DRIVERS AND PREDICTIVE MODELLING OF THUNDERSTORMS AND PRECIPITATION OVER WEST AFRICA

BY

AKUM, ROBERT AYUEBONING

B.Sc., MPhil. (Kumasi)

MCS/20/7729

A Thesis of the Doctoral Research Programme of the West African Climate Systems, under the West African Science Service Centre on Climate Change and Adapted Land Use, in the Department of Meteorology and Climate Science submitted to the School of Postgraduate Studies in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy in Meteorology and Climate Science of The Federal University of Technology, Akure, Nigeria.

JULY, 2025

DECLARATION

I declare that this is my work and that it has not previously been submitted as a dissertation or thesis at any other University for the award of a degree.

Candidate's Name: Akum, Robert Ayueboning

Signature……………………………….. **Date**……………..………………….

CERTIFICATION OF ORIGINALITY

We certify that this Dissertation entitled "Drivers and Predictive Modelling of Thunderstorms and Precipitation over West Africa" is the outcome of the research carried out by Robert Ayueboning Akum under the WASCAL DRP-WACS in the Department of Meteorology and Climate Science of the Federal University of Technology, Akure.

Supervisor's Name:

Major Supervisor: Prof. Philip G. Oguntunde

Signature……………………………….. **Date**……………..…………………

Co-supervisor: Prof. Dr. Gunnar Lischeid

Signature……………………………….. **Date**……………..…………………

Co-supervisor: Dr. R. A. Balogun

Signature……………………………….. **Date**……………..…………………

Director of WASCAL DRP: Prof. I. A. Balogun

Signature……………………………….. **Date**……………..…………………

Head of Department: Prof. V. O. Ajayi

Signature……………………………….. **Date**……………..…………………

ABSTRACT

Thunderstorms provide essential precipitation for agriculture and other economic activities in the tropics. Accurate forecasting of thunderstorm variability and precipitation forecasting over West Africa is vital for agricultural planning, water resource management, hydropower generation, and disaster risk reduction. This study aims to identify the trends and drivers of monthly thunderstorm frequency from 1990 to 2013 across 22 synoptic stations in Ghana using a machine learning approach. Additionally, for the first time, monthly thunderstorm frequencies over Ghana were forecast one year in advance using machine learning models to offer actionable insights for planning and disaster preparedness. Moreover, this study pioneers the application of machine learning to predict precipitation a year ahead over West Africa at a spatial resolution of 0.10° × 0.10°, filling a significant gap in long-term seasonal forecasting systems. LightGBM, XGBoost, Random Forest (RF), and ensemble models were developed utilizing ERA5 reanalysis data, climate indices, and geographic coordinates as input features. LightGBM and XGBoost models were selected for precipitation prediction due to their ability to efficiently process large datasets on standard computing infrastructure. To better understand model behaviour and the influence of features on monthly precipitation over West Africa, SHapley Additive exPlanation (SHAP) plots were employed. Recursive feature elimination was used to select the most important features from over 60 candidates to model thunderstorm variability, which the models explained at a rate of 79%. SHAP analysis revealed that convective available potential energy (CAPE) was the primary driver of thunderstorm variability, followed by top-of-atmosphere incident solar radiation (TISR), convective inhibition (CIN), 10 m wind speed (SI10), and longitude. The frequency of thunderstorms in Ghana declined significantly, coinciding with a notable decrease in CAPE and an increase in CIN, providing strong evidence of a causal relationship. The resulting LightGBM, XGBoost, RF, and ensemble models achieved R^2 values of 0.74, 0.75, 0.75, and 0.76, respectively.

This research introduces a forecasting scheme for thunderstorm frequencies in Ghana, offering a valuable tool for seasonal prediction. The precipitation models developed to forecast rainfall one year ahead across West Africa explained between 77% and 84% of the variance. Effectively capturing temporal and spatial rainfall patterns, including the ‘monsoon jump,’ ‘little dry season,’ and the Ghana-Benin dry zone, highlighting areas with high and low precipitation on a monthly basis. Analysis of model behaviour indicated that different features influence monthly precipitation across various zones of West Africa. These models can be adapted and integrated into the seasonal forecasts of national meteorological agencies to enhance accuracy. Better seasonal rainfall forecasts could assist West African farmers in anticipating rainfall variability, preparing for droughts and floods to safeguard crop yields. Countries within the study area that rely on hydroelectric power or surface water resources can also utilise these precipitation prediction models for planning and mitigating drought and flood risks.

DEDICATION

This thesis is dedicated to you, Lord Jesus Christ, lover and saviour of my life—the source of all good and perfect gifts, and knower of what is to come – the Chief Meteorologist and Modeller.

ACKNOWLEDGEMENTS

I thank God Almighty for His guidance and protection during this PhD journey and throughout my life. Many thanks to the Federal Ministry of Education and Research (BMBF), Germany for sponsoring the WASCAL program at the Federal University of Technology, Akure (FUTA DRP-WACS).

I am grateful to my Supervisor, Prof. Philip G. Oguntunde, for his mentorship and allowing me to explore, learn, and advance as a researcher. I also thank my German supervisor, Prof. Dr. Gunnar Lischeid, for hosting me at Leibniz-Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany. I am grateful for his guidance and for introducing me to advanced time series analysis and machine learning. I have learnt a lot under you and I am very grateful. I am also grateful to my co-supervisor, Dr. Richard Balogun, for helping supervise my thesis. I also thank the Director and staff of WASCAL Head Office, Accra, Ghana, and the Director, Prof. I. A. Balogun and staff of WASCAL DRP-WACS, FUTA, Nigeria, for their support during this study.

I also acknowledge the contributions of the management and staff of Leibniz-Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany, for hosting, mentoring, and providing resources during and after my stay in Germany.

I also acknowledge the contribution of the management and staff of SDD-UBIDS, especially colleagues in the Geography department, for their support and guidance during this study, especially Prof. Raymond Aabeyir and Prof. Issaka Kanton Osumanu. I also acknowledge the guidance and trust of my former head of department, Prof. Kenneth Peprah.

I acknowledge the support, love, and care of my family and friends, my mother, Mrs. Cynthia Agalisilie Akum, and my late father, Mr. Akum Adumzuai, for their love, prayers, and support. I also thank my siblings, Peter, for his love and generosity, Philip, Amanda, Betty, and Beatrice for their love, care and prayers. Thank you, Clothilda Zaabe, Joseph Asuik, Raymond Achiyaale, and Daniel Amoatika, for your friendship, prayers, and love, especially during this period of my study. I also acknowledge the contributions of Christian mentors and ministry members of Catholic Charismatic Renewal in Wa, Ghana, especially my brother and friend Prof. Bernard A. A. Akanbang, Dr. Enoch Kosoe.

Finally, I would like to thank all WASCAL comrades: Nehemiah John Gyuk, Francis Oussou, Diarra Diouf, Nadege Millogo, Wilson Illori, Aissatou Sitta, Yeo Kolotioloma, Moussa Maiga, and Dametoti Yamoula.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AEJ –African Easterly Jet

AMM – Atlantic Meridional Mode

AMO – Atlantic Multi-decadal Oscillation

CAPE – Convective Available Potential Energy

CBH – Cloud Base Height

CDIR – Clear-sky Direct Solar Radiation at the Surface

CIN – Convective Inhibition

CP – Convective Precipitation

CRR – Convective Rain Rate

CVH – High Vegetation Cover

CVL – Low Vegetation Cover

D2M – 2 m Dewpoint Temperature

E – Evaporation

EOF – Empirical Orthogonal Function

ERA5 – ECMWF Reanalysis version 5

GMet – Ghana Meteorological Agency

GPM-IMERG – Global Precipitation Measurement – Integrated Multi-satellite Retrievals for GPM

