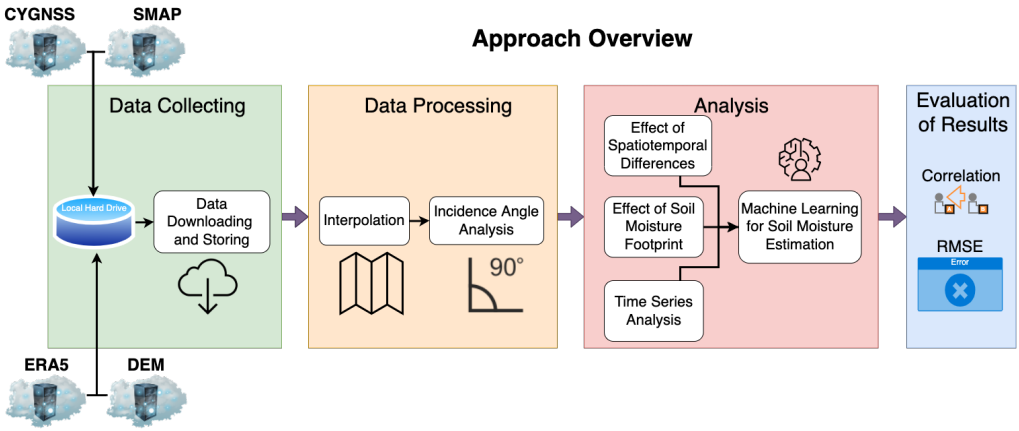


Figure 5.1: Flow chart showing the approach overview.

$$P_{RL}^{coh} = \frac{P_R^t G^t G^r}{(d_{ts}+d_{sr})^2}\left(\frac{\lambda}{4\pi}\right)^2 \Gamma_{RL}, \tag{5.1}$$

where P_{RL}^{coh} is the peak value of the coherently received power, and R and L denote the RHCP transmitting antenna and the LHCP of forward-scattered signals receiving antenna, respectively. Furthermore, P_R^t is the transmitted RHCP power, G^t and G^r are the transmitter and receiver antenna gain, λ is the GPS L1 wavelength ($\sim 0.19m$), while d_{ts} and d_{sr} are the distances from the specular reflection point to the GPS transmitter and the GNSS-R receiver. Lastly, Γ_{RL} represents the surface reflectivity along the incidence angle. Solving Equation 5.1 for surface reflectivity results in Equation 5.2 as presented below.

$$SR = \Gamma_{RL} = \frac{(d_{ts}+d_{sr})^2}{P_R^t G^t G^r}\left(\frac{4\pi}{\lambda}\right)^2 P_{RL}^{coh} \tag{5.2}$$

For easier compatibility during processing, comparisons, and analysis, all parameters in Equation 5.2 are converted to the unit of dB, resulting in Equation 5.3.

$$SR = 20\log(d_{ts}+d_{sr}) + 20\log(4\pi) + 10\log P_{RL}^{coh} - 10\log P_R^t - 10\log G^t - 10\log G^r - 20\log\lambda \tag{5.3}$$

The value of the received power, P_{RL}^{coh}, is calculated from the SNR observations of DDMs available in the CYGNSS L1 data. However, the SNR value has shown to be related, but

not exactly equal, to the received power. This is due to factors such as system noise levels, receiver instrument gain settings, and fluctuations in signal strength due to atmospheric effects [25]. As a result, surface reflectivity is considered proportional to the CYGNSS observations as presented in Equation 5.4 [13].

$$SR \propto SNR - 10\log P_R^t - 10\log G^r - 10\log G^t - 20\log \lambda + 20\log (d_{ts} + d_{sr}) + 20\log (4\pi) \tag{5.4}$$

Even though the obtained SNR values from CYGNSS are solely related to P_{RL}^{coh}, but not equal, Equation 5.4 is used to calculate the surface reflectivity based on the strong relations the values have to the hydrological conditions of the land surface [52].

As a result of some values obtained from the CYGNSS L1 data already being in the unit of dB, combined with the wavelength λ being the GPS L1 carrier, Equation 5.5 presents the final representation of the equation used for surface reflectivity calculations in this study.

$$SR = SNR - P_R^t - G^r - G^t - 20\log (0.19) + 20\log (d_{ts} + d_{sr}) + 20\log (4\pi) \tag{5.5}$$

5.2 Presentation of the Analysed Areas

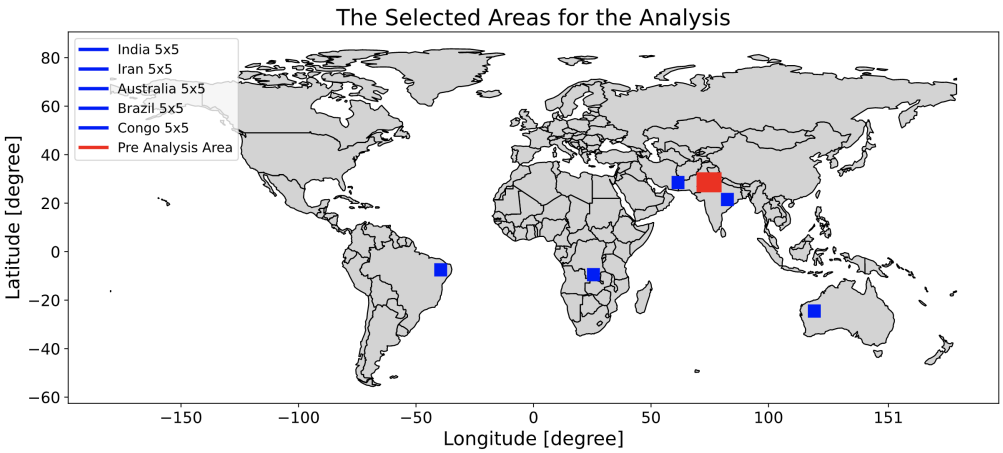


Figure 5.2: Presentation of the areas used in the analysis. The red rectangle shows the area used in the initial analysis and the five blue squares show the latter used areas.

Several areas have been selected to conduct the analysis in this thesis. They are mainly divided into two groups, where the initial research targets a larger region in northern India and eastern Pakistan, shown in red in Figure 5.2. In this area, a case study spanning 7.8° latitude and 9.9° longitude is chosen for the comparison and performance assessment of the three missions CYGNSS, SMAP, and ERA5. Furthermore, this region contains various soil types and areas with different slopes, making it suitable for an initial evaluation of the available data products related to soil moisture.

In addition to the initial exploration of the datasets in northern India and eastern Pakistan, five smaller regions spanning 5° latitude and 5° longitude will be analyzed. These were selected based on their geophysical conditions and are located in India, Iran, Australia, Brazil, and Congo. The areas will be referred to as the country they fall within throughout this thesis and are marked in blue in Figure 5.2. The geophysical conditions deciding which regions to select were scaled surface roughness and vegetation opacity, and the values for each area are presented in Table 5.1. Congo, Iran, and India have a relatively high scaled surface roughness of approximately 0.615, whereas the vegetation opacity varies from close to zero in Iran to 0.43 in Africa. The two final regions in Brazil and Australia have a lower surface roughness, while the vegetation opacity varies similarly to the three former mentioned areas.

Table 5.1: Scaled surface roughness and vegetation opacity for the five 5° x 5° areas used to conduct parts of the analysis in this thesis.

	Scaled Surface Roughness	**Vegetation Opacity**
Congo	0.615	0.431
Brazil	0.465	0.313
Australia	0.407	0.032
Iran	0.614	0.007
India	0.614	0.311

It is also worth noting that each area consists of sub-regions with distinct geophysical conditions. For example, despite possessing a relatively large vegetation opacity value, the 5° x 5° area in India also contains savannas with open landscapes. This means that even smaller sub-regions within the presented areas can be used to isolate different ground conditions, but these will be introduced when describing the concrete results. To summarize, performance assessment of GNSS-R based soil moisture estimates in different geophysical conditions will lead to further insight into how numerous parameters affect the measurements.

5.3 Effects of Surface Roughness and Fresnel Coefficients

One of the reasons why Equation 5.2 and CYGNSS data are used for surface reflectivity computations in this study, in contrast to the surface reflectivity term presented in Equation 5.6, is related to the highly demanding computational power required to perform the estimations.

$$\Gamma = \alpha \, |\Re|^2 \tag{5.6}$$

Equation 5.6 represents the coherent surface reflectivity term along the specular direction, where $\Re$ is the Fresnel reflection coefficient as introduced in subsection 2.2.4, and α denotes the loss of coherent power due to diffused scattering.

$$\alpha = exp(-4k^2h^2\cos^2\theta), \tag{5.7}$$

Equation 5.7 highlights the effects of surface roughness and incidence angle in computations of reflected signal power loss, where k is the electromagnetic wavenumber, h is the RMS surface height, and θ is the incidence angle [79].

To perform accurate computations based on Equation 5.6, detailed information related to surface roughness and topology would be required, such as a high resolution DEM. Measuring surface reflectivity values on a global scale would hence lead to slower, more computational heavy methods compared to the utilization of available CYGNSS data used within a set of assumptions. Surface reflectivity values are dependent on surface roughness, incidence angle, and permittivity [7, 11, 17, 58, 33]. Because this study focuses on changes in permittivity due to the presence of water, power loss caused by surface roughness, permittivity, and incidence angle should preferably be reduced. An assumption used in similar studies is that Fresnel coefficients do not change too much within smaller variations of incidence angles. Figure 5.3, as presented by Hoseini et al., visualize how the power loss due to permittivity of a reflected signal is affected by the elevation angle, which is $90° - \theta$ [33]. The figure shows trends of smaller changes in power loss with incidence angles below 60 degrees, or equivalently, elevation angles larger than 30 degrees.

Another underlying assumption in previous studies worth mentioning is that changes in surface roughness within smaller areas are considered so small that they do not affect the power loss to a large extent. With a power loss less affected by surface roughness and incidence angle, it is reasonable to assume that differences in measured SNR by CYGNSS are solely due to changes in water content, thus forming the foundation of using SNR to estimate soil moisture.

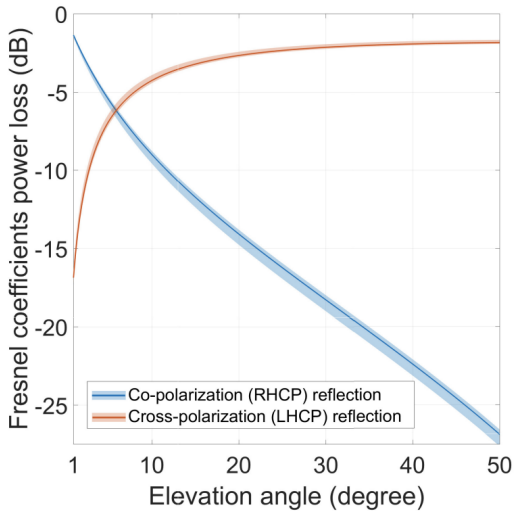


Figure 5.3: Fresnel coefficients calculated using the permittivity of seawater at the Onsala GNSS-R station. Based on the average permittivity, the blue and orange lines denote the magnitude of the copolarization and cross-polarization reflection coefficients in dB, respectively.

Source: [33]

In an attempt to better understand how surface roughness and permittivity actually affect the CYGNSS measurements, the mentioned assumptions are considered insufficient in this study. Instead, the resulting power losses due to the two parameters will be mapped in relation to different incidence angles. The desire of doing this is to obtain optimal incidence angle intervals specific for different areas where the mentioned effects are reduced to a minimum. As a result, the CYGNSS measurements made at incidence angles within a selected range and area would have smaller uncertainties, forming a foundation for more realistic and accurate models and analyses.

5.3.1 Power Loss due to Fresnel Coefficients

Along with the incidence angle parameter, Fresnel coefficient computations include the permittivity of the reflecting medium. Figure 5.3 presented how power loss due to permittivity is affected by the angle of the reflected signal. However, the figure visualizes the behavior of power loss using only the permittivity of seawater measured at the Onsala GNSS-R station. Surface reflectivity measurements over land are performed on various soil types and saturation levels, and the power loss representation changes accordingly. To be able to showcase an overall trend in power loss based on differences in the reflecting surface permittivity, Figure 5.4 presents the power loss per incidence angle for six

different surface permittivities. As introduced in subsection 2.2.4, the Fresnel coefficient is computed using Equation 2.20, inserting permittivity values ranging from dry soil representations with a permittivity of 2, to wet and saturated soil, with permittivity values as high as 30. In addition, the inclusion of a value of 80, representing an arbitrary sea water permittivity, is also illustrated for comparison reasons. The trend in Figure 5.4 is then comparable to that presented in Figure 5.3, with both figures showing clear indications of almost constant power loss independent of incidence angle before a severe increase is experienced at incidence angles higher than 60 degrees.

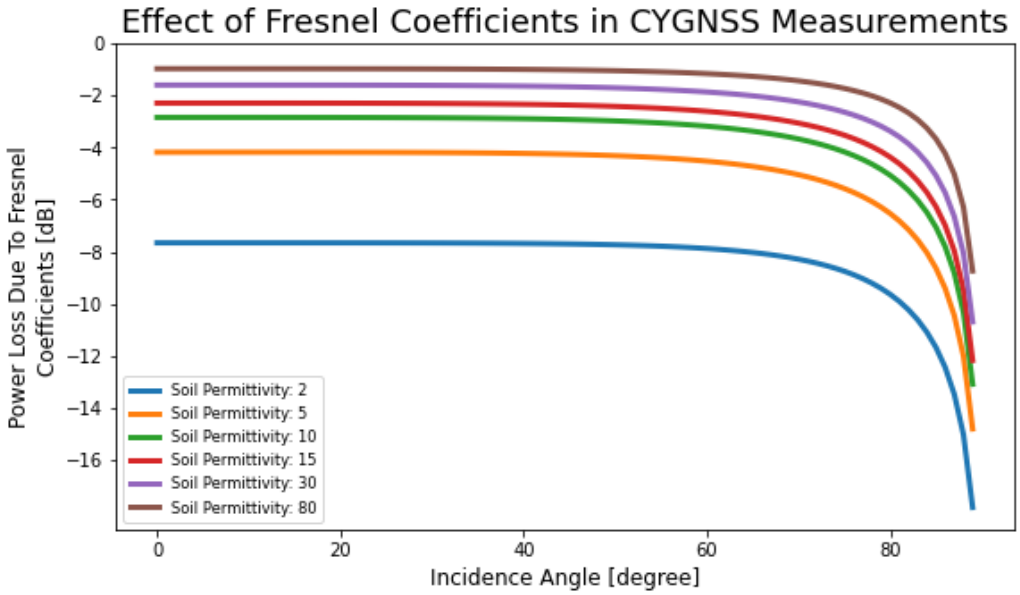


Figure 5.4: Experienced power loss in decibels due to Fresnel coefficients.

Furthermore, Figure 5.4 also highlights how dry soil experience greater power loss than wet and saturated soil, further strengthening the reasoning that an increase in surface reflectivity suggests an increase in soil moisture.

5.3.2 Power Loss due to Surface Roughness

The roughness of different magnitudes in the reflecting surface is already known to cause power loss in CYGNSS measurements. However, in contrast to the common assumptions mentioned in section 5.3, these effects are not circumvented but rather closely studied in an attempt to acquire better data utilization when dealing with CYGNSS measurements. To visualize how surface roughness affects power loss based on different incidence angles, a surface roughness coefficient, S, is first presented in Figure 5.5 using pre-

selected incidence angles ranging from 0 to 60 degrees, with intervals of 10 degrees. The coefficient S relates the standard deviation of surface height, σ, as a measure of surface roughness to the resultant power loss and is defined as in Equation 5.8. A higher value of S would indicate a smoother surface and hence lower power loss than rougher surfaces represented by a smaller values. Additionally, λ represents the signal wavelength, being 19 cm at L-band, and e represents the elevation angle. S is also independent of polarization and is therefore highly representative in the intended use of CYGNSS GNSS-R measurements [54].

$$S = exp\left(-\frac{1}{2}\frac{(2\pi)^2}{\lambda^2}\sigma^2 \sin^2 e\right) \tag{5.8}$$

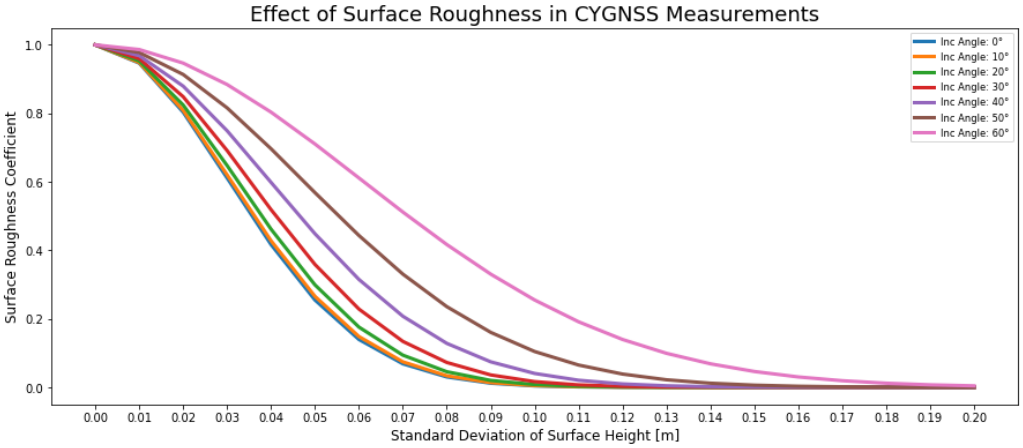


Figure 5.5: The effect of surface roughness on CYGNSS measurements for different incidence angles.

A noticeable feature in Figure 5.5 is how the standard deviation of the surface height only presents values between 0 and 0.2 meters. The reason is simply due to how the seven lines representing different incidence angles all converge towards a surface roughness coefficient of 0. Hence, computations for higher values of standard deviations would not result in more information of interest but highlight a severe power loss. Additionally, Figure 5.5 showcases how the standard deviation of the surface height determines the obtainable range of surface roughness coefficients in measurements made within areas with differences in surface roughness. As introduced in subsection 4.2.1, SMAP consists of measurements made at a constant angle of 40 degrees. Due to the fact that SMAP surface roughness coefficients range from 0 to 0.17, where the scaled representations are presented in section 5.2, a standard deviation of 7-8 cm can be assumed for most of the land within the chosen areas.

Furthermore, in order to more clearly express what differences in surface roughness coefficients means in terms of power loss, Equation 5.8 is converted to the unit of decibels. The resulting expression is presented in Equation 5.9.

$$S_{db} = 10\log_{10}(S) \tag{5.9}$$

Using Equation 5.9 for the same seven incidence angles used to produce Figure 5.5 results in the alternative representation presented in Figure 5.6.

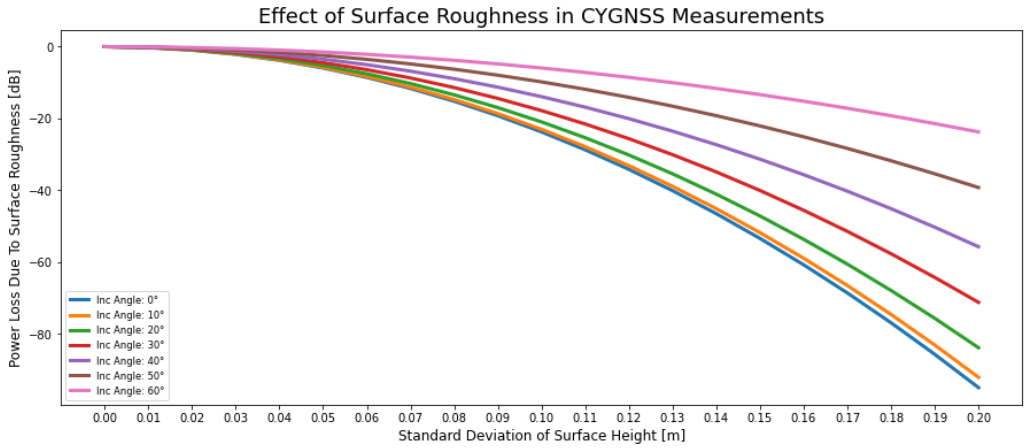


Figure 5.6: Experienced power loss in decibels due to surface roughness.

Figure 5.6 shows clear indications of an increase in power loss with a higher standard deviation of surface height. However, this loss is highly dependent on which incidence angle is used during the measurement. Lower incidence angles tend to experience a significantly higher power loss at higher standard deviations of surface heights than those at higher incidence angles. Given the mentioned standard deviation of 7-8 cm in most parts of the areas chosen for further analysis, these differences become somewhat smaller. Accordingly, combining the results presented in Figure 5.6 with those presented in Figure 5.4 enables data selection based on a minimization in power loss and measurement uncertainty.

5.3.3 Optimizing Incidence Angle Intervals in Data Selection

With a deeper insight into how both Fresnel coefficients and surface roughens affect power loss of CYGNSS measurements, distinct methods to reduce these impacts can be developed based on desired user applications. On a general note, it is desirable to have as many measurements as possible available for analysis purposes. However, it is

also preferable that these experience minimum uncertainties in power loss to ensure the sole utilization of high accuracy measurements. Based on this reasoning, an optimization function to maximize the number of available measurements while minimizing the power loss uncertainty is defined, as presented in Equation 5.10.

$$F(X_0) = \frac{Uncertainty_{[X_0]}}{N\_obs_{[X_0]}} \tag{5.10}$$

The function to be minimized takes an incidence angle interval, represented by the start- and end incidence angle, as input and returns the resulting F-value. The F-value is calculated by dividing the uncertainty by the number of observations. Both the uncertainty and number of observations are computed by dividing the selected 5° x 5° area into smaller 0.2° x 0.2° grid boxes, where the median uncertainty and median number of observations at a certain incidence angle and incidence angle interval are computed. The final values representing the entire 5° x 5° area are the medians of the resulting values, ensuring representative values being returned.

Even though minimization of the F-value provides reasonable estimates of optimal incidence angle intervals when selecting CYGNSS measurements to be used in analyses, these intervals depend on the intended user application. Because the proposed minimization function pays equal weight to the two parameters, different applications can value the parameter accordingly. High accuracy soil moisture estimation studies would prefer to exclusively use measurements of low uncertainty at the expense of the number of observations available. On the contrary, applications focusing on bigger, unnatural changes in soil moisture, such as flood- or drought detection, can sacrifice uncertainty for a higher number of observations. Because such an application-specific formula compensates for the effects of surface roughness or Fresnel coefficients directly, data is filtered based on the computed F-value, in combination with representations of relevant parameters in this study.

Using the area of Iran as an example, such an optimized incidence angle interval selection can be made. Figure 5.7 presents the uncertainty in power loss over the area using different incidence angle ranges. The incidence angle value represents the interval of $\pm$ half the incidence angle range. As the figure is based on real CYGNSS observations, the presence of fluctuations occur. In order to reduce the effect of sudden fluctuations, the graphs are smoothed using a Savitzky-Golay filter with a moving window of 13 degrees and a polynomial order of 3. Accordingly, smoothed representations are visualized as dotted lines of the same color as the initial result.

Notable features in Figure 5.7 include smaller uncertainties at lower incidence angles and that smaller incidence angle intervals reduce the uncertainties. Based on the presented figure, measurements at lower incidence angles would hence be characterized as more

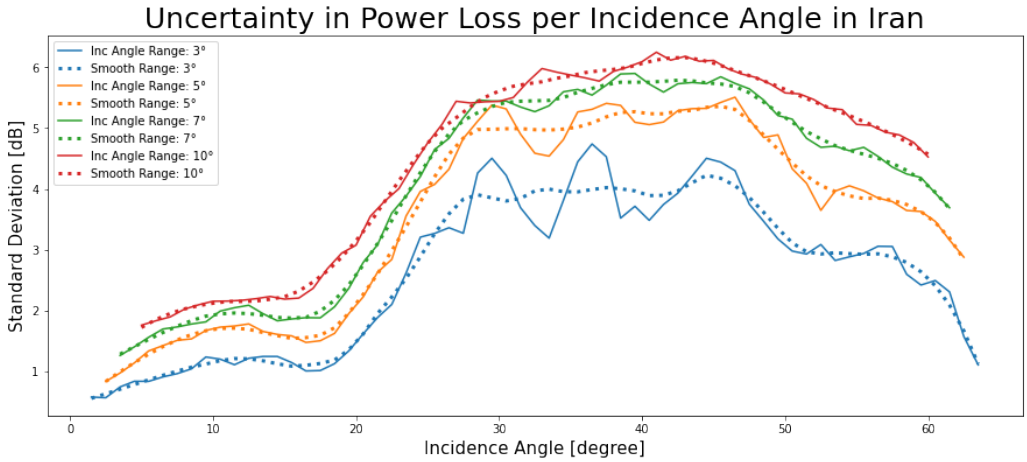


Figure 5.7: Median standard deviation of power loss in decibels due to incidence angle and incidence angle interval for the entire area of Iran.

accurate than those measured above 20 degrees. However, as mentioned above, high accuracy is not always prioritized above the number of available measurements. Figure 5.8 presents the median number of observations available at each incidence angle and incidence angle interval. Smaller fluctuations are still occurring for the same reasons, so smoothed values are computed using the same Savitzky-Golay filter as used in Figure 5.7.

