UNIVERSITY OF KWAZULU-NATAL

MAPPING RAINFALL VARIABILITY IN SOUTH AFRICA'S COMMERCIAL FOREST REGIONS USING COMPETENT ENSEMBLE STATISTICAL INTERPOLATION TECHNIQUES

by

Sarisha Rajen Ramjeawon

216005302

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science (Environmental Science)

In the Discipline of Geography

In the School of Agriculture, Earth and Environmental Sciences

Supervisor: Dr. Kabir Peerbhay

Co-supervisor: Dr. Romano Lottering

February 2023

Declaration 1

The work described in this paper was completed at the University of KwaZulu-Natal, Pietermaritzburg, from March 2020 to February 2023, under the supervision of Dr. K. Y. Peerbhay.

The research represented in this document is original work by the author and has not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others, it is duly acknowledged in the text.

2023-02-09

Signature (Student) Date

Sarisha Rajen Ramjeawon

2023-02-09

Signature (Supervisor) Date

Dr Kabir Yunus Peerbhay

Declaration 2

I, Sarisha Rajen Ramjeawon, declare that;

1. The research reported in this thesis, except where otherwise indicated, is my original work.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other person's data, pictures or graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other person's writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    a. Their words have been re-written, and the general information attributed to them has been referenced.
    b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and reference sections.

Signed  ……..

Dedication

I dedicate this dissertation to my parents, Raj and Asha Ramjeawon, who have always anchored me through every opportunity and obstacle that life has thrown at me, and encouraged me through the highs and lows while being two of my biggest fans! I also dedicate this work to my husband, Nirvikar Bundhoo, for being patient with me as I burnt many midnight oils. He has always supported my passion for science and encouraged me to reach my goals. To my fur babies Gizmo Ramjeawon and Anshi Bundhoo, your cold noses and warm hearts provided me with the enthusiam to continue pursuing another degree. You both sensed the days I needed a friend to hear me and a paw to hold. This work is further dedicated to my sister and brother-in-law, Herusha and Keshav Singh for their unwavering support and limitless motivation over the years. My love and gratitude for the above-mentioned precious souls can never be quantified. Thank you all for believing in me more than I believed in myself, and for encouraging the passion that burned in me to make this dissertation a reality.

"In life, if you ever want to be something, win something, or get something, then always listen to your heart. But if you don't get a signal from your heart, then close your eyes and say your mom and dad's names, then watch, you will achieve every goal, every obstacle will become easy, and the victory will be yours... only yours..."

- Kabhi Khushi Kabhie Gham

Acknowledgements

This entire academic journey would have not been possible without the succour of many instrumental people.

To Dr Kabir Yunus Peerbhay, my supervisor and Dr Romano Lottering, my co-supervisor, without you, this project would not have materialised. Your motivation, assistance and timeless guidance has plunged me into reaching the end of my dissertation.

To the University of KwaZulu-Natal and its enthusiastic staff who have kept me guided throughout and assisted wherever possible to make this a smooth-sailing journey as much as possible.

To my parents, the reason for why I am the person I am today. Your sacrifices and unconditional love has been my source of strength through every curveball and blessing that life has offered me.

To Dr Kabir Yunus Peerbhay, the Institute for Commercial Forestry Research (ICFR), and the South African Weather Service (SAWS) for assisting me with collecting and providing field data for this dissertation.

Finally, a special thanks is given to the National Research Foundation (NRF) for funding my Master's degree and providing financial relief to pursue and complete this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

APPENDICES

ACRONYMS

ADTree – Alternating Decision Tree

Alt - Altitude

ANCOVA – Analysis of Covariance

ANOVA – Analysis of Variance

ANUDEM – Australian National University's Digital Elevation Model

CFP – Citrus Flatid Planthopper

CHIRPS – Climate Hazards Group InfraRed Precipitation with Station

CO_2 – Carbon Dioxide

DEM – Digital Elevation Model

DR – Detection Rate

Elev - Elevation

EOS – Earth Observing Systems

EVI – Enhanced Vegetation Index

FPR – False Positive Rate

GIS – Geographic Information System

GLM – General Linear Model

GPCC – Global Precipitation Climatology Centre

GSR – Greater Sydney Region

ICFR – Institute for Commercial Forestry Research

ID – Identity

IDW – Inverse Distance Weighting

IPCC – Intergovernmental Panel on Climate Change

ITCZ – Inter-Tropical Convergence Zone

KZN - KwaZulu-Natal

LDA – Linear Discriminant Analysis

LSM – Landslide Susceptibility Mapping

MAE – Mean Absolute Error

MAP – Mean Annual Precipitation

MIR – Mid-Infrared

MODIS – Moderate Resolution Imaging Spectroradiometer

MRI – Magnetic Resonance Imaging

NDWI – Normalised Difference Water Index

NDVI – Normalised Difference Vegetation Index

NIR – Near-Infrared

O_2 – Oxygen

OA – Overall Accuracy

OBIA – Object-Based Image Analysis

OK – Ordinary Kriging

OOB – Out-Of-Bag

PCA – Principal Component Analysis

PR – Precipitation Radar

RF – Random Forest

RK – Regression Kriging

RMSE – Root Mean Square Error

RSS – Sum squares of Residuals

RUE – Rainfall Use Efficiency

SA – South Africa

SK – Simple Kriging

TRMM – Tropical Rainfall Measuring Mission

TSS - Total Sum of Squares

WRM – Water Resource Management

Units of Measurement

° – Degrees

% – Percent

°C – Degree Celsius

m – Metre

mm – Millimetre

m^2 – Square metre

m^3 – Cubic metre

nm^3 – Cubic nanometres

ABSTRACT

Rainfall mapping forms an integral part of water resource management (WRM), especially in a water-scare country like South Africa (SA). Further to water scarcity are water-related natural disasters such as floods and droughts which are the effects of climate change. Monitoring rainfall in SA's commercial forests was one of the many disciplines that utilises the assessment of rainfall distributions and mapping. Rainfall can be monitored over different time periods over different areas simultaneously, making it successful at comparing rainfall patterns between two or more regions at any given time. Over the years, monitoring rainfall using ground-based data became more tedious and expensive, even though it is still a popular and successful method. Hence, remote sensing and Geographic Information Systems (GIS) became a more affordable and reasonable alternative method for rainfall monitoring that brought about more reliant information that made the management and understanding of commercial forests much easier and more feasible, covering larger spatial areas.

Rain stations together with remotely sensed data, such as Tropical Rainfall Measuring Mission (TRMM) and Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data are two examples that serve as reliable data for detecting, measuring and mapping rainfall around the world. Rainfall stations also gather rainfall data and are placed in different locations which offer precise measurements, however, they do offer some limitations. These limitations relate to the ability of them being located widespread but not all locations are suitable to collect live-weather data. Rainfall stations also become faulty and do not log events accurately with gaps and missing data usually hampering analytics tasks. The alternate to this limitation is by using recently acquired satellite rainfall data; however the frequency and availability of this is challenging. Therefore interpolating current data that can be used to reliably assess current distribution patterns could be a viable option, even for future rainfall variability predictions.

Traditional interpolation techniques include Inverse Distance Weighting (IDW), kriging and spline; however, do not produce as high accuracies as the ones used in this study. Recent advances in this domain produce high accuracy results. These include the General Linear Model (GLM), Random Forests (RF) regression learning and meta-ensemble regression methods. This study used monthly live-weather data from one hundred and fifteen (115) rainfall stations located in and around commercial forest plantations located in five (5) South African provinces, namely Limpopo, Mpumalanga, KwaZulu-Natal (KZN), Eastern Cape and Western Cape. A clear comparison between the GLM, RF and meta-ensemble was done for

data that was available between July 2018 and July 2019. This study explored these different interpolation techniques by being applied in two different methods with Moderate Resolution Imaging Spectroradiometer (MODIS) ancillary data. The first method was the comparison between the GLM and RF techniques and the second method was the action of stacking the algorithms to demonstrate a meta-ensemble interpolation technique.

The first method demonstrated the GLM and RF being used separately to conduct a constructive comparison and analysis. The findings showed that the GLM had an R^2 value of 0.71 and Root Mean Square Error (RMSE) of 0.1272, while the RF model had an R^2 value of 0.79 and RMSE of 0.0165. This implied that RF was a more successful interpolator for rainfall mapping over the GLM.

The second method of the study used the stacking of the GLM and RF algorithms to demonstrate a meta-ensemble algorithms with a combination of ancillary satellite data to detect and map rainfall patterns, over the same time period. The attribute data that was included were altitude, elevation, Enhanced Vegetation Index (EVI), Normalised Difference Vegetation Index (NDVI) and temperature. This stacking method showed an R^2 value of 0.86 and an RMSE of 0.0453. This implies that the meta-ensemble is a suitable interpolator which combines the strengths of each algorithm to make accurate estimations. However, considering all the results found within chapters two (2) and three (3) combined, the stacking of algorithms to demonstrated a meta-ensemble interpolation technique is the best single interpolator to map rainfall using point information, rather than the GLM and RF techniques.

These two methods generated reliable data that can be used in the management of available surface water and the impacts, trends and shifts of rainfall due to climate change. Hence, there is a need that these interpolation techniques, together with ancillary information such as remotely-sensed variables, to provide reliable results that are able to investigate rainfall variability, especially in countries like SA who need to manage their limited water resources efficiently.

Overall, the results from this study are in support of (i) mapping rainfall patterns in commercial forest regions of Limpopo, Mpumalanga, KZN, Eastern Cape and Western Cape, (ii) the GLM and RF interpolation techniques have high accuracy levels that can be distinctively compared, (iii) meta-ensemble is a successful stacking interpolation technique that can determine good accuracy of rainfall station data and (iv) attribute data together with the meta-ensemble

interpolation technique shows more variations in predicted rainfall maps, however, some of the best rainfall variability mapping and highest accuracy was generated from the meta-ensemble interpolator. The study determined that as much as the meta-ensemble technique is relatively new as compared to the other traditional algorithms, it does not necessarily produce the highest RMSE value; however, it is the most accurate. Therefore, it can be concluded that rainfall station data together with MODIS ancillary data produce the best results using the meta-ensemble interpolation technique.

CHAPTER ONE: Introduction

1.1 Background

Rainfall distributions is an essential part of understanding rainfall variability, such that it allows a more in-depth study of weather-related phenomena and climatic changes. These changes lead to widespread environmental and social impacts and affect various agricultural sectors, such as commercial forestry (Masson-Delmotte *et al.,* 2018). By understanding the spatial variations of rainfall and making deductions from current forecasts, rainfall data is crucial in addressing concerns on planting windows, forest growth related trends and land management matters. Hence, studies like the above-mentioned must be done to help provide and initiate management strategies for water in commercial forests, especially in a country like SA that is water-scare and highly dependent on the commercial forestry sector as a whole for economic growth. However, Rogers *et al.* (2005) stated that water management is a more significant issue than being a 'scarce resource'. As part of the twenty-first century and the climatic changes occurring, national problems lead to a greater global climate crisis (Adger and Agnew, 2004). Hence, water management must be done at a national level first and thereafter expand worldwide. It has been mentioned by Pimentel *et al.* (2004), that once ecosystems are impacted, this results in after effects on the timber and food production industries, which in turn drastically affect our economy.

Different types of rainfall have different climatological drivers that affect climate change. Some of these types include relief rainfall that is explained by Reed *et al.* (2009) as topography being an influential factor that contributes to the types of woodlands and grasslands, such that it grows according to the amount of rainfall on regions that have different reliefs. Another type of rainfall is frontal rain that occurs during the existence of a cyclone that consists of a cold and warm front (Zinevich *et al.,* 2009). They further state that space and time can contribute to large scale frontal rain in different regions that can be studied when the cold air merges with the warm air. Tropical rainfall is another type of rainfall explained by Kang *et al.* (2018) as rainfall influenced due to oceanic dynamics in the inter-tropical convergence zone (ITCZ). These different types of rainfalls can be modelled to indicate storm cell patterns that occur over different places at different times, and can move over land. However, modelling rainfall has limitations to understanding climate change, such that limitations exist with rainfall data gathered from rainfall stations, but are still valuable for explaining these different types of rainfall.

One of the most advanced ways to manage water resources is to study precipitation's spatial and temporal variations (Harrington *et al.*, 2002). According to Pandey and Pandey (2010), monitoring on a spatial and temporal scale can be challenging. For example, monitoring rainfall station data is complex at a national level because they are widespread, which makes it difficult to control all at a given time. Furthermore, controlling these live-weather data systems is costly (Lakshmi, 2004), while others may sometimes be inaccessible (Bayat *et al.*, 2019). Sometimes, rainfall stations can get vandalised, generating errors in raw data that can impact the accuracy of subsequent analyses.

Remote sensing acts as a tool to access inaccessible areas and provides data that can be studied for weather forecasting and natural disaster predictions (Munawar *et al.*, 2022). Although this is an advantage, its disadvantage includes the limitation of not measuring rainfall at a high resolution. Naumann *et al.* (2012) explored the high resolution TRMM data by conducting a study in Africa to assess drought events. This precipitation data can produce a high resolution of 0.25° x 0.25° and have multi-satellite estimates that are computed using re-analysed precipitation data from the Global Precipitation Climatology Centre (GPCC). They have a 3-hourly time scale (Nerini *et al.*, 2015) TRMM data is advantageous such that it allows to investigate rainfall distribution as well as its frequency and intensity. However, TRMM data has limitations such that it can only detect between 50° north and south of the equator and cannot perceive through cloud cover (Zipser *et al.*, 2006). On the other hand, data is compatible with monitoring drought patterns, according to Zambrano *et al.* (2017), and has a resolution of 0.05° x 0.05° and a 6-hourly time scale (Wang *et al.*, 2021). This provides a reason that there certainly is a need to interpolate data by using a finer resolution satellite.

Goovaerts (2000) stated that an estimate could be made using surrounding data to obtain a value of unrecorded localities. To do this, spatial interpolation is needed to obtain estimates of inaccessible points using sample point values. Some traditional spatial interpolation techniques include inverse distance weighting (IDW), spline and kriging. Goovaerts (2000) states that these spatial interpolations assume the sample points found closest to the location already interpolated. The closer the points, the better the interpolated value (Goovaerts, 2000). One way to interpolate precipitation is traditionally using the ‘nearest neighbour’ method, which utilises the closest rainfall station to estimate surrounding areas with unobtainable values (Coulibaly and Becky, 2007).

Goovaerts (2000) investigated the Simple Kriging (SK) and Ordinary Kriging (OK) interpolators of 36 climatic stations in Portugal, covering an area of 5 000 km². Their study yielded high accuracy predictions for OK over linear regressions. A study conducted by Chen and Liu (2012) investigated the IDW interpolator by detecting rainfall in Taiwan and determined how it is distributed. They gathered data from 46 rainfall stations and concluded that drier seasons produce more accurate results than flood seasons. Their overall accuracy was 0.95, which showed that IDW is a good traditional interpolator. A more recent study by Zhang *et al.* (2018) investigated the spline interpolator in Florida to study rainfall variability. They concluded that the IDW and OK methods produced better results than the spline method.

However, this study demonstrates a statistical method that is superior to the traditional interpolators mentioned above, as they can produce higher spatially correlated data of unsampled areas with the best variabilities (Goovaerts, 2000). Furthermore, they can be combined with ancillary data such as elevation, altitude, terrain, vegetation abundance, and temperature to produce more significant results.

Apart from the traditional interpolation techniques mentioned, a more refined approach is required in this study to help determine changes in rainfall patterns. This is done using a more modern interpolator using the General Linear Model (GLM) approach, which can compute precipitation data and compare datasets (Chandler and Wheater, 2002). More importantly, it has provided information on weather-related crises related to flooding and droughts from rainfall patterns. Chandler and Wheater (2002) put the GLM to the test by studying the rainfall records of western Ireland to understand what influenced their historical flood events of the 90s. They deduced that the GLM could be applied in hydrological processes and assist with future predictions.

Further to the understanding of the relationship between GLMs and the hydrological process was a study conducted by Rong *et al*. (2020), who studied the Random Forest (RF) model to determine how rainfall caused landslides in China. They studied landslides because, just like floods and droughts in SA, the landslides in China output severe economic damages. Rong *et al.* (2020) determined Landslide Susceptibility Mapping (LSM) by using RF. They demonstrated that too much rainfall has long-term effects on landslides and that the RF is a suitable method for LSM and can make significant predictions for future management strategies. Like the study conducted by Rong *et al*. (2020), this study investigates the reliability of RF for mapping rainfall variability.

In contrast, Kanavos *et al.* (2001) used remotely-sensed data to make predictions of winter rainfall in weather forecasting, using a meta-ensemble technique (i.e. staking of algorithms) that produced high accuracies but not as high as for accuracies tested on other ground-truthed data. A few years later, Zhang *et al.* (2022) also tested the ensemble method to determine how accurately prediction values can be determined. Like this study, Zhang *et al.* (2022) used the stacking method to determine climate-related parameters. In addition, they also looked at appending bagging and boosting with their remotely sensed data. They deduce that the ensemble successfully predicts accuracies; however, gaps of effective combinations still exist to increase accuracy values. Zhang *et al.* (2022) further state that future studies must be advanced and explore more diverse algorithms. The RF and RK method was also proven to be reliable according to Gia Pham *et al.* (2019) when their precision for OK on soil pH mapping was lower by 1.81%. This study meets their suggestion by determining which interpolator is best for rainfall variability mapping, i.e., the GML, RF or stacking of algorithms in the meta-ensemble.

Accurate rainfall mapping and predictions feed into informed decision-making that enhances the resilience of agricultural systems. By studying these rainfall changes over different regions and periods, algorithms can be performed and analysed to determine how accurately they contribute to making a solid foundation for decision-making. This study utilised interpolation techniques, GLM, RF and meta-ensemble, to compare with each other to determine which interpolator is most successful at making accurate predictions. This research contributes to national climate change understanding by producing comparable and reliable rainfall maps. Furthermore, past research did not explore the stacking of the GLM, RF and meta-ensemble algorithms to determine rainfall variability in SA. Hence, this study is significant to the scientific industry as, further to the algorithms, ancillary satellite data was added to better understand rainfall found in the water-scare country of SA. Vegetation indices and altitude, elevation and temperature were combined with the meta-ensemble technique to provide a more versatile approach to understanding rainfall distribution and mapping.