I10FG – Instantaneous 10 m Wind Gust

IEWS – Instantaneous Eastward Turbulent Surface Stress

INSS – Instantaneous Northward Turbulent Surface Stress

IOD – Indian Ocean Dipole

KX – K-index

LAI_HV – Leaf Area Index, High Vegetation

LAI_LV – Leaf Area Index, Low Vegetation

LCC – Low Cloud Cover

LightGBM – Light Gradient Boosting Machine

Lon – Longitude

LSRR – Large-Scale Rain Rate

LSP – Large-Scale Precipitation

Lat – Latitude

MAE – Mean Absolute Error

MCC – Medium Cloud Cover

MEI – Multivariate ENSO Index

MER – Mean Evaporation Rate

ML – Monsoon Layer

MTPR – Mean Total Precipitation Rate

MSL – Mean Sea Level Pressure

NAO – North Atlantic Oscillation

NWP – Numerical Weather Prediction

P63.162 – Vertical Integral of Total Energy

P79.162 – Vertical Integral of Divergence of Cloud Liquid Water Flux

P88.162 – Vertical Integral of Eastward Cloud Liquid Water Flux

PDO – Pacific Decadal Oscillation

PEV – Potential Evaporation

PNA – Pacific North American Pattern

RF – Random Forest

SHAP – SHapley Additive exPlanation

SI10 – 10 m Wind Speed

SKT – Skin Temperature

SLHF – Surface Latent Heat Flux

SOI – Southern Oscillation Index

SP – Surface Pressure

SST – Sea Surface Temperature

STL1 – Soil Temperature Level 1

STL2 – Soil Temperature Level 2

STL3 – Soil Temperature Level 3

STL4 – Soil Temperature Level 4

T2M – 2 m Temperature

TCIW – Total Column Cloud Ice Water

TCRW – Total Column Rain Water

TCSLW – Total Column Supercooled Liquid Water

TCC – Total Cloud Cover

TCW – Total Column Water

TCWV – Total Column Vertically-Integrated Water Vapour

TISR – Top of Atmosphere Incident Solar Radiation

TNA – Tropical North Atlantic Index

TOTALX – Total Totals Index

TP – Total Precipitation

TSA – Tropical South Atlantic Index

TVH – Type of High Vegetation

U10 – 10 m U Wind Component

V10 – 10 m V Wind Component

WA – West Africa

WHWP – Western Hemisphere Warm Pool

XGBoost – Extreme Gradient Boosting

CHAPTER ONE

INTRODUCTION

1.1 Background

Thunderstorm frequency is the number of thunderstorms that occur over a given location during a day, month, or year. Thunderstorms are very intense over the tropics and provide much of the precipitation needed for agriculture and other economic activities (Niang *et al*., 2014; Mul *et al*., 2015). Large-scale diagnostics applied to global climate models have suggested that the frequency of thunderstorms and their intensity is likely to increase in the future (Taszarek *et al*., 2021). The Intergovernmental Panel on Climate Change (Solomon *et al*., 2007) predicts that in the short term, the number of extremely dry and wet years will increase during the present century due to climate variability and change.

Cumulonimbus clouds are known to provide the majority of rainfall over West Africa (WA). Maranan *et al*. (2018) reported that organised thunderstorms contribute about 71% of Soudanian rainfall and 56% of coastal West African rainfall, with an increasing contribution northward. The prediction of these thunderstorms over WA is very difficult, as numerical weather prediction (NWP) models are not able to predict accurately thunderstorm precipitation intensity, frequency, and timing (Laing *et al*., 2011), and NWP models are particularly shown to be very poor over WA (Parker and Diop-Kane, 2017) because of limits of parameterisation of physical processes.

The devastating effects of thunderstorms have been experienced in many cities and towns all over the world, particularly in West Africa (Balogun *et al*., 2022). Convective weather often has a high impact, such as strong winds, heavy precipitation, and flooding. These can lead to economic losses such as damage to electric power networks, agriculture, property, and sometimes loss of life (Tinmaker *et al*., 2017).

The agricultural and hydropower sectors in West Africa, among many others, benefit directly from convective activities that provide much-needed rainfall. On the other hand, convection could also negatively affect any of these sectors, as strong winds and heavy precipitation associated with intense convection can destroy farm produce and power infrastructure such as electric poles and cables. Thus, sound information about the timing and frequency of thunderstorms and precipitation is urgently needed for planning and water resource management. There have been attempts to do this by the Ghana Meteorological Agency (GMet). One such attempt is the sub-seasonal to seasonal (S2S) forecast.

Although WA experiences many thunderstorms, its frequency, climatology, trend, and the climate drivers of thunderstorm frequencies are not well understood. There have been a number of studies that have related the occurrence of thunderstorms to global teleconnection patterns. For instance, Ghavidel *et al*. (2016) conducted a statistical analysis of the relationship between daily thunderstorm frequency and teleconnection patterns in Iran. They found that the total monthly frequencies of thunderstorms correlated well with teleconnections during some months (May, June, and October); however, the correlations were weaker and non-significant in other months (January, February, and July).

Campozano *et al*. (2018) found that a significant positive correlation exists between the mesoscale convective systems (MCSs) occurrence in the Plaute basin (Ecuador) and the Trans-Niño index. They also concluded that, at the eastern part of the Plaute basin, there is also a positive correlation between MCS variability and tropical Atlantic Sea surface temperature anomalies (SSTAs) south of the equator. Over Brazil, it has been reported that convective activities are enhanced during the occurrence of El Niño events compared to La Niña phases (Sátori *et al*., 2009; Pinto, 2015). Pinto (2015) also found that thunderstorm activities over Brazil are influenced in some way by the Tropical North Atlantic (TNA) and Tropical South Atlantic (TSA). Pinto (2015) concluded that the influence of ENSO, TNA, and TSA on thunderstorm occurrences is different in different parts of Brazil.

The influence of teleconnections on European thunderstorm activities has also been investigated. For example, Piper and Kunz (2017) investigated how the North Atlantic Oscillation (NAO) impacts convective activities over parts of Europe. They observed an increase in convective activities over most parts of their study area when the NAO is in a strongly negative phase. Thus, the negative phase of NAO generally enhances thunderstorm activities, while positive NAO suppresses convective activities. Piper *et al*. (2019), in assessing the large-scale drivers of thunderstorm activities over Central and Western Europe using lightning data, observed that convection-favoured environments occur ahead of upper-level troughs. Their work shows that convective environments are influenced and impacted by teleconnection patterns and sea surface temperatures (SSTs). For instance, a positive East Atlantic Pattern (EA) increases the frequency of troughs over Western Europe and leads to an increase in convective-favouring environments and, hence, a likely increase in convective activities over the region. West African climate is influenced by sea surface temperature (SST) and climate indices in the Atlantic (Lamb, 1978; Vizy and Cook, 2001; Rowell, 2013), Indian Ocean (Lu and Delworth, 2005), and Pacific Ocean (Janicot *et al*., 2001). Climate indices are included as features to test if they are among the primary drivers of monthly thunderstorm frequencies in Ghana.

Variables other than climate indices, such as ERA5 reanalysis variables and geographic coordinates, were added as features to help improve the performance of the thunderstorm frequency prediction model. Stability indices such as convective available potential energy (CAPE), convective inhibition (CIN), and the K-index have previously been used to predict thunderstorms (Ukkonen and Mäkelä, 2019). Forecasting convective weather often includes CAPE (Brooks *et al*., 2007). CAPE appears to be very sensitive to variations in the temperature profile (Ziv *et al*., 2009). Convective inhibition (CIN) is also used in analysing convection variability (Riemann-Campe *et al*., 2009). The K-Index is a combination of moisture availability and atmospheric instability (Harats *et al*., 2010).

Also, this current work aims at developing an ML model that predicts monthly rainfall amounts in West Africa one year ahead for seasonal forecasting. Machine learning models that handle large datasets efficiently with minimal computing resources are ideal for West African national meteorological offices and individuals without high-performance infrastructure. Then SHapley Additive exPlanation (SHAP) plots were used to understand model behaviour for the purposes of understanding the major drivers of monthly rainfall over West Africa, knowing the interaction among features. The significance of such a model and approach includes: first, the developed model will help in planning for agricultural activities, surface water management, and many other socio-economic activities in West Africa. Second, the use of SHAP will help in understanding the most important features in the model; this will also help in building trust and accountability in the model. Third, the developed model can be used by national meteorological offices in West Africa as part of their sub-seasonal to seasonal (S2S) forecasts.