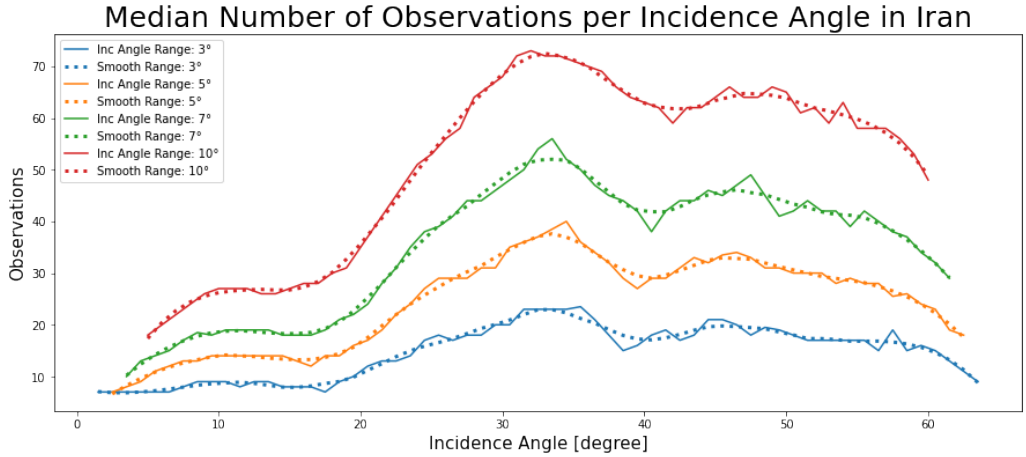


Figure 5.8: Median number of observations measured at each incidence angle and incidence angle interval for the entire area of Iran.

Figure 5.8 highlights how the number of observations increases with higher incidence

angles, with most measurements being made at 30 to 40 degrees. Combined with Figure 5.7, Figure 5.8 confirms the need of an application-specific methodology when selecting data based on incidence angles, as there are less high-accuracy measurements than those being less accurate. Employing the F-function introduced in Equation 5.10 using the smoothed uncertainty- and number of observations values, Figure 5.9 presents the resulting F values in the area of Iran.

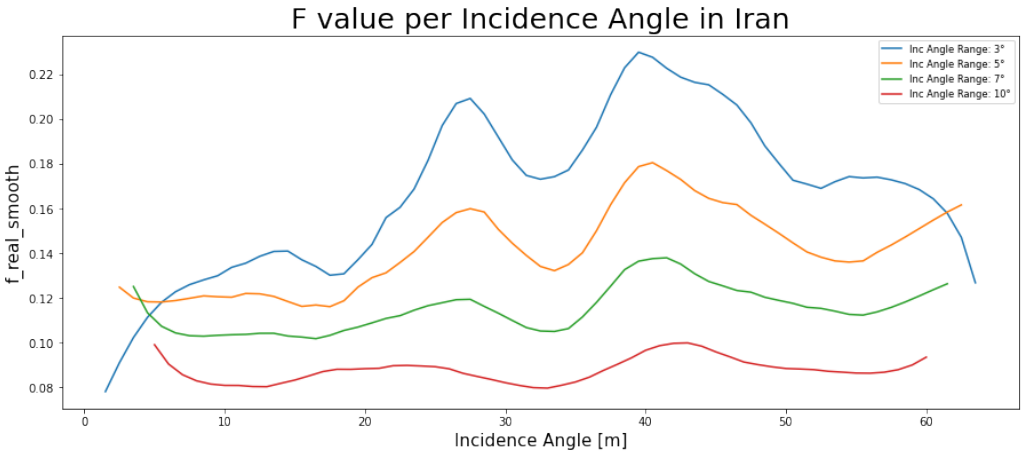


Figure 5.9: The resulting F value at each incidence angle and incidence angle interval for the entire area of Iran. The value is computed using Equation 5.10, with the smoothed representations of uncertainty and number of observations.

In the case related to the area of Iran, incidence angle intervals of 10 degrees seem to provide better data when giving equal weight to accuracy and number of measurements. In addition, incidence angles between 30 and 40 degrees appear to indicate a local minimum for all incidence angle ranges. A proposed interval for use in data selection could therefore be all measurements made at 33 ± 5 degrees. Even though the visualisations presented in Figure 5.7, Figure 5.8, and Figure 5.9 apply for the presented area of Iran only, similar F value visualisations can be used to select data based on incidence angle intervals representing local minimums. An analysis related to the exclusion of less accurate measurements being utilized will be elaborated in section 5.6.

5.4 Gaussian Smoothening Analysis

The spatial resolution of the two space missions CYGNSS and SMAP are different. As explained in chapter 4, CYGNSS possesses a high spatiotemporal resolution compared to SMAP. The higher spatial resolution of CYGNSS enables the possibility to capture small fluctuations in surface reflectivity, thus providing the opportunity of producing soil

moisture estimates with increased spatial resolution. However, this difference can also impact the correlation between the two data sources in a negative manner. Previous approaches to soil moisture estimation using CYGNSS have produced correlations with SMAP that possibly suffer from this fact. In reality, the produced correlations could have been even higher if CYGNSS and SMAP shared the same resolution, hence increasing the reliability of CYGNSS as a source for soil moisture estimation. In order to analyze this effect, a Gaussian filter will be applied to the CYGNSS data. The Gaussian filter in two dimensions is given in Equation 5.11,

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{5.11}$$

where σ is the standard deviation of the Gaussian distribution, x and y is the distance from origin on the horizontal and vertical axis respectively.

This analysis will be performed on five areas in the region covering northern India and Pakistan. The five areas are chosen based on different soil moisture characteristics to evaluate the consequences of different spatial resolutions under different conditions. Measurements from January 2020, both for CYGNSS and SMAP, will be used. Each area is grid boxed with a spatial resolution of 0.1 x 0.1 degrees before each grid box will be assigned the median of the CYGNSS measurements and the mean of the SMAP measurements.

Furthermore, the standard deviation of the Gaussian distribution, σ, will be varied. If applied to an image, more blur will be introduced if a larger σ is used. The sigma determines which points to include in the calculations of the current point. Larger sigmas mean that each point will be affected by points located further away, thus reducing the spatial resolution. Finally, the correlation between CYGNSS and SMAP will be used to evaluate the results.

5.5 Time Series Analysis

In order to analyze the change of surface reflectivity and soil moisture over longer periods, time series analysis will be conducted for the five regions presented in section 5.2. As discussed by Chew and Small, the temporal differences in spaceborne GNSS-R data have the potential to reduce error sources such as vegetation cover and different soil types. This is because such factors remain approximately constant over time, thus leaving changes in soil moisture as the primary factor of variations in surface reflectivity [9].

The time series is performed by re-sampling the CYGNSS and SMAP data within each

area to a spatial resolution of 0.5° x 0.5°. Then, the time series will be computed for each grid cell from the 1st of January 2019 until the 31st of December 2021. Each data point in the time series will be an average of the upcoming three days to match the temporal resolution of SMAP. The temporal step size is one day.

Next, the correlation between CYGNSS surface reflectivity and SMAP soil moisture will be computed for each time series in every grid cell. The time series of the grid cell with the highest correlation and the one with the lowest correlation will be presented and further analyzed, in addition to the time series of the entire 5° x 5° area.

Time series can contain underlying properties, like periodic signals, which are difficult to detect for the human eye. The Lomb-Scargle periodogram is an algorithm developed for this case. The algorithm's objective is to detect and characterize periodicity in unevenly sampled time series data [78]. Because the time series last for a period of three years, they include the same seasons at least three times, thus allowing the Lomb-Scargle periodogram to possibly detect seasonal soil moisture variations. The algorithm accepts the surface reflectivity or soil moisture data and a list of frequencies as input. Then, the periodic power of each frequency is returned, and peaks in this data will reveal seasonal variations in soil moisture.

5.6 A Machine Learning Approach to Soil Moisture Estimation

Lastly, a machine learning approach to create a geophysical model function (GMF) will be conducted. In remote sensing, the variable of interest is often observed through some other observable. Therefore, GMFs are created to convert the observables to estimates of the variable of interest. In this section, this will be done using machine learning models capable of converting reflected GNSS signals, with ancillary data, to soil moisture estimates.

As introduced in section 3.3, previous work within the field has presented promising results. These show indications of the possibility to perform accurate soil moisture estimations on a global scale based on CYGNSS level 1 measurements using machine learning. In contrast to the work presented by Senyurek et al., Lei et al. and Jia et al., this study will focus on estimation within smaller areas [65, 46, 36]. The reason why selective areas are used is the desire to obtain more information on how different parameters and surface conditions affect the estimation performance. Potential findings can then be utilized in further work to develop methods with higher performance and accuracy than those for larger areas existing today.

The CYGNSS L1 data used in the soil moisture estimation analysis will first undergo the

same initial data processing step as that used in the other presented analyses, elaborated in subsection 4.1.3. Furthermore, the data is filtered based the five 5° x 5° areas presented in section 5.2. Soil moisture values from SMAP are then retrieved from the same area, in addition to SMAP vegetation opacity and surface roughness coefficients. SMAP data also include quality flags with quality assessments of successful soil moisture retrievals. Because quality flag values of 0 or 8 indicate high-quality retrievals, SMAP data containing any other flag is intentionally filtered out.

In order to create a colocated dataset of relevant CYGNSS and SMAP parameters at the highest possible spatiotemporal resolution, SMAP values for soil moisture, vegetation opacity, and surface roughness are interpolated. Interpolation is a statistical method by which known values located in sequence are used to estimate unknown values in other subareas [43]. As this study aims to utilize the high spatiotemporal resolution of CYGNSS, the interpolation has to be performed both in space and time. The work presented by Senyurek et al. showed that both linear and natural neighbor interpolation perform well in the three-dimensional interpolation and is therefore tested in this thesis [67]. The linear interpolation is performed using a linear N-dimensional interpolator. Provided by the python library, SciPy, the interpolation is conducted by triangulating input data before a variant of the fast and reliable Lagrange polynomial interpolation is performed on each triangle. Interpolation of CYGNSS points outside the convex hull, being the smallest area that contains all available SMAP values, returns Not a Number (NaN) fill values, and these are filtered out before further use [4].

First proposed by Sibson in 1981, natural neighbour interpolation utilizes Voronoi diagrams to interpolate scattered data, resulting in smoother approximations compared to other methods [70]. As presented in Figure 5.10, a Voronoi diagram can be defined as the partitioning of a plane with distinct points into convex polygons such that each polygon contains a single reference point, and other points within that polygon are closer to its reference point than any other [5]. Natural neighbors can then be defined as two sites whose Voronoi cells share a common edge. To determine the natural neighbors of an interpolation point, one can imagine this point is virtually inserted into the Voronoi diagram, as illustrated in Figure 5.10. This virtual insertion modifies the original Voronoi diagram and creates a new Voronoi cell with a set of points now closer to the inserted reference point. A weighted average of the resulting natural neighbours and respective weights then provide an interpolated value at the inserted reference point [5].

In addition to the interpolated SMAP geophysical parameters, SRTM's DEM is used to generate two additional parameters related to surface topography. As introduced in section 3.1, CYGNSS measurements over land can benefit from additional consideration of terrain variations compared to ocean measurements. Surface roughness affects the coherency of the reflected signals received at CYGNSS spacecrafts. CYGNSS uses the WGS84 ellipsoid to represent the surface of the Earth and does not account for local

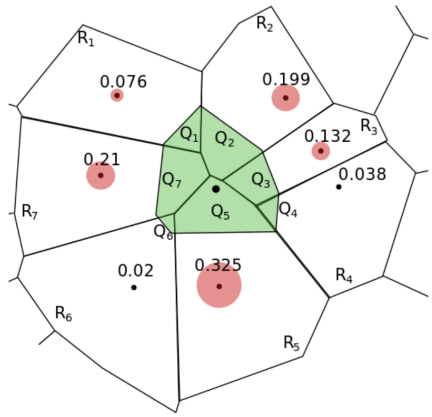
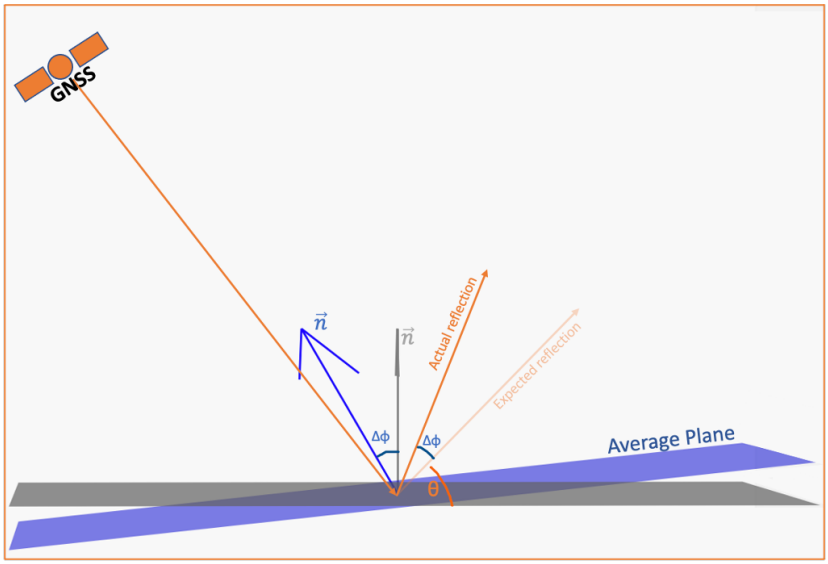


Figure 5.10: An example of a Voronoi diagram, illustrating the insertion of a new point and the resulting Voronoi cell highlighted in green. The points marked in red represents the natural neighbours and their respective weights in natural neighbour interpolation.

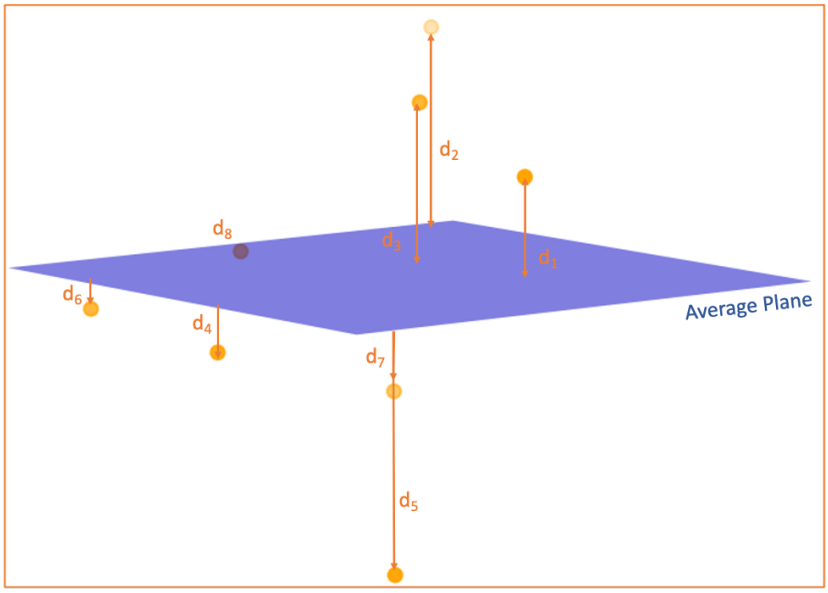
Source: [77]

terrain variations [27]. This results in the provided incidence angle not always being representative of the angle in the reflection geometry. Therefore, a correction for this angle will be computed by utilizing a DEM and 0.01° x 0.01° partitioning of the analyzed area. The angular difference between the normal vector of the partitioned area and the vector to the center point of the same area will be used as correction. This parameter will be referred to as elevation steepness and is represented by the $\Delta\phi$ parameter in Figure 5.11 (a). In addition, the RMSD between all elevation points and the averaged plane is computed for each patch. Higher values represent larger variations in the topography within a single patch, illustrating an additional attribute of the surface. This parameter will be referred to as height deviation and is illustrated in Figure 5.11 (b).

To analyze the effects of the generated elevation steepness and height deviation, and whether or not the inclusion of these improves soil moisture estimation accuracy, the 5° x 5° areas are divided into smaller 0.5° x 0.5° grid cells as done in section 5.5. Based on the obtained results from the time series analysis, grid cells that experienced distinct correlations or surface conditions are selected. The AutoML library H2O is then utilized to generate various soil moisture estimation models based on different combinations of parameters, using data from January 2019 to June 2021 as training data and the latter part of 2021 for testing. In this way, visualization of the estimation performance is also more relatable to the conducted time series in the same areas. Because soil moisture estimation is considered a form of regression, the methods GBM, DRF, and XGBoost will be tested. In addition to these, H2O's deep learning model based on a multi-layered feed-forward



(a)



(b)

Figure 5.11: Illustrations of how the proposed $\Delta\phi$ (a) and RMSD (b) are computed in this thesis. These parameters will be referred to as elevation steepness and height deviation for the remaining study.

artificial neural network trained with stochastic gradient descent using back-propagation is attempted. In addition, a stacked ensemble of the tested methods will be applied.

Lastly, data selection based on proposed optimal incidence angle intervals, as introduced in subsection 5.3.3, is performed. By doing this in areas with differences in topography, vegetation, and climate, this study aims to highlight how different parameters are important in different areas, combined with an assessment of different machine learning methods and performance. Potential findings can then contribute to potential improvements in global soil moisture estimation models in future research.

Chapter 6

Results and Discussion

The previously described methodology was applied to the presented areas to derive information related to soil moisture. These results will be presented in the following section.

6.1 Preliminary Analysis

Initially, the signature of soil moisture in the three datasets described in chapter 4 is investigated. A case study spanning a region of 7.8° latitude and 9.9° longitude in northern India and Pakistan is chosen for the comparison and performance assessment of GNSS-R measurements. An areal photo of the selected region is presented in Figure 6.1. The surface reflectivity and soil moisture for the months of January and August 2020, in addition to the temporal differences between them and standard deviations, are presented and discussed.

Figure 6.2 (a) and (b) plot the average CYGNSS surface reflectivity with a spatial resolution of 0.1° by 0.1° for January and August respectively, in the unit of dB. Figure 6.2 (c) and (d) plots the average soil moisture from SMAP measurements with the same spatial resolution for January and August. The northeast corner of Figure 6.2 (c) is missing because the SMAP constellation was unable to cover that specific area in January, highlighting the difference in temporal resolution between SMAP and CYGNSS. Figure 6.2 (e) and (f) plot the average soil moisture values from ERA5 with the same spatial resolution and period as mentioned above. Common for all plots is that the x-axis shows the longitude coordinates, the y-axis shows the latitude coordinates, and the color bar shows the target value range.

Figure 6.1: An areal photo of the analyzed region covering northern India and Pakistan.

Overall the three datasets show good agreement with each other. In January, the plots illustrate drier regions in the northeast corner and in the western parts. In addition to this, all plots show significantly increased surface reflectivity or soil moisture values in the more lush regions in Figure 6.1.

Table 6.1 shows the average surface reflectivity for CYGNSS and the average soil moisture for SMAP and ERA5. The values are averaged over the whole region for the two months. CYGNSS demonstrates increased soil moisture from January to August as the average surface reflectivity increases. The same applies both for SMAP and ERA5, indicating that the temporal change in soil moisture follows similar patterns for all three datasets. By inspection of Figure 6.2, similar trends occur. This is reasonable as the monsoon season takes place between June and September in India, hence the analysis demonstrates reasonable behavior in terms of temporal changes.

Table 6.1: Average CYGNSS surface reflectivity for January and August 2020 and average soil moisture for the months of January and August 2020 for SMAP and ERA5. The values are calculated over northern India and eastern Pakistan.

	January	**August**	**Relative Change**
CYGNSS	9.36 dB	10.5 dB	+12.2 %
SMAP	0.21 cm^3/cm^3	0.23 cm^3/cm^3	+9.5 %
ERA5	0.21 m^3/m^3	0.25 m^3/m^3	+19.1 %

An important difference between the three datasets is the more smooth properties of

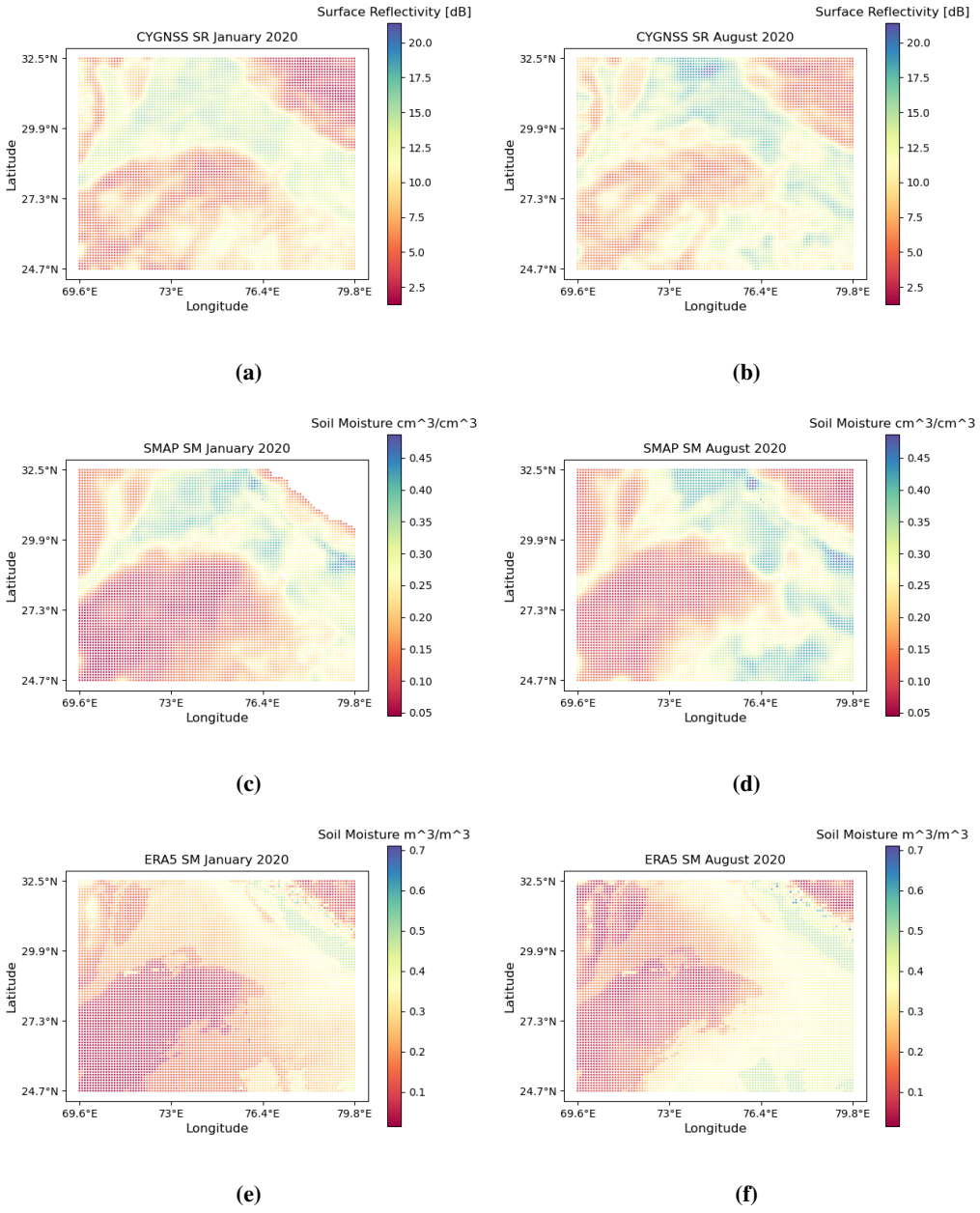


Figure 6.2: Average surface reflectivity for CYGNSS in January (a) and August (b), average soil moisture for SMAP in January (c) and August (d), and average soil moisture for ERA5 in January (e) and August (f).

ERA5. Within small regions of the analyzed area, CYGNSS and SMAP demonstrate more variations than ERA5. This can be seen in the eastern part of the plots between 27.3° and 29.9° longitude in Figure 6.2, where ERA5 shows significantly smaller changes compared to CYGNSS and SMAP. This phenomenon might occur due to ERA5 being a global model produced by the assimilation of different spaceborne or ground-based observations, while CYGNSS and SMAP are based on direct measurements of the Earth's surface. This is also illustrated in Table 6.2, where CYGNSS shows a significantly higher correlation with SMAP than it does with ERA5. CYGNSS has an average correlation with SMAP of 0.774 for January and August, while the correlation with ERA5 is only 0.353 for the same months.

Table 6.2: Correlation between CYGNSS and the two ancillary datasets SMAP and ERA5 for the months of January and August in 2020 in northern India and eastern Pakistan.

	January		**August**		**Temporal Difference**	
	SMAP	ERA5	SMAP	ERA5	SMAP	ERA5
CYGNSS	0.777	0.368	0.770	0.337	0.686	0.463

Despite showing a high correlation, CYGNSS and SMAP differ in local variations. In the southwest parts of Figure 6.2 (a) and Figure 6.2 (c), CYGNSS and SMAP differ in local fluctuations. While SMAP demonstrates the whole southwest part to be uniformly dry with small soil moisture changes, CYGNSS shows several local variations. A reason for this deviation might be the elevation of that specific area. Figure 6.3 shows that several locations in the analyzed region are located at high elevations. CYGNSS measurements where the specular point is located in mountainous areas can be more prone to noise than those measured at lower and less steep elevation [9]. Therefore, it is likely that the correlation would increase further by filtering out high elevation CYGNSS measurements.

Figure 6.4 shows the standard deviations for each CYGNSS grid box presented in Figure 6.2 (a) and (b). There are several noticeable factors from these plots. Firstly, the reflected signal power seems to affect the standard deviation. The drier regions, marked in red in Figure 6.4 (b) and reflect less of the transmitted GPS signal, show a higher standard deviation than elsewhere. The soil type can also affect these results. It might be possible that different soil types generate different responses in terms of standard deviation, which again can form a possible foundation for soil classification.