Chapter two (2) focuses on using the GLM and RF regression algorithms separately to determine the effectiveness of the approach and their accuracies, while chapter three (3) combines the GLM, RF, and ancillary data to provide a meta-ensemble interpolator approach to detecting and mapping rainfall for the period between July 2018 and July 2019. This data was made available by the Institute for Commercial Forest Resources (ICFR) for this project

period and consisted of a network of live daily rainfall stations in and around the forest plantations of South Africa.

In order to address the above focuses the next section provides a briefing to the aims and objectives of this study and how this dissertation has been structured to meet those aims and objectives.

1.2 Aim and Objectives

The main aim of this study was to demonstrate the importance of detecting and mapping rainfall variability in Limpopo, Mpumalanga, KZN, Eastern Cape and Western Cape provinces of SA, from July 2018 to July 2019, using a combination of comparable GIS interpolation techniques. The main objectives were the following:

- To detect and map rainfall in commercial forest regions of SA using the GLM and RF interpolation techniques and MODIS ancillary data.
- To determine whether RF can produce a higher accuracy than the GLM and meta-ensemble techniques.
- To compare the GLM and RF accuracy levels for mapping rainfall in commercial forests.
- Establish whether the meta-ensemble interpolator can be combined with ancillary data to map rainfall variations accurately.

1.3 Outline of thesis

Four (4) chapters are used in this thesis, with Chapters Two (2) and Three (3) being the principal chapters, which are worthy of being publishable papers. These two chapters descriptively explain the study area, literature review and methodology, hence, they are not discussed in Chapter One (1) to avoid repetition of information.

Chapter Two (2) looks at the comparison between the GLM and RF interpolation techniques that were used separately to conduct a constructive analysis. It shows that the General Linear Model (GML) had an R^2 value of 0.71 and Root Mean Square Error (RMSE) of 0.1272, while the RF model had an R^2 value of 0.79 and RMSE of 0.0165. This implies that RF was a more successful interpolator for rainfall mapping than the GLM.

Chapter Three (3) used a stacking of the GLM, RF and meta-ensemble algorithms with a combination of ancillary data to detect and map rainfall patterns during the period between July 2018 and July 2019, to determine the influence of rainfall in commercial forests. This method showed an R^2 value of 0.86 and an RMSE of 0.0453. This implies that the meta-ensemble is a good interpolator. Still, considering the results of chapter 2, the RF is the best single interpolator to map rainfall, rather than the GLM technique.

Chapter Four (4) delivers a summary of the study. The aims and objectives of this research are descriptively explained to highlight the significant findings and the most suitable method for mapping rainfall in SA's commercial forests. This chapter also provided this study's limitations and recommendations for future research.

CHAPTER TWO: Mapping Rainfall in South Africa's Commercial Forests Using the General Linear Model (GLM) and Random Forest (RF) Interpolation Techniques

Abstract

Detecting rainfall within South Africa and mapping its variability is essential in decision-making and water resource management. This study utilized ground truth data to map rainfall across South Africa's commercial forestry regions. The regions focused on were the Western Cape, Eastern Cape, KwaZulu-Natal, Mpumalanga and Limpopo provinces. The General Linear Model (GLM) and Random Forests (RF) interpolation techniques were used separately and then combined to conduct a constructive analysis. This study showed that the GLM had an R^2 value of 0.71 and an RMSE of 0.1272, while the RF model had an R^2 value of 0.79 and an RMSE of 0.0165. This indicated that RF was a more successful interpolator for rainfall mapping in commercial forests. This type of mapping provides better management of water resources. Hence, remote sensing is a critical analysis tool for understanding rainfall patterns in a water-scare country like South Africa.

Keywords: Rainfall mapping, interpolation, commercial forestry, General Linear Model, Kriging, Random Forests

2.1 Introduction

Rainfall is a fundamental weather-related phenomenon that allows vegetation growth in commercial forest plantations. Rainfall variability is associated with climate change and is expected to cause widespread environmental and socio-economic impacts within the commercial forestry sector (Masson-Delmotte *et al.,* 2018). Water is essential for plant growth, and a decline in quantity directly threatens forestry resources and their productivity (Pimentel, 2006). One of the leading causes of tree death arises from extreme rainfall variations, such as flooding and drought (Trenberth, 2011). Hence, understanding the spatial variations of rainfall and the deductions that can be made from useful data is crucial in addressing the global and local climate crisis challenges. Furthermore, the vulnerability of areas to drought and flooding should not be discarded, as it is vital in research related to sustainable water and natural resource management (Feng *et al.,* 2013).

South Africa is located in the subtropics, where atmospheric circulation acts as a driving force to influence rainfall over different temporal and spatial scales (Taljaard, 1996), therefore becoming and displaying an imperative role in managing water. Due to the warm Agulhas current on the east and cold Benguela current on the west, regions in the eastern part of South Africa receive more rainfall than the north-western regions (Landman *et al.,* 2017). However, it is vital to understand how rainfall varies across forestry regions, given the widespread forest plantations within Southern Africa. Forests transpire water due to the high leaf area, canopy height and surface roughness. They have deep root systems that access moisture deep inside the soil in cases whereby the surface soil is dry. However, trunks have a high storage capacity and keep moisture locked in. In addition, forests are a high source of aerosols, where particles in the atmosphere collect water and trap impurities in the air.

Different forestry regions have high rainfall variability. Tracking rainfall variation over time is instrumental in understanding current forest productivity. This allows foresters to plan future field operations, match sites to suitable forest species, help resource planning, and make decisions about optimal forest yields and forecasts. However, the availability of rainfall spatially, along with the technological and methodological limitations, cause inherent difficulties when attempting to model such climatic instances across large spatial scales (King *et al.,* 2011). Remote sensing and satellite-based sensors allow for temporal investigations to be conducted by providing spatial datasets across the landscape and understanding the drivers dependent on the magnitude, time, duration and seasonality (Feng *et al.,* 2013).

Table 2.1 shows key studies that used spatial frameworks to determine rainfall in different regions globally. Interpolation methods such as Inverse Distance Weighting (IDW) and Ordinary Kriging (OK) have been extensively used to study rainfall distributions (Gia Pham *et al.,* 2019). Studies conducted by Zhang *et al.* (2020), Yang *et al.* (2015), and Chen and Liu (2012) show the success of traditional methods, whereby rain gauge data was used for studying rainfall patterns. Zhang *et al.* (2020) showed a negative relationship between Rainfall Use Efficiency (RUE) and rainfall in China's coniferous forests. RUE is used to understand the diversity of vegetation production in areas that experience limited rainfall, such as arid areas. The limitations placed on plant growth due to low rainfall quantities make the analysis of vegetation growth using RUE an essential indicator of the plant's response in different regions. Data were combined from weather stations and peer-reviewed articles. On the other hand, Yang *et al.* (2015) incorporated interpolation techniques such as Spline, IDW, Australian National University's Digital Elevation Model (ANUDEM) and Kriging for data that was generated from weather stations and modelled rainfall. This study successfully illustrated that IDW is the most accurate interpolation technique compared to the other methods. Yang *et al.* (2015) utilized the IDW method to produce a forty-year time series of rainfall data of the Greater Sydney Region, which was studied daily, monthly, and annually at a ground resolution of 100 metres. Their results proved that from the Spline, Kriging, ANUDEM and IDW interpolation techniques, the IDW produced an R-value of 0.55, which was higher than the other methods. Showing similar results of IDW being the best interpolation technique, Chen and Liu (2012) used forty-six weather stations to estimate rainfall distribution in Taiwan. However, satellite-based data is also valuable for rainfall mapping (Naumann *et al.,* 2012).

Table 2.2. Ten main studies between the years 2000–2020 were focused upon in this research towards showcasing spatial modelling techniques in the spatial domain.

	Reference	Target	Method	Accuracy/Result
1	Zhang *et al.* (2020)	Relatedness of Rainfall Use Efficiency (RUE) and rainfall in China's evergreen coniferous forest.	Data were obtained from weather stations and relevant literature. RUE and ecosystem net primary productivity were utilized.	No evidence of RUE. The negative interrelationship between RUE of the coniferous forest and rainfall.

	Reference	Target	Method	Accuracy/Result
2	Bakar (2020)	Analyse daily precipitation using the Bayesian space-time model from 2013 to 2017.	Trends over specific locations are identified. Inferences for the model's parameters were obtained using the Bayesian model and Markov chain algorithm. This was used to interpolate estimates of rainfall.	Overall accuracy was 87.5% for dry and wet days. Overall RMSE was 4.64.
3	Yang *et al.* (2015)	The production of finer-scale rainfall information using spatial interpolation techniques.	Comparison and assessment of ANUDEM, Spline, IDW and Kriging with weather station data and modelled rainfall.	IDW was better than the other three methods. Monthly IDW in Northern beaches had a mean absolute error of 51.29, mean relative error of 0.55 and RMSE of 78.94. Overall, $P < 0.01$.
4	Chen and Liu (2012)	Estimation of the distribution of rainfall in Taiwan using IDW from 1981 to 2010.	46 Rainfall stations were utilised.	Radius of influence of rainfall was from 10 to 30km. IDW shows more accurate results in the dry season rather than in the flooding season. Correlation coefficient was commonly > 0.95.
5	Naumann *et al.* (2012)	Tropical Rainfall Measuring Mission (TRMM) data was used to monitor uncertainties in Africa's drought.	Data was obtained from TRMM and the Global Precipitation Climatology Centre (GPCC). Nonparametric resampling bootstrap was used to estimate the Standardized Precipitation Index (SPI).	TRMM data had a higher spatial resolution than the GPCC, hence better decision-making. Mountainous areas had higher SPI estimates.
6	Coulibaly and	South Africa's annual precipitation was spatially	Variation analyses and cross-validation of IDW, ordinary	Error mean was 11%; The interpolation errors were higher in the coastal and

	Reference	Target	Method	Accuracy/Result
	Becker (2007)	interpolated from 1931 to 1990.	Kriging, universal Kriging and co-Kriging.	mountainous areas than in South Africa's interior.
7	Adeyewa and Nakamura (2003)	Determine how good TRMM Precipitation Radar (PR) data is for Africa.	'Best estimate' data (3B43), data from the rain gauge and TRMM PR data were compared.	Considerable overestimation was seen from the TRMM PR data in specific months. Biasness is lower in dry seasons in Africa's southern climatic regions.
8	Jeffrey *et al.* (2001)	Used ground-based data to understand Australia's rainfall and climate.	Missing data were calculated using algorithms from data recorded daily and continuously. The Spline interpolation method was used for climatic data, whilst ordinary Kriging was used for rainfall.	The interpolated data showed an accurate estimation of errors. The methodology is adaptable to other countries around the world.
9	Kyriakidis *et al.* (2001)	Atmospheric and terrain characteristics are used to interpolate rainfall spatially.	Assimilated data, elevation and gradient were used. Daily precipitation was mapped using Kriging.	A combination of geostatistics, atmospheric and terrain data accurately maps rainfall using rain gauge data.
10	Goovaerts (2000)	Rainfall was predicted using multivariate geostatistical algorithms and a Digital Elevation Model (DEM).	Simple Kriging, Kriging with an external drift and co-located co-kriging showed observations for annual and monthly rainfall.	Results were compared to Theissen polygon, IDW and OK. The three multivariate geostatistical algorithms outdo the other interpolation techniques.

Naumann *et al.* (2012), Adeyewa and Nakamura (2003) and Jeffrey *et al.* (2001) are examples of studies that used satellite-based remote sensing to map rainfall, hence a more 'modern' technique in science. For example, Naumann *et al.* (2012) looked at drought and low rainfall areas using high spatial resolution TRMM 3B43 data, with 0.25° spatial resolution, and GPCC data, with spatial resolution 1.0°, in Africa. Adeyewa and Nakamura (2003) support this study

and used TRMM data to study African rainfall estimations. Jeffrey *et al.* (2001) illustrated that the southern hemisphere, more specifically, Australia, used Kriging of ground-based data to understand rainfall. They believed that traditional observational data is spatially and temporally incomplete. Hence, remotely-sensed data can provide a complete and accurate analysis of processes within the environment.

On the contrary, Bakar (2020) used a different algorithm, called the Markov chain analysis, to model precipitation for four years using the Bayesian model. Coulibaly and Becker (2007) incorporated interpolation techniques to study rainfall variation within South Africa's coastal and mountainous regions. South Africa's rainfall, when spatially interpolated, is accompanied by several challenges (Coulibaly and Becker, 2007). According to these authors, factors such as elevation, wind direction, and proximity of the oceans impact the spatial distribution of rainfall over land. They further stated that IDW is a deterministic model, but not one of the most accurate approaches in spatial interpolation. However, OK shows more reliable results and is suitable for spatially mapping rainfall. This result was deduced using cross-validation and cross-correlation. Coulibaly and Becker (2007) suggested that mean error is a parameter used to compare the IDW and Kriging approaches and displayed that it has a higher value for IDW than most kriging methods. Hence, they concluded that Kriging is a better approach than IDW. On the other hand, Kyriakidis *et al.* (2001) merged atmospheric and terrain characteristics to interpolate rainfall, but rain gauges were used as data sources. However, Goovaerts (2000) introduced the Theissen Polygon along with IDW and OK to study annual and monthly rainfall in Portugal; hence, this supports the deductions made by Coulibaly and Becker (2007) in Australia.

Regardless of developments in research concerning the role of rainfall, the precise deductions of the spatial distribution of rainfall within South Africa's (SA's) agricultural sector, including commercial forestry, using an ensemble of interpolation techniques remain a gap within research and applied techniques. De Anta *et al.* (2020) conducted a study using soil datasets and stated that uncertainty in research exists due to confinements and inconsistency in sampling methods, poor spatial resolution, insufficient data, and different calculation methods. This deduction can be compared and inferred very closely with remotely sensed data and determining the spatial distribution of rainfall. Obtaining adequate results in remote sensing using conventional methods is dependable on factors such as data availability, data sources, spatial resolution, periods, interpretation skills and distortions. As time progressed, the

accuracy of mapping rainfall has increased with technological improvements and advances in statistical methodology.

In summary, there is a need for accurate rainfall mapping techniques for informed decision-making and planning to increase the resilience of agricultural systems, such as in commercial forests, to the impacts of variations in rainfall. Rainfall variations will be characterised by geographic variability; therefore, there is a need to understand regional patterns of rainfall in commercial forests, especially in understudied regions such as South Africa, to allow for accurate future predictions. Furthermore, the anomalies that exist when reproducing data using interpolation techniques in remote sensing show that statistical relationships can provide a strong foundation for decision-making. This study will use an ensemble of techniques to statistically interpolate rainfall over South Africa while encompassing the commercial forestry regions. This research will contribute to national climate change understanding by producing maps of rainfall that can be used to predict rainfall in the future.

2.2 Materials and Methods

2.2.1 Study region

The study was conducted in commercial forests of the Limpopo, Mpumalanga, KwaZulu-Natal (KZN), Eastern Cape and Western Cape provinces (Figure 2.1). Commercial forests occupy approximately 1.2 million hectares in South Africa, with pine and eucalyptus being the dominant types (Forestry South Africa, 2020). The primary purpose of commercial forests is to meet the demand for wood and be self-sufficient for all timber purposes. Of the 1.2 million hectares of commercial forests, pine occupies approximately 49%, eucalyptus occupies approximately 43%, and wattle at about 7% (Forestry South Africa, 2020). Ground-truth data was collected from July 2018 to July 2019 at 115 rainfall stations across the five provinces of South Africa, as mentioned above, between the latitudes 20°S and 33°S and longitudes 15°E and 35°E. Appendix A provides more statistical detail on the rainfall data. Monthly rainfall was summed, to create a total rainfall layer for the available year and subsequently used for statistical analysis.

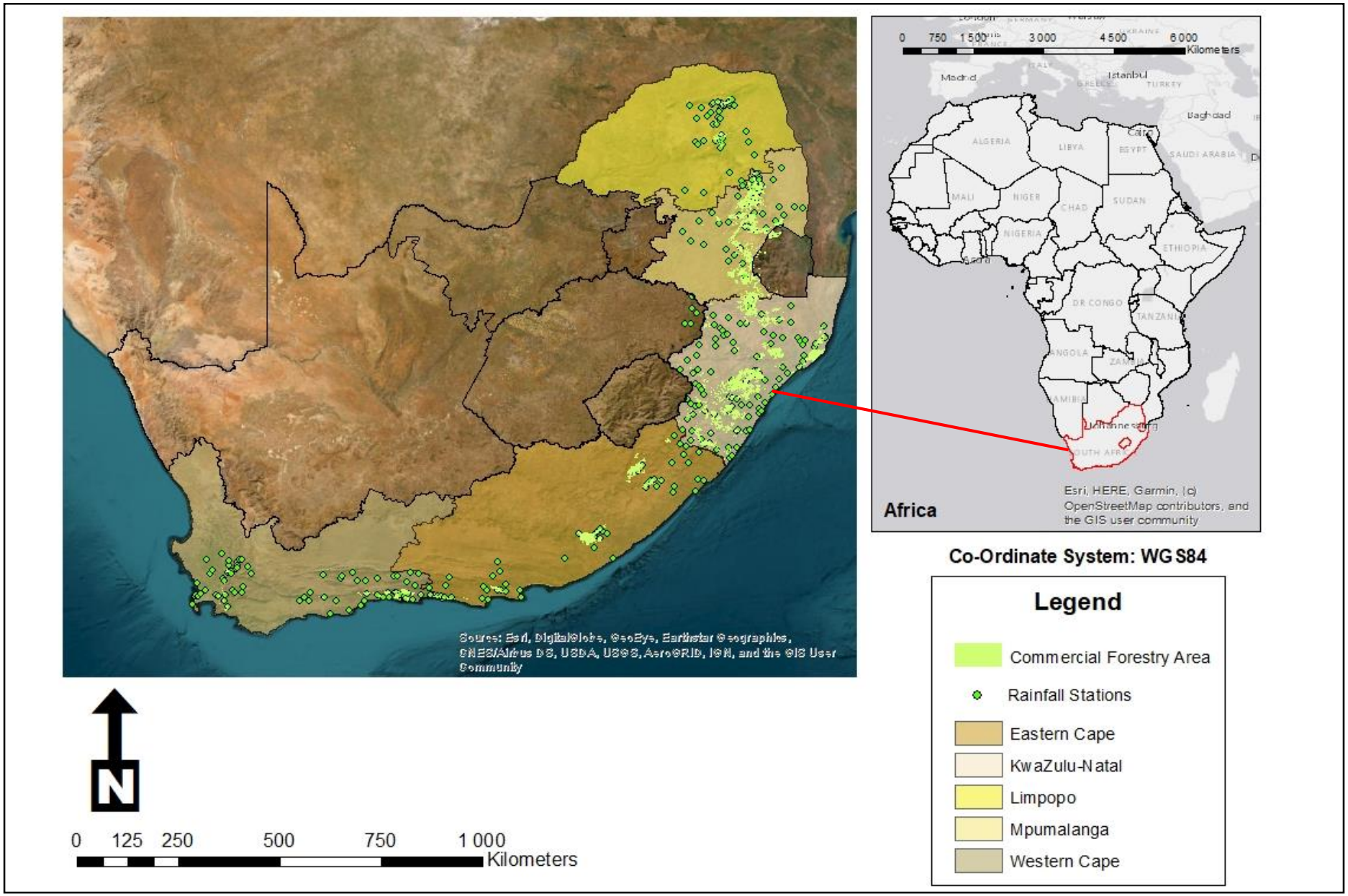

Figure 2.3. Location of the 115 rainfall stations found within and adjacent to South Africa's commercial forests.