Bochenek and Ustrnul (2022) provided an up-to-date perspective on the application of machine learning (ML) in weather and climate. They reported that the most frequent algorithms used are Deep Learning Networks, Random Forests (RF), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and XGBoost. de Burge-Day and Leeuwenburg (2023), making a review of the use of ML in weather and climate modelling, concluded that they have been successfully applied in the field. Some of these applications include the replacement of sub-grid scale parameterizations, the solving of resolved equations, and model replacement. Elsewhere, ML models have successfully been applied in thunderstorm predictions (Ahijevych *et al*., 2016; Ukkonen and Mäkelä, 2019), and the forecast of convection often includes CAPE (Brooks *et al*., 2007).

Features for the thunderstorm and precipitation models are climate indices and ERA5 data, and were lagged by a year. The ERA5 reanalysis atmospheric features were passed through a low-pass filter. The models will serve the following purposes: First, the precipitation and thunderstorm models will help in planning for many sectors in Ghana, like agricultural and hydropower sectors and disaster preparedness. Second, since the GMet's S2S forecasts mostly do not include forecasts of thunderstorms, this work will help by providing this important parameter in the S2S of Ghana's forecasts.

The drivers of thunderstorm variability over West Africa are not well understood, even though they give the majority of the rainfall over the area. The problem is further complicated by a lack of complete observations of thunderstorm frequencies. Therefore, an understanding of the variation of thunderstorms over vulnerable WA and how oceanic and atmospheric interactions, like large-scale flows and global teleconnection patterns, drive these variations is crucial, as this will help improve their forecasts and representation in NWP models. This will help in disaster risk management, for instance, in cases of flash floods, and will help most local farmers and domestic and industrial activities that rely on rainfall amounts and distributions. Lightning data is particularly appealing for flash flood forecasting because of the very short lag times between lightning observation and operational availability (Tapia *et al*., 1998). A better understanding of teleconnection drivers of thunderstorms can also help local forecasters make a more precise prediction of the sub-seasonal to seasonal (S2S) forecast using the indices of the teleconnection patterns identified.

1.2 Problem Statement and Justification

Directly, precipitation and convection weather types are destructive and cost great economic losses and loss of life; they could also provide needed rainfall for many economic activities like rain-fed agriculture, hydropower generation, and surface water management. Indirectly, convective weather affects the earth's climate by producing clouds that affect the energy budget and transport heat and moisture in the vertical. Although organised thunderstorms are known to contribute the majority of rainfall in West Africa, the prediction and drivers of variability in thunderstorm occurrence over West Africa are poorly understood. Furthermore, accurate models that forecast monthly precipitation over West Africa a year ahead of time are nonexistent.

It is therefore important to predict and understand the drivers of thunderstorm and precipitation variability in West Africa, as this will help improve the forecast accuracy of these systems. The forecasts of these systems are further complicated by global and regional climate models' (GCMs and RCMs) inability to represent convection realistically. These models have been shown to have systematic biases in convection intensity distribution (Kyselý *et al.*, 2015) and diurnal cycle of precipitation (Dirmeyer *et al.*, 2011). Such biases impact the overall skill of the models (Sherwood *et al.*, 2014). To overcome such biases, as Laing *et al.* (2011) pointed out, NWP models should be able to reproduce convection precipitation intensity, frequency of occurrence, and duration of occurrence. Another way to overcome this issue is to run models like RCMs at high resolutions, such that deep convection is partially resolved and will lead to improvements in the representation and forecast of deep convection, like the convection-permitting model that was run successfully by Knist *et al*. (2018). However, the computational cost of running such a model is very high. Another way is to have a good parameterisation scheme, which requires a better understanding of convection processes and their variability. A suggested approach used in this work is to model thunderstorms and precipitation using machine learning (ML) algorithms. Especially ML models that require fewer computational resources so that they can easily be used by individuals or national meteorological offices that may not have high-performance computers.

The Intergovernmental Panel on Climate Change (Solomon *et al*., 2007) predicts that in the short term, the number of extremely dry and wet years will increase during the present century due to climate variability and change. The weather phenomena associated with thunderstorms are generally destructive and sometimes accompanied by extreme precipitation. A good understanding and prediction of this will help not only the local forecaster but also government policymaking and local farmers, as thunderstorm frequencies, precipitation amounts, and extreme weather-related hazards can be anticipated beforehand. Monthly thunderstorm frequencies over Ghana and precipitation over West Africa will be used as target variables, and the relationship between these and ERA5 reanalysis and teleconnection patterns will be investigated. Using flash count data from the TRMM satellite, Balogun *et al*. (2022) showed that morning and daytime hours are dominated by stratiform (non-convective type) precipitation and small system sizes, whereas evening and nighttime are dominated by convective and larger-size systems. Many economic activities, like hydropower generation and surface water resource management, depend very much on rainfall. It is therefore important to understand the drivers of thunderstorm activities over the sub-region. As stated earlier, Laing *et al*. (2011) pointed out, one of the most important steps in improving forecasts of convective systems is when NWP models can reproduce their precipitation intensity, their frequency of occurrence, and their duration of occurrence, and a good parameterisation scheme, which requires a better understanding of convection processes and their variability. This proposed work then seeks to help broaden the knowledge of thunderstorm variability over WA and its relationship with global teleconnections. Studying thunderstorm variability and the drivers of this variability will help improve understanding and parameterisation schemes over West Africa.

The first purpose of the present study is to study climatology and trends in thunderstorm frequency over Ghana using station monthly thunderstorm frequency data to investigate how thunderstorm frequencies are related to and impacted by climate variables. First, multivariate non-linear regression analysis was done using a machine learning approach to model thunderstorm frequencies and identify the main drivers of thunderstorms. Second, model behaviour was analysed to understand how the features impact thunderstorm frequencies in Ghana. Third, trend analysis was performed on the thunderstorm data and on the identified drivers. Since the regression model used in this study can handle non-linear relationships (Ke *et al*., 2017), the data was not detrended for the regression model. Identification of the drivers of thunderstorm variability aims at improving modelling and forecasting of thunderstorm frequencies.

Also, this study seeks to develop a machine learning model that can predict monthly thunderstorm frequencies in Ghana one year in advance. This was done by applying three different ML models using the same dataset, comparing the results, and combining them to form an Ensemble model. The study also aims to develop an ML model that predicts monthly rainfall amounts in West Africa one year ahead for seasonal forecasting. Then SHapley Additive exPlanation (SHAP) plots were used to understand model behaviour for the purposes of understanding the major drivers of monthly rainfall over West Africa, knowing the interaction among features, and for model trust, accountability, and model improvement (Lundberg and Lee, 2017). The significance of such a model and approach includes: first, the developed model will help in planning for agricultural activities, surface water management, and many other socio-economic activities in West Africa. Second, the use of SHAP will help in understanding the most important features in the model; this will also help in building trust and accountability in the model. Third, the developed model can be used by national meteorological offices in West Africa as part of their sub-seasonal to seasonal (S2S) forecasts.

An alternative approach that could offer a breakthrough to the difficulty of forecasting thunderstorm occurrences is to learn the complex nonlinear relationships governing subgrid convection directly from observational and numerical modelling data. This can be done using machine learning (ML). The nonlinearity of deep convection initiation makes this problem a good candidate for machine-learnt classification models. This thesis falls in line with WASCAL's priority research theme two (PRT 2), which deals with risks and vulnerability to climate extremes. This work also falls in line with the sustainable development goals (SDG): SDG 3 (good health and wellbeing), SDG 6 (clean water and sanitation), and SDG 13 (climate action).

The following are the research questions:

(i) what are the trends and drivers of thunderstorm variability over Ghana?

(ii) can predictive models be developed and evaluated for thunderstorm frequencies in Ghana for seasonal forecasting?

(iii) can a predictive model be developed and evaluated to predict monthly precipitation over West Africa for seasonal forecasting?

(iv) can machine learning models be used to study oceanic and atmospheric drivers of precipitation over West Africa?

1.3 Aim and Objectives

This study aims to analyse the drivers and develop predictive models for thunderstorms and precipitation over West Africa for sub-seasonal to seasonal (S2S) forecasting.