Furthermore, some CYGNSS tracks possess higher standard deviations than others. One of these tracks is located in the southeast part of Figure 6.4 (b) and is marked in green. Filtering out such tracks can help improve the quality of the surface reflectivity measurements.

Figure 6.5 (a) shows the change in CYGNSS surface reflectivity between January and

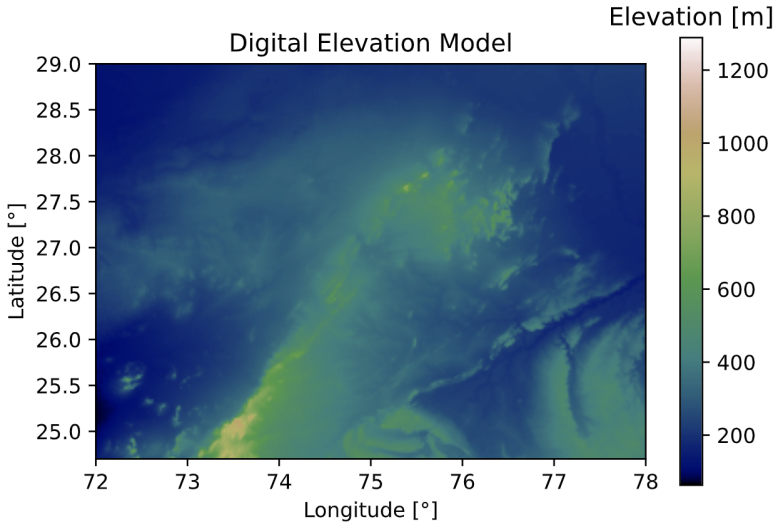
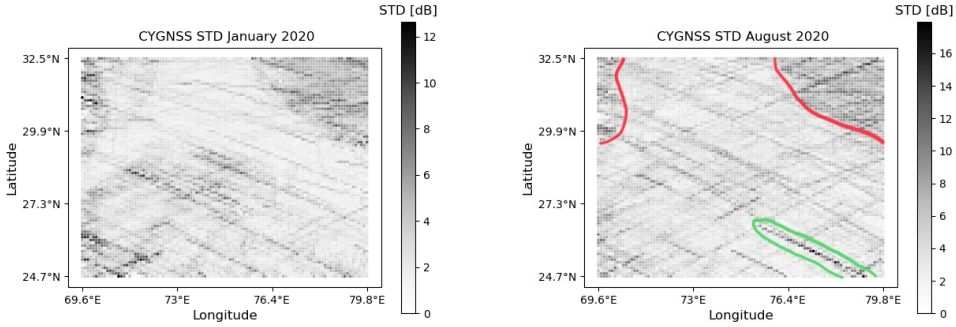


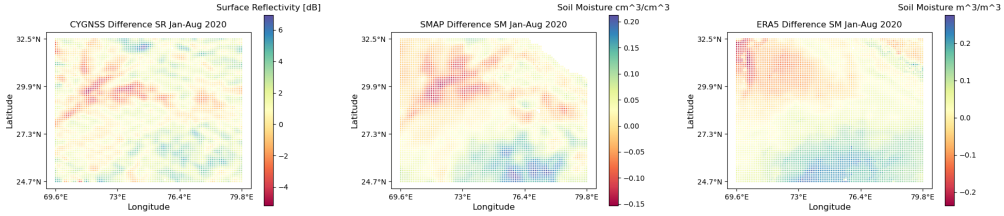
Figure 6.3: Digital elevation model with areas exceeding 600 meters above sea level.



(a) **(b)**

Figure 6.4: Standard deviation of the CYGNSS surface reflectivity measurements for January (a) and August (b). The red lines in (b) show areas with higher standard deviations, whereas the green line shows a CYGNSS track possibly affected by noise.

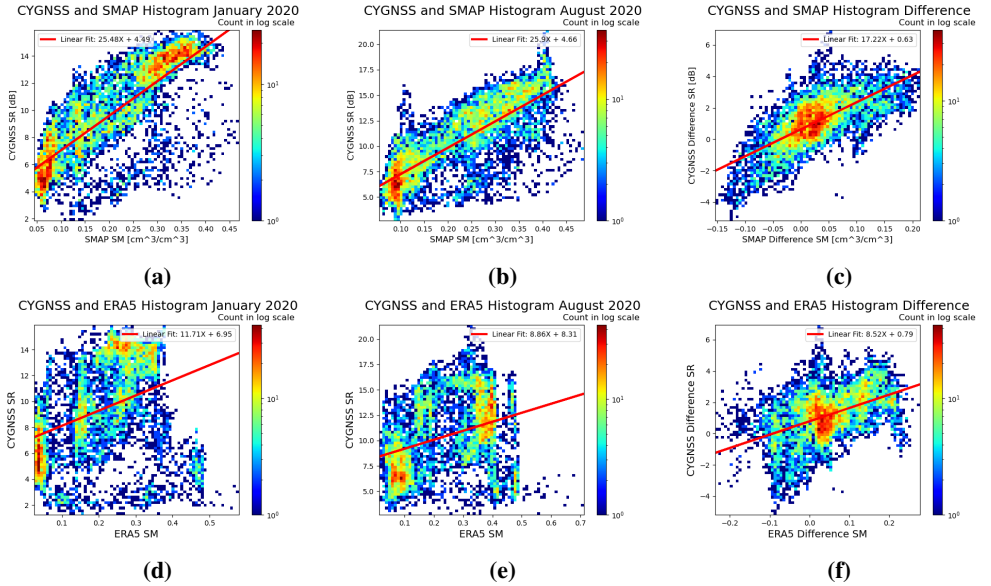
August 2020. Figure 6.5 (b) and (c) show the change in soil moisture for SMAP and ERA5 respectively. Positive values indicate increased surface reflectivity or soil moisture from January to August, and negative values imply the opposite. All three plots show similar global trends over the region, where the northwest part demonstrates lower surface reflectivity and soil moisture, and the southeast parts show increased surface reflectivity and soil moisture.



(a) **(b)** **(c)**

Figure 6.5: Temporal difference in SR and soil moisture between the two months January and August for CYGNSS (a), SMAP (b), and ERA5 (c).

The main difference between the plots in Figure 6.5 is the smooth transitions from increased soil moisture to decreased soil moisture for the three data sources. It appears that ERA5 continuously moves from increased soil moisture in the southeast to decreased soil moisture in the northwest. This overall smooth trend is to a smaller degree present for CYGNSS and SMAP. The spatial resolution of CYGNSS and SMAP can be one of the reasons to why this pattern occurs. Because ERA5 uses reanalysis to create a global model, it struggles to capture the local variations. Furthermore, CYGNSS varies slightly more than SMAP, which might indicate the potential of using CYGNSS for high spatial resolution soil moisture prediction.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 6.6: Correlation between surface reflectivity and soil moisture.

The first row of Figure 6.6 shows two-dimensional histograms of CYGNSS surface reflectivity and SMAP soil moisture for January (a), August (b), and the temporal difference between January and August (c). The bottom row shows similar histogram plots between CYGNSS and ERA5. Each bin in these plots illustrates the number of observations falling into the surface reflectivity from CYGNSS and soil moisture from SMAP using a logarithmic scale. A red color indicates more observations and blue indicate less. The soil moisture, from either SMAP or ERA5, is plotted along the x-axis, and the CYGNSS surface reflectivity is plotted along the y-axis, together forming a plot capable of analyzing the compliance between the measurements.

Figure 6.6 (a), (b), and (c) show a linearly increasing trend between CYGNSS and SMAP. This is reasonable as previous work show that increased soil moisture values lead to an increase in surface reflectivity. However, the plots using ERA5 as ancillary data, shown in Figure 6.6 (d), (e), and (f), do not possess the same clear pattern. The overall trend of these histograms is still somehow linearly increasing but showing a clearly weaker correlation than SMAP. This is in agreement with the results presented in Table 6.2, where CYGNSS and SMAP display a significantly higher correlation than what CYGNSS and ERA5 do. It is worth noting that the linear trend in Figure 6.6 (f) is stronger than in Figure 6.6 (d) and (e), which is also suggested by the values in Table 6.2.

These results suggest that ERA5 is infeasible as an ancillary data source in comparisons with CYGNSS due to its inability to capture local variations. Therefore, ERA5 is excluded in the upcoming analyses, and SMAP will be the primary ancillary data source.

6.2 The Effect of Noise and Spatiotemporal Resolution

CYGNSS and SMAP are two space missions with different orbits, antenna gain, and design. This leads to two different spatiotemporal resolutions and possible deviations in systematic noise affecting the measurements. The effect of these differences will be investigated to gain a better understanding of using the correlation between CYGNSS and SMAP as a performance measure.

Two 3° x 3° areas with different soil moisture distributions were selected to conduct this analysis. These are shown in Figure 6.7, where the average CYGNSS surface reflectivity is plotted for January 2020 on a 0.1° by 0.1° spatial resolution. The black square covers an area where the soil moisture is uniform and contains few large changes. This area will be referred to as the "uniform soil moisture" area. Furthermore, the blue area represents an area with a significant increase in soil moisture going from south to north and will be referred to as the "variable soil moisture" area. The correlation between CYGNSS and SMAP is computed for January 2020 and August 2020 to illustrate the effect the soil

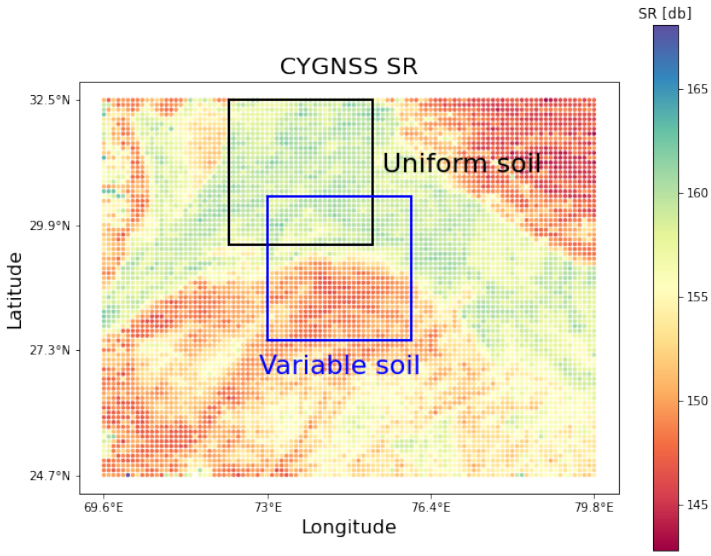


Figure 6.7: The two areas selected to analyse the effect of noise and spatial resolution of CYGNSS and SMAP.

moisture footprint has on the correlation between CYGNSS and SMAP.

Figure 6.8 and Figure 6.9 show the surface reflectivity and soil moisture for January 2020 and August 2020 respectively. The left side of both figures illustrates the variable soil moisture area, and the right side shows the uniform soil moisture area. The top row is CYGNSS surface reflectivity, while the bottom row illustrates SMAP soil moisture.

Table 6.3: Correlation between CYGNSS and SMAP for one area containing uniform soil moisture and one area containing variable soil moisture located in northern India and Pakistan.

Area	Correlation Coefficient
Uniform SM January 2020	0.453
Variable SM January 2020	0.924
Uniform SM August 2020	0.813
Variable SM August 2020	0.914

The correlations between CYGNSS and SMAP for the two analyzed areas in the two selected periods are presented in Table 6.3. In January 2020, the uniform soil moisture area demonstrated a weak correlation between CYGNSS and SMAP. This can result from a lack of general trends in the soil moisture footprint, meaning that factors such as different spatiotemporal resolution and the impact of noise contribute to lowering the correlation

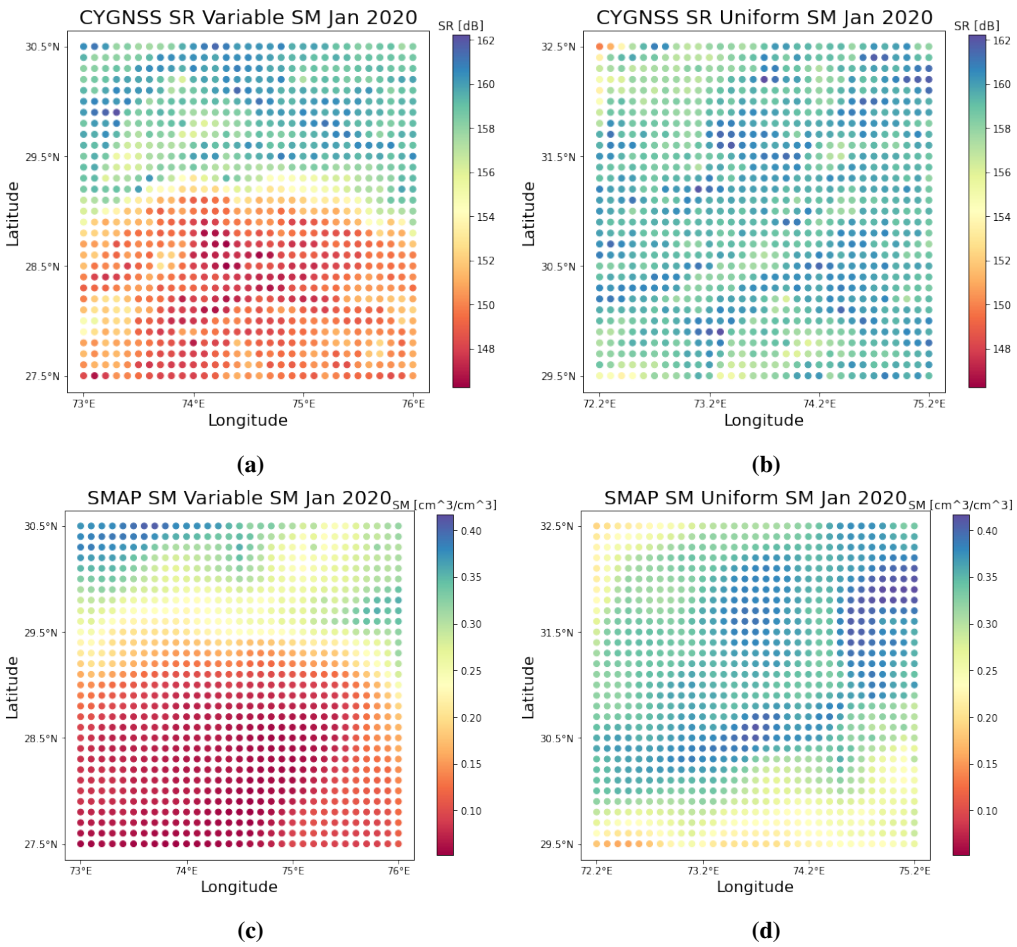


Figure 6.8: Average CYGNSS surface reflectivity (a) and (b), and average SMAP soil moisture (c) and (d) for two 3 by 3 degree areas in January 2020.

between the two data sources. However, in August 2020, the uniform soil moisture area demonstrates larger variations in soil moisture. This can be seen in Figure 6.9 (b) and (d), where the southwest parts are significantly drier in August than in January 2020. As a result, the correlation between CYGNSS and SMAP in the "uniform soil moisture" area increased from 0.453 in January to 0.813 in August, demonstrating the impact of different spatiotemporal resolutions and noise between CYGNSS and SMAP.

CYGNSS and SMAP show a good agreement with each other, both in January 2020 and August 2020, for the variable soil moisture area. Areas like this one are less impacted by the difference in spatiotemporal resolution and noise because the significant variations in soil moisture dominate these other factors.

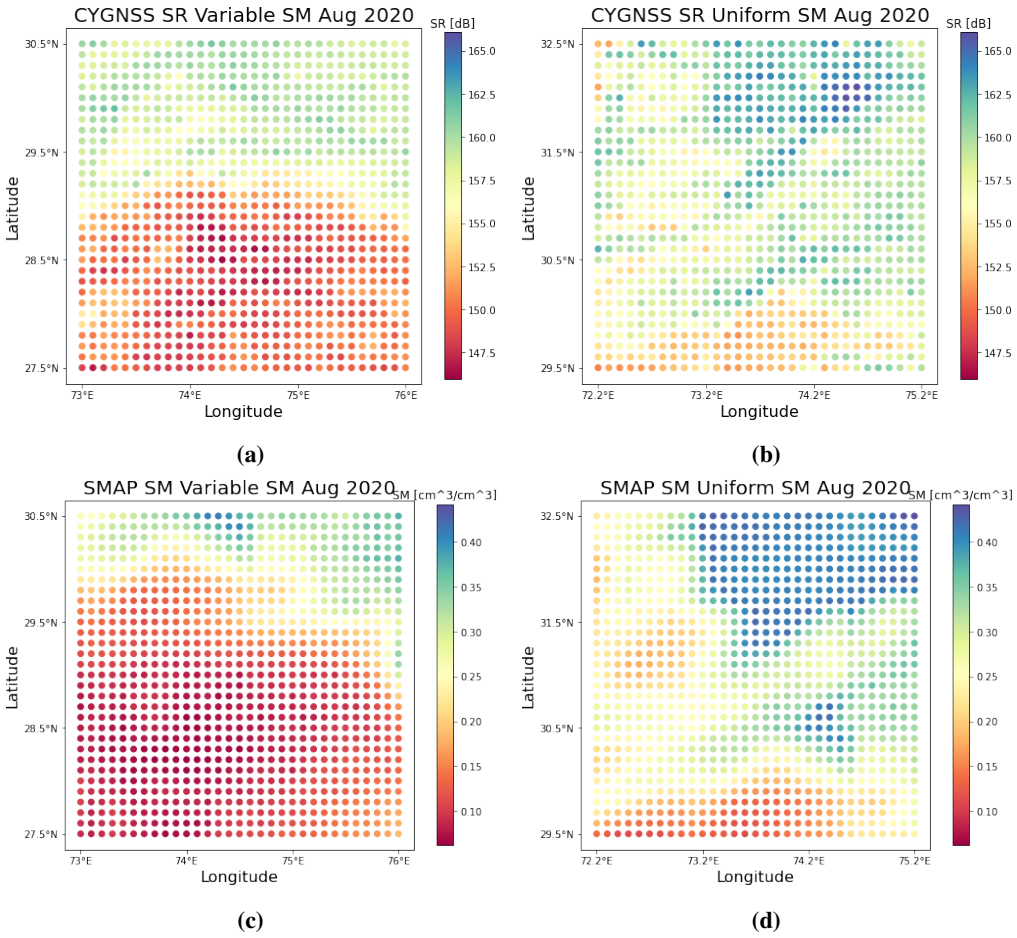


Figure 6.9: Average CYGNSS surface reflectivity (a) and (b), and average SMAP soil moisture (c) and (d) for two 3 by 3 degree areas in August 2020.

These results also enlighten aspects of the correlation between CYGNSS and SMAP as a performance measure. Despite the uniform soil moisture area demonstrating a low correlation in January 2020, the fact that CYGNSS possesses a higher spatiotemporal resolution compared to SMAP means that CYGNSS might capture fluctuations unavailable to SMAP. The low correlation is therefore not necessarily a sign of weak performance but a sign of the improved capabilities of producing high spatiotemporal resolution soil moisture products using the CYGNSS constellation.

6.3 Smoothening Analysis

To demonstrate the effect of smoothening, a Gaussian filter was applied to the CYGNSS measurements in the region covering northern India and Pakistan used in the previous section. This is shown in Figure 6.10, where Figure 6.10 (a) shows the data without performing smoothening. Figure 6.10 (b) shows the same area with smoothening using $sigma = 0.5$, (c) with $sigma = 1$ and (d) with $sigma = 2$. All plots have been grid boxed with a resolution of 0.1 by 0.1 degrees and retrieving the median of each grid box.

As discussed in section 6.2, CYGNSS shows the potential to capture high-frequency fluctuations of surface reflectivity within areas containing uniform soil moisture. This can be seen in Figure 6.10 (a), where small changes in soil moisture occur both in dry and wet regions. Furthermore, applying smoothening reduced these high-frequency fluctuations. The use of larger sigmas in the Gaussian kernel limits the local variations, thus lowering the spatial resolution. Increasing the values for sigma, smoothening of CYGNSS measurements results in lower spatial resolution, approaching SMAP's resolution of 36km x 36km.

Five areas in northern India and Pakistan will be used to conduct this analysis. The regions were selected based on the CYGNSS surface reflectivity footprint and will be referred to as areas one through five. Figure 6.10 (a) shows the average surface reflectivity in January 2020 with the five areas marked in different colors. Area 3, area 4, and area 5 were selected because they possess relatively uniform surface reflectivity. Whereas area 2 has high surface reflectivity values, area 5 covers low values, and area 4 is between the latter regions. Furthermore, area 2 was selected due to its large variation in surface reflectivity moving from south to north. Finally, area 1 also contains large surface reflectivity oscillations, but the variations are more distributed inside the area than in the case for area 2.

Figure 6.11 shows the development of the correlation between the CYGNSS surface reflectivity and the SMAP soil moisture when varying the sigma of the Gaussian kernel. The x-axis shows the different sigmas used in the smoothening, ranging from 0 to 20. The y-axis shows the unitless Pearson correlation coefficient between the smoothened CYGNSS data and SMAP. All areas demonstrate an increase in correlation after the smoothening is performed. Area 1, 3, and 4 demonstrate similar patterns with a rapid increase in correlation using sigmas between 0.2 and 2.7 before slowly decreasing towards the maximum sigma. This pattern is also present for area 5. However the increase is less rapid than the case for the formerly mentioned areas. Area 2 distinguishes itself from the other areas in such a way that the correlation slowly increases during the whole range of applied sigmas.

The analysis was conducted using a spatial resolution of 0.1° x 0.1°. In northern India

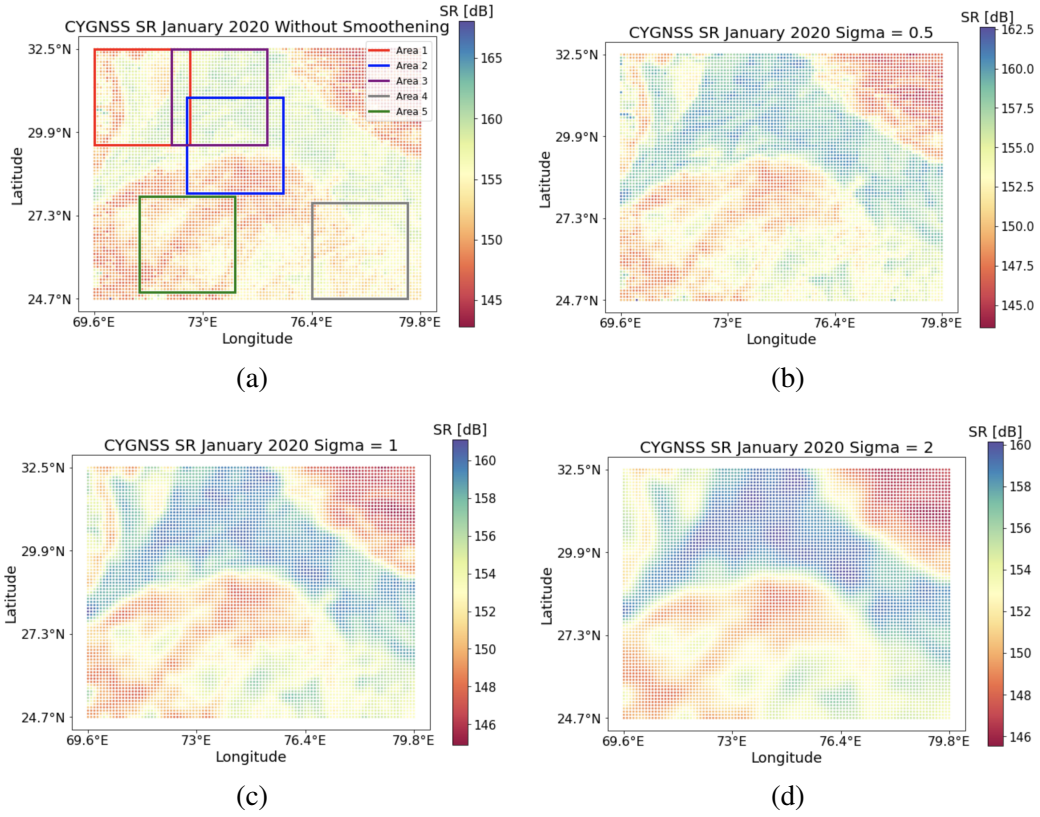


Figure 6.10: Surface reflectivity for the region covering northern India and Pakistan without smoothening and the five areas used to conduct the spatial resolution analysis (a). The same area with smoothening using sigma = 0.5 (b), sigma = 1 (c) and sigma = 2 (d).

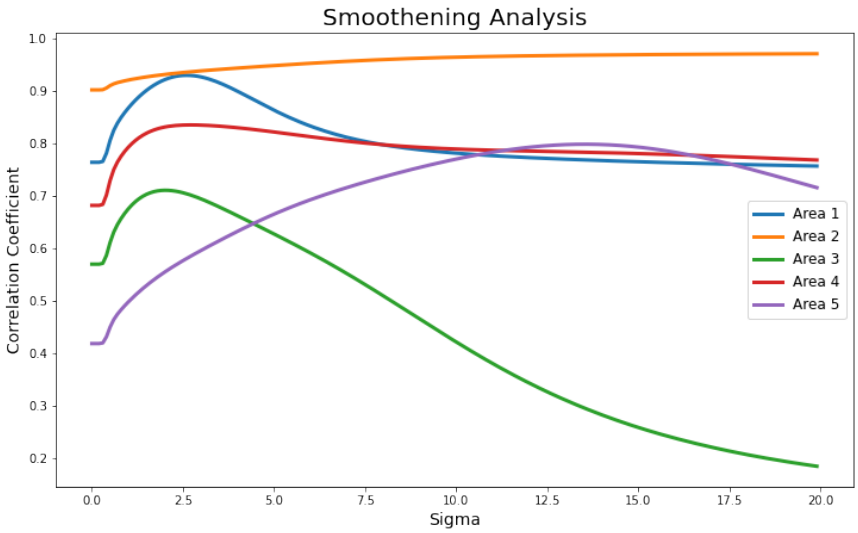


Figure 6.11: The distribution of correlation between CYGNSS and SMAP when applying Gaussian filters with different sigmas on the CYGNSS data.