2.2.2 Databases and field data

This study used environmental data obtained from ground-based data for July 2018 to July 2019, obtained from rainfall stations, together with satellite-based Moderate Resolution Imaging Spectroradiometer (MODIS) data. Each rainfall station was identified by its latitude, longitude and station Identity (ID). The rainfall stations fell between the latitudes of ± 18°–30° S and longitudes of ± 22°–34° E. The study only shows parameters of rainfall gathered by rainfall stations and no other secondary reliable data collector. Some forestry regions may have had faults within the rainfall stations, while others may have had rainfall stations in perfect working conditions. One of the most significant Earth Observing Systems (EOS) is MODIS, which is essential in monitoring the atmosphere, land and ocean surfaces. This observation is done through the visible, near-infrared (NIR), mid-infrared (MIR) and thermal spectrum channels. MODIS consists of 36 spectral bands, of which bands 1-2 are 250 m, bands 3-7 are 500 m and bands 8-36 are 1000 m, giving a high temporal resolution of one (1) to two (2) days. This is beneficial for tracking changes over a period of time. MODIS data was obtained from USGS Earth Explorer, whereby the study site was selected with zero percent cloud cover for July 2018 to July 2019.

2.3 Statistical analysis

2.3.1. General Linear Model (GLM)

The GLM is a statistical approach used to make predictions using dependent (continuous response) variables obtained by collecting independent (continuous predictor) variables (Johnson, 1998). The GLM is a multiple regression whereby a sequence of numbers is estimated (dependent variable) by the weighted sum of the other numbers (independent variables). Linear regression and Analysis of Variance (ANOVA) make up the GLM, comprising multivariate methods (>1 column in the Y matrix) (Johnson, 1998). According to Madsen and Thyregod (2010), GLMs represent a common statistical framework that allows users to analyse the importance of combining continuous and categorical predictors to a response variable. They further state that it subsumes a continuous prediction, including regression, while the categorical predictors include ANOVA and Analysis of Covariance (ANCOVA). This paper assesses the regression subsume to make continuous predictions. Parameters are utilised to minimize errors such as squared deviations, which can be calculated given the observed data using linear algebra.

2.3.2. Random Forests (RF) unsupervised learning

RF Regression refers to an accumulation of multiple iterations of decision trees (Xiao and Segal, 2009). In other words, the model works on its own to discover information that may not be visible to the human eye. It uses machine learning algorithms that conclude unlabelled data. Trees are assembled randomly to form an RF. This data analysis tool has become vital compared to single iteration classification and regression tree analysis. Robust error estimation of the remaining test set is permitted by bootstrapping the original data set in each tree, more commonly known as the Out-Of-Bag (OOB) sample (Ahmed *et al.,* 2017). This type of machine learning is called ‘random’ because it utilises two random processes: ‘Bootstrapping’ and ‘Random Feature Selection’ (Duangsoithong and Windeatt, 2010). Using the bootstrap samples, the excluded OOB samples can be foreseen. OOB predictions from all the trees can then be pooled.

Linear regression can then be performed using the RF algorithm, and RF does not need criteria of the probability density function of the target variable (Hengl *et al.,* 2018). RF enables the estimation of variable importance. This is done by measuring the mean decrease in prediction

accuracy before and after interchanging OOB variables. Thereafter, the difference between the two is averaged over all trees and then normalised by the standard deviation of the differences (Ahmed *et al.,* 2017):

1. The RF model will be fitted with an efficient environmental co-variate.
2. That was the computation and modelling of the variogram of the RF model residuals, followed by Simple Kriging (SK) on the residuals. This was done to generate an estimation of the spatial predictions.
3. The RF regression predicts results, and the SK residuals are added to estimate the interpolated rainfall.

RFs are known to be used in remote sensing studies because of their known high accuracy levels and reduced training time. Using multiple trees reduces the risks of over-fitting, and it can run on large databases. It produces highly accurate predictions for large databases and can maintain accuracy when a large proportion of data is missing (Pal, 2005). Hence, in this study, we will use RF as an interpolation method (Ahmed *et al.,* 2017).

2.3.3 Accuracy Assessment

2.3.3.1 R Squared (R^2)

R^2 is the coefficient of determination between the dependent and independent variables. The dependent variable has a proportional variation that is predictable from the independent variable. The R^2 value was calculated using equation (2):

$$R^2 = 1 - \frac{RSS}{TSS} \qquad \ldots (2)$$

Where:

- *RSS* refers to the sum squares of residuals; and
- *TSS* refers to the total sum of squares.

2.3.3.2 Root Mean Square Error (RMSE)

Model performance was based on an independent test dataset. Model performance was tested using the RMSE between the measured and predicted rainfall distribution. Models that attained the lowest RMSE were retained for predicting rainfall distribution. The RMSE was calculated using equation (3):

$$\text{RMSE} = \sqrt{\frac{\text{SSE}^2}{\text{n}}} \qquad \ldots (3)$$

Where:

- *SSE* is the sum of errors (measured-predicted values); and
- *n* refers to the number of samples.

2.4 Results

2.4.1 Frequency Plot

Figure 2.2 shows a frequency plot to determine an analysis of the rainfall distributions in SA during the period between July 2018 and July 2019. The rainfall distribution is between 0-200 mm with a frequency of greater than 20 but less than 40. Rainfall distributions between 200 mm and 600 mm have the highest frequency of > 50 to ≤ 70. A frequency of 35 is shown for rainfall distributions between 100-200mm and 500-600mm. Between 700-1000 mm of rainfall distribution, the frequencies are less than and equal to 20 but more than and equal to 5. Rainfall below 600 mm to ≥ 1200 mm has the lowest frequency of below 5.

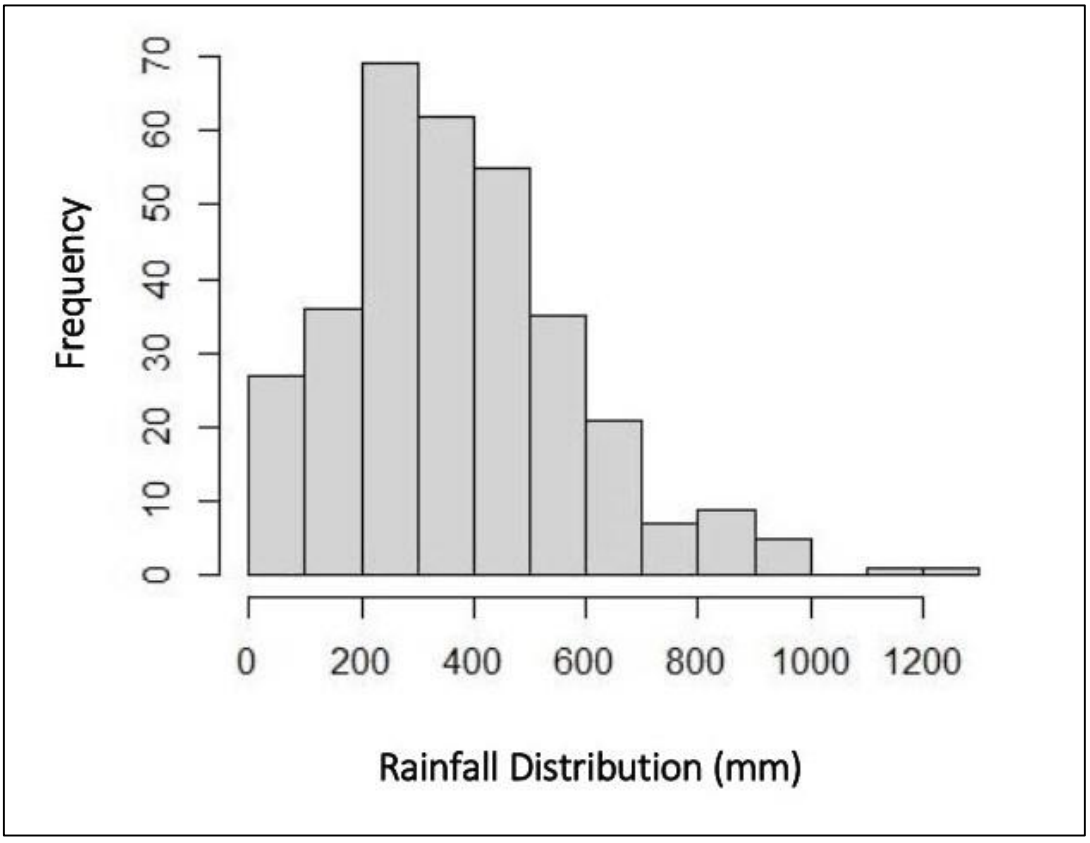

Figure 4.2. Rainfall distributions and their frequencies between July 2018 and July 2019 in South Africa.

2.4.2 General Linear Model (GLM)

The GLM algorithm was run to determine rainfall variability in South Africa, as shown in Figure 2.2. The GLM used to make predictions had an R^2 value of 0.71 and an RMSE value of 0.1272, expressing the accuracy.

Figure 2.3 shows the interpolation results obtained from July 2018 to July 2019, using the GLM algorithm. This map shows the variability of rainfall ranging from 958.6 mm to 159.2 mm. This map relates that the northeastern and eastern regions of the country receive the most amount of rainfall, reaching approximately 958.6 mm, while other interior regions receive relatively low rainfall reaching approximately 159.2 mm. This ranges from 800.2 mm of rainfall between July 2018 and July 2019.

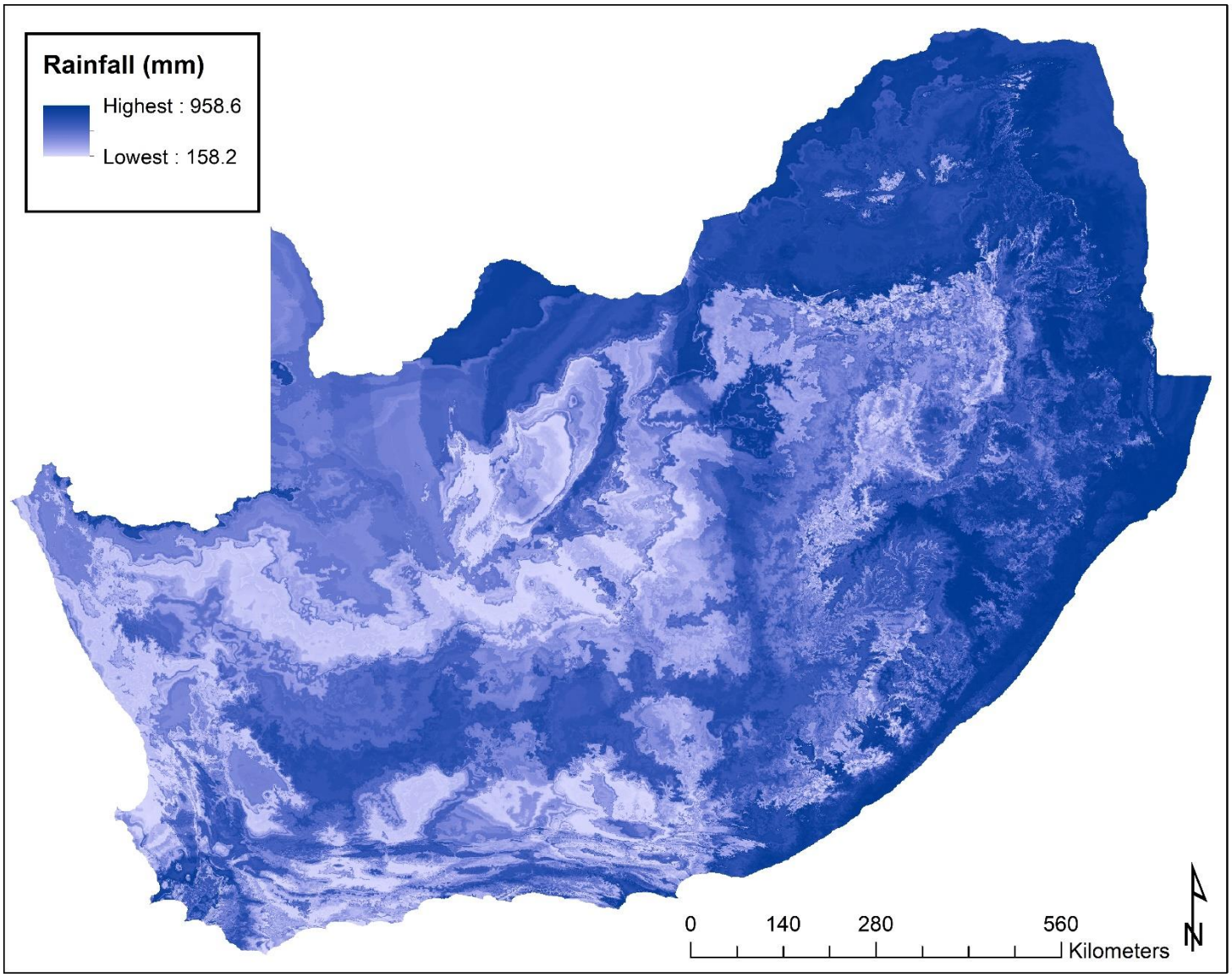

Figure 2.3. Rainfall variability mapped in South Africa using the GLM interpolation technique for the period between July 2018 and July 2019.

2.4.3 Random Forests (RF)

The RF algorithm was run to determine rainfall variability in South Africa, as shown in Figure 2.3. The RF algorithm used to make predictions had an R^2 value of 0.79 and an RMSE value of 0.0165, expressing the accuracy.

Figure 2.4 shows the interpolation results obtained from July 2018 to July 2019 using the RF algorithm. This map shows a rainfall variability ranging from 1188.6 mm to 158.2 mm. This map relates that the country's north-eastern and eastern regions receive the most rainfall, reaching approximately 1188.6mm. In contrast, the other interior regions, the Western Cape and a majority of the Eastern Cape, receive relatively low rainfall reaching approximately 158.2 mm. This provides a rainfall range of 1030.4 mm between July 2018 and July 2019.

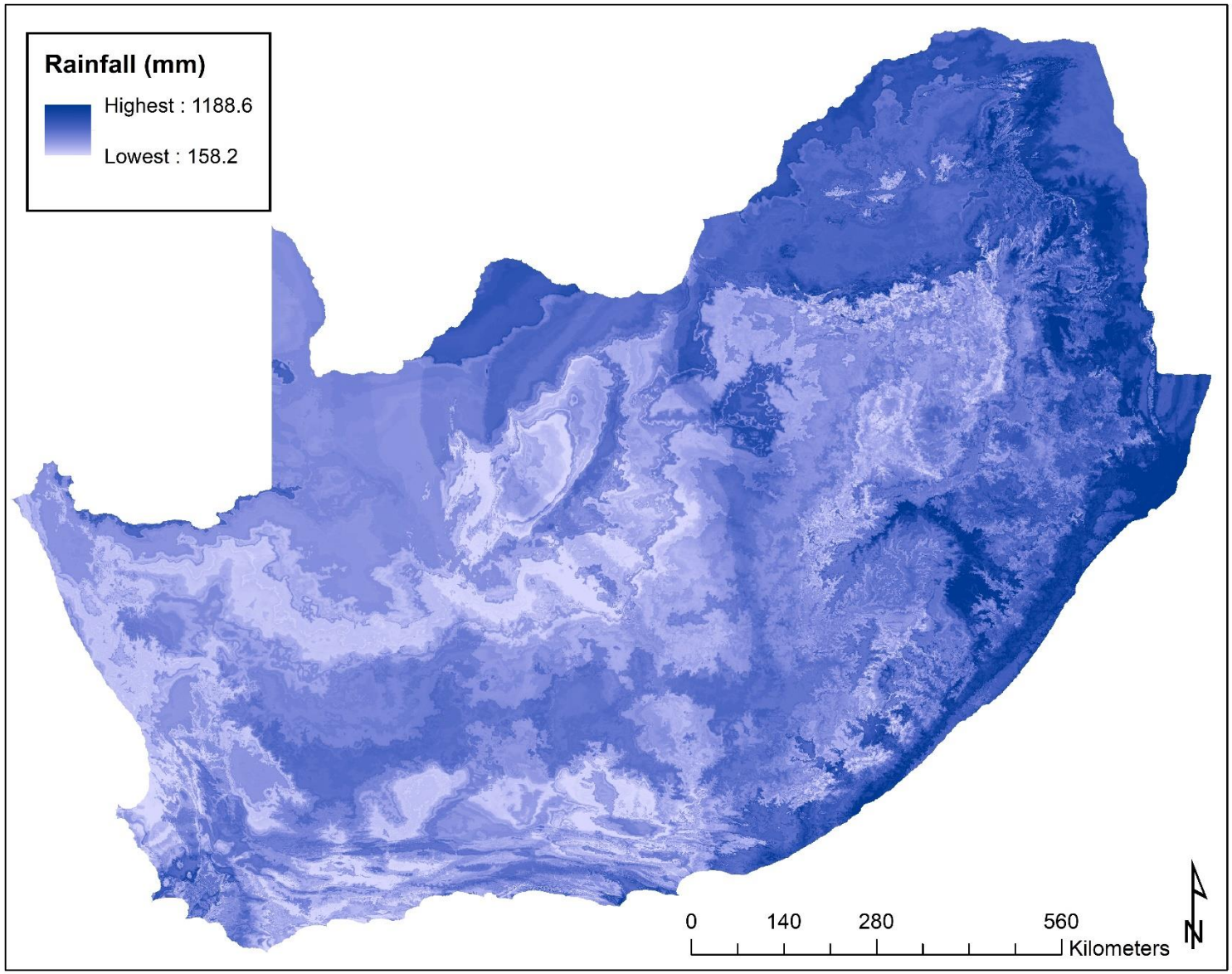

Figure 2.4. Rainfall variability mapped in South Africa using the RF interpolation technique for the period between July 2018 and July 2019.