The specific objectives of the research are to:

(i) analyse trends and drivers of thunderstorm variability over Ghana;

(ii) develop and evaluate predictive models for thunderstorm frequencies in Ghana for seasonal forecasting;

(iii) develop and evaluate a predictive model that predicts monthly precipitation over West Africa for seasonal forecasting; and

(iv) analyse oceanic and atmospheric drivers of precipitation over West Africa.

1.4 Innovation

This study develops, for the first time, predictive models for thunderstorm frequencies over Ghana and precipitation over West Africa for seasonal forecasting using machine learning. The application of machine learning allows for the use of partial dependence plots (PDPs) and Shapley additive explanation plots (SHAP) in the study of the nonlinear drivers of precipitation and thunderstorms.

1.5 Structure of the thesis

This thesis is organised into five chapters. Chapter one gives the background of the study, a statement of the problem, and the aim and objectives of the research. The second chapter reviews the relevant literature. Discussed in this chapter are the known drivers of the West African climate, how global teleconnections relate to the West African climate, and how machine learning algorithms have been used in climate studies. Chapter three discussed the data and methods used to achieve the results of this thesis. Chapter four contains the results and the discussion of these results. Chapter five concludes the research with recommendations.

CHAPTER TWO
LITERATURE REVIEW

2.1 WACS and their drivers / Large scale flow over West Africa

The West African Climate Systems (WACS) is dominated by the West African monsoon (WAM). Therefore, a good understanding of the WAM will greatly improve the forecast of weather phenomena over the region. The WAM occurs when the boreal wintertime dry north-easterly surface winds are replaced by south-westerly surface winds during the boreal summertime over the region. The south-westerly winds emanate from the tropical Atlantic and are therefore moisture-leading winds and greatly influence the annual rainfall over the West African region (Fink *et al*., 2017). Thus, the annual rainfall variability of West Africa depends on the variation of the WAM. With much of the region's agriculture being rain-fed agriculture, the rainfall variability could greatly influence agricultural productivity and other economic activities.

Hamilton *et al*. (1945) were part of the pioneers of the four weather zones concept over West Africa. As Fink *et al*. (2017) pointed out, there have been some advances in our understanding of the West African monsoon, but the 'concept of the four weather zones' is still a basic approach to the West African monsoon. Figure 2.1 shows the traditional four zones over West Africa during the boreal summer month of July, adopted from chapter one of the forecasters' handbook.

The northernmost zone is zone A. This zone is north of the intertropical convergence zone (ITCZ) and is therefore mostly dry. The ITCZ is a zone that separates, to the north, the dry north-easterlies, and to the south, the moist south-westerlies. Zone A is mostly dry with rare precipitation, and the temperature is mostly high in this zone and could exceed 40 °C. The West African surface heat low (area of low pressure) is found in this zone.

To the immediate south of zone A is zone B. The latitudinal coverage of zone B is very narrow. Although there is monsoon flow within this zone, the monsoon layer is very shallow. Precipitation could occur here, but the amounts are very low.

Zone C is the most weather-active zone, with a high frequency of weather activities like thunderstorms and precipitation. Present in this zone are the mid-level African easterly jet (AEJ) and upper-level tropical easterly jet (TEJ).

Zone D mostly has shallow and mid-level clouds with fewer convective clouds and convective activities, except in some parts of the Guinea Coast that have rainfall. In this zone, the 'little dry season' occurs during July or August. Of course, the four zones are not stationary. They migrate southward into the Atlantic, leaving only zones A and B on the continent around January (Fink *et al*., 2017). The thermodynamic differences between the hot and dry Sahara to the north and the cool and moist Atlantic to the south determine the seasonal variation of the WAM. The climate of West Africa is influenced by many synoptic features. Some of which are described below.

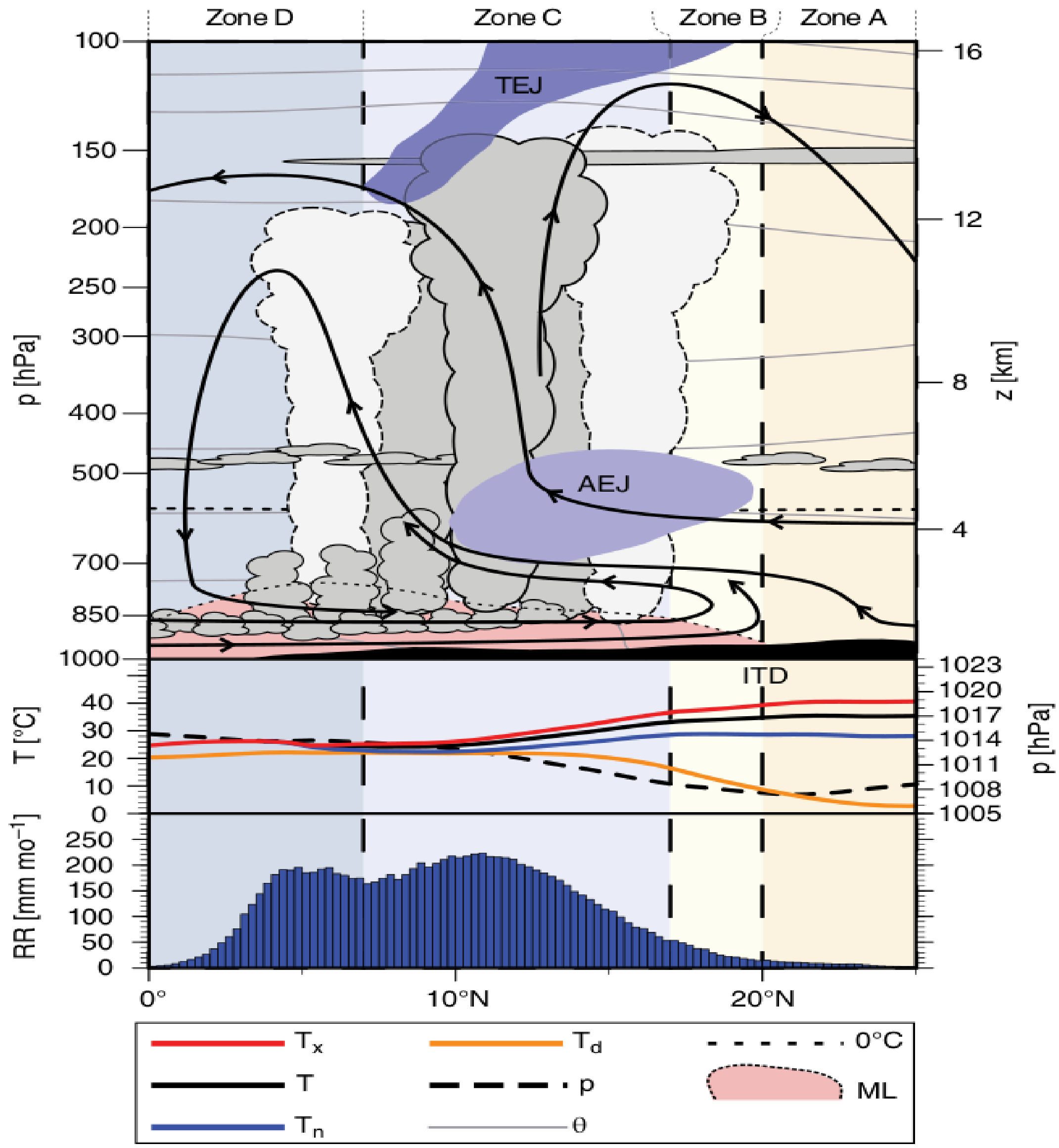

Figure 2.1: An illustration of the cross-section of the atmosphere between 10^0W and 10^0E. Showing the latitudinal positions of the four zones over West Africa. Adopted from chapter one of Meteorology of Tropical West Africa: The Forecasters' Handbook.