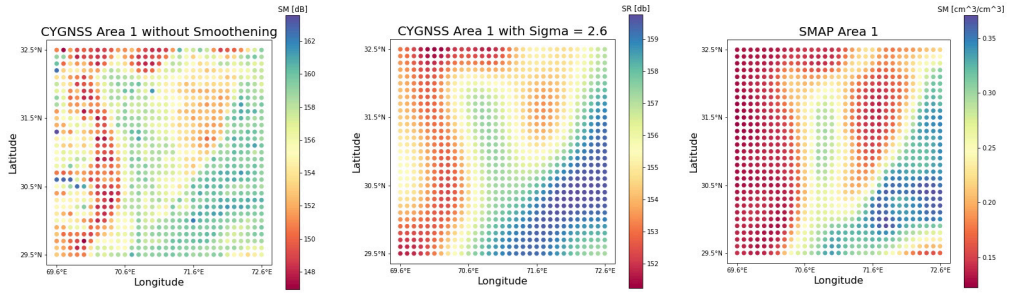
and Pakistan, 0.1° corresponds to approximately 11.1 km. This means that using a sigma of 2.7 results in a spatial resolution of $11.1km * 2.7 = 30km$. This is close to the original spatial resolution of the SMAP dataset, which is 36km x 36km.

Table 6.4: Correlation between CYGNSS and SMAP after applying Gaussian smoothening for five areas in northern India and Pakistan.

Area	**Start Correlation**	**Peak Correlation**	**Relative Change [%]**	**Best Sigma**
Area 1	0.764	0.929	21.6	2.6
Area 2	0.902	0.970	7.5	19.9
Area 3	0.570	0.710	24.6	2.0
Area 4	0.682	0.835	22.4	2.7
Area 5	0.418	0.798	47.6	13.5

The start correlation, peak correlation, relative correlation change, and the best sigma for the five analyzed areas are presented in Table 6.4. Areas with initially lower correlation, such as area 5, show a more considerable relative change than the latter areas. The opposite is true for e.g. area 2, experiencing the lowest relative change, but the initial start correlation is highest. Area 1, 3, and 4 show smaller changes, but the same pattern

is also present here.



(a) **(b)** **(c)**

Figure 6.12: CYGNSS area 1 without smoothening (a), with the best smoothening sigma = 2.6 (b), and SMAP (c).

Figure 6.12 (a) shows Area 1 without smoothening, (b) shows Area 1 when smoothening using the highest correlated sigma was performed, and (c) shows SMAP soil moisture for the same region. The best sigma was 2.6, which illustrates the increased similarities between CYGNSS and SMAP as their spatial resolution approaches each other.

6.4 Time Series Analysis

Time series were produced for all five regions presented in section 5.2. However, some of them demonstrate similar patterns and do not contribute significantly to increased insight into the use of CYGNSS for soil moisture estimation. Therefore, only the time series analysis for India, Iran, and the Democratic Republic of the Congo is presented and discussed in this section. The results from Australia and Brazil are shown in Appendix B.

6.4.1 Time Series Analysis in India

Figure 6.13 (a) shows the correlation between the CYGNSS time series and the SMAP time series for each grid cell in India. Darker green colors indicate higher correlation, and the exact correlation is written inside each cell. The blue cell has the highest correlation, and the red possesses the lowest correlation. The figure shows a trend where high and low correlation areas form clusters with similarly performing cells. This can result from geophysical conditions at the surface affecting the scattering of the signals. Figure 6.13 (b) shows a satellite image of the same region. The marked red box covers parts of the Eastern Ghats, a discontinuous chain of mountains along the east coast of India. The

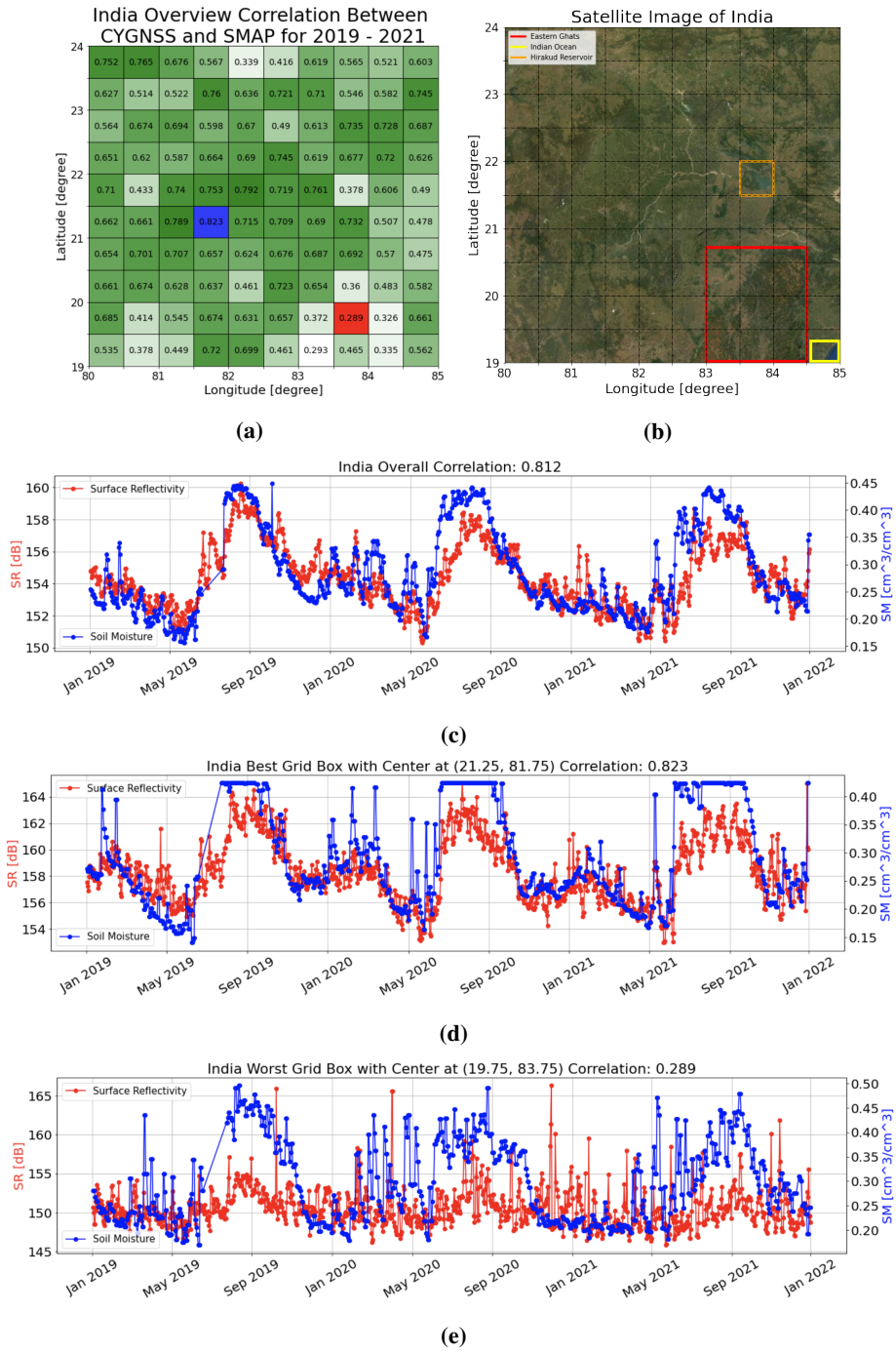
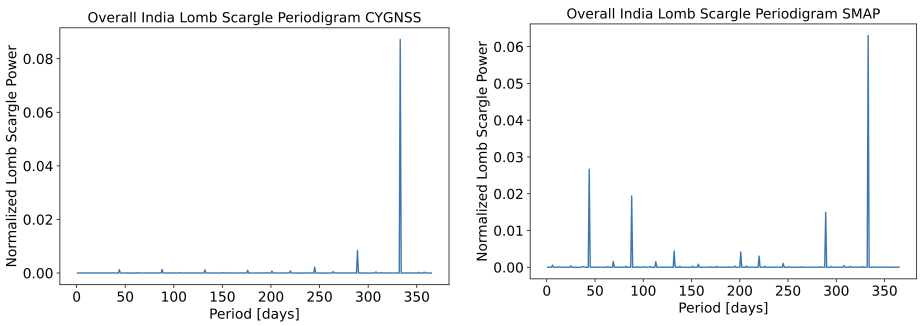


Figure 6.13: Overview of time series correlation for each grid cell in India (a). Overall Time series for India (b). Time series for the best (c) and worst (d) correlated grid cell.



(a) **(b)**

Figure 6.14: Lomb-Scargle periodogram in India for CYGNSS (a) and SMAP (b).

tallest mountains in this area exceed 1700 meters above sea level and contribute to significantly increased surface steepness. This can be one reason for the lower correlation in these grid boxes. Furthermore, the yellow box in Figure 6.13 (b) shows that the Indian Ocean is present in the south-easternmost grid box, which can also contribute to inaccurate GNSS-R measurements. This can happen if e.g. parts of the satellite footprint cover water during a single measurement. It can also occur when large inland water bodies are present. The Hirakud Reservoir is a 55 km long lake, which is marked in orange in the satellite image. This area corresponds to the grid cell with a center point at 21.75 latitude and 83.75 longitude, with a relatively low correlation of 0.378, where the presence of the Hirakud Reservoir can explain this performance. On the other hand, the cluster of grid cells with high correlation surrounding the blue grid cell share similar geophysical surface conditions, with relatively low steepness and little vegetation cover.

Furthermore, the three years dual-axis time series for the entire 5° x 5° area in India is shown in Figure 6.13 (c), where the red line shows the CYGNSS surface reflectivity and the blue line shows the SMAP soil moisture. CYGNSS has its corresponding y-axis to the left and SMAP to the right. The x-axis shows the day after the 1st of January 2019. This region demonstrates a correlation of 0.812 and displays clear seasonal variations. During the summer months between June and September, India is affected by heavy precipitation from the southwest monsoon. This results in periods where increased soil moisture emerges, thus creating the pattern shown in the time series.

Figure 6.13 (d) and (e) plots the time series for the blue and red cell respectively. Because the spatial resolution is higher for these plots, they demonstrate larger variance compared to Figure 6.13 (c). This is partly due to the fewer measurements located within a small grid cell than for the whole region. However, the same seasonal pattern occurs for the highest correlated cell as in the overall time series. The larger differences between

CYGNSS and SMAP takes place in Figure 6.13 (e). SMAP still demonstrates the same seasonal pattern for this area, which is not equally present in the CYGNSS time series.

The seasonality also reveals itself in Figure 6.14, which plots the Lomb-Scargle periodogram in the entire India-region for CYGNSS (a) and SMAP (b). Figure 6.14 (a) suggests that CYGNSS follows a clear seasonal pattern with a period of 333 days, which is slightly less than one year. The sharp peak at a period of 333 days is also present for SMAP in Figure 6.14 (b). However, the SMAP time series appears to possess other periodic signals with different periods, although they are less significant than the close to annual signal.

6.4.2 Time Series Analysis in Iran

Figure 6.15 shows the time series analysis between 2019 and 2021 in Iran. The altitude in this area varies greatly, with the lowest areas being located at sea level and the highest points exceeding 1900 meters above sea level. Furthermore, there is little to no vegetation cover in the entire region. The correlation between CYGNSS and SMAP for the overall area is 0.347, being considerably lower compared to India. The SMAP soil moisture is usually low, with a few peaks occurring during extreme precipitation. This means that the general seasonal trend in soil moisture seen in India is absent. In this case, the differences between the CYGNSS and the SMAP space missions affect the measurements to a larger extent, similar to what was discussed in section 6.2.

Between the mountains in this region, there exist several sand basins. These basins are prone to flooding during extreme rainfall due to the lower altitude location compared to their surroundings. The water flows down the mountains and accumulates in the sand basins. One such sand basin is marked in the blue rectangle in Figure 6.15 (b). This is also the grid cell with the highest correlation, marked in blue in Figure 6.15 (a). Between the 8th and 14th of January 2020, this region experienced a flood after days of heavy precipitation [58]. The same reason caused another flood to occur at the beginning of May 2021. These flooding events are present in the time series in Figure 6.15 (d), where both the surface reflectivity and the soil moisture increase significantly during the flooding periods. The presence of major changes in soil moisture is captured by both CYGNSS and SMAP, which leads to a correlation of 0.753. This is significantly better than for the region as a whole. Finally, in the worst-performing time series shown in Figure 6.15 (e), CYGNSS and SMAP seem to be uncorrelated. This can again result from a lack of clear soil moisture patterns, leading to noise and different mission configurations dominating the measured fluctuations.

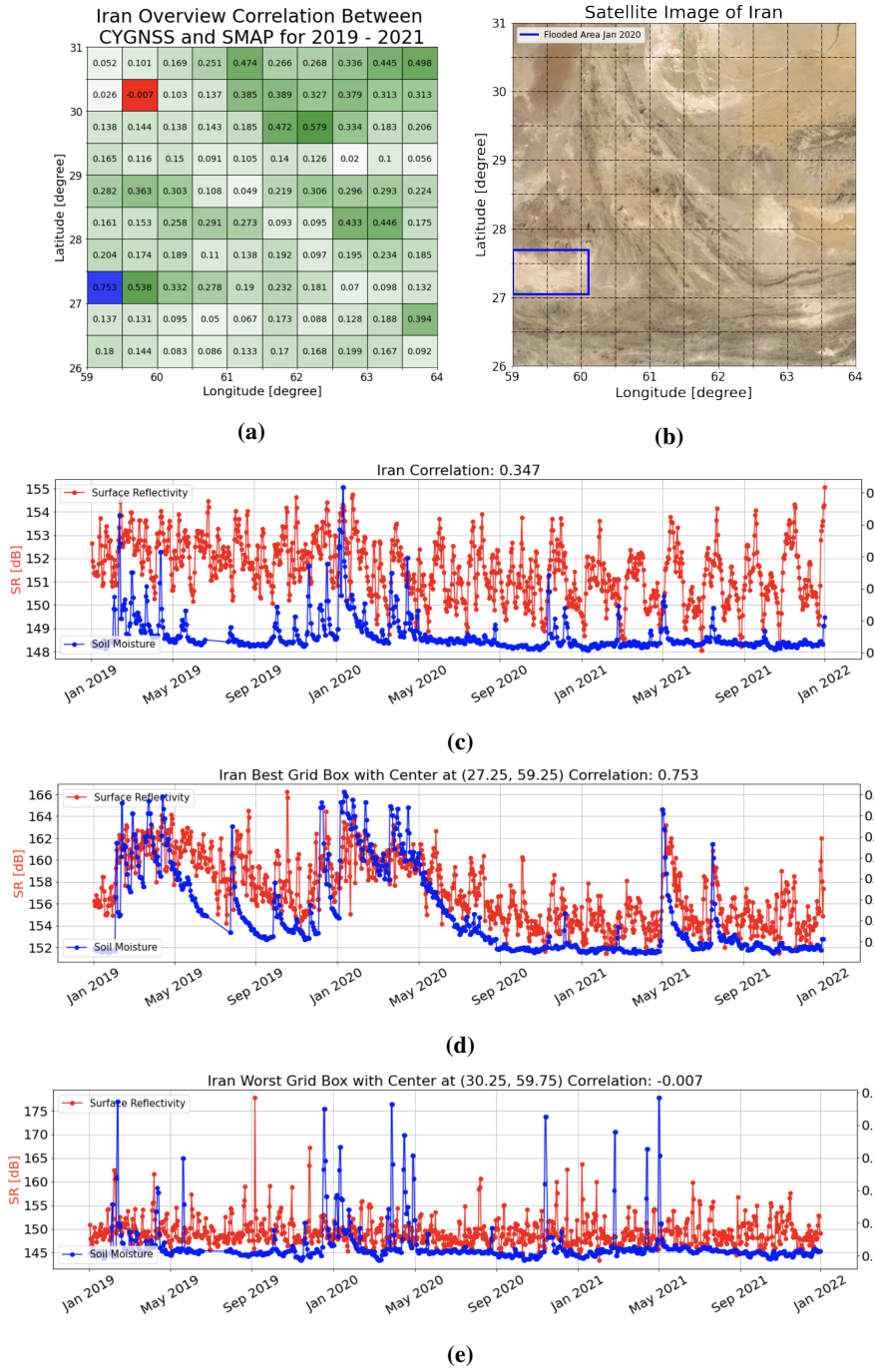
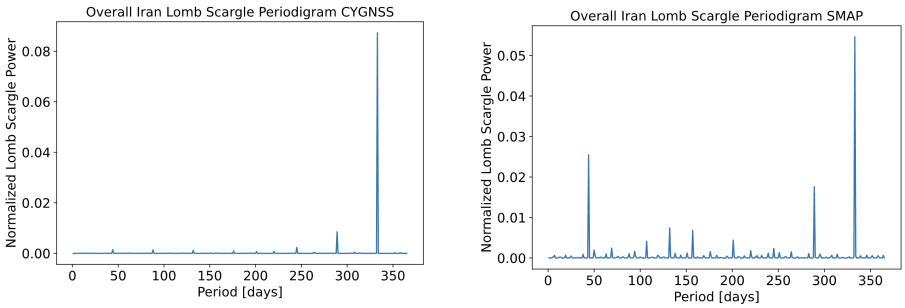


Figure 6.15: Overview of time series correlation for each grid cell in Iran (a). Satellite image of the same region (b). Overall Time series for Iran (c). Time series for the best (d) and worst (e) correlated grid cell.



(a) **(b)**

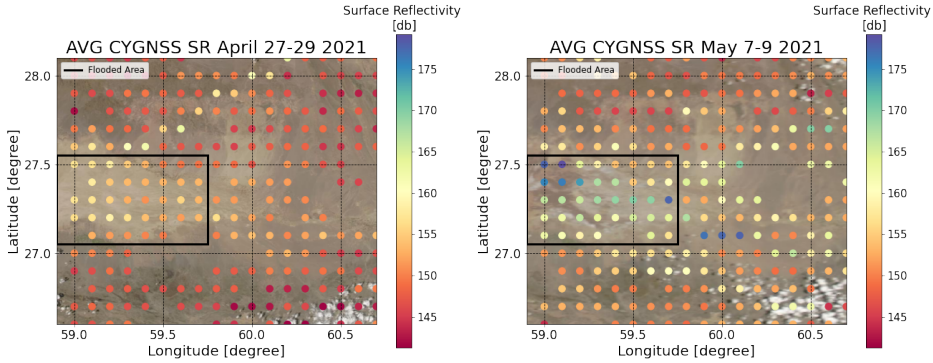
Figure 6.16: Lomb-Scargle periodogram in Iran for CYGNSS (a) and SMAP (b).

The Lomb-Scargle periodogram for the entire Iran area is shown in Figure 6.16, based on the CYGNSS (a) and SMAP (b) time series respectively. Similar to the same periodogram for India, CYGNSS reveals a seasonal period of 333 days. However, the 333 days period is less visible in the time series in Figure 6.15 (c) compared to the India time series. SMAP shows three significant peaks. The first peak is located at 44 days, the second at 290 days, and the final and most dominant at 333 days. This means that both CYGNSS and SMAP share the same overall periodicity of 333 days also in Iran.

Figure 6.17 takes a closer look on the surface reflectivity during the flood in May 2021, using a spatial resolution of 0.1° x 0.1°. Figure 6.17 (a) shows the average surface reflectivity from the 27th to the 29th of April 2021 and Figure 6.17 (b) shows the same values for the 7th to 9th of May 2021. The flooded area is marked in the black rectangle and shows an increase in surface reflectivity after the heavy precipitation. The surface reflectivity is overlaid MODIS images, where the extensive water is visible as a darker color in the flooded area. CYGNSS appears to be able to capture the flooded area with an increase of between 5 and 10 dB.

6.4.3 Time Series Analysis in the Democratic Republic of Congo

The time series analysis conducted in DR Congo is shown in Figure 6.19. DR Congo introduces a new pattern in terms of clustering of similar performing grid cells. The grid cell with the lowest correlation, marked in red in Figure 6.19 is positioned close to a cell with a correlation of 0.612. The adjacency of differently correlated grid cells was not present in India or Iran. Figure 6.19 (b) shows that this region is characterized by two major inland water bodies, namely the Upemba Lake and the Kisale Lake, and is marked in orange. In India, regions containing lakes caused a lower correlation between CYGNSS and SMAP. However, despite the grid cell with a correlation of 0.612 covering



(a) **(b)**

Figure 6.17: CYGNSS surface reflectivity before flooding in Iran 2021 (a) and during flood in Iran 2021 (b).

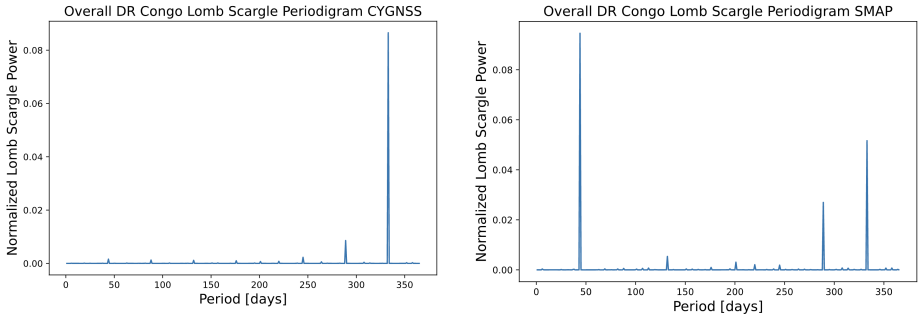
parts of the Kisale Lake, the geophysical conditions in the remaining cell is a relatively flat swamp with little to no vegetation cover. These conditions appear to be favorable for GNSS-R measurement, leading to a high correlation. Inland water bodies still seem to create unpredictable behavior in terms of CYGNSS and SMAP correlation and need to be taken into account when performing remote sensing of soil moisture.

Furthermore, the 5° x 5° region in DR Congo contains several forests characterized by dense vegetation cover. One such forest is marked in purple in Figure 6.19 (b), which leads to a low correlation between CYGNSS and SMAP. The low correlation indicates that vegetation cover is unfavorable for GNSS-R measurements. African savannas are ecosystems with mostly grassland and small or dispersed trees, creating favorable conditions to avoid interrupting the reflection of GNSS signals. Two such regions are marked in yellow in the satellite image of DR Congo. These areas also correspond to a higher correlation, again highlighting the importance of vegetation cover. Less vegetation cover tends to increase the correlation between CYGNSS and SMAP.

The overall time series shown in Figure 6.19 (c) demonstrates a correlation of 0.554. Both CYGNSS and SMAP appear to change with seasonality. The surface reflectivity varies between 148 dB and 154 dB, and SMAP varies between 0.15 cm^3/cm^3 and 0.35 cm^3/cm^3. The same pattern is present in Figure 6.19 (d), which is the time series for the highest correlated grid cell. However, in the time series for the lowest correlated cell shown in Figure 6.19 (e), the same pattern is only present in the SMAP data. This led to the two datasets being close to uncorrelated in this region, as also occurred in the lowest

correlated grid cell in India.

The Lomb-Scargle periodogram for the CYGNSS surface reflectivity in DR Congo is shown in Figure 6.18 (a) and for SMAP in Figure 6.18 (b). The periodogram for CYGNSS is similar to the likes of India and Iran, with a seasonal pattern of 333 days. However, SMAP has its highest peak at 44 days, revealing a different dominating period than India and Iran.



(a) **(b)**

Figure 6.18: Lomb-Scargle periodogram in DR Congo for CYGNSS (a) and SMAP (b).