2.4.4 GLM versus RF

This study compared two interpolation techniques, GLM and RF. For the two interpolation techniques, the RF performed better at predicting rainfall values than the GLM between July 2018 and July 2019. The first interpolation technique demonstrated the GLM to predict rainfall, while the second interpolation technique showed the RF performance of predicting rainfall. Results of these two interpolation techniques had similar performances; however, RF was the most successful at rainfall predictions. It was proved to be more accurate by providing a statistical framework with a high accuracy value of 0.79 and R^2 of 0.52, while GLM had an accuracy value of 0.71 and R^2 of 0.49. The RF model used a large independent test dataset that was not utilised in the OOB samples but used a collection of random decision trees.

2.5 Discussion

One of the critical factors influencing commercial forestry and its primary purpose of meeting the demand for wood is the rainfall variability in SA's commercial forestry areas. South Africa has two (2) dominant commercial exotic tree types, pine and eucalyptus, followed by wattle. Pine, eucalyptus and wattle have a dominance percentage of 49%, 43% and 7%, respectively (Forestry South Africa, 2020). Hence, the study of rainfall in commercial forests is crucial. This chapter aims to discuss the dynamics involved in interpolating rainfall and look at conclusions of interpolation techniques that previous studies demonstrated individually. Furthermore, it discusses a comparison between the GLM and RF algorithms in this study, in comparison with previous studies.

2.5.1 The relationship between GLM and RF Interpolation Techniques and rainfall mapping

An advantage of spatial interpolation is that different techniques are available with different variations of complexities and advancements (Stewart Fotheringham, 1993). Hence, for the aim of this study, the GLM and RF interpolation techniques were used to understand their respective complexities and success relevance to this study. Understanding and analysing precipitation's spatial variability, specifically rainfall, is important to water resource planners and managers. Scientific studies and analyses like this study help manage water resources, especially for water-scare countries like SA. Different interpolation techniques have been expedient within the hydrological field and are highly noteworthy to the Remote Sensing and GIS industries.

This study showed that the GLM and RF both had high accuracy values. The GLM had an R^2 value of 0.71 and an RMSE of 0.1272, while the RF model had an R^2 value of 0.79 and an RMSE of 0.0165. The difference between the GLM and RF RMSE was only 0.1107, which is significantly small but very precise to indicate that RF is the better one of both interpolators. This supports the theory that the eastern part of SA receives more rainfall than the western part (McBride *et al.*, 2022).

2.5.2 Correlation with past studies

This study shows that rainfall predictions can be made across the country using two different interpolation techniques. This is accordant with findings such as the ones conducted by

Coulibaly and Becker (2007) and Kyriakidis *et al.* (2001), who researched SA using kriging interpolation techniques. Coulibaly and Becker (2007) looked at SA's annual interpolating precipitation between the years 1931 and 1990. The interpolators used were IDW, OK, universal kriging and co-kriging. They obtained an error mean of 11%, with coastal areas showing higher interpolation errors than mountainous areas. However, they concluded that the IDW performed the best and still better than the results of this study obtained for RF, which had an R^2 of 0.79. Coulibaly and Becker (2007) and Kyriakidis *et al.* (2001) were able to make rainfall predictions and also acknowledged the gaps within their research, such that they did not use the GLM to make predictions, as they were studies done years ago. This provided motivation as to do this study to fill in the gaps of previous studies. Hence, in contrast to Coulibaly and Becker (2007) and Kyriakidis *et al.* (2001), this study looked at demonstrating GLM against RF at rainfall predictions within SA's commercial forests.

Within the Greater Sydney Region (GSR), Yang *et al.* (2015) used spatial interpolation techniques to model regional climate. They have utilised and compared the ANUDEM, Spline, IDW and Kriging techniques to model the rainfall within this area. The accuracy assessment methods demonstrated were Mean Absolute Error (MAE) and mean RMSE. Similarly, my study utilised RMSE. Results from Yang *et al.* (2015) show that the IDW method was the best of all techniques when used in GIS. This IDW method firstly generated rainfall data for daily, monthly and annual predictions within forty years, with results that were significant in the predictions of erosion associated with rainfall, and secondly, for the management of impacts related to climate change on a local scale. In comparison, this smaller-scale study obtained monthly data for one year.

Coulibaly and Becker (2007) used rainfall stations to obtain SA's annual rainfall data. Their study is in close comparison to this study, which also resorted to using rainfall stations to obtain SA's rainfall data. However, in contrast to this study, Coulibaly and Becker (2007) explored the methodology of wielding IDW, OK, universal kriging and co-kriging, which were compared to determine to each other to determine which one was the most suitable for precipitation predictions. They demonstrated that OK had the best results. Furthermore, a circular semivariogram model was used using thirteen samples to obtain results. The overall kriging was authenticated to be better than IDW, while the circular semivariogram model was deemed the most acceptable. There were low interpolation errors over the interior of South Africa and high for regions along the coast and mountainous areas; hence, their methodology

was proven successful at determining rainfall within the country. Similarly, this study shows that the correlation plot is essential for determining positive and negative correlations to define errors through the study and is also an acceptable model to determine SA's rainfall.

Kyriakidis *et al.* (2001) studied spatial rainfall interpolation using lower-atmosphere and terrain features in conjunction with a DEM. Their study, similar to this study, used rain station observations to map rainfall estimates for seasonal averages of precipitation accounted for daily. The spatial interpolation technique that was used was kriging; however, their interpolation of rain station data was deemed as having the worst results, while the atmospheric and terrain data showed good cross-validation. A regression model was used with precipitation data to estimate with cross-validations and jackknife tests. RMSE values were from 9% for cross-validations to 25% for the jackknife test. Kyriakidis *et al.* (2001) concluded that mapping could be better done using other variables, such as humidity and temperature. This is where research became useful and bridged the gaps in science by using altitude, EVI and elevation to make mapping predictions better.

By furthering this study and filling in the current gaps in research, accurate rainfall mapping techniques will allow more informed decisions to be made, and greater planning will enable more resilience of agricultural systems. Commercial forests will be further studied with additional anomalies that can be tackled and provide enhanced data for interpolation. This means that more accurate data will contribute to understanding national and international climate changes and produce rainfall maps that sanction better rainfall predictions for the future.

2.5.3 Spatial distribution of rainfall over the study area

In any systematic investigation of rainfall seasonality over a large region, especially one which is climatologically heterogeneous, it is essential that the region be divided into a number of smaller homogeneous regions so that each can be treated as a separate entity for application into research (Keen, 1971). This applies to South Africa, where there is remarkable diversity in the total annual rainfall amount and the seasonal distribution of rainfall. Since there are many ways in which rainfall seasonality at a particular location may be characterized, there are many ways of achieving a classification of a region into smaller rainfall regions (Schumann and Hofmeyr, 1938). The GLM and RF algorithms applied to create a map of rainfall seasonality across South Africa have been evaluated in terms of their ability to provide a complete

description of rainfall seasonality. A complete description captures measurements of the magnitude, timing and duration of wet and dry seasons (Pascale *et al.*, 2016). These maps have typically been created to characterize rainfall seasonality into homogeneous rainfall regions as is experienced on the ground (Herrmann and Mohr, 2011), based on either point weather stations, remote-sensing products or a combination thereof.

The results show that both the GLM and RF are capable of detecting and mapping variations of rainfall in SA during the period of July 2018 and July 2019. From the results maps, it can be seen that rainfall of higher ranges in the north-eastern and eastern regions of SA. Lower rainfall variabilities can be seen towards the Western Cape, Eastern Cape and interior of SA. This could be due to the reality of collecting data during peak winter in July 2019 and just before the next year's peak winter in July 2019. This result aligns with the study done by Roffe *et al.* (2019), who stated that SA has great diversity in rainfall distribution and that some regions experience more rainfall than others, depending on their season. Strydom and Savage (2016) stated that the Western Cape is more prone to winter rainfall than summer, while KZN is more prone to summer rainfall than winter.

2.6 Conclusion

Commercial forest plantations have become an essential source of job creation and the invention of efficient timber businesses. Private businesses find investing in exotic species such as the pine, eucalyptus and wattle beneficial. In conjunction with job creation and improving the country's economy, it must not be forgotten that commercial forests need sustainable management. According to Forestry South Africa (2020), forestry is more than the science of just planting trees and managing and caring for them. It includes the best practices that create the least adverse environmental, economic and social problems. This study has demonstrated that interpolation techniques can produce rainfall estimates for mapping such that:

- The detection and mapping of rainfall in commercial forest regions of South Africa were successful;
- A comparison of the GLM and RF accuracy levels for mapping rainfall across commercial forestry areas was successfully done;
- RF was a more successful interpolation technique for predicting rainfall.

However, more research is required to determine more recent rainfall maps for analysis in our drought-prone country and its rainfall variations over the last few years. More research should be put into other variables such as vegetation indices, altitude, elevation, temperature and humidity to extend the study into more versatile and in-depth research. Using the latest imagery, such as TRMM data, could bring a more precise analysis for predictions and estimations. Since South Africa is a large country, each coastal province can be studied independently with its vast data.

In light of the accuracy values of both interpolation techniques, RF and GLM, it is confirmed that SA has wetter areas towards the eastern region of the country, more particularly the northern KZN. For this reason, this study indicated that areas with higher rainfall serve well for commercial forests that must also be managed sustainably in a drought-prone country like SA. The attribute data used were desirable to study the different rainfall values across the country over a period of time. Overall, this study showed reliable indicators for rainfall patterns and mapping them, which can further enhance the studies related to commercial forests and how the changing environment and climate may or may not affect them.

This study is the first demonstration to look at rainfall data obtained from rainfall stations from July 2018 to July 2019, using the GLM and RF interpolation techniques in SA's commercial forests. It offers a more detailed understanding of rainfall patterns and their effects on drought-like conditions within the country. This decrease in rainfall over a period of time is associated with climate change which the next chapter focuses on. Overall, the results of this research show the potential need to refine data and further map commercial forests at a more provincial level rather than a national level. In addition, access to satellite data will enable a more illustrative analysis of data to uncover further gaps in research that will allow an overall improved understanding of the complexities associated with commercial forests in the face of SA's changing weather and climate.

CHAPTER THREE: A Meta-Ensemble Interpolation Technique for Mapping Rainfall Distribution across South Africa's Commercial Forests

Abstract

Mapping rainfall variability allows researchers to understand climate-related phenomena such as floods and droughts. South Africa is one of southern Africa's countries undergoing several drought periods, associated with more than just '*low rainfall'*, but rather *'low rainfall over a long period of time'*. This study used ground-based rain station data and ancillary data, such as MODIS image variables, to map rainfall across the commercial forest region of South Africa. Stacking of the General Linear Model (GLM) and Random Forest (RF) provided a meta-ensemble approach to undertake detections and map rainfall variability found within the commercial forests which spanned across five provinces. This study has shown that MODIS ancillary data is proficiently combined with a meta-ensemble interpolation technique and has shown an R^2 value of 0.86 and RMSE of 0.0453. When the GLM algorithm was used, it had an R^2 value of 0.71 and Root Mean Square Error (RMSE) of 0.1272, while the RF model used had an R^2 value of 0.79 and RMSE of 0.0165.Based on this study's high accuracy and the low-cost implications, further studies can be derived from this approach for detecting and mapping rainfall. Furthermore, it provides insight into understanding climate-related phenomena and bridges the gap of previous studies that did not use the stacking method to develop a significant analysis of rainfall patterns in a drought-prone country like South Africa.

Keywords: Rainfall, forestry, mapping, detecting, commercial forestry, meta-ensemble, General Linear Model, Random Forest

3.1 Introduction

Rainfall mapping is an essential segment in the study of climate change, such that it can help one understand the changes of an area's rainfall over time. Studying rainfall also enables the assessment of the quantity of water below and above the earth's surface. The Intergovernmental Panel on Climate Change (IPCC, 2014) acknowledges the role of forests in mitigating climate change and the adaptive strategies for this phenomenon. Furthermore, the quantitative studies of rainfall mapping is an integral part of forest management and climate change mitigation. Climate change adaptation can be supported by forest management through the establishment of ecosystem services; these ecosystems are such that man depends on them for his livelihood.

Rainfall measurements are essential for determining hydrological processes that occur in nature and improving our understanding of it (Beusch *et al.*, 2018). Quantitative studies utilise a combination of temporal and spatial data to improve accuracy. Besides people's livelihoods, forests provide important ecosystem goods and services such that it is vital for the wellbeing of society, in the sense that forests are pivotal in providing employment in many developing countries, both directly and indirectly (Dlamini, 2014). However, for all the circumstances mentioned above, rainfall is needed to enable the growth of forests, which is why they must be managed well.

In South Africa, forests play a significant role in households' survival in terms of providing timber, firewood, and thatch grass to make and strategically run their homes (Chidumayo *et al.*, 2011). Many African countries have climate change policies but need to guide managing forests and climate interventions. Nhamo (2015) states that access to relevant climate information is needed to support climate interventions. This helps to adapt to risks modelled by climate change and plan strategically. Over the years of studying climate change, it is evident that one climatic change occurrence will lead to another, for which impacts and consequences can be either positive or negative. In many cases, as explained by Keenan (2015), climate change can modify forest disturbances which may lead to wildfires, invasive species and even storms. Thus, resulting in decreasing productivity of forests and alters the distributions of plant, tree and animal species.

Chandler and Wheater (2002) studied rainfall variability in Ireland using the GLM. One of the reasons for their study was the 1990s flood events and their need to understand the drastic changes in rainfall patterns. They indicate that the GLM was a successful indicator of historical

rainfall records that helped understand probable climate change developments. They posed that working with daily rainfall data is more significant than utilising monthly data.

Hagmann *et al.* (2021) stated that forested landscapes have evolved due to historical events, different species composition and even fire regimes that alter ecosystems. Fire and fire frequency in forests play an essential role in the rapidly changing forestry ecosystems. It acts as a management strategy that has resulted in ecosystem structure changes and even composition over the last 200 years (Hessburg *et al.*, 2015). In more severe cases, climate changes in forested landscapes are also vulnerable to disturbances such as droughts, species outbreaks, outbursts of diseases and even natural fires (Allen *et al.*, 2010). Table 3.1 explains different studies conducted worldwide using GIS-based and remote sensing-based techniques to study the changes in forestry caused by direct and indirect alterations in ecosystems.

Kanavos *et al.* (2021) used remotely-sensed data to make predictions of winter rainfall in weather forecasting, using a meta-ensemble technique (i.e. staking of algorithms) that produced high accuracies but not as high as for accuracies tested on other ground-truthed data. They utilised the OzaBag and OzaBagAdwin meta-algorithms with small datasets that resulted in low accuracies. The OzaBag and OzaBagAdwin meta-algorithms produced the highest accuracies of all the algorithms tested and also suggested that classifiers perform better if there is a larger dataset, and if data is obtained from a real-time system. Thus, this study improved performance by using a larger dataset from 115 live-weather rainfall stations. The meta-ensemble approaches are not only used in rainfall mapping, but also in a study of spatially predicting gully erosion in Iran. This refers to a study done by Tien Bui *et al.* (2019) who had a large sample size of 915 locations and 22 gully conditioning factors. Their model ran an RF with alternating decision trees (ADTree) that gave a high prediction accuracy of 0.882. However their findings are that the meta-ensemble approach would improve the accuracy significantly. The accuracy value obtained by Tien Bui *et al.* (2019) showed a higher accuracy than the meta-ensemble approach used in this study.

Studying tree growth changes is common in literature; however, in contrast was a study conducted by McNellis *et al.* (2021), who were able to study tree mortality. They proved that tree mortality in the western USA was greatly influenced by insects and diseases catalysed by climatic variations and the responses produced by species. However, variability within soils did not show much of an influence on tree mortality. Hence, RF can present both the practical and ineffective attributes found within nature that influence the growth of forests.

The RF models are shown to be effective in making predictions within ecosystems. For example, concerning critical environmental variables, Lee *et al.* (2019) used RF to map Citrus Flatid Planthopper (CFP), an invasive species. These variables include altitude, distance to road, slope, minimum temperature, maximum temperature, annual mean temperature and land cover types. In addition, RF showed the influences of human activities and allowed to make strategies of monitoring and surveillance further. In another study, Peerbhay *et al.* (2016) utilised WorldView-2 imagery to map the occurrence of bugweed (*Solanum mauritianum*) within forest margins, open areas and riparian areas. An unsupervised RF technique was used to analyze the species' occurrence, demonstrating the success of remotely sensed data for studies in South Africa. However, this study was a build-up of the study conducted by Peerbhay *et al.* (2015), whereby the capabilities and effectiveness of unsupervised RF were also investigated. In the study conducted by Peerbhay *et al.* (2015), the RF proximity matrix was used to detect bugweed using a principle component approach (PCA).