2.1.1 Saharan Heat Low

An area is said to have a heat low or thermal low when there is a low surface atmospheric pressure at the area, which is a result of the heating of the area; this results in aloft air divergence. The flow pattern around heat lows is generally weak cyclonic flow. During the summertime, the West African heat low is located over the Sahara, hence the name Saharan heat low (SHL). The SHL, which is a low mean sea level pressure, is centred around 22°N, 3°W. The SHL in West Africa is a major component of the West African monsoon system and is known to influence the West African monsoon (Parker *et al*., 2005; Peyrillé and Lafore, 2007; Lavaysse *et al*., 2009, 2016). Being a major component of the WAM system, the SHL has been found to influence rainfall over the Sahel region of West Africa (Lavaysse *et al*., 2010a; Evan *et al*., 2015). Like most climate systems, the SHL interacts with other components of the WAM system. For instance, when the SHL is intensified, it enhances low-level south-westerly monsoon flow and north-easterly flow from the desert (Lavaysse *et al*., 2016). This kind of flow could enhance convection and cloud cover (Parker *et al*., 2005). An intensified African easterly jet (AEJ) is seen during the intensification of the SHL (Lavaysse *et al*., 2010b). Evan *et al*. (2015) examined the role of SHL in Sahel drought recovery in the 1980s. In their work, a positive correlation was found between Sahel rainfall and the surface temperatures over the Sahara. Their work showed that the Sahel rainfall recovery is consistent with an intensified SHL. Lavaysse *et al*. (2016) made a study of the SHL trends and their impact on the West African climate, and they observed that the increase in temperature over the SHL area is higher than in the surrounding areas. They suggest that the influence of the SHL on the WAM system might have increased due to the high temperature gradient between the Gulf of Guinea and the Sahara.

2.1.2 Intertropical Discontinuity or Intertropical Boundary

The intertropical discontinuity (ITD) or intertropical boundary (ITB) is a zone of confluence of the dry and warm north-easterly winds from the desert and the moist and cool south-westerly winds from the Atlantic Ocean during the spring and summer months. The physical processes that control the SHL are the same that control the ITD on a day-by-day basis (Fink *et al*., 2017). It is known that during nighttime, the monsoonal flow is strengthened and flows into the SHL, thereby shifting the ITD poleward, making the ITD move several kilometres poleward during the night (Fink *et al*., 2017). Seasonally, the ITD migrates northward and southward, mimicking the apparent 'movement' of the sun, somewhat lagging the zenith angle of the sun (Lélé and Lamb, 2010). The position of the ITD and its accompanying features play very important roles in the seasonal to inter-annual climate variability over West Africa.

Some studies have tried to establish ITD latitude and rainfall relationship over West Africa. For instance, Ilesanmi (1971) observed a statistically significant positive relationship between the ITD latitudinal location and rainfall on monthly and annual time scales. Lélé and Lamb (2010), making a study of the meridional movement of the ITD and the role it plays in Sudan–Sahel rainfall, made the following observations: The northward movement of the ITD in the Sudan–Sahel from April to early August averages about 8.8 km per day. However, its southward retreat from mid-August to mid-November is faster, at 15.5 km per day. They observed that as ITD advances northward, the monsoon rainbelt also advances northward. The northern boundary of the rain belt, however, lags 100–250 km south of the ITD. They also noticed that during both the ITD northward advancement and retreat, the Sudan–Sahel rainfall is positively related to the ITD latitude.

The ITD is therefore an important component and driver of the climate system of West Africa. This has been shown to drive the rainfall of the region and influence convection as its position determines the type of wind flow.

2.1.3 African Easterly Jet

The WAM has been linked to the mid-tropospheric (600–700 hPa level) easterly flow that could reach a speed of 20 ms^{-1} around 15^0N (Parker *et al*., 2005; Balogun, 2024a; Balogun, 2024b). This mid-tropospheric easterly flow is termed the African easterly jet (AEJ). The flow of the AEJ is such that the zonal flow is essentially geostrophic while the meridional flow is mostly ageostrophic (Cook, 1999). West African wave development and propagation has a link with the AEJ (Burpee, 1972) and plays a role in the interannual variability of rainfall over West Africa (Newell and Kidson, 1984; Fontaine and Janicot, 1992). The AEJ has a jet core, and south of that core is where the African easterly waves (AEWs) form; here squall lines and maximum precipitation occur (Wu *et al*., 2009).

Cook's (1999) study of the AEJ using the global climate model (GCM) and reanalysis data showed that the two conditions that are necessary for the formation of AEJ over WA are insolation during summer and soil moisture gradients. Cook (1999) also observed that the AEJ's height is determined by the level at which the positive temperature gradient (induced by the surface) is replaced by the negative temperature gradient (as a result of the free atmosphere). In their study, they eliminated insolation, cloud, and SST distribution as possible sources of AEJ; they, however, observed that variations in SST and surface roughness influence the jet's position and intensity.

Wu *et al*. (2009) made a continuation of Cook (1999), using a model with higher resolution. In their work, Wu *et al*. (2009) acknowledge the contribution of soil moisture gradient in producing the AEJ, as earlier reported by Cook (1999). However, other factors that contribute to producing the AEJ other than soil moisture gradient are vegetation distribution, moisture gradients, and orography (Wu *et al*., 2009).

2.2 Global thunderstorm variability/ convective rainfall variability

The latter part of the 19th century saw an increase in the research on lightning when photography became popular (Uman, 2001). Lightning strikes can occur in thunderstorms, convective clouds, or stratiform clouds, and sometimes too from clear skies. Lightning in clear skies is possible because lightning strikes do not always take a vertical path to the ground. So, a lightning strike can occur some kilometres from the cloud of origin. However, the majority of lightning occurs in thunderstorms. Lightning can generally be divided into two types: cloud discharge (CD) and cloud-to-ground (CG). CG lightning causes much of the damage as it can destroy power plants and cause injuries and deaths.

Globally, lightning frequencies as estimated by satellites are 45 flashes/sec (Christian *et al*., 2003). The global lightning hot spots over the tropical landmasses, from the most active to the least active regions, are Africa, South America, and South Asia (Price, 2008). In trying to examine the differences in lightning activity between the two most active lightning regions of the world (Africa and South America), Williams and Satori (2004) found that Africa had higher lightning because it is hotter and drier than South America.

Figure 2.2 shows the global annual distribution of lightning flashes from the Global Hydrology Resource Centre. Generally, lightning flashes are more over the continents than over the ocean.

Most of the intense lightning occurs over continental tropics compared to other regions, as global lightning observations indicate that 90% of global lightning occurs over the continental tropics and the summer hemisphere (Price, 2009). Over continental tropics, the diurnal cycle of lightning is such that there are low lightning flashes during the local morning, but it reaches a peak during the local late afternoon (Price, 2009). This diurnal cycle of thunderstorms is consistent with the diurnal cycle of surface temperature over the tropics, however, with a time lag. Over the ocean, the diurnal cycle of lightning activity is less pronounced (Price, 2009).

Convective clouds from the oceans have moderate updraft velocities of less than 10 ms^{-1}; however, those that emanate from land may have updraft velocities of up to or more than 15 ms^{-1} (Stith *et al*., 2004). The strong updraft over land is enough to produce more lightning (Zipser and Lutz 1994). These differences in updraft velocities between the land and ocean are one of the reasons responsible for the high lightning flashes over land as compared to the ocean; this is because updraft intensity has been found to play a role in the electrification of thunderstorms and lightning frequencies (Baker *et al*., 1995; 1999). Land and ocean have different thermal capacities; this makes the land heat faster than the ocean, resulting in longitudinal differences in the orientation of the intertropical convergence zone (ITCZ) over the land and ocean in the tropics. This difference is also a reason for the differences in convective activities observed between continents and oceans over the tropics (Lemone and Zipser, 1980; Lucas *et al*., 1994).

Much of tropical rainfall is associated with deep convection (Houze, 1993, chapter 7). Newly formed convective clouds contain air that has been lifted to its “level of free convection” and therefore begin to ascend freely. As it ascends, it draws on the convective available potential energy (CAPE) that is available in the conditionally unstable midtroposphere (Houze, 1993). Over the tropics, about half of rainfall is heavy and short-lived from the updrafts from convective clouds. The other half of rainfall in the tropics is from spreading anvil clouds that fall gently in mesoscale rain areas (Schumacher and Houze, 2003). The frequency of the occurrence of lightning can be a proxy for the occurrence of strong updrafts over land and ocean (Virts *et al*., 2013) and associated phenomena, like flash floods (Tapia *et al*., 1998). Lightning is modulated by the diurnal cycle because daytime heating of the boundary layer air over land can increase the CAPE, and if there is sufficient lifting to trigger convection (Virts *et al*., 2013).