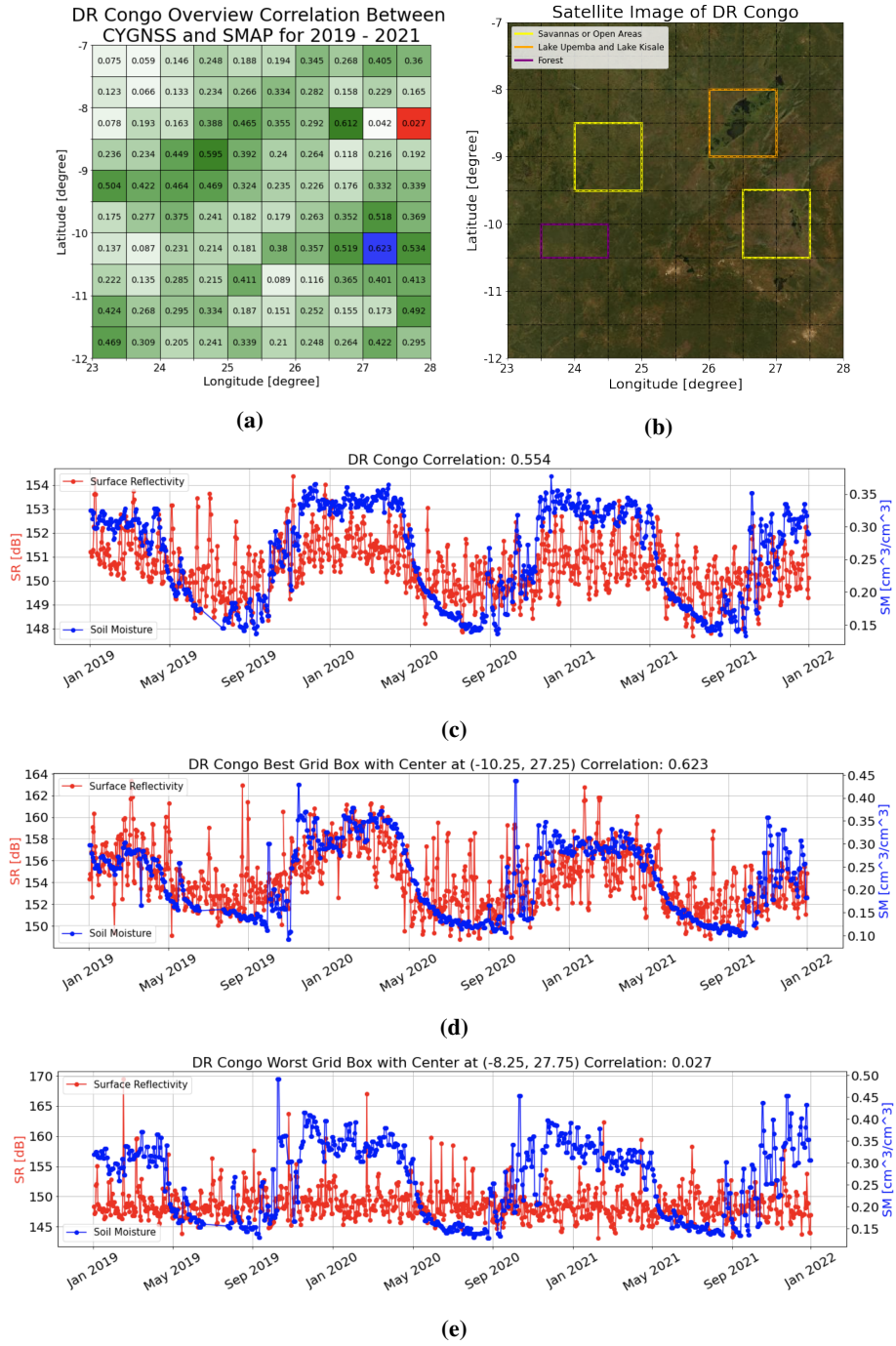


Figure 6.19: Overview of time series correlation for each grid cell in DR Congo from 2019-2021 (a). Satellite image of the area (c). Overall Time series for DR Congo (b). Time series for the best (c) and worst (d) correlated grid cell.

6.5 Machine Learning for Soil Moisture Estimation

A machine learning approach for the estimation of soil moisture from CYGNSS surface reflectivity was conducted using grid cells in India, Iran, and DR Congo. The cells were selected based on distinct surface characteristics or stand-out correlation values, which were presented in the previous section.

Initially, the effect of using different interpolation methods was investigated by applying AutoML using solely surface reflectivity as input. In the data preprocessing, both linear and natural neighbor interpolation was utilized to colocate the CYGNSS and SMAP measurements. The implementation of the natural neighbor algorithm was restricted to return a regular 3D grid, with interpolated values for all combinations of latitude, longitude, and time. To interpolate a region spanning 1.5° x 1.5°, using a spatial resolution of 0.5km x 0.5km and a temporal resolution of 12 hours, more than 238 million interpolated values would have to be generated. In addition, nearest neighbor interpolation would then have to be performed in order to fully colocate the CYGNSS measurements to the regular grid. These limitations for natural neighbor interpolation, both with respect to computation complexity and spatiotemporal resolution, caused linear 3D interpolation to outperform natural neighbor. Consequently, linear interpolation is used in the remaining study.

6.5.1 Soil Moisture Estimation Using Solely Surface Reflectivity

Firstly, the performance of soil moisture estimation, solely using surface reflectivity as model input, is investigated. Table 6.5 presents the RMSE obtained in each area of interest. The grid cells with the highest and lowest time series correlation are inspected for each area. In addition, a densely vegetated region in DR Congo, denoted *Forest* and represented by the left-most grid cell highlighted in purple in Figure 6.19 (b), is selected. The table illustrates how some areas with a low correlation between CYGNSS and SMAP achieve smaller RMSEs than areas with a higher correlation. Particularly the lowest correlated grid cell in Iran shows a significant reduction in RMSE compared to those obtained in other areas. One of the reasons why Iranian areas indicate high model performance is the absence of significant soil moisture fluctuations, in contrast to India and DR Congo.

Figure 6.20 (a) presents the estimated soil moisture values in the grid cell with the highest correlation in India using XGBoost for the latter half of 2021. The values are plotted against SMAP soil moisture, highlighting how the model struggles to capture the magnitude of the seasonal patterns in soil moisture. A similar visualization for the lowest correlated area in Iran is presented in Figure 6.20 (b). Scaled to represent the same soil

Table 6.5: Obtained RMSE when solely using surface reflectivity in soil moisture estimation. The values are calculated for grid cells obtaining the highest and lowest correlation in the time series analyses, in addition to a densely vegetated area in DR Congo.

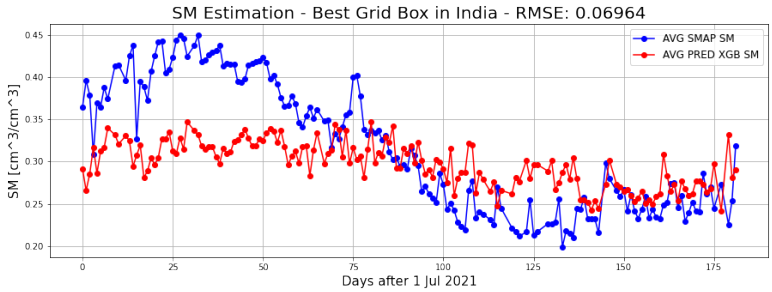
Area	XGB	GBM	DRF	Stacking	DL
India - High Corr	0.06964	0.06958	0.07568	0.06974	0.07807
India - Low Corr	0.08622	0.08559	0.09441	0.08546	0.13314
Iran - High Corr	0.05222	0.05184	0.05994	0.05065	0.03827
Iran - Low Corr	0.01206	0.01206	0.01937	0.01163	0.00634
DR Congo - High Corr	0.09593	0.09604	0.10514	0.09700	0.12299
DR Congo - Low Corr	0.09613	0.09575	0.10429	0.09564	0.09236
DR Congo - Forest	0.09013	0.09011	0.09462	0.09007	0.10687

moisture variations as in (a), Figure 6.20 (b) shows how the absence of large soil moisture fluctuations results in a lower RMSE. The small RMSE does not necessarily indicate that the XGBoost would perform well in the event of sudden precipitation, as this phenomenon is absent in the training data. However, Figure 6.20 enlightens the effects of periodicity, especially in areas with seasonal variations in climate and precipitation.

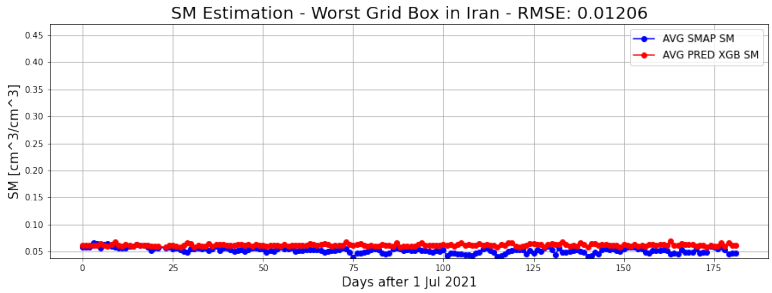
6.5.2 Soil Moisture Estimation with Ancillary Data

In addition to the inclusion of time, represented by the day of year, the relevance of other ancillary parameters were investigated. Therefore, the inclusion of geophysical parameters such as SMAP vegetation opacity and surface roughness coefficients, incidence angle, time of day, and the transmitting satellite block were used as input to the AutoML.

Table 6.6 presents the resulting RMSE within the different areas when the mentioned parameters were included. The table shows a decrease in RMSE in India, DR Congo, and the highest correlated cell in Iran, except the neural network. Figure 6.21 shows the predicted soil moisture utilizing the additional input parameters and XGBoost for the highest correlated grid cell in India. As indicated by the obtained RMSEs, the XGBoost model is now able to capture seasonal fluctuations in soil moisture, thus halving the RMSE compared to the model solely based on solely surface reflectivity. The feature importance of the XGBoost model, presented in Figure 6.22, highlights the contribution of the additional parameters in soil moisture estimation. The categorical features are one-hot-encoded, resulting in these values representing single instances of the specific features. On the contrary, the Iranian areas show no clear improvements when adding the ancillary data. This might result from noise being responsible for the majority of the RMSE, which is not connected to surface-related terms.



(a)



(b)

Figure 6.20: The predicted soil moisture in the highest correlated grid cell in India (a) and the lowest correlated grid cell in Iran (b) for the latter half of 2021. The predictions, colored in red and solely based on CYGNSS surface reflectivity, are compared to SMAP soil moisture shown in blue.

Table 6.6: Obtained RMSE using eight relevant features in soil moisture estimation within the grid cells obtaining the highest and lowest correlation in the time series analysis. A densely vegetated area in DR Congo is also studied.

Area	**XGB**	**GBM**	**DRF**	**Stacking**	**DL**
India - High Corr	0.03772	0.03255	0.03936	0.03252	0.04249
India - Low Corr	0.05032	0.05106	0.06063	0.04839	0.05260
Iran - High Corr	0.04705	0.03983	0.03495	0.03975	0.04815
Iran - Low Corr	0.02291	0.02711	0.01994	0.02417	0.02185
DR Congo - High Corr	0.04823	0.05397	0.04995	0.04936	0.05378
DR Congo - Low Corr	0.05934	0.06366	0.06504	0.06239	0.06922
DR Congo - Forest	0.04926	0.05137	0.05650	0.04872	0.06233

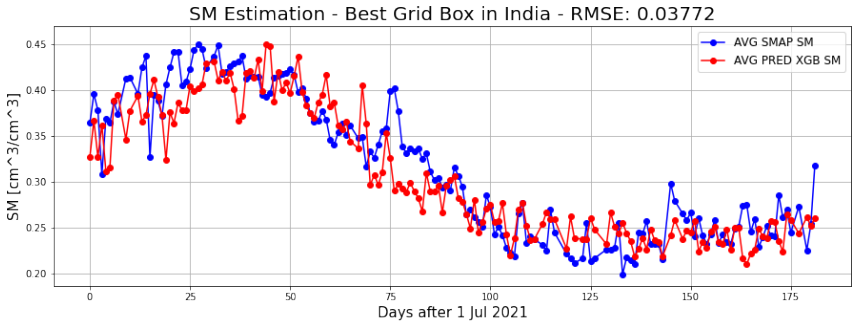


Figure 6.21: A visualization of the predicted soil moisture estimations in the best correlation grid cell in India for the last half of 2021 using XGBoost. The predictions, colored in red and based on eight relevant parameters, are compared to the true SMAP values, colored in blue.

6.5.3 Correction for Topographical Effects

As introduced in section 5.6, elevation steepness and height deviation were computed. The inclusion of such parameters in soil moisture estimation was based on observed fluctuations in CYGNSS data, believed to be caused by steep topography. Feeding the two parameters as additional input to the AutoML models already utilizing ancillary data reduced the RMSEs in both the highly correlated cell and the densely vegetated cell in DR Congo. The two cells obtained respective RMSEs of 0.04550 cm^3/cm^3 and 0.04890 cm^3/cm^3, despite the remaining areas experiencing a decrease in estimation accuracy.

The change in elevation steepness can contribute to reducing or strengthening the reflection of GNSS signals. This depends on the geometry between the transmitting GPS and the receiver and the direction of the slope. However, the elevation steepness proposed in this study only accounts for the magnitude of the angular difference. As a result, the parameter provides no information on wether or not it contributes to strengthening or reducing the reflections. This can be improved by including the positions of the GPS and CYGNSS satellites so that the complete geometry is taken into account. In this way, it is possible to determine the actual effect of different slopes.

6.5.4 Evaluation of the Incidence Angle Filtration Algorithm

As introduced in subsection 5.3.3, this thesis proposed a novel method to perform optimized data selection based on incidence angles. This algorithm aims to reduce the measurement uncertainties while preserving a high amount of CYGNSS measurements. By minimizing the F-value presented in Equation 5.10, two optimal incidence angles were retrieved for each of the areas in India, Iran, and DR Congo. These are presented in

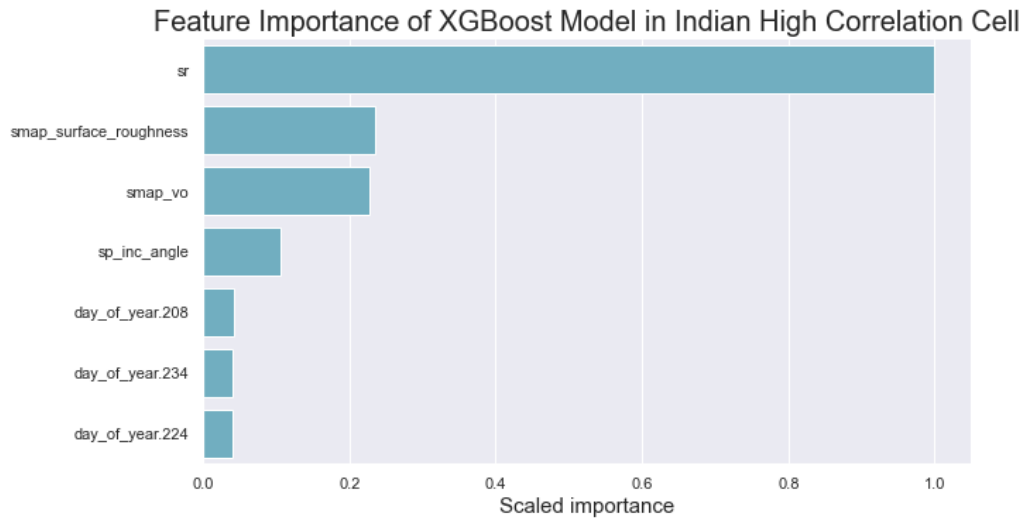


Figure 6.22: Feature importance of the XGBoost model for the highest correlated grid cell in India with ancillary data included as input parameters.

Table 6.7. In addition, Figure C.1 and Figure C.2 in Appendix C show how the F-value fluctuate using different incidence angle intervals, from which the optimal intervals are selected in India and DR Congo respectively.

Table 6.7: Proposed incidence angle intervals in India, Iran and DR Congo based the optimization method introduced in this thesis.

	Incidence Angle Interval 1	**Incidence Angle Interval 2**
India	22° - 32°	44° - 54°
Iran	7° - 17°	28° - 38°
DR Congo	11° - 21°	45° - 55°

The same machine learning approach, including all input parameters, was used to evaluate the change in model performance when filtering on the proposed incidence angles. Whereas the RMSE in the two grid cells in Iran decreased, the cells in India and DR Congo obtained larger RMSEs, as shown in Table 6.8.

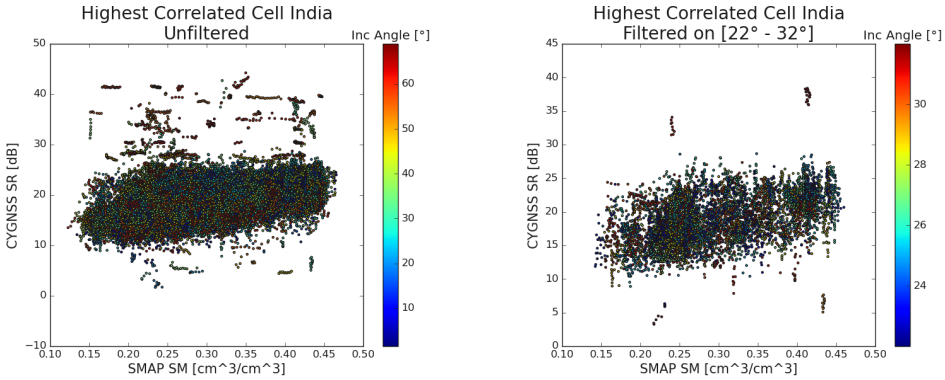
A challenge with the proposed method is to obtain accurate approximations of the uncertainties for the areas. As previously mentioned, Iran is mostly arid, meaning that the uncertainty of the measurements can be approximated by computing the median standard deviation of CYGNSS measurements made within uniformly dry periods. This leads to an accurate estimate of the uncertainty for the region. However, in India and DR Congo,

Table 6.8: Comparison of RMSEs with and without optimized data selection based on incidence angle filtration.

	RMSE Without Filtration	**RMSE With IA Interval 1**	**RMSE With IA Interval 2**
India - High Corr	0.03289	0.04384	0.04925
India - Low Corr	0.04870	0.06400	0.06447
Iran - High Corr	0.03645	0.03253	0.03353
Iran - Low Corr	0.02056	0.01281	0.00627
DR Congo - High Corr	0.04550	0.06203	0.06073
DR Congo - Low Corr	0.06034	0.06851	0.08163
DR Congo - Forest	0.04848	0.06458	0.06905

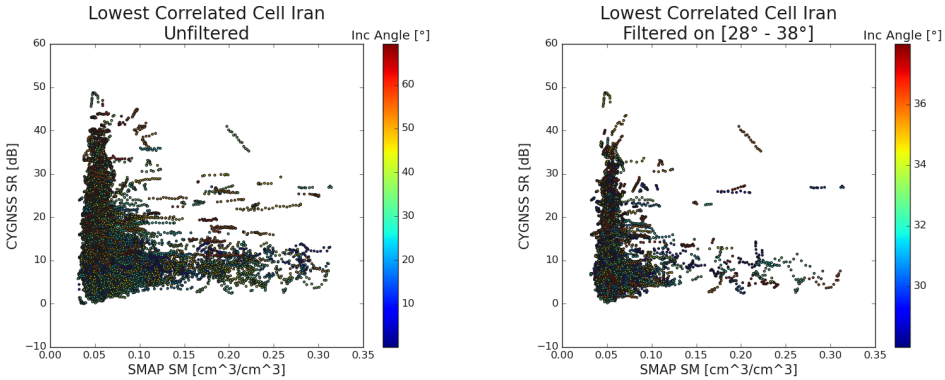
the seasonal variations in soil moisture make the same approach less accurate. This is because fluctuations in soil moisture contribute to increasing the uncertainties instead of noise being the primary factor. This can explain why the models performed better after filtration in Iran but not in India and DR Congo.

Another possible limitation of the proposed algorithm is the reduction in the amount of CYGNSS measurements fed to the machine learning models after filtration. Data-driven methods, including machine learning, highly rely on large amounts of training data to capture the trends. The Indian grid cells show a reduction in data of around 80%, while the areas in Iran and DR Congo neglect up to 90% of the dataset. Figure 6.23 and Figure 6.24 shows the data distribution before (a) and after (b) incidence angle filtration in the grid highest correlated grid cell in India and the lowest correlated grid cell in Iran respectively. The figures present how the proposed optimization of data filtration based on incidence angle removes a significant amount of observable outliers.



(a) **(b)**

Figure 6.23: Scatter plots of CYGNSS surface reflectivity and SMAP soil moisture values at different incidence angles in the highest correlated cell in India measured between 2019 and 2021. The figures visualize the measurements before (a) and after (b) applying optimized data filtration based on incidence angle.



(a) **(b)**

Figure 6.24: Scatter plots of CYGNSS surface reflectivity and SMAP soil moisture values at different incidence angles in the lowest correlated cell in Iran measured between 2019 and 2021. The figures visualize the measurements before (a) and after (b) applying optimized data filtration based on incidence angle.

This page is intentionally left blank

Chapter 7

Conclusion

In this thesis, the feasibility of soil moisture remote sensing using the Global Navigation Satellite System (GNSS) Reflectometry (GNSS-R) technique is investigated. We used a dataset of spaceborne GNSS-R observations from the NASA Cyclone GNSS (CYGNSS). The data includes information about the strength of the reflected GNSS signals, enveloped in Delay Doppler Maps (DDMs) as one of the main CYGNSS data products. Signal-to-noise ratio (SNR) observations from the DDMs were used in the calculation of a parameter called surface reflectivity, which can reveal soil moisture variations.

For evaluation of the results, two ancillary data sources, namely European Centre for Medium-Range Weather Forecast (ECMWF) ReAnalaysis-5 (ERA5) and Soil Moisture Active Passive (SMAP), were colocated with the CYGNSS data. ERA5 uses retrospective analysis and physical models to produce global estimates of soil moisture, and SMAP utilizes both active radio and passive radiometer instruments to produce soil moisture estimates based on observations.

We have developed a software package to download, process, store, and analyze CYGNSS Level 1 data in combination with the ancillary data sources. The software, written in Python, is capable of handling large amounts of data, enabling the production of CYGNSS surface reflectivity time series over longer periods. Additionally, it includes machine learning models functioning as Geophysical Model Functions (GMFs), converting observed surface reflectivity to soil moisture estimates.

Several case studies are included in this thesis. A case study in northern India and Pakistan for January and August 2020 was analyzed to evaluate the performance of CYGNSS compared to SMAP and ERA5. The average CYGNSS surface reflectivity values for January and August demonstrated a significant correlation with the colocated SMAP soil moisture dataset, reaching a Pearson's correlation coefficient of 0.777 in Janu-

ary 2020. The comparison between CYGNSS and ERA5 showed a significantly lower correlation of 0.368 in the same time period. This can be due to ERA5's inability to capture small fluctuations in soil moisture and resulted in ERA5 being removed from the remaining analysis.

Furthermore, we showed that CYGNSS is able to capture more detailed variations in soil moisture compared to SMAP by applying Gaussian smoothening to the CYGNSS surface reflectivity. Five smaller areas in northern India and Pakistan all demonstrated an increase in correlation when CYGNSS approached the spatial resolution of SMAP.

Five areas of 5° x 5° in terms of latitude and longitude with different conditions related to vegetation and topography were selected to analyze the geophysical parameters affecting remote sensing of soil moisture. Time series of CYGNSS surface reflectivity and SMAP soil moisture were computed for each 0.5° x 0.5° grid cell in all five areas from 2019 until the end of 2021. The time series in Iran displayed a significant increase in surface reflectivity and soil moisture during two known flooding events, occurring in January 2020 and May 2021. This was evaluated against Moderate Resolution Imaging Spectroradiometer (MODIS) images, visually displaying the flood. The results highlight the potential of spaceborne GNSS-R for detecting and mapping floods even under severe weather conditions

Time series of the individual grid cells also revealed that mountainous areas, densely vegetated areas, and inland water bodies contributed to reducing the correlation between CYGNSS and SMAP. We then utilized machine learning for soil moisture estimation, which substantiated these results by obtaining higher root mean square errors (RMSEs) in the same areas. The importance of seasonality and topography in soil moisture estimation was also highlighted. The developed machine learning models in areas with significant soil moisture variations and steep topography experienced an increase in performance when these factors were taken into account.

We also presented a novel approach to perform optimal incidence angle filtration on the CYGNSS data. The proposed method showed promising results in areas where an accurate approximation of CYGNSS measurement uncertainties was producible. An area in Iran highlighted this promising result, where the obtained RMSE was reduced from 0.0204 cm^3/cm^3 to 0.0063 cm^3/cm^3 after data filtering on the optimal incidence angle interval was applied. This model used CYGNSS surface reflectivity, incidence angle, day of year, time of day, and SMAP ancillary data as input from 2019 to 2021. However, areas with larger soil moisture variation proved more difficult to produce accurate uncertainty approximations, thus enlightening an area of improvement for the proposed method.

7.1 Scientific Contributions

One of the objectives of this thesis was to make the most important findings from the analysis available to the science community and public as an attempt to raise awareness of this new remote sensing application. This goal accumulated in two different contributions, a web page allowing users to display time series of selected areas and a scientific paper aimed at a Norwegian audience.

We developed a web page where users can obtain CYGNSS- and SMAP time series based on a desired location and time frame. The graphical user interface (GUI) of the web page enables users to either physically draw a region on a map or feed specific coordinates in order to choose an area to analyze. Then, passing on the user-selected area, along with the time frame chosen by the user, the backend software conducts the data processing before delivering the final visualization to the users.

In addition, a Norwegian paper was written to introduce the possibility of using remote sensing and GNSS-R to produce soil moisture estimates. This technique is relatively new in Norway, and the goal of the paper was to give the readers an overview of the technology and describe two possible use cases. Initially, the article addresses the comparison between CYGNSS, SMAP, and ERA5, which was presented in section 6.1. Next, time series of the floods in Iran in 2020 and 2021 are presented to highlight the behavior of surface reflectivity during extreme weather conditions.

The paper is submitted to the Norwegian journal *Kart og Plan* and is currently under review for possible publication in the journal. *Kart og Plan* is a scientific journal established in 1908, focusing on geomatics, including satellite mapping and remote sensing. The Norwegian paper is attached in Appendix D.

Finally, in collaboration with the co-supervisor of this thesis, another article presenting high spatiotemporal resolution time series and utilization of machine learning for soil moisture estimation is in progress.

7.2 Future Work

The time series presented in this thesis showed significant increases in surface reflectivity in the event of past floods. In order to generate relevant information from time series analyses to be used in future flood detection, further research is advised to investigate the possibility of developing an algorithm capable of automatically detecting anomalies in time series. This can be utilized to indicate and alert regions endangered by extreme weather hazards.