Silva *et al.* (2017) also investigated forest plantations to make deductions about Brazil's forest structures and the volumes to which they grow. Airborne lidar data were used with field data for this study and was to be applied by pulp and paper companies within the country. Prior to the study conducted by Peerbhay *et al.* (2016), was a study done by Mutanga *et al.* (2012) who also utilized WorldView-2 data to study a vegetated wetland's biomass. Here, RF regression was applied to make accurate predictions beyond the ones provided within the training data set. Here too, the RF algorithm was successful in making predictions. In addition, Ismail *et al.* (2010) explored how GIS and the RF model can be useful in managing forest pests. A high infestation of *S. noctilio* was detected in Mpumalanga. Therefore, the RF model applied was able to provide deductions on environmental conditions within the region and how it allows managers to monitor, adopt strategies, and intervene in pest infestations within forests. Below is Table 3.1 which shows a summary of studies showcasing the benefit of using such modelling techniques in the spatial domain with reliable applicability when dealing with rainfall datasets.

Table 3.1. Ten main studies between years 2010–2021 focused upon this research towards showcasing the benefit of using modelling techniques in the spatial domain.

	Reference	Target	Method	Accuracy/Result
1	Mo and Cao (2021)	Classify land covers.	RF classifier was used with Orbita Hyperspectral Satellite images.	OA= 93.21 and Kappa= 0.91.
2	Zhang *et al.* (2021)	Map the distribution of Mangrove forest (MF) for conservation and restoration.	Object-based image analysis (OBIA) and RF algorithm approaches were applied together with Gaofen-2 (GF-2) and Sentinel-2 imagery.	GF-2 had an OA of 96% and Sentinel-2 had an OA of 94%.
3	McNellis *et al.* (2020)	Study climate change in western United States.	RF classifier model was used to show levels of mortality.	Low levels of mortality was observed in many species. High levels of mortality shown in Middle and Southern rocky mountains.
4	Mngadi *et al.* (2019)	Mapping commercial forest species.	Sentinel 1 and 2 imagery was utilized. Linear Discriminant Analysis (LDA) algorithm was applied.	Sentinel-2 OA was 84% (kappa= 0.81). Sentinel-2 infused with Sentinel-1, OA was 87% (kappa= 0.83) and 88% (kappa= 0.85) respectively.
5	Lee *et al.* (2019)	Study the occurrence patterns of citrus flatid planthopper (CFP) in South Korea.	RF model was applied to study occurrence patterns.	CFP occurred mostly in low altitudes, near roads and urban areas.
6	Silva *et al.* (2017)	Estimate the volume of *Pinus taeda* forest plantation in southern Brazil.	Field and lidar data was combined for the study.	The total pulp volume had R^2 of 0.96 and RMSE of 8.63%.

7	Peerbhay *et al.* (2016)	Detect the occurrence of *Solanum mauritianum* (bugweed) found in forest margins, open areas and riparian zones.	The RF proximity matrix was applied to WorldView-2 imagery.	Forest margins, open areas and riparian zones had an orruence of 91.33%, 85.08% and 67.90%, respectively.
8	Peerbhay *et al.* (2015)	Detect and map S*olanum mauritianum* to stop ecological and economic damage this plant has on forest ecosystems.	RF was used with Anselin Moran's I.	Detection rate (DR) was 89%, False Positive Rate (FPR) was 9.26%. PCA resulted in a DR of 95% and lower FRP of 6.39%.
9	Mutanga *et al.* (2012)	Biomass was estimated in densely vegetated wetland areas.	NDVI from WorldView-2 imagery was used and applied with RF regression algorithm and multiple linear regression was compared to predict biomass.	RMSE of prediction was 0.441kg/m² vs. RMSE of 0.5465kg/m²
10	Ismail *et al.* (2010)	Study pine forests and how they are distributed and susceptible to *S. noctilio* infestations in Mpumalanga.	The RF model was used for the study.	KHAT value of 0.84 and F>0.87, therefore RF produced accurate results.

Rainfall distribution mapping of commercial forests is important for decision-making due to the variations across South Africa. Many geographical attributes, such as location and height above sea level, change the variability of rainfall; therefore, studying the rainfall patterns across different regions and their climatic influences is an essential phenomenon to be studied. Combined with satellite data information, ground-based data could effectively show commercial forestry rainfall variations reliably. In this study, an ensemble of interpolation

techniques and ancillary data are studied together with climatic datasets. This research will provide an overall understanding of producing rainfall maps and predictions of the future. This research aims to map the variability of rainfall across the commercial forest region spanning over the Western Cape, Eastern Cape, KwaZulu-Natal, Mpumalanga and Limpopo provinces in South Africa, using spatial interpolation techniques.

3.2 Methodology

3.2.1 Study region

Field data was collected at different rainfall stations throughout commercial forests of South Africa's coastal provinces, namely Limpopo, Mpumalanga, KwaZulu-Natal (KZN), Eastern Cape and Western Cape (See Appendix A**).** Data were made available for 1 year between July 2018 and July 2019 based on availability from the Institute for Commercial Forestry Research (ICFR). This was conducted between latitudes of 20°S and 33°S and longitudes 15° and 35°W (Figure 3.1). Commercial forests have two main tree types, such as pine and eucalyptus, which occupy a large area (Forestry South Africa, 2010). For this study, data was obtained via ground-truthing using the rainfall stations, which are indicated by the red dots in Figure 3.1. For all rainfall stations that showed evidence of no recordings, non-activity and errors were eliminated; hence the final number of rainfall stations were 115. Monthly rainfall was summed, to create a total rainfall layer for the available year and subsequently used for statistical analysis.

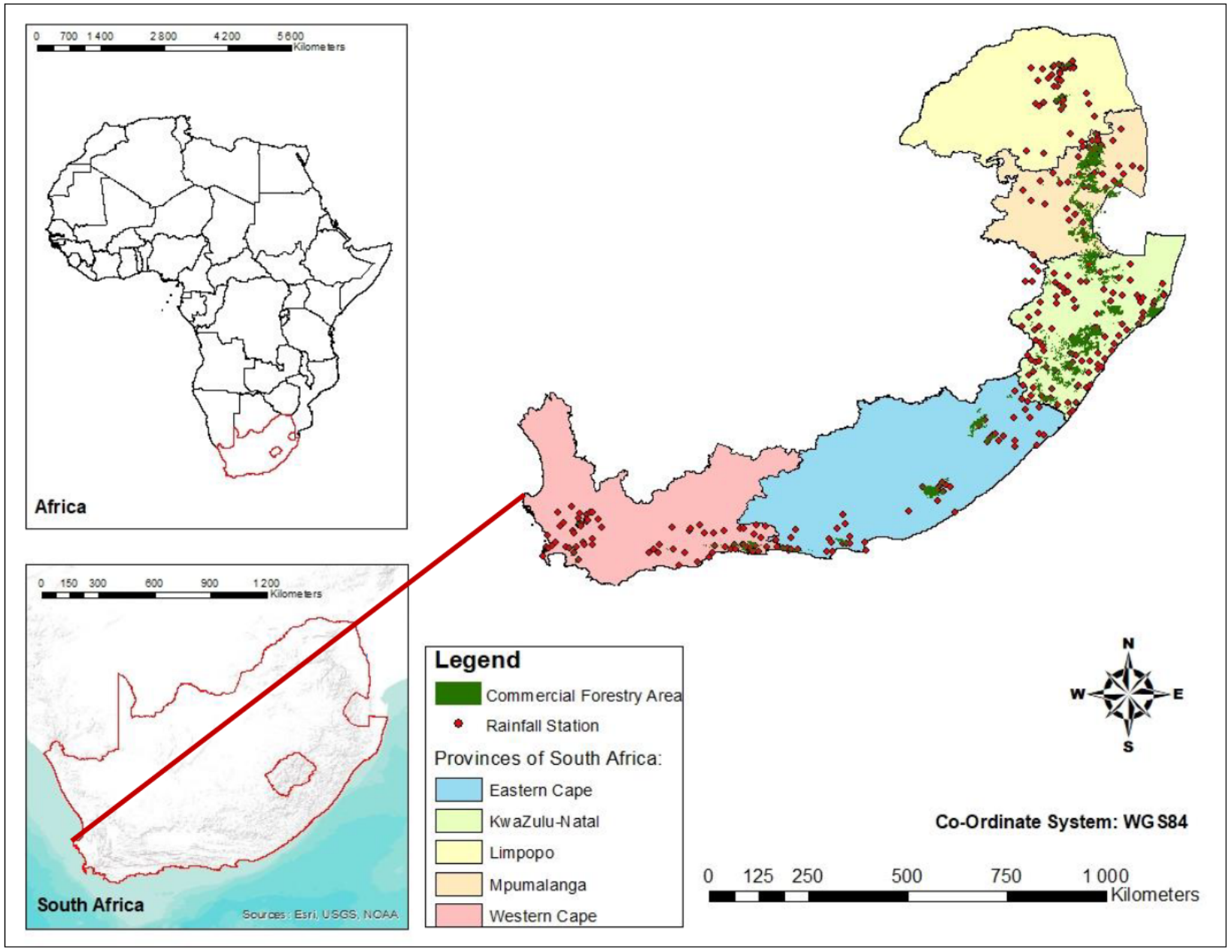

Figure 3.1. Rainfall stations in commercial forests with five provinces of South Africa.

3.2.2 Databases and Field Data

Ground-based data was obtained for this study using rainfall stations in conjunction with Moderate Resolution Imaging Spectroradiometer (MODIS) data. The latitude, longitude and station identity (ID) details of each rainfall station was identified and rainfall readings were recorded for each. The rainfall stations fell between the latitudes of ± 18°–30°S and longitudes of ± 22°–34°E. Monitoring is essential to understand rainfall patterns across the country, hence, MODIS imagery was used due to its high temporal resolution having one (1) or two (2) days. Furthermore, MODIS bands have different resolutions, as described in Table 3.2 below, and the visible, near-infrared (NIR), mid-infrared (MIR) and thermal spectrum channels. Hence, it is a successful method at collecting data related to changes in rainfall over a period of time. For example, vegetation abundance is a good indicator of rainfall distribution and could be used for determine rainfall patterns over time.

3.2.3 MODIS Data

MODIS data is essential in studying environmentally-related phenomena such as atmospheric changes, land cover changes and oceans. This study adopted the 250 m resolution bands to calculate specific indices to examine their contribution for when interpolating rainfall distribution. Below shows Table 3.2 with the variations in MODIS bands:

Table 3.2. MODIS bands and their respective wavelengths.

Band number	Band width (nm^3)	Ground resolution	Visible Range
Band 1	620 – 670	250 m	Red
Band 2	847 – 876	250 m	NIR 1
Band 3	459 – 479	500 m	Blue
Band 4	545 – 565	500 m	Green

Therefore, MODIS data is said to have high temporal resolution but low spatial resolution, thus making predictions over a length of time can be done.

3.2.4 Ancillary Data

Ancillary data was used to supplement the ground-based rainfall data. These attributes included Altitude (Alt), Elevation (Elev), Enhanced Vegetation Index (EVI), Normalised Difference Vegetation Index (NDVI) and average Temperature. According to USGS (2022), elevation refers to a topographic depiction of what is found on the bare ground and excludes infrastructure, trees and plants but includes aspect and altitude. Height of singularities on the Earth's surface refers to the altitude which helps develop aspect, slope and surface gradients. NDVI was used to detect the healthiness level of vegetation found on the Earth's surface, dependent on 'how green' a plant is. High NDVI values indicate healthy vegetation whilst low NDVI values show unhealthy vegetation, a possible indication of how rainfall maybe distributed on land. Temperature was also studied as an attribute since it is one of the most changing variables during climate change. Hence, all the above-mentioned attributes contributes to the significance of detecting rainfall patterns and studying its variability across the coastal provinces of South Africa from July 2018 to July 2019. All climate variables were extracted from the freely available bioclimatic variables (Fick and Hijmans, 2017).

3.3 Statistical analysis

3.3.1 General Linear Model (GLM)

Poline and Brett (2012) state that for the past two decades and more, Magnetic Resonance Imaging (MRI) was based around the General Linear Model (GLM). This method persists in research despite new and other methods. The GLM operates by relying on assumptions and uses appropriate regressors, however, by using too few or too many regressors, it impacts on the sensitivity. According to Johnson (1998), continuous responses are given by making predictions using dependent variables and the continuous predictors are independent variables. Dependent variables and independent variables are used to conduct a multiple regression for number estimation. Both an ANOVA and linear regression is combined to form the GLM.

3.3.2 Random Forest (RF) unsupervised learning

Random Forest (RF) unsupervised learning is the method of regression that is used in this study. RF is formed by the random gathering of trees, called decision trees. There is multiple iterations of gathering of these trees. This system of data analysis is fundamental in research. Bootstrapping of the original data set allows for an estimation of error in the test set. This method is known as the Out-Of-Bag sample (OOB). Hence, predictions can be grouped from all the trees.

On the contrary, RF has its disadvantages such that it is a challenge to interpret response and predictor variables. However, as a positive, it allows for the estimation of variables and how important they are. To do this, the mean value of the decreasing prediction accuracy is measured with the OOB variables that interchange before and after. An average of the two differences is calculated and then the standard deviation (Ahmed *et al.*, 2017). The RF model is fitted with a covariate and then a variogram is generated via computing and modelling. Thereafter, simple kriging (SK) is performed on the residuals to make spatial predictions. The RF regression allows for prediction of results to make an estimation of the rainfall that is interpolated. Therefore, this study RF is a conductive method of interpolation to utilize.

3.3.3 Meta-Ensemble Algorithms

Machine learning algorithms have been done to make improvements in accuracies. This has been done by classification combinations which have algorithms applied to them (Sesmero *et*

al., 2015). However, stacking is a new approach that is able to generate ensembles of classifiers; however according to Sesmero *et al.* (2015), there are limitations as to what attributes are used with this technique that impacts results and their accuracies. This study used the action of stacking the GLM and RF algorithms to demonstrate a meta-ensemble interpolation technique, used to study rainfall in SA commercial forests, and for which their results will be evaluated.

3.3.4 Correlation plot/matrix

Correlation coefficients between different variables is what a correlation matrix is. One of its main functions is to summarize data in order for people make forward-thinking analyses. Typically, it is a 'square' shape and uses colour as a characteristic to plot values. Interrelated values are associated with each other and the matrix has the ability to rearrange variables according to their degree of association. Blue is the colour generally used to indicate positive correlations and red shows correlations that are negative. Coefficients that are correlated are displayed on the right side of the correlogram in accordance to their colours whilst the inconsequential values are left blank with a p-value > 0.05.

3.3.5 Accuracy Assessment

3.3.5.1 R squared (R^2) Coefficient

The dependent variable and independent variable has a coefficient of determination which is referred to as 'R^2' or 'R-squared' (Kasuya, 2019). The dependent variable has variability of proportionality that can be predicted from the independent variable. The R^2 value is calculated using the equation below:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where, '*RSS*' refers to the sum squares of residuals; and '*TSS*' refers to the total sum of squares.

3.3.5.2 Root Mean Square Error (RMSE)

RMSE is the method used to statistically interpolate how rainfall is spatially distributed in South Africa's commercial rainforests. 'Known locations' and 'interpolated or digitized locations' are measured and the difference of these two variables gives us the RMSE. Squaring of this difference, adding them, then dividing that number by the number of test points and

finally finding the square root of the value will give you the data plotted around a line of best fit. RMSE values of the test set greater than the RMSE values of the training set show that the sample is tested well but not better than the predicted value out of the sample that is tested. If predictions are accurate, RMSE values lie between 0.2 and 0.5. RMSE is calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{SSE}^2}{\mathrm{n}}}$$

Where, '*SSE*' is sum of errors (measured-predicted values); and '*n*' refers to the number of samples.

3.4 Results

3.4.1 Correlation Plot

The correlation plot is coloured according to the value of the coefficients. The variables associated with each other is another way the matrix can be coloured. The blue colour is used to indicate the positive correlations whilst the red colour is used to indicate the negative correlations. Looking closely at Figure 3.2, the colour intensity and sizes of the circles vary. There is a proportional relationship between the colour and the correlation coefficient, such that it displays a stronger correlation with a darker and bigger circle. Meaning, the stronger the correlation, the larger the circle. This can be a value that is closer to -1 or 1, as per the colour legend on the y-axis, on the right side of the correlogram. This displays the correlation coefficients and their conforming colours. Hence, both the colour intensity and size of the circle show proportionality of the correlation coefficients. When *p*-values are > 0.05, the correlations are regarded as insignificant and are left blank within the matrix.

Figure 3.2 below indicates the correlogram values for this study. It shows the most significant variables for elevation, altitude, temperature, EVI and NDVI.

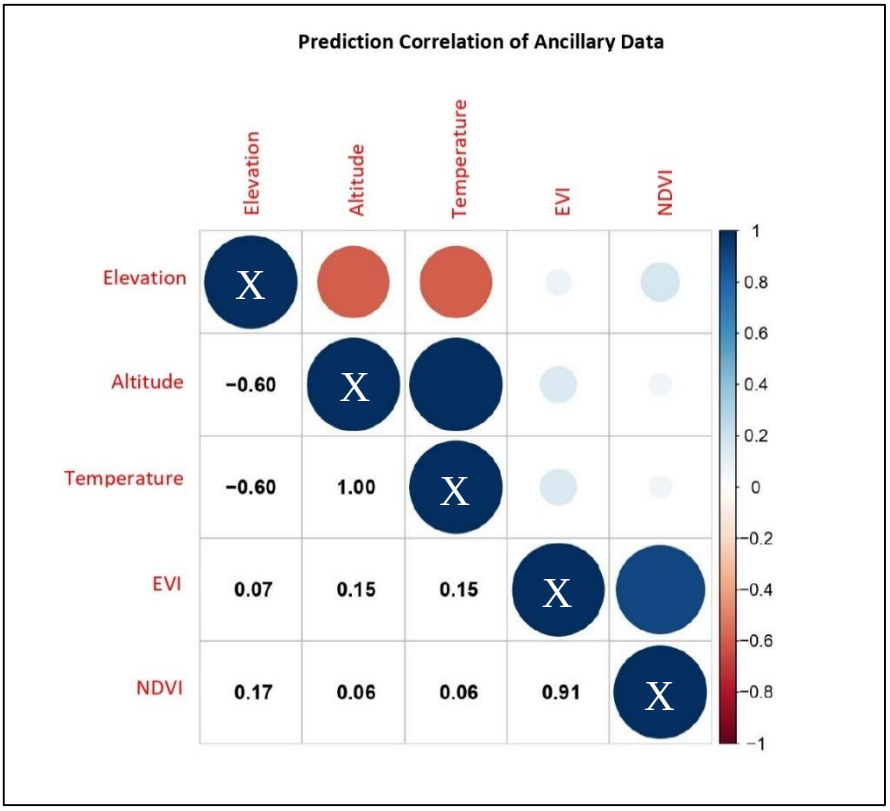

Figure 3.2: Correlation plot/matrix showing the values of ancillary data according to their positive and negative correlations.