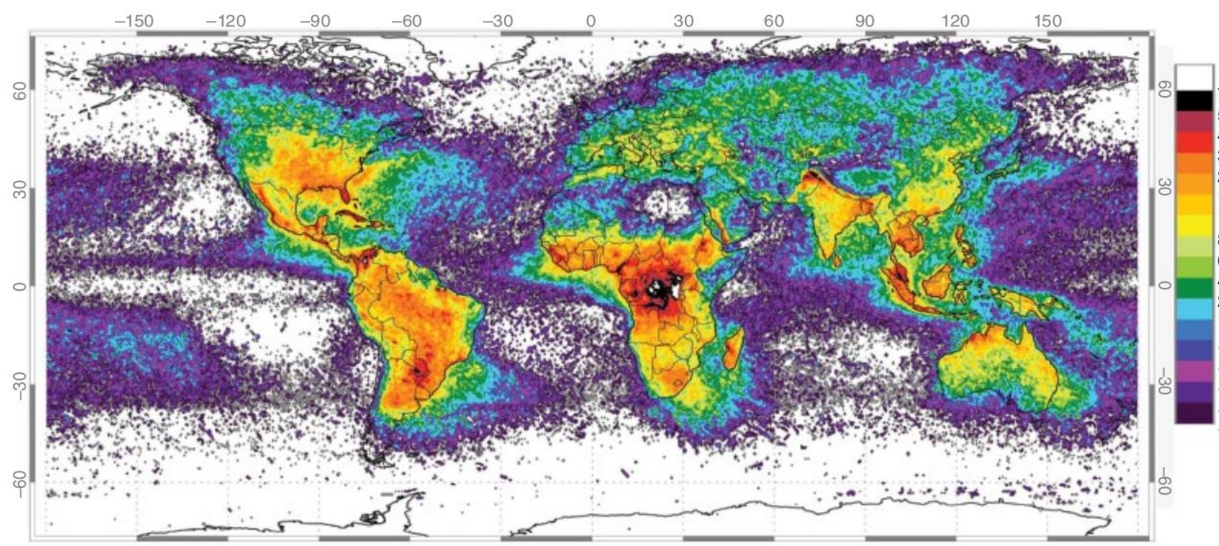

Figure 2.2: Annual number of lightning flashes by pixel over the 1995–2003 period. Source: Global Hydrology Resources Centre.

To overcome the lack of historical data on lightning and thunderstorms over Africa, Harel and Price (2020) constructed an empirical model of thunderstorm clusters over the region using lifted index and specific humidity from 1948-2016. Their results show spikes of thunderstorm clusters from 1948 to the 1990s over Africa. However, from the mid-1990s to the mid-2000s, there has been a general increase in thunderstorm clusters; this rise continued and reached a peak level in 2010.

Over West Africa, convective clouds give the majority of rainfall. These convective clouds over West Africa are part of a global tropical circulation and are also driven locally by the surface and atmospheric conditions (Parker and Diop-Kane, 2017). Unstable conditions are almost easily generated over the tropics, including West Africa. This is because of the availability of moist energy that is gained from heated surfaces and the free atmosphere losing energy through radiative processes.

Roehrig *et al*. (2013), in examining the autocorrelation of Global Prediction Climatology Project (GPCP) precipitation anomalies that lagged at one (1) day, observed that Africa has a unique precipitation probability. Some places in Africa had negative lag correlation values. This indicates that if there is rainfall today, it is less likely that it will rain again tomorrow. There is no rainfall persistence over the region (Parker and Diop-Kane, 2017), unlike many other regions of the world (Roehrig *et al*., 2013). Sahel rainfall probability density functions show a non-Gaussian distribution with strong variability (Parker and Diop-Kane, 2017).

2.3 Machine learning algorithms in climate studies

Machine learning (ML) can be thought of as both an art and a science, where computers are programmed to learn from a set of data (Geron, 2017). The "programmed computer" can then describe the relationship that exists in the data (between the predictor and response) and make future predictions. Based on the amount of supervision the machine receives, ML can be classified into supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning.

In supervised machine learning, the data used in the training processes include a label (a desired outcome). Thus, for every observation made, there is an associated response. For example, predicting whether a particular day will have thunderstorms or not using past thunderstorm data (response) and surface or reanalysis data (predictors) to train the model. Examples of these types are logistic regression, linear regression, neural networks, k-nearest neighbours, decision trees, etc. However, unsupervised learning has no labels. Thus, the algorithms learn with no supervision, making this more challenging than supervised learning. An example of unsupervised learning is principal component analysis (PCA), k-means, kernel PCA, Apriori, etc. Semi-supervised learning is when the data used in training the computer has much of it being unlabelled and the other being labelled; an example is the deep belief networks (DBNs). In reinforcement learning, "the learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return" (Geron, 2017, pp. 13).

Generally, in ML data, labels (responses) that are quantitative are often referred to as regression problems, and those with qualitative responses are classification problems. Thus, when we are more interested in the amount of rainfall on a particular day, that will be a regression problem. But if we want to know whether it will rain on a particular day, that will be a classification problem.

Most atmospheric phenomena are chaotic and non-linear; this makes their prediction very difficult. For instance, deterministic methods and models are unable to capture well the nonlinearity of air pollutant concentration and their dispersion (Rybarczyk and Zalakeviciute, 2018). Machine learning approaches are increasingly becoming good tools for forecasting weather phenomena (Sønderby *et al*., 2020). Therefore, machine learning (ML) algorithms are now increasingly being used to study atmospheric or climate processes (Shabib Aftab *et al*., 2018, and the references therein).

Rybarczyk and Zalakeviciute (2018), making a review of ML in outdoor air quality modelling, concluded that most articles that focused on pollutant concentration estimation used ensemble methods or regression algorithms. However, those articles that focus on forecasting air quality use neural networks (NN) and support vector machines (SVM). They found that the precision of estimation methods is better and less variable than the forecasting methods. Chen *et al*. (2020), making a review of ML forecasts of tropical cyclones, concluded that ML has made some advances in the prediction of tropical cyclones' genesis, path, and intensity and in improving numerical forecast models. Foresti *et al*. (2019), using machine learning to nowcast the growth and decay of precipitation, found that artificial neural networks were able to learn and effectively reproduce long-term growth and decay of precipitation.

Azad *et al*. (2021) proposed two new ensemble empirical mode decomposition (EEMD) models, namely EEMD-ANN and EEMD-SVM, for the prediction of thunderstorm frequency over Bangladesh. They compared these two new models with ANN, auto-regressive integrated moving average (ARIMA), and SVM. Their study found that the proposed EEMD-ANN and EEMD-SVM had better root mean square error (RMSE) compared to ANN, ARIMA, and SVM. They then concluded that hybridising EEMD with conventional ML models could increase monthly thunderstorm frequency prediction. Shi *et al*. (2022), using machine lightning data and learning algorithms to simulate deep convection, concluded that lightning data, machine learning algorithms, and ERA5 reanalysis data could greatly enhance deep convection forecasting.

Mostajabi *et al*. (2019) used readily available meteorological parameters (surface wind, air temperature, relative humidity, and station-level air pressure) to train machine-learning algorithms for the prediction of lightning activities in a time frame from the present to 30 minutes into the future. The machine learning algorithm was able to learn from the data and make a good prediction of the occurrence of lightning within the time frame. Their results show that the ML model had at least a probability of detection (POD) of 71% at the different stations and different lead times (0-10 min, 10-20 min, and 20-30 min) of the study. The ML model also outperformed the persistence forecasting method, the electrostatic field method, and a model developed based on the CAPE threshold.

Different ML techniques have been shown to be good at different situations. For example, Zhang *et al*. (2016), comparing different ML models (Artificial Neural Networks (ANN), Adaptive Neuro Fuzzy Inference System (ANFIS), and SVM) in predicting rainfall, found that ANN performed better when trained with heavy rainfall data only, while ANFIS was better in typhoon events that are extreme. In another comparative study of rainfall prediction using Random Forest, Support Vector Machine, Naive Bayes, Neural Network, and Decision Trees over Malaysia, Random Forest has been shown to classify instances correctly with a small amount of training data (Zainudin *et al*., 2016). Therefore, there might not be one single ML model that is better than the others in all situations.