Furthermore, accounting for the effect of topography and surface slopes can improve the GMF performance. For this purpose, the vector of incident signal can be compared to the normal vectors of local surface slopes for generating a correction term. The comparison would help machine learning models to infer if the angular change contributes to reduced or strengthened signal reflection. This can be done by utilizing the geometry of the transmitting GPS, the receiving CYGNSS satellite, and the specular point. These modifications can improve the soil moisture estimation and continue to lower the RMSE in future applications.

In addition, further investigation on how to approximate CYGNSS measurement uncertainties used in the proposed incidence angle optimization method is advised. The analysis in this thesis shows promising results when uncertainties can be approximated, thus highlighting the need for a good retrieval algorithm.

A possible addition to the generated software package would be extending the web page with a map presenting the most recent surface reflectivity values available. Combined with a geophysical model function for soil moisture conversion, generated using machine learning, this extension could provide the public with close to real-time monitoring of soil moisture on a global scale.

References

[1] Mohammad M. Al-Khaldi and Joel T. Johnson. Soil moisture retrievals using cygnss data in a time-series ratio method: Progress update and error analysis. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/LGRS.2021.3086092.

[2] Mohammad M. Al-Khaldi, Joel T. Johnson, Andrew J. O'Brien, Anna Balenzano, and Francesco Mattia. Time-series retrieval of soil moisture using cygnss. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7):4322–4331, 2019. doi: 10.1109/TGRS.2018.2890646.

[3] Milad Asgarimehr. *Spaceborne GNSS Reflectometry: Remote Sensing of Ocean and Atmosphere*. dissertation, Technische Universität Berlin, 2020.

[4] Jean-Paul Berrut and Lloyd N. Trefethen. Barycentric lagrange interpolation. *SIAM Review*, 46(3):501–517, 2004. doi: 10.1137/S0036144502417715. URL https://doi.org/10.1137/S0036144502417715.

[5] Priyadarshi Bhattacharya and Marina L. Gavrilova. Voronoi diagram in optimal path planning. In *4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD 2007)*, pages 38–47, 2007. doi: 10.1109/ISVD.2007.43.

[6] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

[7] Adriano Camps, Hyuk Park, Miriam Pablos, Giuseppe Foti, Christine P. Gommenginger, Pang-Wei Liu, and Jasmeet Judge. Sensitivity of gnss-r spaceborne observations to soil moisture and vegetation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(10):4730–4742, 2016. doi: 10.1109/JSTARS.2016.2588467.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016. doi: 10.1145/2939672.2939785.

[9] C. Chew and E. Small. Soil moisture sensing using spaceborne gnss reflections: Comparison of cygnss reflectivity to smap soil moisture. *Geophysical Research Letters*, 45, 04 2018. doi: 10.1029/2018GL077905.

[10] C. Chew and Eric Small. Description of the ucar/cu soil moisture product. *Remote Sensing*, 12:1558, 05 2020. doi: 10.3390/rs12101558.

[11] CC Chew and EE Small. Soil moisture sensing using spaceborne gnss reflections: Comparison of cygnss reflectivity to smap soil moisture. *Geophysical Research Letters*, 45(9):4049–4057, 2018.

[12] Clara Chew. Spatial interpolation based on previously-observed behavior: a framework for interpolating spaceborne gnss-r data from cygnss. *Journal of Spatial Science*, 0(0):1–14, 2021. doi: 10.1080/14498596.2021.1942253. URL https://doi.org/10.1080/14498596.2021.1942253.

[13] Clara Chew, John T. Reager, and Eric Small. Cygnss data map flood inundation during the 2017 atlantic hurricane season. *Scientific Reports (Nature Publisher Group)*, 8:1–8, 06 2018. URL https://www.proquest.com/scholarly-journals/cygnss-data-map-flood-inundation-during-2017/docview/2057071880/se-2?accountid=12870.

[14] Raffaele Cioffi, Marta Travaglioni, Giuseppina Piscitelli, Antonella Petrillo, and Fabio De Felice. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability*, 12(2), 2020. ISSN 2071-1050. doi: 10.3390/su12020492. URL https://www.mdpi.com/2071-1050/12/2/492.

[15] Gabriela Civeira. *Introductory Chapter: Soil Moisture*. 01 2019. ISBN 978-1-78985-103-8. doi: 10.5772/intechopen.83603.

[16] Maria Paola Clarizia, Christopher Ruf, Paolo Cipollini, and Cinzia Zuffada. First spaceborne observation of sea surface height using gps-reflectometry. *Geophysical Research Letters*, 43(2):767–774, 2016. doi: https://doi.org/10.1002/2015GL066624. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL066624.

[17] Maria Paola Clarizia, Nazzareno Pierdicca, Fabiano Costantini, and Nicolas Floury. Analysis of cygnss data for soil moisture retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2227–2235, 2019. doi: 10.1109/JSTARS.2019.2895510.

[18] Maria Paola Clarizia, Valery Zavorotny, Darren McKague, and Chris Ruf. Algorithm theoretical basis document level 2 wind speed retrieval, 2020. URL https://podaac-tools.jpl.nasa.gov/drive/files/allData/cygnss/L2/docs/148-0138-6_ATBD_L2_v3.0_Wind_Speed_Retrieval.pdf.

[19] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. *Supervised Learning*, pages 21–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-75171-7. doi: 10.1007/978-3-540-75171-7_2. URL https://doi.org/10.1007/978-3-540-75171-7_2.

[20] Alejandro Egido. *GNSS Reflectometry for Land Remote Sensing Applications*. PhD thesis, 07 2013.

[21] Alejandro Egido Egido. *GNSS reflectometry for land remote sensing applications*. PhD thesis, Universitat Politècnica de Catalunya Barcelona, Spain, 2014.

[22] Dara Entekhabi, Simon H. Yueh, Peggy E. O'neill, Kent H. Kellogg, Angela Morin Allen, Rajat Bindlish, Molly E. Brown, Steven Tsz K. Chan, Andreas Colliander, Wade T. Crow, Narendra Narayan Das, Gabrielle De Lannoy, Roy Scott Dunbar, Wendy N. Edelstein, Jared K. Entin, Vanessa M. Escobar, Shawn D. Goodman, Thomas J. Jackson, Benhan Jai, Joel T. Johnson, Edward J. Kim, Seungbum Kim, John S. Kimball, Randal D. Koster, Amanda Leon, Kyle McDonald, Mahta Moghaddam, Priscilla N. Mohammed, Susan Moran, Eni G. Njoku, Jeffrey R. Piepmeier, Rolf H. Reichle, Francois Rogez, Jiancheng Shi, Michael W. Spencer, Sam W. Thurman, Leung Tsang, Jakob J. van Zyl, Barry H. Weiss, and Richard D. West. Smap handbook–soil moisture active passive: Mapping soil moisture and freeze/thaw from space. 2014.

[23] Andrew Feldman, Ruzbeh Akbar, and Dara Entekhabi. Characterization of higher-order scattering from vegetation with smap measurements. *Remote Sensing of Environment*, pages 324–338, 10 2018. doi: 10.1016/j.rse.2018.10.022.

[24] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[25] Giuseppe Foti, Christine Gommenginger, Philip Jales, Martin Unwin, Andrew Shaw, Colette Robertson, and Josep Roselló. Spaceborne gnss reflectometry for ocean winds: First results from the uk techdemosat-1 mission. *Geophysical Research Letters*, 42(13):5435–5441, 2015. doi: https://doi.org/10.1002/2015GL064204.

[26] Jianghui Geng, Peng Jiang, and Jingnan Liu. Integrating gps with glonass for high-rate seismogeodesy. *Geophysical research letters*, 44(7):3139–3146, 2017.

[27] Scott Gleason, Andrew O'Brien, Anthony Russel, Mohammad M. Al-Khaldi, and Joel T. Johnson. Geolocation, calibration and surface resolution of cygnss gnss-r land observations. *Remote Sensing*, 12(8), 2020. ISSN 2072-4292. doi: 10.3390/rs12081317. URL https://www.mdpi.com/2072-4292/12/8/1317.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Machine learning basics. *Deep learning*, 1(7):98–164, 2016.

[29] Mohinder S. Grewal, Lawrence R. Weill, and Angus P. Andrews. *Global Positioning Systems, Inertial Navigation and Integration*. John Wiley Sons, Inc., 2001. ISBN 0-471-20071-9.

[30] Mohinder S. Grewal, Lawrence R. Weill, and Angus P. Andrews. *Global Positioning Systems, Inertial Navigation, and Integration*. Wiley-Interscience, USA, 2007. ISBN 0470041900.

[31] C.D. Hall and R.A. Cordey. Multistatic scatterometry. In *International Geoscience and Remote Sensing Symposium, 'Remote Sensing: Moving Toward the 21st Century'.*, volume 1, pages 561–562, 1988. doi: 10.1109/IGARSS.1988.570200.

[32] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[33] Mostafa Hoseini, Maximilian Semmling, Hossein Nahavandchi, Erik Rennspiess, Markus Ramatschi, Rüdiger Haas, Joakim Strandberg, and Jens Wickert. On the response of polarimetric gnss-reflectometry to sea surface roughness. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7945–7956, 2021. doi: 10.1109/TGRS.2020.3031396.

[34] Dong Huang, Ricardo Cabral, and Fernando De la Torre. Robust regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):363–375, 2016. doi: 10.1109/TPAMI.2015.2448091.

[35] V.J. Inglezakis, S.G. Poulopoulos, E. Arkhangelsky, A.A. Zorpas, and A.N. Menegaki. Chapter 3 - aquatic environment. In Stavros G. Poulopoulos and Vassilis J. Inglezakis, editors, *Environment and Development*, pages 137–212. Elsevier, Amsterdam, 2016. ISBN 978-0-444-62733-9. doi: https://doi.org/10.1016/B978-0-444-62733-9.00003-4. URL https://www.sciencedirect.com/science/article/pii/B9780444627339000034.

[36] Yan Jia, Shuanggen Jin, Haolin Chen, Qingyun Yan, Patrizia Savi, Yan Jin, and Yuan Yuan. Temporal-spatial soil moisture estimation from cygnss using machine

learning regression with a preclassification approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4879–4893, 2021. doi: 10.1109/JSTARS.2021.3076470.

[37] Shuanggen Jin, Estel Cardellach, and Feiqin Xie. *GNSS Remote Sensing*. Springer Netherlands, 2014. ISBN 978-94-007-7481-0. doi: 10.1007/978-94-007-7482-7.

[38] Cheng Jing, Yang Yu, Xiaofeng Yang, Wentao Ma, Di Dong, and Ziwei Li. Studies on delay-doppler mapping of gnss-r signals. In *Remote Sensing of the Environment: 19th National Symposium on Remote Sensing of China*, volume 9669, page 966904. International Society for Optics and Photonics, 2015.

[39] ADNAN Kavak, WJ Vogel, and Guanghan Xu. Using gps to measure ground complex permittivity. *Electronics Letters*, 34(3):254–255, 1998.

[40] Eugene F Knott, John F Schaeffer, and Michael T Tulley. *Radar cross section*. SciTech Publishing, 2004.

[41] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2014.11.005. URL https://www.sciencedirect.com/science/article/pii/S2001037014000464.

[42] Herbert J. Kramer. Cygnss (cyclone global navigation satellite system). https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/cygnss#spacecraft, 2021. Accessed April 30th 2022.

[43] Nina Lam. Spatial interpolation methods: a review. *American Cartographer*, 10: 129–149, 01 1983.

[44] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

[45] Erin LeDell and Sebastien Poirier. H2o automl: Scalable automatic machine learning. 2020.

[46] Fangni Lei, Volkan Senyurek, Mehmet Kurum, Ali Gurbuz, Dylan Boyd, and Robert Moorhead. Quasi-global gnss-r soil moisture retrievals at high spatio-temporal resolution from cygnss and smap data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 6303–6306. IEEE, 2021.

[47] Chor Pang Lo. Applied remote sensing. 1986.

[48] Manuel Martin-Neira. A passive reflectometry and interferometry system (paris): Application to ocean altimetry. *ESA Journal*, 17:331–355, 01 1993.

[49] Loris Michel, Jeffrey Näf, Peter Bühlmann, Nicolai Meinshausen, and Domagoj Ćevid. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. 09 2021.

[50] Vimal Mishra, Amar Deep Tiwari, Saran Aadhar, Reepal Shah, Mu Xiao, DS Pai, and Dennis Lettenmaier. Drought and famine in india, 1870–2016. *Geophysical Research Letters*, 46(4):2075–2083, 2019.

[51] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[52] Mary Morris, Clara Chew, John T. Reager, Rashmi Shah, and Cinzia Zuffada. A novel approach to monitoring wetland dynamics using cygnss: Everglades case study. *Remote Sensing of Environment*, 233:111417, 2019. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2019.111417.

[53] M.S. Neiman. The principle of reciprocity in antenna theory. *Proceedings of the IRE*, 31(12):666–671, 1943. doi: 10.1109/JRPROC.1943.233683.

[54] Felipe G. Nievinski and Kristine M. Larson. Forward modeling of gps multipath for near-surface reflectometry and positioning applications. *GPS Solutions*, 18: 309–322, 2013.

[55] Clement Ogaja. *Applied GPS for engineers and project managers*. 08 2011. ISBN 978-0-7844-1150-6. doi: 10.1061/9780784411506.

[56] Linda Paulson. Developers shift to dynamic programming languages. *Computer*, 40:12 – 15, 03 2007. doi: 10.1109/MC.2007.53.

[57] Bohdan Pavlyshenko. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 255–258, 2018. doi: 10.1109/DSMP.2018.8478522.

[58] Mahmoud Rajabi, Hossein Nahavandchi, and Mostafa Hoseini. Evaluation of cygnss observations for flood detection and mapping during sistan and baluchestan torrential rain in 2020. *Water*, 12(7), 2020. ISSN 2073-4441. doi: 10.3390/w12072047. URL https://www.mdpi.com/2073-4441/12/7/2047.

[59] R. Rew and G. Davis. Netcdf: an interface for scientific data access, 1990.

[60] Glenn Robb. Circularly polarized antennas explained, without the math. 2017.

[61] Thomas Roberts, Ian Colwell, C. Chew, Rashmi Shah, and S. Lowe. A deep learning approach to soil moisture estimation with gnss-r. 08 2021. doi: 10.36227/techrxiv.15068031.v1.

[62] Christopher Ruf, Derek Posselt, Sharanya Majumdar, Scott Gleason, M.P. Clarizia, Derek Starkenburg, Damen Provost, Valery Zavorotny, John Murray, Stephen Musko, Zorana Jelenak, Paul Chang, and Mary Morris. *CYGNSS Handbook*. 04 2016. ISBN 978-1-60785-380-0.

[63] Christopher S Ruf, Clara Chew, Timothy Lang, Mary G Morris, Kyle Nave, Aaron Ridley, and Rajeswari Balasubramaniam. A new paradigm in earth environmental monitoring with the cygnss small satellite constellation. *Scientific reports*, 8(1): 1–13, 2018.

[64] A. Maximilian Semmling, Anja Rösel, Dmitry V. Divine, Sebastian Gerland, Georges Stienne, Serge Reboul, Marcel Ludwig, Jens Wickert, and Harald Schuh. Sea-ice concentration derived from gnss reflection measurements in fram strait. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10350–10361, 2019. doi: 10.1109/TGRS.2019.2933911.

[65] Volkan Senyurek, Fangni Lei, Dylan Boyd, Ali Gurbuz, Mehmet Kurum, and Robert Moorhead. Evaluations of machine learning-based cygnss soil moisture estimates against smap observations. *Remote Sensing*, 12, 10 2020. doi: 10.3390/rs12213503.

[66] Volkan Senyurek, Fangni Lei, Dylan Boyd, Mehmet Kurum, Ali Gurbuz, and Robert Moorhead. Machine learning-based cygnss soil moisture estimates over ismn sites in conus. *Remote Sensing*, 12:1168, 04 2020. doi: 10.3390/rs12071168.

[67] Volkan Senyurek, Ali Gurbuz, Mehmet Kurum, Fangni Lei, Dylan Boyd, and Robert Moorhead. Spatial and temporal interpolation of cygnss soil moisture estimations. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 6307–6310, 2021. doi: 10.1109/IGARSS47720.2021.9553900.

[68] Jyoti Sharma, Rajenda Prasad, Prashant Srivastava, Shubham Singh, Suraj Yadav, and Dr Yadav. Roughness characterization and disaggregation of coarse resolution smap soil moisture using single-channel algorithm. *Journal of Applied Remote Sensing*, 15, 03 2021. doi: 10.1117/1.JRS.15.014514.

[69] Yajie Shi, Yueji Liang, Chao Ren, Jianmin Lai, Qin Ding, and Xinmiao Hu. Investigating the effects of meteorological data rainfall and temperature on gnss-r

soil moisture inversion. pages 97–100, 2021. doi: 10.1109/GNSSR53802.2021.9617574.

[70] Robin Sibson. A brief description of natural neighbor interpolation. 1981.

[71] Merill I. Skolnik. An analysis of bistatic radar. *IRE Transactions on Aerospace and Navigational Electronics*, ANE-8(1):19–27, 1961. doi: 10.1109/TANE3.1961.4201772.

[72] Benjamin Southwell and Andrew Dempster. A new approach to determine the specular point of forward reflected gnss signals. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–8, 01 2018. doi: 10.1109/JSTARS.2017.2775647.

[73] I. Stančin and A. Jović. An overview and comparison of free python libraries for data mining and big data analysis. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 977–982, 2019. doi: 10.23919/MIPRO.2019.8757088.

[74] Fengyu Tang and Songhua Yan. Cygnss soil moisture estimations based on quality control. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/LGRS.2021.3119850.

[75] Peter J.G. Teunissen and Oliver Montenbruck. *Springer Handbook of Global Navigation Satellite Systems*. Springer International Publishing AG, 2017. ISBN 978-3-319-42926-7.

[76] Omar Torres. *Analysis of reflected global positioning system signals as a method for the determination of soil moisture*. The University of Texas at El Paso, 2004.

[77] Alexander G. Tsidaev. Parallel algorithm for natural neighbor interpolation. 2016.

[78] Jacob T VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, 2018.

[79] Simon H. Yueh, Rashmi Shah, M. Julian Chaubell, Akiko Hayashi, Xiaolan Xu, and Andreas Colliander. A semiempirical modeling of soil moisture, vegetation, and surface roughness impact on cygnss reflectometry data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. doi: 10.1109/TGRS.2020.3035989.

[80] Paul Zandbergen. Applications of shuttle radar topography mission elevation data. *Geography Compass*, 2(5):1404–1431, 2008. doi: https://doi.org/10.1111/j.1749-8198.2008.00154.x. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2008.00154.x.

[81] V Zavorotny, Dallas Masters, A Gasiewski, B Bartram, S Katzberg, Penina Axelrad, and R Zamora. Seasonal polarimetric measurements of soil moisture using tower-based gps bistatic radar. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, volume 2, pages 781–783. IEEE, 2003.

[82] Valery Zavorotny and Chris Ruf. Algorithm theoretical basis document level 2 mean-square slope retrieval, 2016. URL https://podaac-tools.jpl.nasa.gov/drive/files/allData/cygnss/L2/docs/148-0139-L2-MSS-ATBD-R3.pdf.

[83] V.U. Zavorotny and A.G. Voronovich. Bistatic gps signal reflections at various polarizations from rough land surface with moisture content. In *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, volume 7, pages 2852–2854 vol.7, 2000. doi: 10.1109/IGARSS.2000.860269.

This page is intentionally left blank

Appendix A

Missing CYGNSS Satellites Information

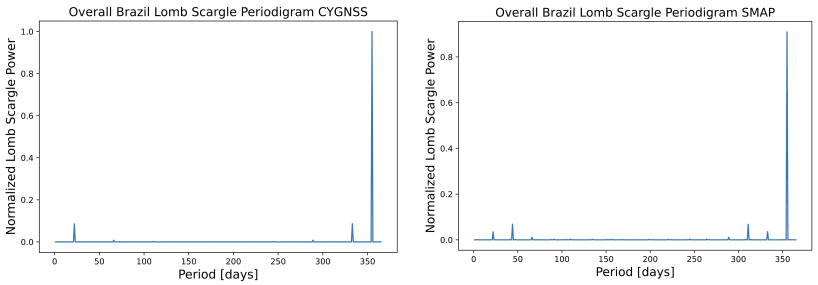
Table A.1: A complete list of satellites failing to provide daily CYGNSS measurements during the downloading of data for the years of 2019, 2020 and 2021. The values of format X(Y, Z) indicates missing data from satellites Y and Z on day X in that respective month.

Month	2019	2020	2021
Jan	12(6), 13(6), 14(6), 15(6), 21(7), 22(7)	23(3), 24(4), 26(5, 6), 29(3), 30(4), 31(4)	1(4), 2(4), 5(2), 6(2), 7(2), 12(2), 13(2), 14(2), 19(7), 20(1, 2, 3, 5, 6, 7), 21(5), 22(5), 23(5), 24(5), 25(5), 26(5), 27(5)
Feb	27(7), 28(4, 7)	1(4), 18(2), 20(7), 21(7, 8), 22(6, 8), 23(6), 24(1), 25(1), 28(4, 5)	24(7)
Mar	1(4), 12(8), 13(8)	31(6)	10(3), 11(3), 21(6), 22(6), 23(6, 7)
Apr	7(2), 8(2)	14(2), 15(2)	27(3), 28(3), 29(3)
May	No satellites missing	15(6, 8), 16(6, 8), 21(2, 3), 30(7, 8)	13(7), 16(1), 17(1)

Jun	1(5, 6, 7, 8), 22(7, 8), 25(1, 2, 3, 4)	1(4), 5(1), 6(1), 7(1, 4), 8(4)	15(6), 16(5), 17(5), 18(5), 28(6), 29(6), 30(6)
Jul	12(8), 13(8), 28(3, 7), 29(3, 7), 30(5), 31(5)	No satellites missing	6(1), 7(1), 8(1), 9(1), 10(1), 11(1), 12(1), 17(5), 18(5), 25(1)
Aug	1(5), 16(3), 17(3), 19(4), 20(4), 27(2), 28(2), 29(2)	6(4), 7(4), 10(4), 11(1), 22(2), 25(2)	No satellites missing
Sep	7(5), 8(5), 9(5), 11(3), 12(3), 18(1), 19(1), 20(1), 21(1), 24(1)	14(6), 15(6), 16(6), 17(6), 25(4), 26(4)	No satellites missing
Oct	4(7), 9(7), 15(8), 16(8), 17(8), 18(8)	3(4), 4(4, 7), 5(4, 7), 6(4, 7)	No satellites missing
Nov	14(7)	4(5), 5(5), 6(5, 6), 7(5), 27(1)	16(2), 17(2, 5), 18(5), 19(5)
Dec	14(7), 15(7), 16(7)	7(4), 8(4), 14(6), 15(2, 6), 16(2, 6), 17(2), 18(2), 20(7), 21(7), 22(7), 23(7), 24(1), 25(1), 28(1), 29(1), 30(1), 31(4)	1(2), 9(6, 7), 13(8), 23(5), 24(5), 28(8), 29(8), 31(8)

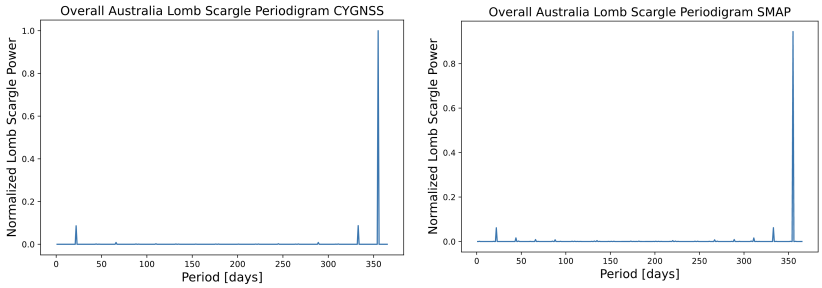
Appendix B

Time Series in Australia and Brazil



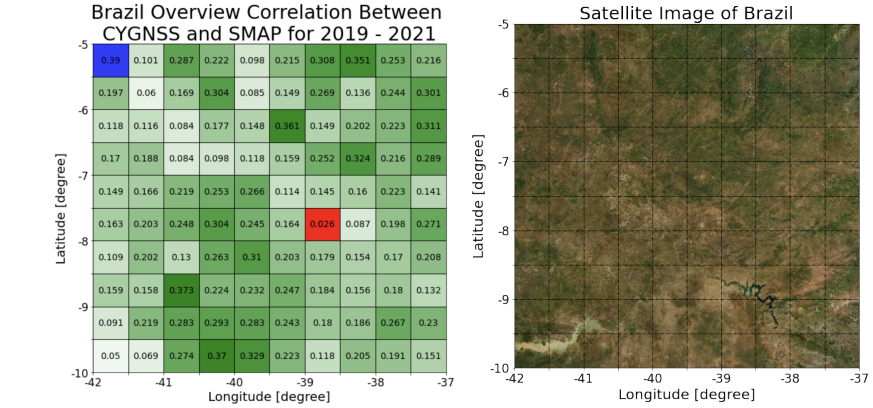
(a) **(b)**

Figure B.1: Lomb-Scargle periodogram in Brazil for CYGNSS (a) and SMAP (b)