The correlation matrix in Figure 3.2 shows that there are strong correlation values of r = 1 illustrated by the large blue circles. The temperature-altitude coefficients have shown the strongest relationship of r = 1, followed by NDVI-EVI with r = 0.91. Following these correlations are NDVI-elevation with an r value of 0.17. EVI-altitude and EVI-temperature, both with r = 0.15. The EVI-elevation coefficients have an r value of 0.07; and both the NDVI-altitude and NDVI-temperature coefficients have an r = 0.06. All these abovementioned relationships are positive correlations. However, the altitude-elevation and temperature-elevation coefficients have r = -0.06 which is a strong negative relationship, represented by the red circles.

Figure 3.3 below shows rainfall variability (mm) of Limpopo, Mpumalanga, KZN, Eastern Cape and Western Cape using the RF algorithm. The darkest blue colour represents rainfall of approximately 989.6 mm, as seen in eastern KZN, Eastern Cape, and parts of the Western Cape. The lightest blue colour represents rainfall of approximately 235.5 mm, as seen in the Western Cape and interior of the Eastern Cape.

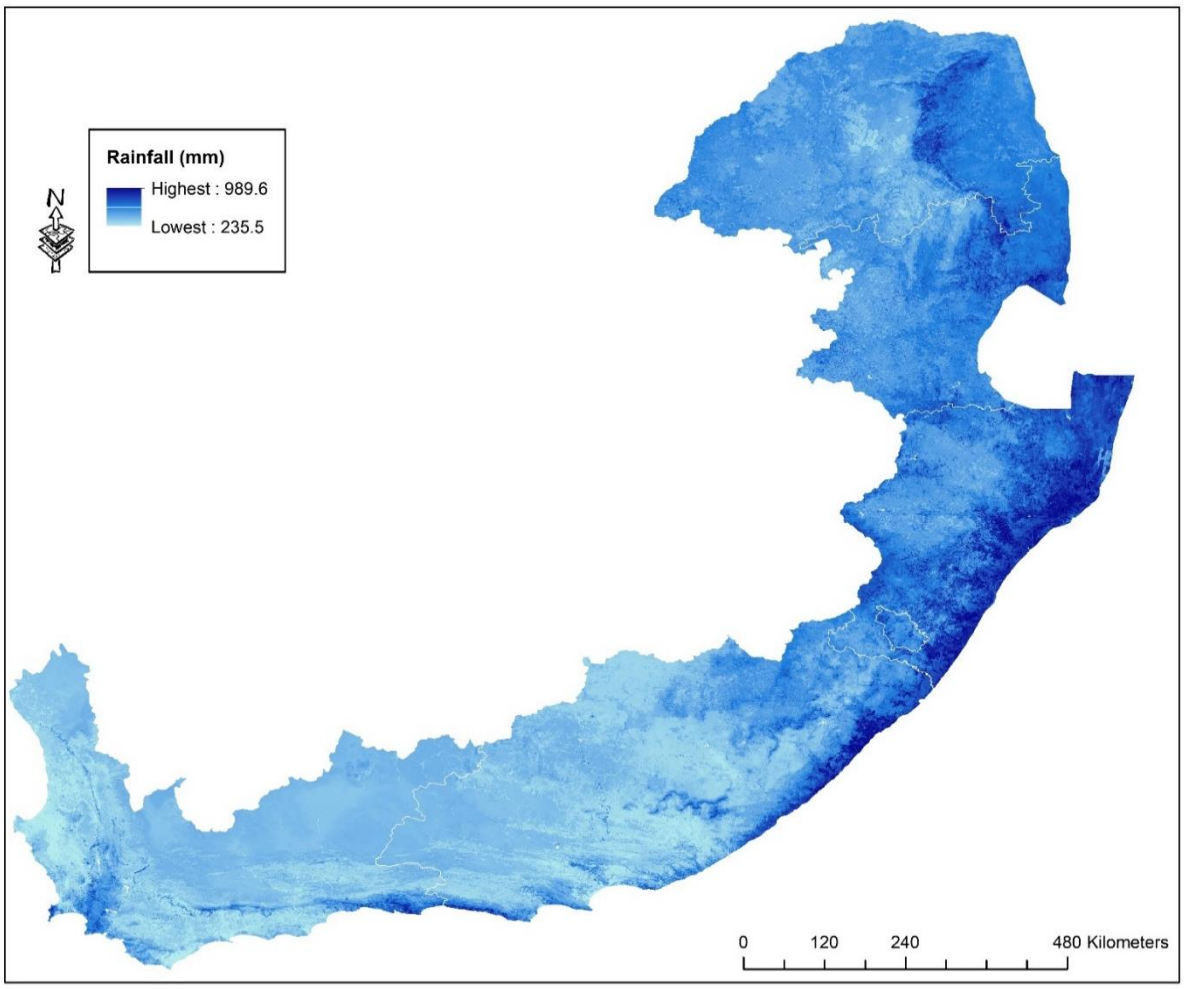

Figure 3.3: Rainfall variability of Limpopo, Mpumalanga, KZN, Eastern Cape and Western Cape using the RF algorithm.

When the stacking of the algorithms were done with ancillary data to demonstrate the meta-ensemble interpolation, the following ancillary data was included:

- Altitude;
- Elevation;
- Temperature;
- EVI; and
- NDVI.

These layers are illustrated in Figure 3.4 below for a clear representation. Elevation had a range of -2 m to 2968 m, whereby the interior of the five provinces showed the highest elevation. Altitude had a range of 0 m to 3170 m, where the interior of the fiver provinces also showed the highest altitude, similar to the elevation results. The EVI values had a range of -1 and 1, showing that Mpumalanga and KZN had the highest EVI values, similar to the NDVI values that ranged from -1 to 1. Lastly, the average temperature were calculated to have a range of 7.08°C to 24.93°C, with Limpopo, Mpumalanga and north-eastern KZN having the highest temperatures.

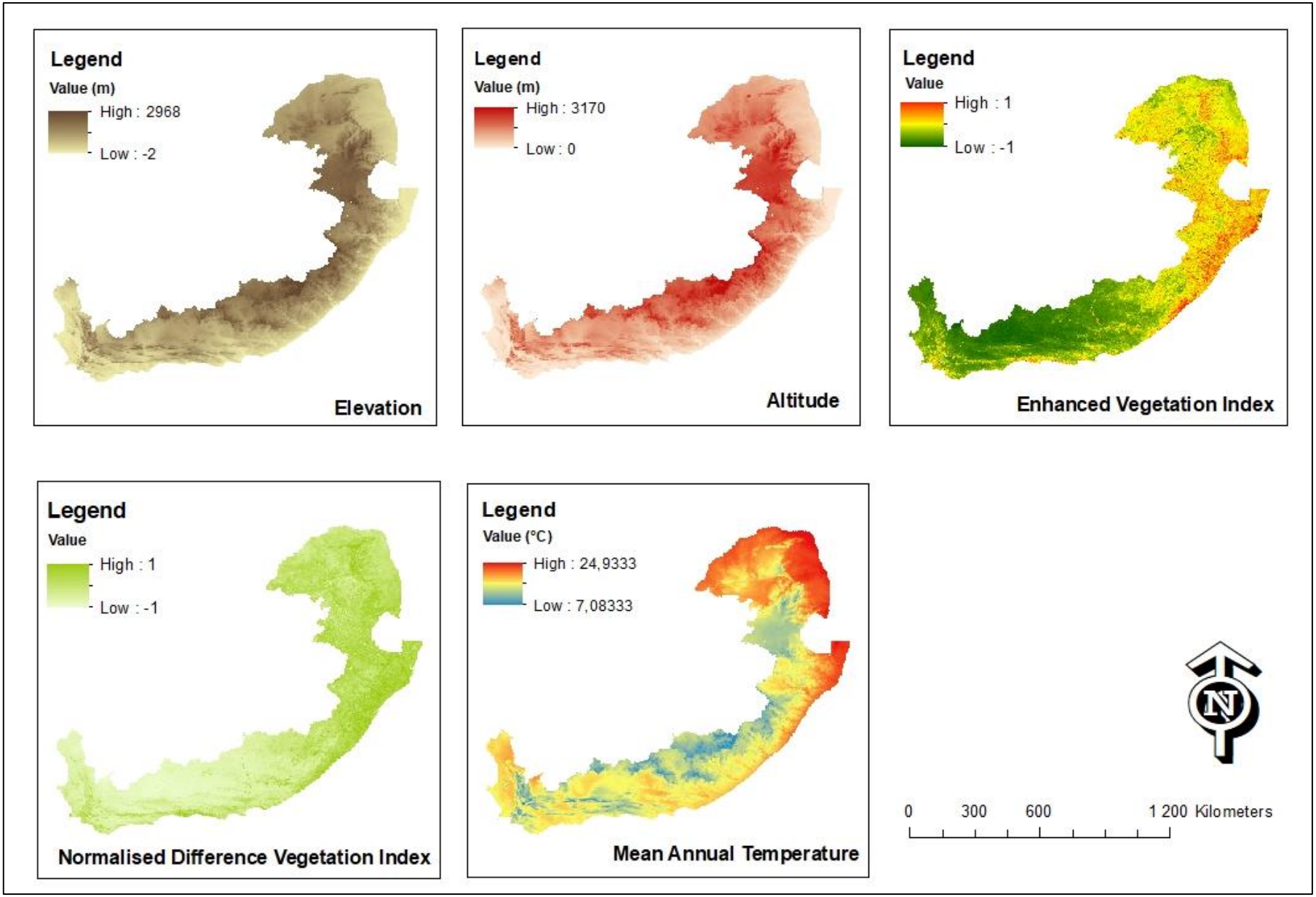

Figure 3.4: Ancillary data used in the stacking process.

Figure 3.5 indicates the stacking of the algorithms to demonstrate a meta-ensemble interpolation technique. The darkest blue colour represents rainfall of approximately 1149.08 mm, as seem in the Limpopo, Mpumalanga and KZN provinces, and parts of the coastal regions of Western Cape and Eastern Cape. The lightest blue colour represents rainfall of approximately 295.4 mm, as seen in the interior of Western and Eastern Cape. Thus, by calculating its accuracy, it has an R^2 value of 0.86 and RMSE of 0.0453.

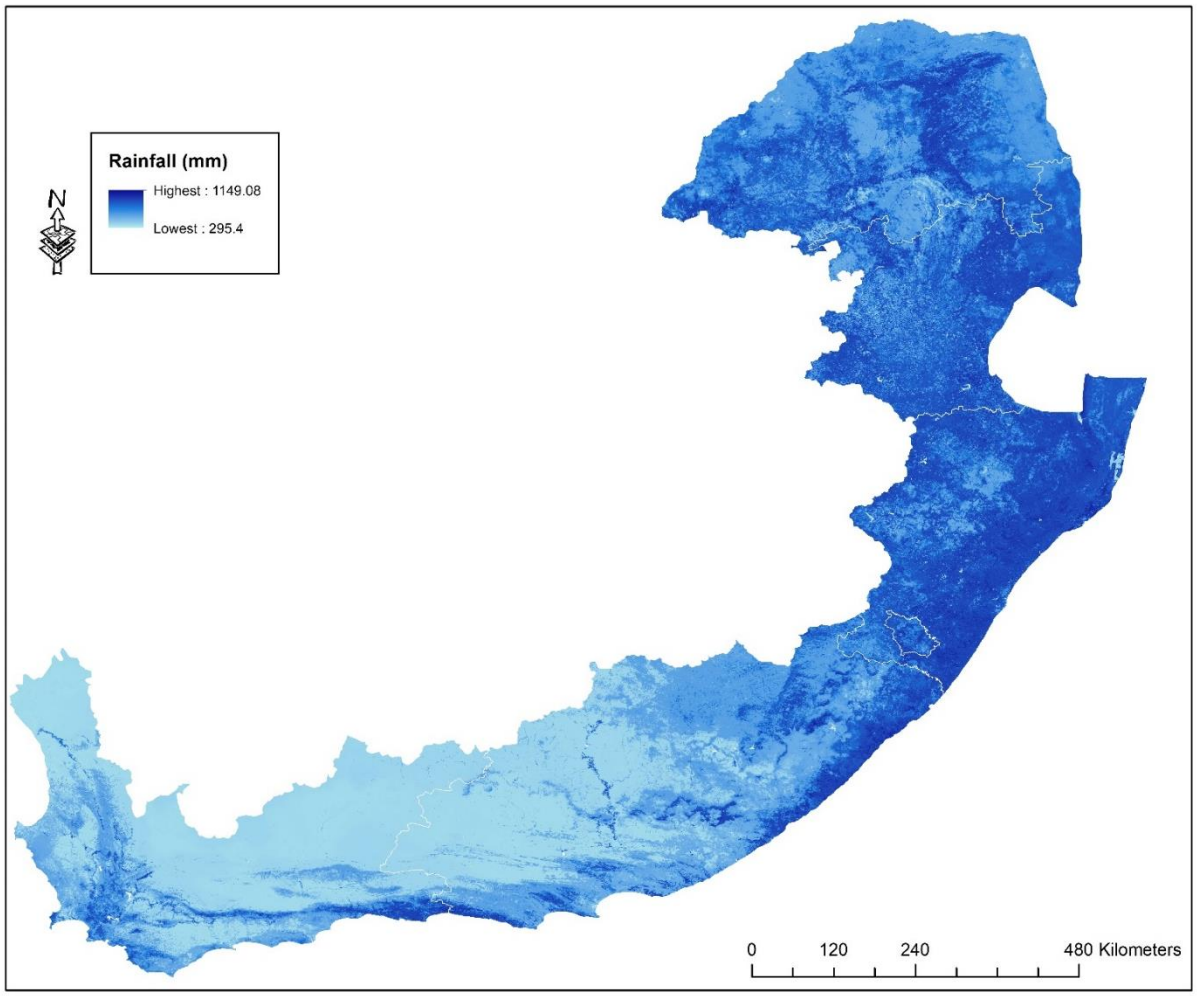

Figure 3.5. Rainfall variability of Limpopo, Mpumalanga, KZN, Eastern Cape and Western Cape using the meta-ensemble interpolation techniques.

Figure 3.4 and 3.5 clearly represent the rainfall variability and indicate that the meta-ensemble is able to detect and map more rainfall variability, producing better results and more variability.

3.5 Discussion

3.5.1 General Discussion

One of the many challenges of managing rainfall in commercial forests is understanding the rainfall variability that exist across different plantation regions of SA, which influence the growth and mortality rates of these forests. SA is wetter on the eastern regions than on the western regions of the country (van Wilgen *et al.*, 2016). In order to understand this natural trend in rainfall, mapping of its variability in commercial forests plays a pivotal role in the management of them. This study serves the purpose of helping manage commercial forests that are affected my climate changes such as droughts and floods. By detecting and mapping its rainfall variability, predictions can be made for the future to help combat climatic variations and prevent negative impacts from these extreme weather events.

By using ground-truth data and remotely-sensed image variables, such as this study, annual rainfall is able to tell a story that helps fill in the gaps of previous studies by providing essence to more accurate and trustworthy data. This study confirms the effectiveness and reliability of using a meta-ensemble of interpolation techniques via the stacking method, to effectively detecting and mapping rainfall variability in SA's commercial forests. More specifically, results have demonstrated that imagery is effective when applied with interpolation techniques to get a clearer understanding of the rainfall found within the five provinces of SA. Furthermore, the utilisation of the GLM and RF algorithms gives an ultimate outline for improving commercial forest management and determining ways to combat the climate change crisis on these ecosystems.

The additional attribute data such as elevation, altitude, elevation, temperature, EVI and NDVI enhanced a more detailed study of rainfall predictions and even possible drought-like areas. The results of this study showed a positive and strong correlation with temperature-altitude and NDVI-EVI.

3.5.2 Correlation with Past Studies

This chapter aims to discuss the role of the meta-ensemble interpolation technique used to understand rainfall variability in SA's commercial forests. It has been deduced that by interpreting these rainfall patterns, climate change can be understood better to give rise to more appropriate management of these forests. This in turn, develops management strategies that

can be used locally, nationally and globally. Commercial forests play a vital role in the production of timber and the money that is generated from it. Hence, the growth rates of these trees are heavily influenced on the amount of photosynthesis it undergoes using the right amounts of carbon dioxide (CO_2) and oxygen (O_2). These two gases are dependent on the rates of global warming which leads to climate change. At the rate of the current climatic changes and intense global warming, CO_2 and O_2 levels of these trees are at a rate that forest production may not be at its optimal, due to CO_2 levels being much higher than before (Change, 2018). In addition, rainfall variability is also a result of climate change as the frequency and intensity of rain has been ever-changing within the past few decades (Stevenson *et al.*, 2022). Changes in rainfall variability impacts the photosynthesis rate of trees, their growth and mortality rates and overall development, hence, directly impacting the forest industry and by-product productions.

Predictions of rainfall variability cannot entirely be accurate due to the constantly changing outcome factors of climate change, such as increase in greenhouse gases, global warming and CO_2 levels. Furthermore, timber/wood quality is not only affected by rainfall variations but also the chemical composition found within the trees that are directly impacted upon my fire events, droughts, floods, historical events, disease outbreaks and surrounding species composition such as alien and invasive species. These contributing factors impact the manner in which forests are managed and over what time periods they occur. Further research is needed on a larger scale over a longer period of time to understand the fluctuations in rainfall and forest responses to these changes.

Zhang *et al.* (2022) is one of the studies that utilised the ensemble method. They investigated accurate prediction values by using the stacking method to determine climate-related parameters. Furthermore, they demonstrated appending bagging and boosting with their remotely sensed data and deduced that the ensemble was successful at predicting accuracies. However, gaps of effective combinations still exist to increase accuracy values, thus, they state that future research must be advanced and explore more diverse algorithms. Go meet the recommendation by Zhang *et al.* (2022), this chapter meets their suggestion by determining which interpolator is best for rainfall variability mapping, i.e., the GML, RF or stacking of algorithms in the meta-ensemble. Studies with rainfall and meta-ensemble approaches may be limited. Thus, this study improved accuracy by stacking the algorithms to demonstrate a meta-ensemble interpolation technique.