2.4 Empirical Orthogonal Functions

Empirical orthogonal functions (EOF) or principal components analyses (PCA) are unsupervised machine learning techniques or methods and are great tools for dimension reduction. This technique can reduce the dimension of an $n \times k$ data matrix X to a smaller dimension and still maintain the variance of the matrix. Thus, EOF is able to find a relatively small number of independent variables from the original data (say, matrix X) that convey much of the original information as possible without redundancy. The direction of the first principal component shows the direction in which the data matrix X varies most. The second principal component has the second highest variability within the data matrix or observation. However, the first principal component and the second principal component are said to be orthogonal to each other and the other components.

EOF can be used to produce variables that can now be used in supervised regression (an example is principal component regression), and it can also be used as a great tool for data visualisation (James *et al*., 2013). The principal component loading vector is unique, and so is the score vector, both "up to a sign flip" (James *et al*., 2013). Thus, two different programming packages should give the same loading vectors of the principal components, but the signs of the loading vectors might change.

After a PCA or EOF, one will not be interested in keeping or retaining all the principal components. There must be a rule that can be used to decide which principal components to retain and which to reject, such that the small number of principal components retained should be able to explain much of the variability within the data set. Generally, this is done by observing the scree plot and cutting off from the elbow of the scree plot. The elbow of the scree plot is the point where the proportion of variance explained by the principal components drops off (James *et al*., 2013). Another criterion for deciding the number of principal components to retain is what has been proposed by North *et al*. (1982): "If the sampling error of a particular eigenvalue is comparable to or larger than the spacing between the eigenvalue and the neighbouring eigenvalue, then the sample error, which is the EOF, will be comparable to the size of the neighbouring EOF.". Another way of deciding the number of principal components to retain is by deciding the percentage variance (e.g., 90%) and accepting the first principal components that altogether explain that percentage variance.

2.5 Large Scale Environments and Teleconnections relation with Thunderstorms

Early studies have proposed strong wind shear between lower and upper levels, lower tropospheric moisture, dynamical lifting, and conditional instability as environmental conditions that are favourable for convection or thunderstorm formation (Fulks, 1951; Miller, 1972). Mesoscale and synoptic scale processes are known to help in thunderstorm initiation (Cotton *et al*., 1983; Doswell, 1984; Rockwood and Maddox, 1988). However, wind shear plays an important role in determining the severity and longevity of an initiated storm (Weisman and Klemp, 1982; Schoenberg Ferrier *et al*., 1996; Takemi, 2007).

Liu *et al*. (2020), in examining the environments that are associated with the most intense thunderstorms over some thunderstorm hot spots (south-central United States, Sahel, Congo, Colombia, Himalayas, Argentina, and northwest Mexico) in the world, made the following findings: the low-mid troposphere wind shear is highest in the sub-tropics (south-central United States, Himalayas, and Argentina) and Sahel as compared to the other hot spots. They also found that of all the selected tropical hot spots, only the Sahel has a combined high CAPE and large wind shear; however, in relative terms, lower convective inhibition (CIN), higher CAPE, lower helicity, and weaker low-level wind shear are found in intense thunderstorms over the tropics compared to the sub-tropics. In comparing the environments for weak and intense thunderstorms over Sahel hot spots, they observed that intense thunderstorms over these hot spots have higher CAPE and are stronger with shear as compared to the weak thunderstorms. In the other tropical regions, intense thunderstorms have higher CAPE values as compared to the weaker ones, but the value of the differences is smaller compared to the differences in the subtropical hot spots and Sahel.

There have been a number of studies that have related the occurrence of thunderstorms and global teleconnection patterns. For instance, Ghavidel *et al*. (2016) made a statistical analysis on the relationship of daily thunderstorm frequency and teleconnection patterns over Iran. They found that the total monthly frequencies of thunderstorms correlated well with teleconnections during some months (May, June, and October); however, the correlations were weaker and non-significant in some months (January, February, and July). Campozano *et al*. (2018) found that a significant positive correlation exists between the mesoscale convective systems (MCSs) occurrence in the Plaute basin (Ecuador) and the Trans-Niño index. They also concluded that, in the eastern part of the Plaute basin, there exists also a positive correlation of MCSs variability with tropical Atlantic Sea surface temperature anomalies (SSTAs) south of the equator.

Over Brazil, it has been reported that convective activities are enhanced during the occurrence of El Niño than during La Niña events (Sátori *et al*., 2009; Pinto, 2015). Pinto (2015) also found that thunderstorm activities over Brazil are influenced in some way by the Tropical North Atlantic (TNA) and Tropical South Atlantic (TSA). Pinto (2015) then concluded that the influence of ENSO, TNA, and TSA is different on thunderstorm occurrence in different parts of Brazil. The influence of teleconnections on European thunderstorm activities has also been investigated. For example, Piper and Kunz (2017) looked at how the North Atlantic Oscillation

(NAO) impacts convective activities over parts of Europe. They observed an increase in convective activities over most parts of the study area when NAO is in a strongly negative phase. Thus, the negative phase of NAO generally enhances thunderstorm activities in the study area, while positive NAO suppresses convective activities.

Piper *et al*. (2019), in assessing the large-scale drivers of thunderstorm activities over central and western Europe using lightning data, observed that over Europe, convection-favouring environments occur ahead of upper-level troughs. Their work shows that convective environments are influenced and impacted by teleconnection patterns and sea surface temperatures (SSTs). For instance, a positive East Atlantic Pattern (EA) increases the frequency of troughs over western Europe and leads to an increase in a convective-favouring environment and hence a likely increase in convective activities over the region.

Sátori *et al*. (2009) used the multivariate ENSO index (MEI) and global lightning data to study how lightning activities vary during warm El Niño and cold La Niña phases. Although, on the diurnal time scale, dominant subsidence occurs over land (ocean) during the nighttime (daytime), on the ENSO time scale, the dominant subsidence occurs over land (ocean) during the El Niño (cold La Niña) phase (Sátori *et al*., 2009). Thus, on the ENSO time scale, lightning is enhanced over land during the El Niño phase of ENSO but rather enhanced over the ocean during the La Niña phase (Sátori *et al*., 2009).

The influence of different teleconnections on convective activities is diverse. However, there has not been a comprehensive study of teleconnection patterns influence on the variability of thunderstorms over West Africa, even though convective rainfall gives much of the rainfall over the sub-region.

2.6 GPM-IMERG usage over West Africa

Precipitation over West Africa is essential to many sectors of the West African economy, including but not limited to agriculture, hydropower generation, and water resources planning and management. Extreme and/or continuous precipitation could lead to flooding, loss of property, and sometimes even death. It is therefore necessary to predict the rainfall amount in the region in advance.

One of the main challenges to forecasting and research in the region is the issue of insufficient data. West Africa has a sparse rain gauge network, and this greatly hampers rainfall prediction modelling. One way to overcome this challenge is to use satellite data to monitor rainfall in the region. The Integrated Multi-satellite Retrieval for Global Precipitation Measurement (GPM-IMERG) (Skofronick-Jackson *et al*., 2017) is one of such useful satellite datasets over West Africa. Although GPM-IMERG has been reported to have some biases (Maranan *et al*., 2020), it has been extensively used over West Africa. It has been shown to perform well on monthly and seasonal scales (Boluwade, 2020; Nwachukwu *et al*., 2020) and has also been shown to give realistic estimations of extreme precipitation events over Burkina Faso (Sahel) (Sanogo *et al*., 2022). Nwachukwu *et al*. (2020), in evaluating satellite-based precipitation products over Nigeria (West Africa), observed that on a monthly time scale, GPM-IMERG-F v.6 scored one of the best performances in Kling-Gupta efficiency (KGE), correlation, and bias in all the zones of Guinea, Savanah, and the Sahel.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study area and data

3.1.1 Thunderstorm data

To achieve the first two objectives, thunderstorm data from the 22 synoptic stations in Ghana (Figure 3.1) were used. The stations are divided into three zones in this study: coastal, middle, and northern. Monthly thunderstorm frequency data for all synoptic stations in Ghana were obtained from the Ghana Meteorological Agency (GMet) from 1990 to 2013. The total number of months with missing data over the 22 stations is 118 out of the 6336 total months. The Wa station had the highest number of missing data points in 13 months.