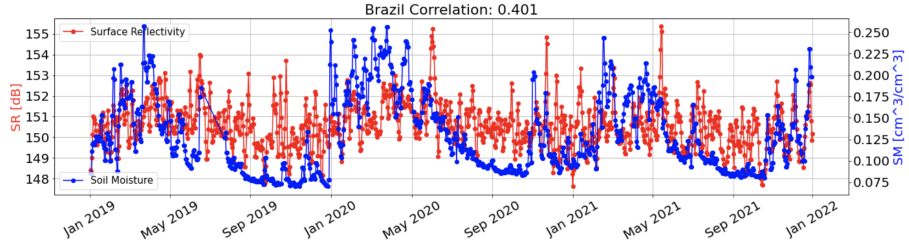


(a) **(b)**

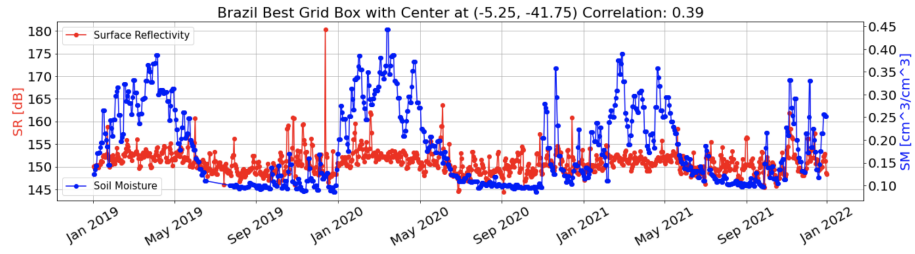
Figure B.2: Lomb-Scargle periodogram in Australia for CYGNSS (a) and SMAP (b)



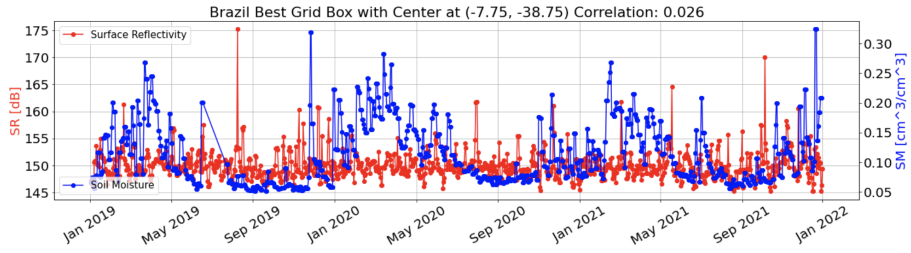
(a) **(b)**



(c)

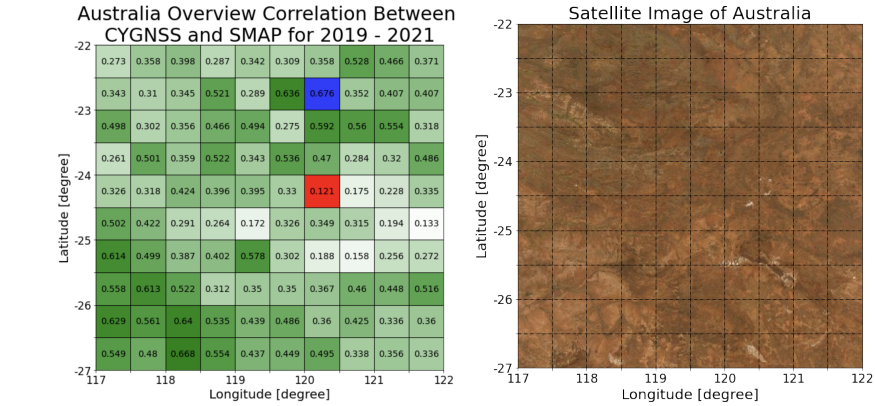


(d)

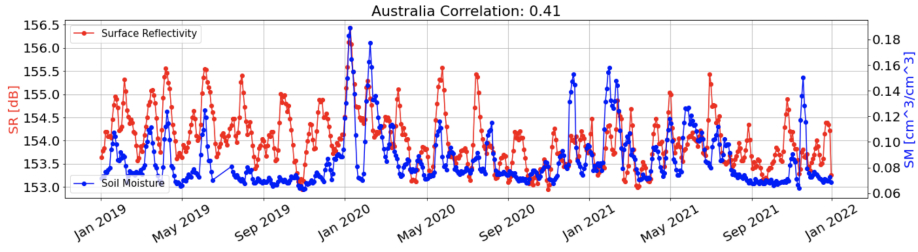


(e)

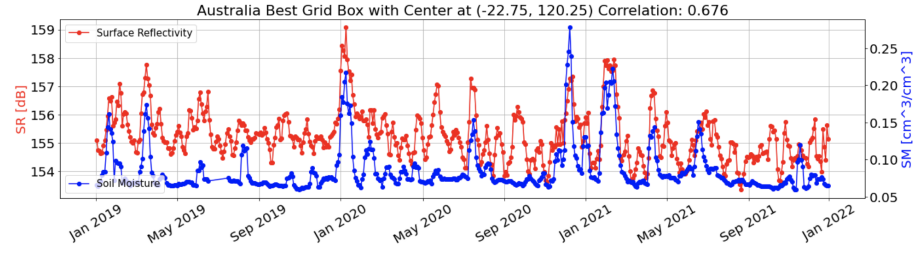
Figure B.3: Overview of time series correlation for each grid cell in Brazil from 2019-2021 (a). Satellite image of the area (c). Overall Time series for Brazil (b). Time series for the best (c) and worst (d) correlated grid cell.



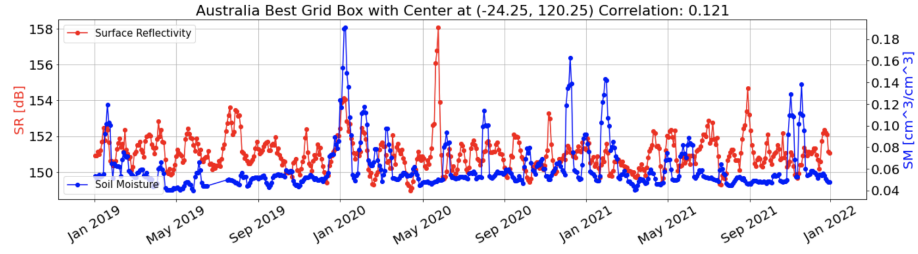
(a) **(b)**



(c)



(d)



(e)

Figure B.4: Overview of time series correlation for each grid cell in Australia from 2019-2021 (a). Satellite image of the area (c). Overall Time series for Australia (b). Time series for the best (c) and worst (d) correlated grid cell.

Appendix C

Computed F-values in India and DR Congo

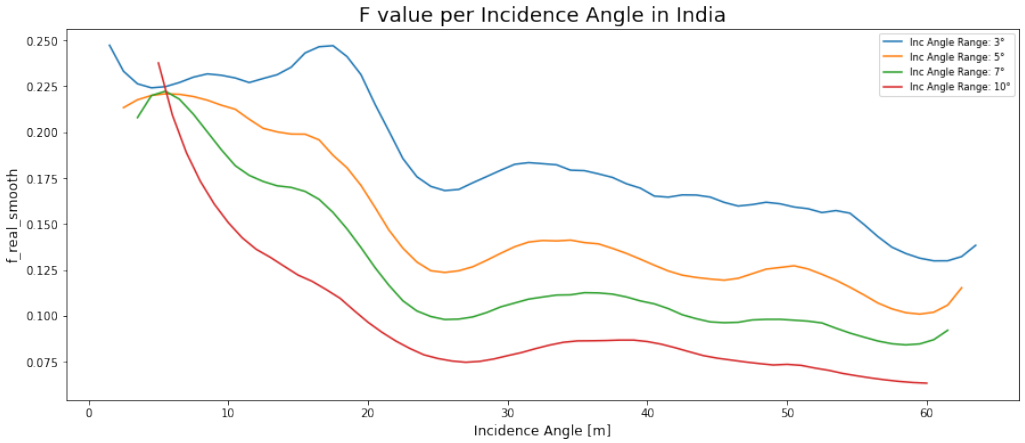


Figure C.1: The generated F-value in India using different incidence angles and incidence angle intervals.

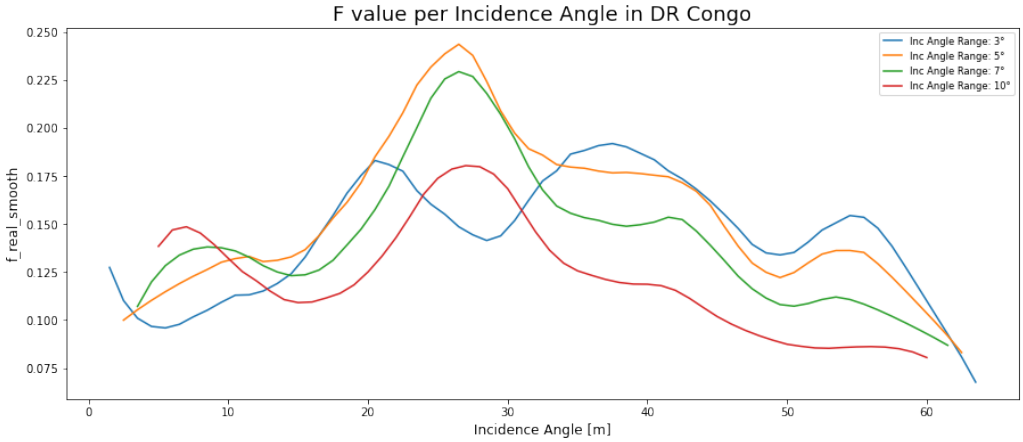


Figure C.2: The generated F-value in DR Congo using different incidence angles and incidence angle intervals.

Appendix D

The Norwegian Paper Publication

The submitted Norwegian paper is shown below. The layout is made for submission purposes, and will undergo further improvements.

En ny fjernmålingsteknikk for naturfarer og miljøovervåking ved bruk av mikrosatellitter

Vegard Haneberg, Mads Engja Rindal

June 10, 2022

Sammendrag

De stadig økende konsekvensene av klimaendringer fremhever behovet for å overvåke jordens vannsykler over land, spesielt med tanke på økningen av naturkatastrofer som flom og tørke. Denne studien viser at en ny fjernmålingsteknologi som kan overvåke variasjoner i jordfuktighet og flomhendelser ved bruk av mikrosatellitter. Denne teknologien er basert på konseptet Global Navitagion Satellite System Reflektometri (GNSS-R). GNSS-R analyserer variasjoner i signalstyrke etter refleksjon over land for å utlede informasjon om endringer i bakkens vanninnhold. To casestudier i India og Iran ble gjennomført for å undersøke fjernmåling av henholdsvis jordfuktighet og flomdeteksjon. Det benyttede datasettet i våre undersøkelser er hentet fra oppdraget kalt NASA Cyclone GNSS (CYGNSS), som består av åtte mikrosatellitter med en gjenbesøkstid på ca. 7 timer. I tillegg brukes to datasett med jordfuktighetsmålinger for valideringsformål. Våre resultater viser at målinger av overflatereflektion utledet fra GNSS-R-dataene effektivt kan fange opp romlige og tidsmessige variasjoner av jordfuktighet over casestudiet i India. Videre har studien av to flomhendelser i Sør-Iran vist potensialet til GNSS-R-målinger for å oppdage flom. Dette lovende resultatet antyder at fremtidige konstellasjoner av små GNSS-R-satellitter kan bidra til monitoreringen naturkatastrofer i nær sanntid.

Abstract

Climate change has caused an increase in the occurrence of natural disasters and hazards, including flooding events and droughts. This study investigates the utilization of Global Navigation Satellite System (GNSS) Reflectometry (GNSS-R) over land for remote sensing of soil moisture. The NASA Cyclone GNSS (CYGNSS) mission provides a publicly available GNSS-R dataset counting signal-to-noise ratio (SNR) observations of the reflected GNSS signal, with a revisit time of approximately 7 hours. We use these observations to calculate the surface reflectivity as an indicator of soil moisture variations. Two ancillary datasets providing soil moisture estimates are spatiotemporally co-located with the GNSS-R measurements and used for performance assessment. The ancillary datasets are the data products of the European Center for Medium-Range Weather Forecast (ECMWF) Re-Analysis-5 (ERA5) and the Soil Moisture Active Passive (SMAP) mission. We conducted two case studies in India and Iran to assess the performance of remote sensing for soil moisture estimation and flood detection. Our results show that CYGNSS-derived surface reflectivity capture variations in soil moisture in the case study in Iran. Furthermore, time series analysis demonstrates a significant increase in surface reflectivity during two known floods in Iran in January 2020 and May 2021. These promising results suggest that microsatellites functioning as receivers of reflected GNSS signals can contribute to the monitoring of natural hazards close to real-time.

Nøkkelord: fjernmåling, GNSS-R, jordfuktighet, naturkatastrofer

Forfattere:
Mads Engja Rindal. Master i Ingeniørvitenskap og IKT med spesialisering i geomatikk og geodesi ved NTNU Trondheim.
Institutt for bygg- og miljøteknikk
mads1997rindal@gmail.com

Vegard Haneberg. Master i Ingeniørvitenskap og IKT med spesialisering i geomatikk og geodesi ved NTNU Trondheim.
Institutt for bygg- og miljøteknikk
vegard.haneberg@live.no

Medforfattere:
Mostafa Hosieni. PhD fra departementet for bygg- og miljøteknikk ved NTNU med tittel *On the Remote Sensing of the Atmosphere and Ocean Using Direct and Reflected GNSS Signals*.
Institutt for bygg- og miljøteknikk
mostafa.hoseini@ntnu.no

Forfatterene erklærer ingen interessekonflikter ved denne artikkelen.

English title:
A novel remote sensing technique for monitoring of natural hazards

1 Introduksjon

Livet på jorden er tilknyttet ulike klimasykler der vann ofte innehar en viktig rolle. Vann er en livsviktig entitet for alle levende organismer, og god vanntilgang er essensielt for helsen til blant annet oksygen- og matproduserende planter. Mengden vann som er tilgjengelig i jorden avhenger av faktorer som årstid, jordtype og geografisk plassering. Likevel har målinger av jordfuktighet lenge vært en viktig parameter da de kan si mye om klimasituasjonen i området der målingen er gjort. Markante globale klimaendringer er blitt en stort utfordring dette århundret, som blant annet har medført hyppigere tilfeller av flom, tørke, skogbrann og andre hendelser som følge av ekstremvær [Trenberth, 2005]. Kontinuerlige målinger av jordfuktighet har derimot vist seg å bistå med relevant informasjon i monitoreringen av klimaendringene, og er vurdert til å være en viktig parameter i utviklingen av prognoser og skadeforebyggende arbeid relatert til vær og klima [Kerr, 2007, Entekhabi et al., 2014].

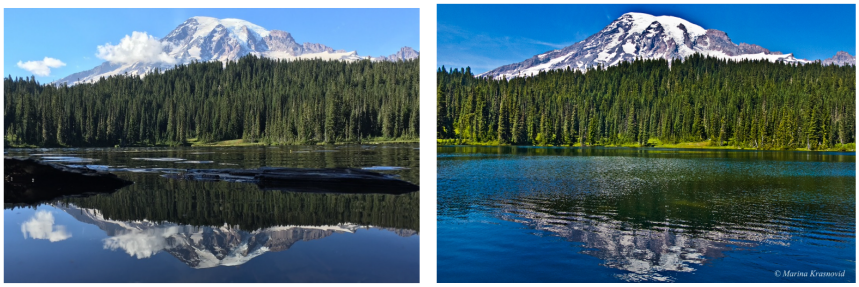
De tidligste metodene for innsamling av data tilknyttet jordfuktighet bestod av utstasjonerte måleinstrumenter for overvåkning av mindre områder som for eksempel jorder eller åkre. Slike metoder gjorde det mulig å måle jordfuktighet med høy romlig og tidsmessig oppløsning, men fungerte kun for begrensede områder. Videre ble muligheten for å gjøre tilsvarende undersøkelser ved hjelp av fjernmåling forsket på. Fly ble utstyrt med teknologi for både transmittering og måling av elektromagnetiske bølger som muliggjorde observasjoner av større områder enn utstasjonerte måleinstrumenter [Dobson and Ulaby, 1986]. Denne tankegangen ble deretter utvidet til å utnytte allerede transmitterte Global Positioning System (GPS) signaler for å utføre observering av jordfuktighet globalt [Kavak et al., 1998]. De nyeste metodene bruker dedikerte satellitter til å motta reflekterte GPS signaler og anvende disse til å utlede egenskaper om det reflekterte medium. Dette gjør det mulig å foreta globale målinger av jordfuktighet med høyere tidsmessig og romlig oppløsning enn tidligere metoder.

Global Navigation Satellite System (GNSS) satellitter transmitterer signaler med frekvenser mellom 1.2 GHz og 1.6 GHz, tilhørende L-båndet. L-båndet ble tatt i bruk for dette formålet på grunn at dets gunstige oppførsel ved penetrering av jordens atmosfære, samt lav påvirkningsgrad av ulike værforhold. For å sikre at GNSS skal kunne operere med ønsket nøyaktighet uavhengig av brukerplassering er det avgjørende at de transmitterte signalene ikke påvirkes av ulike værforhold i for stor grad. Anvendelsen av signaler med frekvenser innenfor L-båndet er dermed med på å minimere disse feilkildene [Teunissen and Montenbruck, 2017, Ogaja, 2011].

En annen kjent feilkilde ved GNSS-posisjonering er flerveisinterferens. Dette fenomenet oppstår når GNSS-signalene reflekteres før de treffer mottakeren. Da vil ikke signalet ha traversert korteste vei fra satellitt til mottaker, noe som kan bidra negativt til posisjoneringens nøyaktighet. Ved nyere forskningsprosjekter gjennomført av blant annet NASA, har det kommet frem at disse reflekterte signalene kan brukes til andre formål enn å være en feilkilde i posisjonering. Geofysiske forhold ved den reflekterende overflaten kan påvirke signalet på ulike måter, og brukes til å utlede egenskaper om jordens overflate [Martin-Neira, 1993].

GNSS-Reflektometri (GNSS-R) er en teknikk som kan utlede overflateegenskaper ved å analysere reflekterte GNSS-signaler. I motsetning til monostatiske radarsystemer der transmitter og mottaker er kolokalisert, benytter GNSS-R seg av bistatiske systemer hvor disse er separert. Refleksjon av elektromagnetiske (EM) signaler kan skje koherent eller inkoherent, avhengig av blant annet overflateøffheten. Jevnere overflater, som defineres basert på signalets bølgelengde, fører til koherent refleksjon hvor alle reflekterte EM-signaler sendes ut i én retning. På den annen side inntreffer inkoherent refleksjon ved ujevne overflater og dette fører til en spredning av de reflekterte EM signalene [Asgarimehr, 2020, Egido, 2014]. Figur 1 viser hvordan synlig lys reflekteres over vann med ulik overflateøffhet. Illustrasjon (b) i figuren demonstrerer hvordan en røff vannoverflate leder til inkoherent refleksjon av lysstråler og dermed hindrer dannelsen av klare refleksjonsbilder.

Fordelene med å bruke L-båndet er tilstedeværende også i GNSS-R. Ved å utnytte allerede opererende GNSS-satellitter kan GNSS-R systemer benyttes til global overvåkning av jordoverflaten med forbedret romlig og temporal oppløsning enn det som tidligere har vært mulig. NASA har engasjert romoppdrag hvor både monostatiske og bi-statiske GNSS-R systemer har vært i bruk. I desember 2016 ble Cyclone GNSS (CYGNSS) lansert med den hensikt å samle inn vær- og vinddata i forbindelse med dannelsen og oppførselen til tropiske sykloner over hav ved bruk av GNSS-R [Ruf et al., 2016]. Anvendelsen av et flertall mikrosatellitter opererende som mottakere av reflekterte GPS signaler i lav jordbane vil bidra til lavere gjenbesøkelsestid, og følgelig forbedret oppløsning på vitenskapelige målinger, sammenlignet med tidligere metoder. I tillegg til bedre oppløsning vil bruken av mikrosatellitter redusere kostnadene tilknyttet oppdrag som omhandler innsamling av vitenskapelig data ved



(a) (b)

Figur 1: Lysrefleksjon over en jevn vannoverflate (a) og over en ujevn vannoverflate (b).

Kilde: [Asgarimehr, 2020]

bruk av GNSS-R. De lavere utvikling- og brukskostnadede baseres blant annet på at mikrosatellittene er mindre og lettere å utvikle enn tradisjonelle satellitter, kombinert med muligheten for å utplassere flere satellitter på samme rakettoppskyting [Clarizia et al., 2016].

I denne artikkelen presenteres en ny anvendelse av rombåren GNSS-R. Først vil utnyttelsen av GNSS-R målinger brukes til å overvåke variasjoner i jordfuktighet og detektere flom over to utvalgte landområder. Denne bruken av GNSS-R er betraktet så nyskapende og viktig at Den europeiske romfartsorganisasjonen (ESA) i mars 2021 finansierte HydroGNSS, et oppdrag basert på samme teknologi, verdt 30 millioner euro. Oppdraget, som vil levere målinger av viktige hydrologiske parametere som blant annet jordfuktighet, vil dermed komplementere eksisterende ESA oppdrag som SMOS og Biomass, samt NASA sitt SMAP [Unwin et al., 2021]. I neste seksjon følger en beskrivelse av datasettene brukt i denne studien. I tillegg til GNSS-R målinger inneholder de målinger fra to andre datakilder brukt for validering.

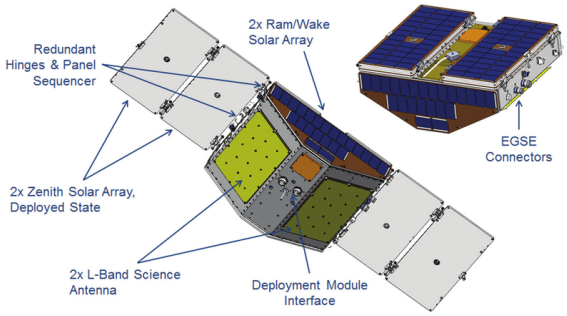
2 Data

En oversikt over de ulike datasettene som er brukt i denne studien, samt relevante parametere, oppløsning, gjenbesøkelsestid og ventetiden fra målinger er gjennomført til data er tilgjengelig for offentligheten er presentert i Tabell 1 senere i seksjonen.

2.1 CYGNSS

CYGNSS-konstellasjonen, som er utviklet med et budsjett på omtrent 150 millioner amerikanske dollar, opererer som GNSS-R mottakere ∼500km over jordoverflaten og består av åtte ∼25 kg lettvekts mikrosatelliter [Ruf et al., 2018]. Disse er omtrent 1.6 meter lange og er illustrert i Figur 2. I motsetning til tyngre, tradisjonelle satellitter som går i bane over polene er CYGNSS-satellittene designet til å benytte seg av en banehelling på 35 grader fra ekvator. Dette fører til at CYGNSS kan tilby målinger med en romlig oppløsning på 25x25 km og en gjennomsnittlig gjenbesøkelsestid på 7.2 timer mellom ± 40 breddegrader og ± 180 lengdegrader. Hver satellitt innehar et Delay Doppler Mapping Instrument (DDMI) ansvarlig for å generere relevant data fra målingene gjort av satellitten med en frekvens på 2Hz. Et DDMI er kapabelt til å motta fire reflekterte signaler samtidig, som på tvers av CYGNSS-konstellasjonen resulterer i at 32 målinger blir gjort hvert halve sekund.

CYGNSS produserer fire ulike dataprodukter som alle er offentlig tilgjengelig. Produktene, som er referert til som nivå 0 til nivå 3 data, er publisert gjennom NASA's Physical Oceanography Data Active Archive Center (PO.DAAC) med en maksimal forsinkelse fra gjennomførte målinger til tilgjengelighet på seks dager. Data på nivå 1, som benyttes i denne studien, inkluderer kalibrerte DDMer basert på den målte signalstyrken, konvertert fra rådata samlet inn av CYGNSS-satellitter (nivå 0). I tillegg til DDMer inneholder også dataproduktet metadata som er brukt i konverteringen fra nivå 0



Kilde: [Ruf et al., 2016]

Figur 2: Illustrasjon av en CYGNSS-satellitt

til nivå 1 data. Nivå 2 data og dets representasjon i et rutenett levert i nivå 3 data, er de resterende dataproduktene produsert av CYGNSS. Disse inneholder geofysiske parametere som blant annet vindhastighet og konsentrasjon av mikroplastikk [Ruf et al., 2016]

2.2 SMAP

Soil Moisture Active Passive (SMAP) er et romoppdrag utført av NASA med mål om å måle jordfuktighet over hele verden. Oppdraget ble startet i slutten av 2015 og er fortsatt i drift ved dags dato. Det består av én satellitt som ble designet for å utnytte både aktive og passive målemetoder, med den hensikt å utnytte det beste av hver metode. Det er derimot kun den passive metoden som fungerer per dags dato. Uansett, SMAP gjør det mulig å produsere jordfuktighets-estimater med en tidsmessig oppløsning på to dager ved polene og tre dager ved ekvator, og romlig oppløsning på 36 km. Budsjettet til SMAP er rett i underkant av én milliard amerikanske dollar [Entekhabi et al., 2014], og satellitten er vesentlig større enn de benyttede CYGNSS-satellittene. SMAP satellitten er illustrert i Figur 3.

En sammenlikning av oppløsningen til CYGNSS og SMAP er vist i Tabell 1, hvor CYGNSS er vist (a) og SMAP er vist i (b).