Kanavos *et al.* (2001) also used meta-ensemble technique (i.e. staking of algorithms) that produced high accuracies but not as high as for accuracies tested on other ground-truthed data. They made predictions of winter rainfall in weather forecasting using remotely-sensed data and the OzaBag and OzaBagAdwin meta-algorithms that produced the highest accuracies of all the algorithms tested. Their datasets were small which resulted in low accuracies; hence, suggested that classifiers perform better if there is a larger dataset, and if data is obtained from a real-time system. Thus, this study improved performance by using a larger dataset from 115 live-weather rainfall stations. Monthly rainfall was summed, to create a total rainfall layer for the available year and subsequently used for statistical analysis.

As per Appendix A, rainfall varies across the country from winter to summer months. Some regions such as the Western Cape experiences higher rainfall in winter while other regions such as KwaZulu-Natal experiences higher rainfall in summer (Strydom and Savage, 2016). Such variations stand ground to assessing which plants are able to survive in different regions during different climates. As assessed in this study, temperature plays a pivotal role in the production of commercial forests and is not only reliant on rainfall (Saunders *at al.*, 2012). However, further studies should focus on the role of temperature and rainfall ranges over a long period of time to assess their impacts on commercial forests throughout of South Africa. These studies can help manage commercial forests, improve timber production and further understand different species compositions.

In this study, the GLM and RF models were stacked as a Meta-Ensemble approach to make an accurate prediction of rainfall detection. The meta-ensemble used real-time data obtained via rainfall stations in order to withdraw information from the MODIS imagery. Consecutive to this, the rainfall predictions from July 2018 to July 2019 aims at determining how successful the meta-ensemble is at detecting these predictions from the Earth Observing Systems (EOS).

The meta-ensemble interpolation technique is helps perform data and handle it for processing to improve accuracy of the rainfall detection. Furthermore, advancements in remote-sensing technology and groundtruthing has been improved by algorithms like meta-ensemble to make precipitation predictions (Kim *et al.*, 2009). However, as mentioned by Straka *et al.* (2000), precipitation evaluations can be difficult and challenging due to the reliability of measuring, such that the types of precipitations, and how it is measured is different in many countries (Kanavos *et al.*, 2021)

This study has shown that the MODIS ancillary data has good proficiency of being combined with a meta-ensemble interpolation techniques and has shown an R^2 value of 0.86 and RMSE of 0.0453. Based on the high accuracy this study has obtained, and the low cost implications, further studies can be derived from this approach for detecting and mapping rainfall.

3.5.3 Recommendations

There is much uncertainty that arises from this study due to the following reasons:

- The study was limited to commercial forests found in five provinces of the nine, of South Africa.
- The study only shows parameters of rainfall gathered by rainfall stations and no other secondary reliable data collector.
- Future research should include complete datasets with each year, however, based on availability of data, it is difficult to do this for live-weather data.
- Attributes such as atmospheric contributors, such as pollutants, were not accounted for.
- Some forestry regions may have had faults within the rainfall stations whilst other regions may have had rainfall stations in perfect working conditions.
- One of the drivers of climate change is the direct insolation received from the sun, which should be accounted for under a climatic studies.

3.6 Conclusion

In this study, the meta-ensemble interpolation technique and MODIS image variables were visually represented with rainfall variability maps to understand climate change and related weather phenomena. The GLM and RF models were stacked with ancillary data such as altitude, elevation, temperature, EVI and NDVI. In order to determine how accurately the data was represented the R^2, RMSE and correlation matrix was used. The results were compared and analysed to show that the meta-ensemble interpolation technique could detect and map rainfall. The training of the data was able to smoothen out outliers and create finer maps to improve the overall visual representation of the predicted rainfall. Furthermore, monthly rainfall was summed, to create a total rainfall layer for the available year and subsequently used for statistical analysis.

The performance of the interpolator is promising and this encourages an improvement of research for water management in a water-scare country like South Africa. However, one of the main limitations of this study is that rain stations are spread around the country and as such, it is hard to determine and monitor whether or not they are active. Hence, it is recommended that future research should include complete datasets each year. This could be achieved by obtaining data much earlier and analysing different seasons over many years to see an evident variation of climate changes.

Climate changes occur at global and regional scales; however, the extent and timing of it are inexact. However, the severity associated with the impacts of climate change impacts the commercial forestry sector of South Africa, especially economically. In addition, these trees have climatic drivers that initiate changes, making them susceptible to changes. However, some species, such as eucalypts, cannot undergo changes and can barely withstand climatic variations, unlike the *Pinus* family.

Improvements in future studies and management strategies of commercial forests must be implemented in South Africa, such as:

- Study-specific species reactions and growth rates concerning the rainfall variations on the receiving end;
- Determine which areas are at higher risk of climatic regimes;
- Initiation of commercial forest species that can withstand drought-like conditions;
- Study the heat waves and frost days that impact the rates of rainfall received; and

- Determine the influence of soil moisture on commercial trees' growth and mortality rates.

CHAPTER FOUR: Conclusion

4.1 Introduction

Mapping can be done using remote sensing, a valuable tool to obtain data from inaccessible places and provides data that can be studied for weather forecasting and natural disaster predictions (Munawar *et al.,* 2022). Interpolation methods are used to estimate values of points that are inaccessible. Some of the traditional spatial interpolation techniques include IDW, spline and kriging. However, this study utilised a more sophisticated method of interpolation rather than the traditional methods. This study explored different interpolation techniques by being applied in two different methods. The first method was the comparison between the GLM and RF techniques and the second method was the action of stacking the algorithms to demonstrate a meta-ensemble interpolation technique.

The study showed that the monthly data obtained (Appendix A**)** was substantial to work with algorithms and provide high accuracy levels; however, there are limitations associated with collecting data from live-weather systems such as rainfall stations in and around remote forest plantations. Rainfall stations also become faulty and do not log events accurately with gaps and missing data usually hampering analytics tasks. The alternate to this limitation is by using recently acquired satellite rainfall data; however the frequency and availability of this is challenging. Due to this, all rainfall stations that showed gaps in data obtaining were eliminated to provide a more reliable dataset for this study. Lakshmi (2004) and Bayat *et al.* (2019) provide insight into these limitations to which this study clearly gave rise, such that some rainfall stations in this study exhibited gaps and/or no recordings. The most common way for collecting data has been through ground-truthing and surveying, which is understood by Ustin *et al.* (2004) that it is an accurate method; however, it is time-consuming and not cost-effective.

4.2 Aims and Objectives Reviewed

4.2.1 Aim

The main aim of this study was to demonstrate the importance of the detection and mapping of rainfall variability in commercial forests, from July 2018 to July 2019, using a combination of comparable GIS interpolation techniques.

4.2.2 Objectives Reviewed

To make a comparison between the GLM and RF accuracy levels for mapping rainfall in commercial forests, four objectives were set for the aim mentioned above. This section provides a discussion for how these objectives were met.

- **A comparison of the GLM and RF accuracy levels for mapping rainfall across commercial forestry**

The study showed that the GLM and RF both had high accuracy values. The GLM had an R^2 value of 0.71 and RMSE of 0.1272, while the RF model had an R^2 value of 0.79 and RMSE of 0.0165. The difference between the GLM and RF RMSE was only 0.1107, which is significantly small but very precise to indicate that RF is the better one of both interpolators. This supports the theory that the eastern part of SA receives more rainfall than the western part (McBride *et al.*, 2022).

When rainfall was mapped the more traditional way, Simple Kriging (SK), Ordinary Kriging (OK) and Thiessen Polygon were traditional interpolators studied by Goovaerts (2020). They investigated 36 climatic stations in Portugal which covered an area of 5 000 squared kilometres and concluded high accuracy predictions for OK over linear regressions. They also deducted that the algorithms used excluded elevation and the Thiessen polygon produced the worst results. The OK had smaller predictions errors than when rainfall was predicted against elevation in linear regression.

On the contrary, the traditional IDW interpolator was investigated by Chen and Liu (2012) in Taiwan to determine how rainfall was distributed. They gathered data from 46 rainfall stations to produce a high OA of 0.95, implying that IDW is a good traditional interpolator. They further deduced that drier seasons produce more accurate results over flood seasons. A more recent study conducted by Zhang *et al.* (2018) investigated the spline interpolator in Florida to study rainfall variability. They concluded that the IDW and OK methods produced better results than the spline method.

Bakar (2020) also analysed precipitation data, but for data that was obtained on a daily basis. They used a Bayesian space-time model for the duration between 2013 and 2017. With this model, they applied the Markov chain analysis to interpolation rainfall estimates. Both the dry and wet days had an overall accuracy of 87.5% and RMSE of 4.64. Thus, by having daily data

and not monthly data it creates a larger dataset and accuracy can be improved. This study had a large dataset of 115 rainfall stations, and compliments Bakar (2020) and Kanavos *et al.* (2001) findings that larger datasets improve overall accuracy.

Coulibaly and Becker (2007) looked at SA's annual interpolating precipitation between the years 1931 and 1990. The interpolators used were IDW, OK, universal kriging and co-kriging. They obtained an error mean of 11%, with coastal areas showing higher interpolation errors over mountainous areas. All the above-mentioned examples show that of all traditional interpolators, the IDW performed the best and still better than the results of this study obtained for RF that had an R^2 of 0.79. OK is a reliable interpolator but has its limitations that do not produce the lowest errors. Spline is the least preferred method as the other interpolators outdo it. Hence, it can be concluded that of these three traditional interpolators, IDW produces the highest accuracies and is the most reliable method.

Furthermore, the 115 rainfall stations make a suitable sample size to study commercial forests' rainfall patterns that have given rise to high and reliable accuracy levels. The traditional studied previously done had a relatively smaller sample size, not exceeding 50. However, this study uses a much larger sample size. This allows suitable mapping that provides better management of water resources. Hence, ancillary data is a critical analysis component for understanding rainfall patterns in a water-scare country like South Africa.

- **Determine whether using a meta-ensemble technique can provide better results than single interpolation approaches for rainfall mapping**

When the meta-ensemble interpolator was applied with MODIS ancillary data, maps showed good variations in each attribute, such as altitude, elevation, EVI, average temperature and NDVI. Chapter Three (3) used the meta-ensemble stacking of the algorithms with ancillary data and produced an R^2 value of 0.86 and RMSE of 0.0453. When comparing the GLM, RF and meta-ensemble RMSE values, the meta-ensemble still produced the highest accuracy; thus it is the most successful interpolator from this study, and not any of the single approaches.

When previous studies were done, Zhang *et al.* (2022) used the ensemble method to determine how accurately prediction values can be determined. Like this study, Zhang *et al.* (2022) used the stacking method to determine climate-related parameters. In addition, they also looked at appending bagging and boosting with their remotely sensed data. They deduced that the ensemble was successful at predicting accuracies; however, gaps of effective combinations still

exist to increase accuracy values. Zhang *et al.* (2022) further state that future research must be advanced and explore more diverse algorithms. This study meets their suggestion by determining which interpolator is best for rainfall variability mapping, i.e., the GML, RF or stacking of algorithms in the meta-ensemble. Studies with rainfall and meta-ensemble approaches may be limited. Thus, this study improved accuracy by stacking the algorithms to demonstrate a meta-ensemble interpolation technique.

Kanavos *et al.* (2001) used remotely-sensed data to make predictions of winter rainfall in weather forecasting, using a meta-ensemble technique (i.e. staking of algorithms) that produced high accuracies but not as high as for accuracies tested on other ground-truthed data. They utilised the OzaBag and OzaBagAdwin meta-algorithms that produced the highest accuracies of all the algorithms tested. Their datasets were small which resulted in low accuracies; hence, suggested that classifiers perform better if there is a larger dataset, and if data is obtained from a real-time system. Thus, this study improved performance by using a larger dataset from 115 live-weather rainfall stations.

Meta-ensemble approaches are not only used in rainfall mapping, but also in a study of spatially predicting gully erosion in Iran, done by Tien Bui *et al.* (2019). They had a large sample size of 915 locations and 22 gully conditioning factors. Their model ran an RF with alternating decision trees (ADTree) that gave a high prediction accuracy of 0.882; however, stated that a meta-ensemble approach would improve the accuracy significantly. The accuracy value obtained by Tien Bui *et al.* (2019) showed a higher accuracy than this study, thus, improvement on this study are given in Section 4.4.

- **Detecting and mapping rainfall in commercial forest regions of SA using the GLM and RF interpolation techniques with MODIS ancillary data; and**
- **Establishing whether the meta-ensemble interpolator can be combined with satellite ancillary data to map rainfall variations accurately.**

Attribute data such as available vegetation indices, altitude, elevation, and temperature were combined with the meta-ensemble technique to provide a more versatile approach to understanding rainfall mapping. The combinations of these attributes were shown in the correlation matrix. NDVI showed a weak correlation with the EVI ($R^2 = 0.2$). However, there was a strong, negative correlation between elevation and altitude, and average temperature and altitude, with values $R^2 = >0.8$, showing a strong negative relationship between elevation and

altitude, and temperature and altitude, while average temperature and DEM250 showed a strong correlation of $R^2 = 0.6$. Similarly, average temperature and elevation have a robust negative correlation with $R^2 = 0.6$. Furthermore, there is a strong association between EVI and NDVI. When the meta-ensemble interpolator was used, KZN showed the highest predicted rainfall of millimetres per metre with values > 0.40. Limpopo and the Mpumalanga provinces are showing rainfall < 0.35 mm. Thus, by calculating its accuracy, it has an R^2 value of 0.86 and an RMSE of 0.0453.

4.3 A synthesis

The GLM, RF and meta-ensemble interpolation techniques were used in mapping rainfall to determine which regression interpolator was most accurate at mapping rainfall variability. Rainfall mapping can be expressed more efficiently by having daily data and not only monthly data over a more extended period of time. GLM detected and mapped rainfall well, but less significant to RF when they were compared. GLM and RF produced R^2 values of 0.71 and 0.79, respectively, showing a slight difference of 0.08. However, the meta-ensemble approach had the highest R^2 value of 0.86; hence, the meta-ensemble detected and mapped rainfall for the time period most accurately.

Complexities arise from monthly rainfall data as it needs to be a larger data set to study rather than for only one year. Thus, by testing the interpolation techniques, they can be compared to reveal the best performing for South Africa. The results have proven that SA has a wetter eastern region than the western (McBride *et al.*, 2022). Because rainfall is unpredictable and inconsistent, water must be managed well to benefit the present and future generations (van Koppen and Schreiner, 2014). Many studies have been done to understand rainfall patterns in SA, but over the past few years, more developed due to water scarcity being experienced. This study is the first to use interpolation methods to detect and map rainfall variability in SA's commercial forest regions using competent ensemble statistical interpolation techniques. The interpolation accuracy results show that the widely used meta-ensemble approach is the most preferred and over the GLM and RF. However, more studies using these algorithms must be done to improve the statistical approach and better understand the limitations of those interpolators. It can be concluded that rainfall mapping is an integral part of WRM in SA. Rainfall mapping must be done with a better geostatistical approach to determine new parameters to understand the hydrological processes involved in WRM.

4.4 Limitations and Recommendations for Future Research

There is much uncertainty that arises from this study due to the following reasons:

- The study was limited to commercial forests found in five provinces of the nine of South Africa. Future studies can use all nine provinces of SA to gather a larger dataset.
- The study only shows parameters of rainfall gathered by rainfall stations and no other secondary reliable data collector.
- High-resolution imagery can be used and compared to better understand variations in spatial resolution, such as TRMM or CHIRPS data.
- Future research should include complete datasets each year; however, based on data availability, it is sometimes difficult to do this for live-weather data.
- Attributes such as atmospheric contributors and pollutants were not accounted for, which can be studied in future research for their influence on rainfall recordings.
- Future research can make a comparison with rainfall data and Normalised Difference Water Index (NDWI), which can study the water/moisture base.
- Some forestry regions may have had faults within the rainfall stations, while others may have had rainfall stations in perfect working conditions. Try to detect the most reliable rainfall stations, and use only those.
- One of the drivers of climate change is the direct insolation from the sun, which should be accounted for in climatic studies by studying temperature patterns more closely.
- The forestry industry can be further researched to elaborate on planted exotic species and native species which are key species in understanding the growth patterns that are influenced by rainfall. This adds essence to improving yield projections and planning to counteract climate change.

4.5 Concluding remarks

The main aim of this study was to demonstrate the importance of the detection and mapping of rainfall variability in South Africa's commercial forests, from July 2018 to July 2019, using a combination of comparable GIS interpolation techniques. The research showed that the GLM, RF and meta-ensemble techniques are compatible with MODIS ancillary data to do the detection and mapping of rainfall. The final concluding remarks are based on the following interpretations as made in this dissertation:

1. MODIS ancillary data can be applied with algorithms to run interpolators to demonstrate rainfall detection and mapping.
2. The study used the GLM, RF and meta-ensemble interpolation techniques to produce accurate information based on available data. The Meta-ensemble approach produced the highest accuracy of R^2 value of 0.86, coming close to the RF approach of an R^2 value of 0.76 and lastly was the GLM with an R^2 value of 0.7.
3. When compared individually, RF has the lowest RMSE value of 0.0165, followed by meta-ensemble with 0.0453 and lastly, the GLM with RMSE of 0.1272. Hence, GLM is the least accurate and preferred interpolator.
4. Algorithms and ancillary data that were stacked served well in identifying areas with high rainfall variability.

References

Adeyewa, Z.D. and Nakamura, K., 2003. Validation of TRMM radar rainfall data over major climatic regions in Africa. *Journal of Applied Meteorology*, *42*(2), 331-347.

Adger, W.N. and Agnew, M., 2004. *New indicators of vulnerability and adaptive capacity* (Vol. 122). Norwich: Tyndall Centre for Climate Change Research.

Ahmed, Z.U., Woodbury, P.B., Sanderman, J., Hawke, B., Jauss, V., Solomon, D. and Lehmann, J., 2017. Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks. *Journal of Geophysical Research: Biogeosciences*, *122*(2), 354-369.

Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.T. and Gonzalez, P., 2010. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, *259*(4), 660-684.

Bakar, K.S., 2020. Interpolation of daily rainfall data using censored Bayesian spatially varying model. *Computational Statistics*, *35*(1), 135-152.

Bayat, B., Hosseini, K., Nasseri, M. and Karami, H., 2019. Challenge of rainfall network design considering spatial versus spatiotemporal variations. *Journal of Hydrology*, *574*, 990-1002.

Beusch, L., Foresti, L., Gabella, M. and Hamann, U., 2018. Satellite-based rainfall retrieval: From generalized linear models to artificial neural networks. *Remote Sensing*, *10*(6), 939.

Change, P.C., 2018. Global warming of 1.5 C. Geneva, Switzerland: World Meteorological Organization.

Chandler, R.E. and Wheater, H.S., 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, *38*(10), 10-1.

Chen, F. and Liu, C., 2012. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy Water Environment*, *10*, 209-222.

Chidumayo, E., D. Okali, G. Kowero, and M. Larwanou., 2011. *Climate change and african forest and wildlife resources*. Nairobi, Kenya: African Forest Forum.

Coulibaly, M. and Becker, S., 2007. Spatial interpolation of annual precipitation in South Africa-Comparison and evaluation of methods. *Water International*, *32*(3), 494-502.

De Anta, R.C., Luís, E., Febrero-Bande, M., Galiñanes, J., Macías, F., Ortíz, R. and Casás, F., 2020. Soil organic carbon in peninsular Spain: Influence of environmental factors and spatial distribution. *Geoderma*, *370*, 114365.

Dlamini, C. S., 2014. "African Forests, People and Climate Change Project: Forest and Climate Change Policies, Strategies and Programmes in the SADC and COMESA Regionz." Nairobi: African Forest Forum, Working Paper Series, 2 (17).

Duangsoithong, R. and Windeatt, T., 2010. Bootstrap feature selection for ensemble classifiers. In *Industrial Conference on Data Mining*, 28-41

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Feng, X., Porporato, A., and Rodriguez-Iturbe, I. 2013. Changes in rainfall seasonality in the tropics. *Nature Climate Change*, *3*, 811-815.

Fick, S.E. and R.J. Hijmans, 2017. WorldClim 2: New 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37 (12), 4302-4315.

Forestry South Africa, 2020, *Forestry explained*, viewed 17 October 2021, <https://www.forestrysouthafrica.co.za/info-graphics/homepage/introducing-commercial-forestry/>.

Gia Pham, T., Kappas, M., Van Huynh, C. and Hoang Khanh Nguyen, L., 2019. Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of Central Vietnam. *ISPRS International Journal of Geo-Information*, *8*(3), 147.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, *228*(1-2), 113-129.

Hagmann, R.K., Hessburg, P.F., Prichard, S.J., Povak, N.A., Brown, P.M., Fulé, P.Z., Keane, R.E., Knapp, E.E., Lydersen, J.M., Metlen, K.L. and Reilly, M.J., 2021. Evidence for widespread changes in the structure, composition, and fire regimes of western North American forests. *Ecological Applications*, *31*(8), 2431.

Harrington, G.A., Cook, P.G. and Herczeg, A.L., 2002. Spatial and temporal variability of ground water recharge in central Australia: A tracer approach. *Groundwater*, *40*(5), 518-527.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *Journal of Life, Biological, Environmental and Health Sciences, 6*, 26693.

Herrmann, S.M. and Mohr, K.I., 2011. A continental-scale classification of rainfall seasonality regimes in Africa based on gridded precipitation and land surface temperature products. *Journal of Applied Meteorology and Climatology*, *50*(12), 2504-2513.

Hessburg, P.F., Churchill, D.J., Larson, A.J., Haugo, R.D., Miller, C., Spies, T.A., North, M.P., Povak, N.A., Belote, R.T., Singleton, P.H. and Gaines, W.L., 2015. Restoring fire-prone Inland Pacific landscapes: Seven core principles. *Landscape Ecology*, *30*(10), 1805-1835.

IPCC (Intergovernmental Panel on Climate Change)., 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. In *Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, (Eds.) Field C.B., Barros V.R., Dokken D.J., Mach K.J., Mastrandrea M.D., Bilir T.E., M. Chatterjee M., *et al.*, 1132. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Ismail, R., Mutanga, O. and Kumar, L., 2010. Modeling the potential distribution of pine forests susceptible to sirex noctilio infestations in Mpumalanga, South Africa. *Transactions in GIS*, *14*(5), 709-726.

Jeffrey, S.J., Carter, J.O., Moodie, K.B. and Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling & Software*, *16*(4), 309-330.

Johnson, D. E., 1998. Applied multivariate methods for data analysts. Pacific Grove, California, USA: Duxbury Press.

Kanavos, A., Trigka, M., Dritsas, E., Vonitsanos, G. and Mylonas, P., 2021. A regularization-based big data framework for winter precipitation forecasting on streaming data. *Electronics*, *10*(16), 1872.

Kang, S.M., Shin, Y. and Xie, S.P., 2018. Extratropical forcing and tropical rainfall distribution: energetics framework and ocean Ekman advection. *Npj Climate and Atmospheric Science*, *1*(1), 20172.

Kasuya, E., 2019. *On the use of R and R squared in correlation and regression*. Vol. 34, No. 1, 235-236). Hoboken, USA: John Wiley & Sons, Inc.

Keenan, R. J. 2015. Climate change impacts and adaptation in forest management: A review. *Annals of Forest Science*, 72, 145-167.

Kim, D., Nelson, B. and Seo, D.J., 2009. Characteristics of reprocessed Hydrometeorological Automated Data System (HADS) hourly precipitation data. *Weather and Forecasting*, *24*(5), 1287-1296.

King, J., Mitchell, S., and Pienaar, H., 2011. Water supply and demand. In King, J. and Pienaar, H. (Eds.), Sustainable use of South Africa's inland waters. *Pretoria: Water Research Commission.*:1-6.

Kyriakidis, P.C., Kim, J. and Miller, N.L., 2001. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. *Journal of Applied Meteorology*, *40*(11), 1855-1877.

Lakshmi, V., 2004. The role of satellite remote sensing in the prediction of ungauged basins. *Hydrological Processes*, *18*(5), 1029-1034.

Landman, W.A., Malherbe, J., Engelbrecht, F., Mambo, J. and Faccer, K., 2017. South Africa's present-day climate. *Understanding the social and environmental implication of global change,* Stellenbosch: African Sun Media.

Lee, D.S., Bae, Y.S., Byun, B.K., Lee, S., Park, J.K. and Park, Y.S., 2019. Occurrence prediction of the citrus flatid planthopper (Metcalfa pruinosa (Say, 1830)) in South Korea using a random forest model. *Forests*, *10*(7), 583.

Madsen, H. and Thyregod, P., 2010. *Introduction to general and generalized linear models*. Boca Raton, Florida: CRC Press.

Masson-Delmotte, V., Zhai, P., Pörtner, H.O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R. and Connors, S., 2018. Global warming of 1.5 C. *An IPCC Special Report on the impacts of global warming*, *1*(5), 43-50.

McBride, C.M., Kruger, A.C. and Dyson, L., 2022. Changes in extreme daily rainfall characteristics in South Africa: 1921–2020. *Weather and Climate Extremes*, *38*, 100517.

McNellis, B.E., Smith, A.M., Hudak, A.T. and Strand, E.K., 2021. Tree mortality in western US forests forecasted using forest inventory and Random Forest classification. *Ecosphere*, *12*(3), 3419.

Mngadi, M., Odindi, J., Peerbhay, K. and Mutanga, O., 2021. Examining the effectiveness of Sentinel-1 and 2 imagery for commercial forest species mapping. *Geocarto International*, *36*(1), 1-12.

Munawar, H.S., Hammad, A.W. and Waller, S.T., 2022. Remote sensing methods for flood prediction: A review. *Sensors*, *22*(3), 960.

Mutanga, O., Adam, E. and Cho, M.A., 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression

algorithm. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 399-406.

Naumann, G., Barbosa, P., Carrao, H., Singleton, A. and Vogt, J., 2012. Monitoring drought conditions and their uncertainties in Africa using TRMM data. *Journal of Applied Meteorology and Climatology*, *51*(10), 1867-1874.

Nhamo, G. 2015. Policy Coherence in Tackling Climate Change in Africa. Germany: The Heinrich Böll Foundation. Fact Sheet 2.

Nerini, D., Zulkafli, Z., Wang, L.P., Onof, C., Buytaert, W., Lavado-Casimiro, W. and Guyot, J.L., 2015. A comparative analysis of TRMM–rain gauge data merging techniques at the daily time scale for distributed rainfall–runoff modeling applications. *Journal of Hydrometeorology*, *16*(5), 2153-2168.

Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.

Pandey V. and Pandey P.K., 2010. Spatial and temporal variability of soil moisture. *International Journal of Geosciences*, 2010, 1, 87-98.

Pascale, S., Lucarini, V., Feng, X., Porporato, A. and ul Hasson, S., 2016. Projected changes of rainfall seasonality and dry spells in a high greenhouse gas emissions scenario. *Climate Dynamics*, *46*, 1331-1350.

Peerbhay, K., Mutanga, O., Lottering, R. and Ismail, R., 2016. Mapping Solanum mauritianum plant invasions using WorldView-2 imagery and unsupervised random forests. *Remote Sensing of Environment*, *182*, 39-48.

Peerbhay, K.Y., Mutanga, O. and Ismail, R., 2015. Random forests unsupervised classification: The detection and mapping of Solanum mauritianum infestations in plantation forestry using hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 3107-3122.

Pimentel, D., Berger, B., Filiberto, D., Newton, M., Wolfe, B., Karabinakis, E., Clark, S., Poon, E., Abbett, E. and Nandagopal, S., 2004. Water resources: Agricultural and environmental issues. *BioScience*, *54*(10), 909-918.

Pimentel, D., 2006. Soil erosion: A food and environmental threat. *Environment, development and sustainability*, *8*(1), 119-137.

Poline, J.B. and Brett, M., 2012. The general linear model and fMRI: Does love last forever?. *Neuroimage*, *62*(2), 871-880.

Reed, D.N., Anderson, T.M., Dempewolf, J., Metzger, K. and Serneels, S., 2009. The spatial distribution of vegetation types in the Serengeti ecosystem: the influence of rainfall and topographic relief on vegetation patch characteristics. *Journal of Biogeography*, *36*(4), 770-782.

Roffe, S.J., Fitchett, J.M. and Curtis, C.J. 2019. Classifying and mapping rainfall seasonality in South Africa: A review. *South African Geographical Journal*, 101:158-174.

Rogers, P.P., Llamas, M.R. and Cortina, L.M. (Eds.), 2005. *Water crisis: Myth or reality?*. Spain: CRC Press.

Rong, G., Alu, S., Li, K., Su, Y., Zhang, J., Zhang, Y. and Li, T., 2020. Rainfall induced landslide susceptibility mapping based on Bayesian optimized random forest and gradient boosting decision tree models—A case study of Shuicheng County, China. *Water*, *12*(11), 3066.

RMSE (Root Mean Square Error)., 2022. Available at: <https://c3.ai/glossary/data-science/root-mean-square-error-rmse/#:~:text=To%20compute%20RMSE%2C%20calculate%20the,square%20root%20of%20that%20mean>. [Accessed 2 August 2021].

Saunders, M., Tobin, B., Black, K., Gioria, M., Nieuwenhuis, M. and Osborne, B.A., 2012. Thinning effects on the net ecosystem carbon exchange of a Sitka spruce forest are temperature-dependent. *Agricultural and Forest Meteorology*, *157*, 1-10.

Schumann, T.E.W. and Hofmeyr, W.L., 1938. The partition of a region into rainfall districts: With special reference to South Africa. *Quarterly Journal of the Royal Meteorological Society*, *64*(276), 482-488.

Sesmero, M.P., Ledezma, A.I. and Sanchis, A., 2015. Generating ensembles of heterogeneous classifiers using stacked generalization. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(1), 21-34.

Silva, C.A., Klauberg, C., Hudak, A.T., Vierling, L.A., Jaafar, W.S.W.M., Mohan, M., Garcia, M., Ferraz, A., Cardil, A. and Saatchi, S., 2017. Predicting stem total and assortment volumes in an industrial Pinus taeda L. forest plantation using airborne laser scanning data and random forest. *Forests*, *8*(7), 254.

Stevenson, S., Coats, S., Touma, D., Cole, J., Lehner, F., Fasullo, J. and Otto-Bliesner, B., 2022. Twenty-first century hydroclimate: A continually changing baseline, with more frequent extremes. *Proceedings of the National Academy of Sciences*, *119*(12), p.e2108124119.

Stewart Fotheringham, A. and Rogerson, P.A., 1993. GIS and spatial analytical problems. *International Journal of Geographical Information Science*, *7*(1), 3-19.

Straka, J.M., Zrnić, D.S. and Ryzhkov, A.V., 2000. Bulk hydrometeor classification and quantification using polarimetric radar data: Synthesis of relations. *Journal of Applied Meteorology and Climatology*, *39*(8), 1341-1372.

Strydom, S. and Savage, M.J., 2016. A spatio-temporal analysis of fires in South Africa. *South African Journal of Science*, *112*(11-12), 1-8.

Taljaard, J.J., 1996. *Atmospheric circulation systems, synoptic climatology and weather phenomena of South Africa. Part 6, Rainfall in South Africa*. South Africa: Department of Environmental Affairs and Tourism.

Trenberth, K.E., 2011. Changes in precipitation with climate change. *Climate Research*, *47*(1-2), 123-138.

Tien Bui, D., Shirzadi, A., Shahabi, H., Chapi, K., Omidavr, E., Pham, B.T., Talebpour Asl, D., Khaledian, H., Pradhan, B., Panahi, M. and Bin Ahmad, B., 2019. A novel ensemble artificial intelligence approach for gully erosion mapping in a semi-arid watershed (Iran). *Sensors*, *19*(11), 2444.

USGS (United States Geological Survey)., 2022. USGS Science for a Changing World. Available at: < https://www.usgs.gov/#:~:text=A%20Look%20at%202022,climate%20and%20land%2Duse%20change.> [Accessed 04 February 2022.]

Ustin, S.L., Roberts, D.A., Gamon, J.A., Asner, G.P. and Green, R.O., 2004. Using imaging spectroscopy to study ecosystem processes and properties. *BioScience*, *54*(6), 523-534.

van Koppen, B. and Schreiner, B., 2014. Moving beyond integrated water resource management: Developmental water management in South Africa. *International Journal of Water Resources Development*, *30*(3), 543-558.

van Wilgen, N.J., Goodall, V., Holness, S., Chown, S.L. and McGeoch, M.A., 2016. Rising temperatures and changing rainfall patterns in South Africa's national parks. *International Journal of Climatology*, *36*(2), 706-721.

Wang, M., Rezaie-Balf, M., Naganna, S.R. and Yaseen, Z.M., 2021. Sourcing CHIRPS precipitation data for streamflow forecasting using intrinsic time-scale decomposition based machine learning models. *Hydrological Sciences Journal*, *66*(9), 1437-1456.

Xiao, Y. and Segal, M.R., 2009. Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS Computational Biology*, *5*(6), 1000414.

Yang, X., Xie, X., Liu, D. L., Ji. F. and Wang, L., 2015. Spatial Interpolation of Daily Rainfall Data for Local Climate Impact Assessment over Greater Sydney Region. *Advances in Meteorology*.

Zambrano, F., Wardlow, B., Tadesse, T., Lillo-Saavedra, M. and Lagos, O., 2017. Evaluating satellite-derived long-term historical precipitation datasets for drought monitoring in Chile. *Atmospheric Research*, *186*, 26-42.

Zhang, M., Leon, C.D. and Migliaccio, K., 2018. Evaluation and comparison of interpolated gauge rainfall data and gridded rainfall data in Florida, USA. *Hydrological Sciences Journal*, *63*(4), pp.561-582.

Zhang, R., Jia, M., Wang, Z., Zhou, Y., Wen, X., Tan, Y. and Cheng, L., 2021. A comparison of Gaofen-2 and Sentinel-2 imagery for mapping mangrove forests using object-oriented analysis and random forest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 4185-4193.

Zhang, X.J., Jun-Bang, W.A.N.G., Chu, W.U. and Kamil, K.U.Č.A., 2020. The spatial distribution patterns of rainfall use efficiency (RUE) of evergreen coniferous forests in Chinese subtropical zone. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca*, *48*(1), 492-502.

Zhang, Y., Liu, J. and Shen, W., 2022. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, *12*(17), 8654.

Zinevich, A., Messer, H. and Alpert, P., 2009. Frontal rainfall observation by a commercial microwave communication network. *Journal of Applied Meteorology and Climatology*, *48*(7), 1317-1334.

Zipser, E.J., Cecil, D.J., Liu, C., Nesbitt, S.W. and Yorty, D.P., 2006. Where are the most intense thunderstorms on Earth?. *Bulletin of the American Meteorological Society*, *87*(8), 1057-1072.

APPENDIX A: MEAN ANNUAL PRECIPITATION (mm)

	JUL 2018	AUGT 2018	SEPT 2018	OCT 2018	NOV 2018	DEC 2018	JAN 2019	FEB 2019	MAR 2019	APR 2019	MAY 2019	JUN 2019	JULY 2019
MEAN	38,378	49,303	63,960	35,957	44,123	40,207	43,071	47,231	60,947	63,089	29,492	45,488	52,603
MEDIAN	14,5	21	4,2	28,8	28,1	15,8	69	11,4	25,8	41,3	19,8	30,8	6,7
MAX	28,2	41,8	6,8	53,6	43,6	19,2	123,8	19,2	44,4	65,4	28,4	61	13
MIN	0,8	0,2	1,6	4	12,6	12,4	14,2	3,6	7,2	17,2	11,2	0,6	0,4
RANGE	27,4	41,6	5,2	49,6	31	6,8	109,6	15,6	37,2	48,2	17,2	60,4	12,6
STDEV	19,374	29,415	3,676	35,072	21,920	4,808	77,498	11,030	26,304	34,082	12,162	42,709	8,909