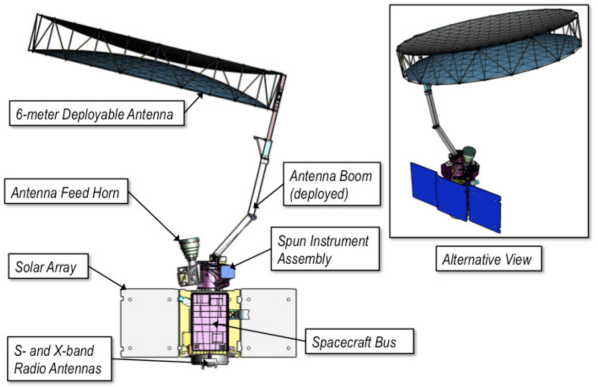
2.3 ERA5

The European Centre for Medium-Range Weather Forecasts (ECMWF) bruker retrospektiv analyse for å produsere et dataset bestående av informasjon om den globale atmosfæren, landoverflater og havbøler fra 1950 til dags dato. Dette datasettet skiller seg fra CYGNSS og SMAP ved at rombårene observeringer ikke er den eneste kilden til deres dataset. I stedet brukes en kombinasjon av direkte observasjoner og globale modeller til å produsere et datasett som blant annet inneholder globale estimater av bakkens fuktighet i de øverste fem cm av overflaten. Den romlige oppløsningen er 0.25° x 0.25° og den tidsmessige oppløsningen en en time, men gitt at det ikke utelukkende er et romoppdrag opereres det ikke med gjenbesøkelsestid som vist i Tabell 1.

3 Metode

3.1 Estimering av jordfuktighet

En fremgangsmåte for å oppnå relevant informasjon relatert til jordfuktighet er å analysere overflaterefleksjon. Ved å anta koherent refleksjon når GNSS-signaler reflekteres av jordens overflate kan den koherente komponenten i uttrykket for reflektert signalstyrke beskrives basert på målinger utledet fra satellittenes DDMer [Rajabi et al., 2020]. Styrken på reflekterte signaler er beregnet basert på verdier

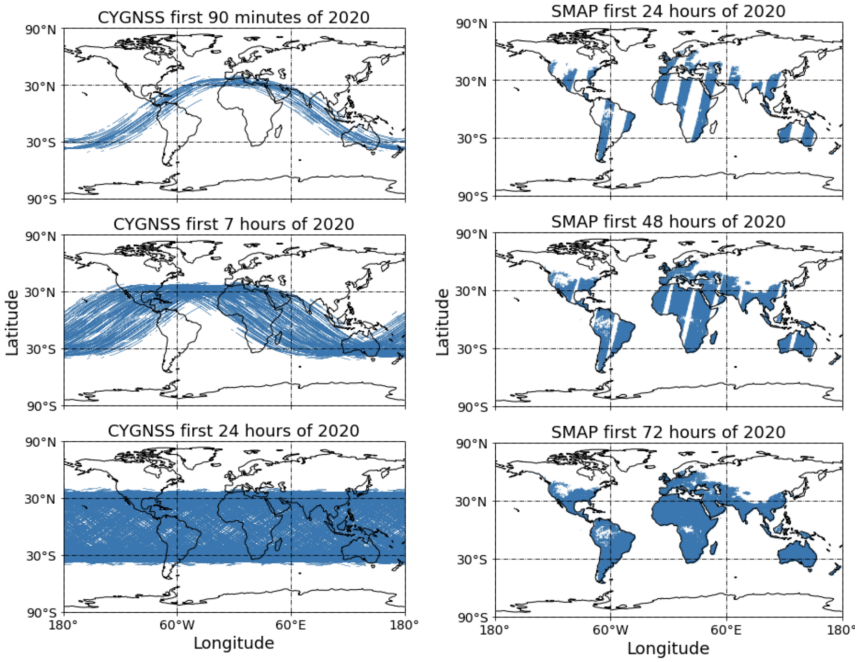


Kilde: [Entekhabi et al., 2014]

Figur 3: Illustrasjon av en SMAP satellitt

Tabell 1: En oversikt over de ulike datasettene som er brukt i denne studien, samt relevante parametere, oppløsning, gjenbesøkelsestid og ventetiden fra målinger er gjennomført til data er tilgjengelig for offentligheten

	CYGNSS (nivå 1)	**SMAP (nivå 3)**	**ERA5**
Romlig dekningsgrad	40° N/S	Global	Global
Romlig oppløsning	25km x 25km	36km x 36km	0.25° x 0.25°
Tidsmessig dekningsgrad	01.08.2018 - d.d.	31.03.2015 - d.d.	1979 - d.d.
Gjenbesøkelsestid	7.2 timer	2-3 dager	N/A
Dataforsinkelse	6 dager	50 timer	5 dager
Parameter i fokus	SNR	Jordfuktighet (SM)	Jordfuktighet (SM)



(a) (b)

Figur 4: CYGNSS sin romlige dekningsgrad i løpet av de første 24 timene i januar 2020 (a) og SMAP romlig dekningsgrad i løpet av de første tre dagene av januar 2020 (b)

som representerer forholdet mellom signal og støy (SNR) tilhørende DDMene som er tilgjengelig i CYGNSS nivå 1 data. Ligning 1 beskriver hvordan overflaterefleksjon målt i desibel (dB) er utregnet basert på SNR verdier og andre relevante variabler. Disse representerer signalstyrken på transmitterte GNSS-signaler (P_R^t), avstandene mellom henholdsvis GNSS- og CYGNSS-satellitter og refleksjonspunktet på jordoverflaten (d_{ts} og d_{sr}) og bølgelengden på GNSS-signalene som transmitteres (0.19m) med en frekvens innenfor L-båndet. Som også andre studier har påpekt, er SNR kun proporsjonal med P_R^t, da SNR også er påvirket av andre faktorer som systemstøy [Chew et al., 2018]. Da enkelte variables i CYGNSS nivå 1 dataproduktet er oppgitt i ulike måleenheter, ble disse konvertert til dB for å enklere kunne benyttes i Ligning 1.

$$SR\ SNR - P_R^t - G^r - G^t - 20\log(0.19) + 20\log(d_{ts} + d_{sr}) + 20\log(4\pi) \tag{1}$$

Tilstedeværelsen av fuktighet i jorden er med på å endre dens elektromagnetiske egenskaper, noe som fører til endringer i landoverflatens respons på innkommende GNSS-signaler. Denne responsen kan representeres ved en parameter for overflaterefleksjon. På denne måten kan GNSS-R målinger av overflaterefleksjon tilføre verdifull informasjon om mengden vann som befinner seg i jorden. I forskning relatert til bruken av CYGNSS-målinger for å oppnå estimater på jordfuktighet er det nødvendig med enkelte underliggende antagelser for at parametere, som den presentert i Ligning 1, skal kunne knyttes til jordfuktighet. Det innebærer blant annet at refleksjonsområdet der målingen er gjort antas å ikke påvirkes bemerkelsesverdig av jordens overflateøffhet. Den viktigste antagelsen som underbygger sammenhengen mellom overflaterefleksjon og jordfuktighet er derimot at endringer i målt overflaterefleksjon utelukkende er et resultat av endringer i andelen vann i jorden. Dersom dette viser seg å stemme vil monitoring av overflaterefleksjon kunne bidra til generering av data og informasjon av høyere presisjon og oppløsning for flere bruksområder som benytter målinger av jordfuktighet.

3.2 Detektering av flom

Flom kan forekomme som et resultat av stor nedbør, snø og issmelting og andre årsaker. Felles for alle årsakene er at vannstanden tar seg over sitt vanlige nivå, som kan skape store farer og ødeleggelser for nærliggende bebyggelse og infrastruktur. Ved forekomst av flom vil den økte mengden vann føre til muligheter for økt overflaterefleksjon. Hvorvidt CYGNSS evner å fange opp slike hendelser kan analyseres i tidsserieanalyser.

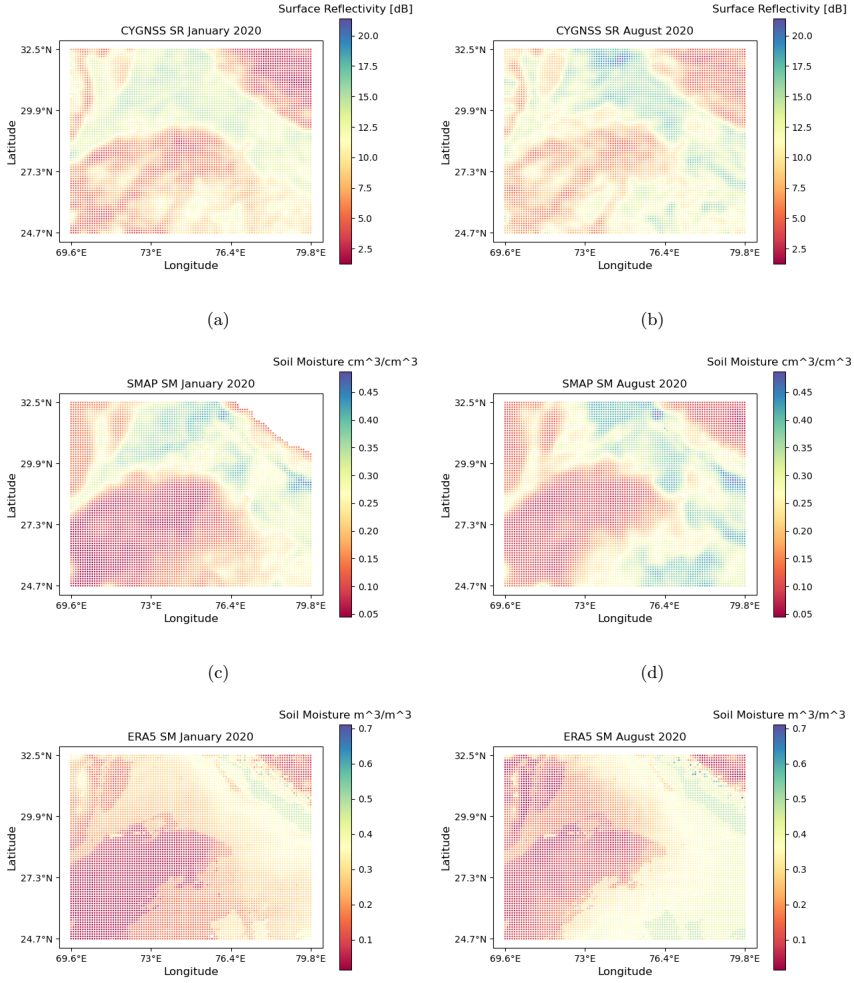
I både januar 2020 og mai 2021 forårsaket kraftig nedbør flom i Sistan- og Baluchestan-provinsen i Iran. Dette området består av flere fjellkjeder med sandlagre mellom seg som akkumulerer vann fra de ovenforliggende områdene. Dette gjør de lavtliggende områdene utsatt for flom. En tidsserieanalyse i dette området vil bli gjort for å evaluere om CYGNSS fanger opp slike naturkatastrofer. Her vil den gjennomsnittlige overflaterefleksjonen bli regnet ut for hver dag fra begynnelsen av 2019 til slutten av 2021. Disse verdiene blir sammenliknet med jordfuktighetsverdier tilgjengelig i SMAP sitt nivå 3 dataprodukt for å evaluere resultatene. I tillegg vil resultatene evalueres basert på om flommen gir et avtrykk på CYGNSS-tidsserien i form av økt overflaterefleksjon.

4 Resultater og diskusjon

4.1 En sammenligning av overflaterefleksjon og jordfuktighet i India

Overflaterefleksjon utledet fra CYGNSS-målinger vil bli sammenliknet med jordfuktighetsestimater produsert av både SMAP og ERA5 over et område som dekker nord-India og Pakistan i januar og august 2020. Alle målingene har blitt konvertert til en oppløsning på 0.1° x 0.1° ved å ta gjennomsnittet av alle målingene i hver celle for hver respektive måned. Resultatet presenteres i Figur 5, hvor den venstre kolonnen viser målingene i januar 2020 og den høyre kolonnen viser det samme for august. Figuren illustrerer en ulempe ved bruk av SMAP når det gjelder romlig dekningsgrad. Det nordøstlige hjørnet av Figur 5 (c) har ikke blitt dekket av SMAP i januar 2020 og demonstrerer den forbedrede dekningsgraden til CYGNSS. Videre kommer det fram at ERA5 har glatte overganger fra tørre til våte områder, der det forekommer større variasjoner i CYGNSS-dataen. Dette kan tyde på at den romlige oppløsningen til CYGNSS er bedre og derfor fanger opp variasjoner som ikke plukkes opp av ERA5.

Tabell 2 viser gjennomsnittlig overflatereflektivitet for CYGNSS, samt gjennomsnittlig jordfuktighet for SMAP og ERA5 i januar og august 2020. I tillegg presenteres den relative endringen til alle de



(a) (b)

(c) (d)

(e) (f)

Figur 5: Gjennomsnittlig overflatereflektivitet for CYGNSS i januar 2020 (a) og august 2020 (b), gjennomsnittlig jordfuktighet for SMAP i januar 2020 (c) og august 2020 (d) og gjennomsnittlig jordfuktighet for ERA5 i januar 2020 (e) og august 2020 (f)

Tabell 2: Gjennomsnittlig overflatereflektivitet for januar og august 2020 for CYGNSS og gjennomsnittlig jordfuktighet i de samme månedene for SMAP og ERA5. Verdiene er kalkulert over nord-India og Pakistan.

	Januar	**August**	**Relativ endring**
CYGNSS	9.36 dB	10.5 dB	+12.2 %
SMAP	0.21 cm^3/cm^3	0.23 cm^3/cm^3	+9.5 %
ERA5	0.21 m^3/m^3	0.25 m^3/m^3	+19.1 %

Tabell 3: Korrelasjon mellom CYGNSS og SMAP og ERA5 for januar og august i 2020 i nord-India og Pakistan.

	Januar		**August**		**Tidsmessig differanse**	
	SMAP	ERA5	SMAP	ERA5	SMAP	ERA5
CYGNSS	0.777	0.368	0.770	0.337	0.686	0.463

tre datakildene. Tabell 3 viser den romlige korrelasjonen mellom de tre datakildene for de to månedene. Her kommer det tydelig frem at CYGNSS og SMAP korrelerer i mye høyere grad enn det CYGNSS gjør med ERA5. Dette er et resultat av de globale numeriske metodene som benyttes i ERA5, i motsetning til de direkte observasjonene hos SMAP og CYGNSS.

4.2 Analyse av to nylige flommer i Iran

For hver 0.5° x 0.5° celle i det valgte området i Iran ble det produsert tidsserier fra 2019 til og med 2021. Figur 7 (a) viser en oversikt over korrelasjonen mellom CYGNSS og SMAP for hver enkelt celle. Mørkheten i grønnfargen til hver celle angir hvor høy korrelasjon cellen besitter. Figur 7 (b) viser et satellittbilde av det samme området. I begge figurene representerer den blå boksen området som opplevde flom i januar 2020 og mai 2021. Figur 7 (c) viser tidsserien til den blå cellen, med de to flomperiodene markert i grått. Videre viser henholdsvis Figur 7 (d) og (e) tidsseriene til den første og den andre flomen. Alle disse tidsseriene har en rød linje som viser CYGNSS overflatereflektivitet og en blå linje som viser SMAP jordfuktighet. Korrelasjonen mellom de to linjene står oppført i tittelen til hver figur.

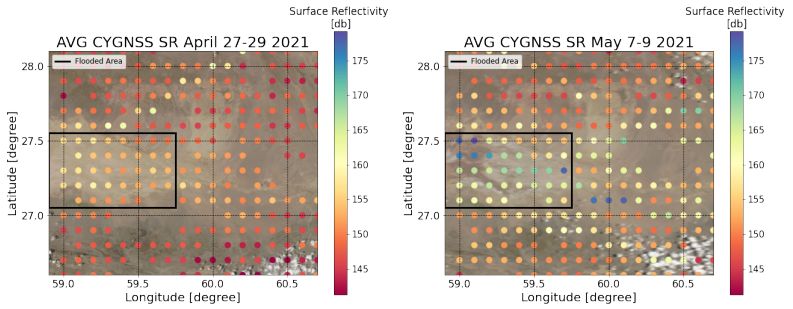
Under begge flommene viser Figur 7 (c) en markant økning i både overflatereffleksjon og jordfuktighet. Det samme mønsteret er synlig i Figur 7 (d) og (e), hvor SMAP sine verdier for jordfuktighet øker raskere enn CYGNSS sin overflatereffleksjon.

4.3 Konklusjon

Vi har presentert en ny metode for å produsere kvasi-globale estimater av jordfuktighet med høy romlig og tidsmessig oppløsning ved bruk at Global Navigation Satellite System Reflektometri (GNSS-R). NASA sitt nye romoppdrag Cyclone GNSS (CYGNSS) tar i bruk denne teknikken for å produsere Delay Doppler Maps (DDMs) med Signal to Noise Ratio (SNR) verdier av GNSS-signaler. Disse kan brukes til å regne ut overflatereffleksjon og videre estimere jordfuktighet.

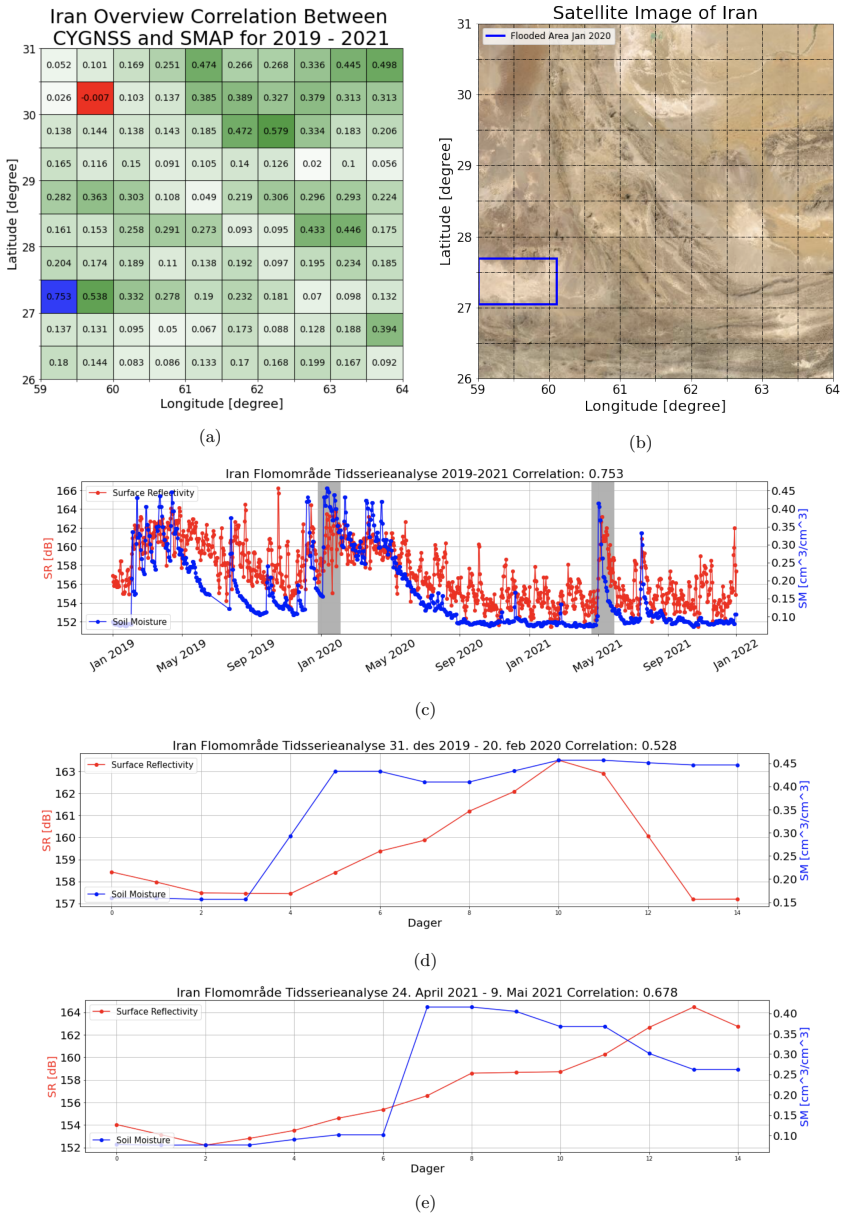
En studie i nord-India og Pakistan viste at CYGNSS overflatereklektivitet korrelerer sterkt med Soil Mostire Active PAssive (SMAP) sine estimater av jordfuktighet. Deler av differansene kan komme av at CYGNSS demonstrerer høyere romlig og tidsmessig oppløsning, noe som kan gjøre at CYGNSS fanger opp mindre variasjoner i jordfuktighet enn SMAP er kapabel til. ERA5 korrelerte dårlig med CYGNSS sammenliknet med SMAP i samme område.

Videre ble analyser av tidsserier i Iran gjennomført for å evaluere muligheten til å fange opp flom ved bruk av CYGNSS-målinger. Det analyserte området opplevde to flommer i henholdsvis januar 2020 og mai 2021. I disse periodene fremhevet tidsseriene signifikante økninger i målt overflatereflektivitet. Det er derfor rimelig å anta at resultatet bekrefter muligheten til å bruke CYGNSS-målinger i fremtidige prognoser og deteksjon av flom, og det til reduserte kostnader sammenlignet med eksisterende metoder.



(a) (b)

Figur 6: CYGNSS overflatereflektsjon før flommen i Iran 2021 (a) og under flommen i Iran i 2021 (b)



Figur 7: Oversikt over korrelasjon mellom CYGNSS og SMAP i hver 0.5° x 0.5° celle (a). Satellittbilde over det analyserte området i Iran (b). Tidsserie for den blå ruten (c). Tidsserie over den første flomen (d) og den andre flomen (e).

Bibliografi

Milad Asgarimehr. Spaceborne gnss reflectometry: Remote sensing of ocean and atmosphere. 2020.

C. Chew, John Reager, and Eric Small. Cygnss data map flood inundation during the 2017 atlantic hurricane season. *Scientific Reports*, 8, 06 2018. doi: 10.1038/s41598-018-27673-x.

Maria Paola Clarizia, Christopher Ruf, Paolo Cipollini, and Cinzia Zuffada. First spaceborne observation of sea surface height using gps-reflectometry. *Geophysical Research Letters*, 43(2):767–774, 2016. doi: https://doi.org/10.1002/2015GL066624. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL066624`.

M Craig Dobson and Fawwaz T Ulaby. Active microwave soil moisture research. *IEEE Transactions on Geoscience and Remote Sensing*, (1):23–36, 1986.

Alejandro Egido Egido. *GNSS reflectometry for land remote sensing applications*. PhD thesis, Universitat Politècnica de Catalunya Barcelona, Spain, 2014.

Dara Entekhabi, Simon H. Yueh, Peggy E. O'neill, Kent H. Kellogg, Angela Morin Allen, Rajat Bindlish, Molly E. Brown, Steven Tsz K. Chan, Andreas Colliander, Wade T. Crow, Narendra Narayan Das, Gabrielle De Lannoy, Roy Scott Dunbar, Wendy N. Edelstein, Jared K. Entin, Vanessa M. Escobar, Shawn D. Goodman, Thomas J. Jackson, Benhan Jai, Joel T. Johnson, Edward J. Kim, Seungbum Kim, John S. Kimball, Randal D. Koster, Amanda Leon, Kyle McDonald, Mahta Moghaddam, Priscilla N. Mohammed, Susan Moran, Eni G. Njoku, Jeffrey R. Piepmeier, Rolf H. Reichle, Francois Rogez, Jiancheng Shi, Michael W. Spencer, Sam W. Thurman, Leung Tsang, Jakob J. van Zyl, Barry H. Weiss, and Richard D. West. Smap handbook–soil moisture active passive: Mapping soil moisture and freeze/thaw from space. 2014.

ADNAN Kavak, WJ Vogel, and Guanghan Xu. Using gps to measure ground complex permittivity. *Electronics Letters*, 34(3):254–255, 1998.

Yann Kerr. Soil moisture from space : Where are we? *Hydrogeology Journal*, 15:117–120, 02 2007. doi: 10.1007/s10040-006-0095-3.

Manuel Martin-Neira. A passive reflectometry and interferometry system (paris): Application to ocean altimetry. *ESA Journal*, 17:331–355, 01 1993.

Clement Ogaja. *Applied GPS for engineers and project managers*. 08 2011. ISBN 978-0-7844-1150-6. doi: 10.1061/9780784411506.

Mahmoud Rajabi, Hossein Nahavandchi, and Mostafa Hoseini. Evaluation of cygnss observations for flood detection and mapping during sistan and baluchestan torrential rain in 2020. *Water*, 12(7), 2020. ISSN 2073-4441. doi: 10.3390/w12072047. URL `https://www.mdpi.com/2073-4441/12/7/2047`.

Christopher Ruf, Derek Posselt, Sharanya Majumdar, Scott Gleason, M.P. Clarizia, Derek Starkenburg, Damen Provost, Valery Zavorotny, John Murray, Stephen Musko, Zorana Jelenak, Paul Chang, and Mary Morris. *CYGNSS Handbook*. 04 2016. ISBN 978-1-60785-380-0.

Christopher S Ruf, Clara Chew, Timothy Lang, Mary G Morris, Kyle Nave, Aaron Ridley, and Rajeswari Balasubramaniam. A new paradigm in earth environmental monitoring with the cygnss small satellite constellation. *Scientific reports*, 8(1):1–13, 2018.

Peter J.G. Teunissen and Oliver Montenbruck. *Springer Handbook of Global Navigation Satellite Systems*. Springer International Publishing AG, 2017. ISBN 978-3-319-42926-7.

Kevin E Trenberth. The impact of climate change and variability on heavy precipitation, floods, and droughts. *Encyclopedia of hydrological sciences*, 17, 2005.

Martin J. Unwin, Nazzareno Pierdicca, Estel Cardellach, Kimmo Rautiainen, Giuseppe Foti, Paul Blunt, Leila Guerriero, Emanuele Santi, and Michel Tossaint. An introduction to the hydrognss gnss reflectometry remote sensing mission. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6987–6999, 2021. doi: 10.1109/JSTARS.2021.3089550.