3.1.2 GPM-IMERG Precipitation and ERA5 reanalysis data

The GPM-IMERG final run version 7 monthly precipitation over West Africa (Figure 3.2a) from 2000-2023 is used as the target variable in this study, with a spatial resolution of 0.1^0 by 0.1^0. The IMERG precipitation data are available at https://disc.gsfc.nasa.gov/. To help predict the monthly thunderstorm frequency over Ghana and monthly rainfall amount over West Africa, ECMWF Reanalysis version 5 **(**ERA5) reanalysis data with a spatial resolution of 0.25^0 by 0.25^0 (Hersbach *et al*., 2020) was used. The ERA5 data are also available at https://cds.climate.copernicus.eu/. Climate indices were also used as part of the features in predicting the monthly rainfall amount over West Africa. The National Oceanic and Atmospheric Administration’s (NOAA) physical sciences laboratory website contains the climate indices used in this study: https://psl.noaa.gov/data/climateindices/list/. The features used are from 1999–2022. All the features used are detailed in Table 3.1.

Table 3.1: List of all features used.

Features	Abbreviation
Atlantic Meridional Mode	AMM
Atlantic Multi-decadal Oscillation	AMO
El Niño/Southern Oscillation (ENSO) region 3.4	Niño 3.4
El Niño/Southern Oscillation (ENSO) region 1+2	Niño 1+2
El Niño/Southern Oscillation (ENSO) region 3	Niño 3
El Niño/Southern Oscillation (ENSO) region 4	Niño 4
Multivariate ENSO index	MEI
Indian Ocean Dipole	IOD
North Atlantic Oscillation	NAO
Pacific North American Pattern	PNA
Tropical North Atlantic Index	TNA
Tropical South Atlantic Index	TSA
Southern Oscillation Index	SOI
Multivariate ENSO Index	MEI
Western Hemisphere Warm Pool	WHWP
Pacific Decadal Oscillation	PDO
Longitude	Lon
Latitude	Lat
Clear-sky direct solar radiation at the surface	CDIR
Convective precipitation	CP
10 m wind speed	SI10
Convective available potential energy	CAPE
Convective inhibition	CIN
Top of atmosphere incident solar radiation	TISR
2 m dewpoint temperature	D2M

Features	Abbreviation
2 m temperature	T2M
Cloud base height	CBH
Vertical integral of eastward cloud liquid water flux	P88.162
k-index	KX
10 m U wind component	U10
10 m V wind component	V10
Convective rain rate	CRR
High vegetation cover	CVH
Instantaneous 10 m wind gust	I10FG
Instantaneous eastward turbulent surface stress	IEWS
Instantaneous northward turbulent surface stress	INSS
Large-scale precipitation	LSP
Large-scale rain rate	LSRR
Leaf area index, high vegetation	LAI_HV
Leaf area index, low vegetation	LAI_LV
Low cloud cover	LCC
Low vegetation cover	CVL
Mean sea level pressure	MSL
Mean total precipitation rate	MTPR
Medium cloud cover	MCC
Total cloud cover	TCC
Total column cloud ice water	TCIW
Total column rain water	TCRW
Total column supercooled liquid water	TCSLW
Total column water	TCW
Total column vertically-integrated water vapour	TCWV

Features	Abbreviation
Total precipitation	TP
Total totals index	TOTALX
Type of high vegetation	TVH
Vertical integral of divergence of cloud liquid water flux	P79.162
Vertical integral of eastward cloud liquid water flux	P88.162
Vertical integral of total energy	P63.162
Evaporation	E
Mean evaporation rate	MER
Potential evaporation	PEV
Skin temperature	SKT
Soil temperature level 1	STL1
Soil temperature level 2	STL2
Soil temperature level 3	STL3
Soil temperature level 4	STL4
Surface latent heat flux	SLHF
Surface pressure	SP

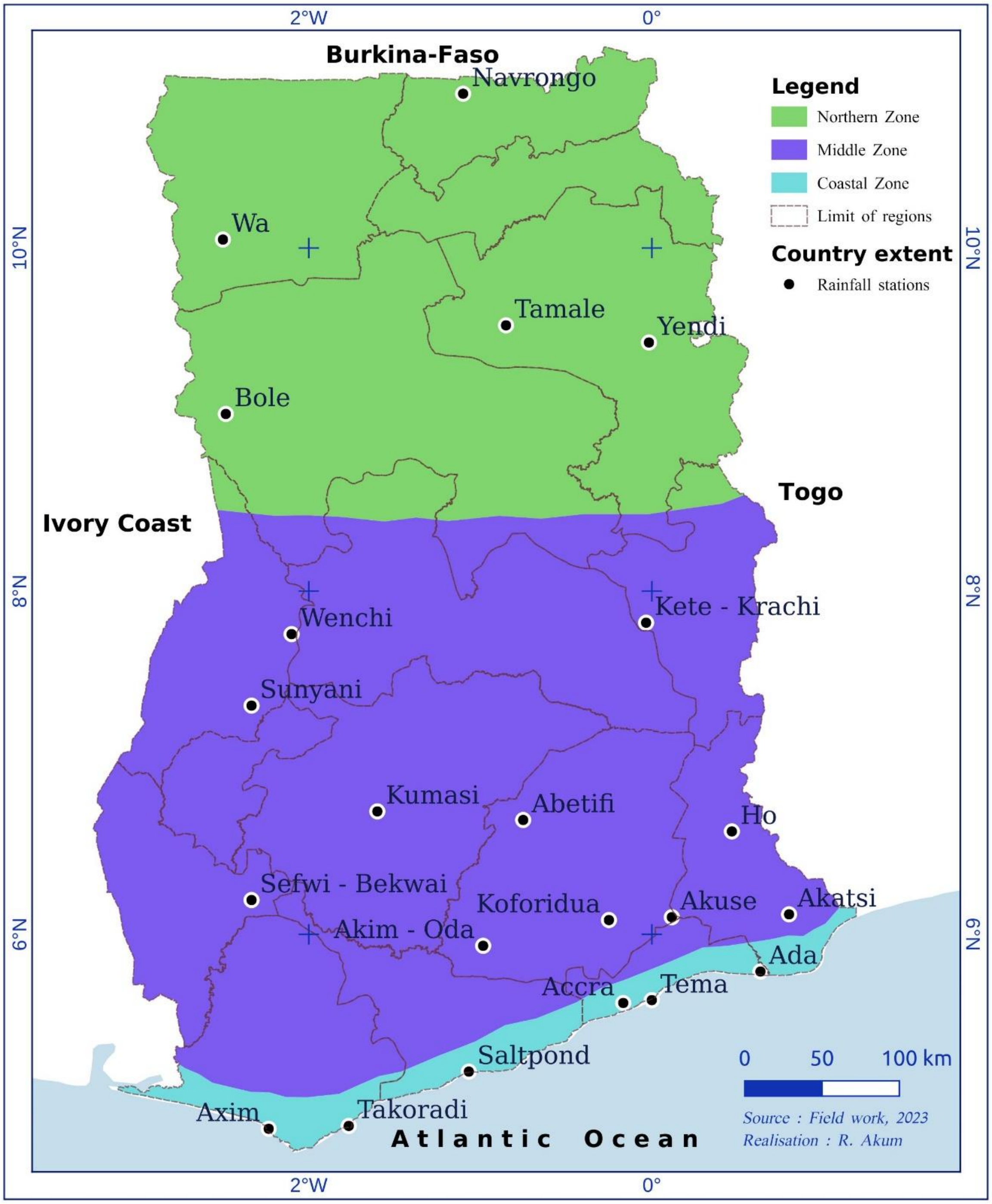

Figure 3.1. Map showing the locations of the stations used in this study. Colours indicate different climate zones of Ghana.

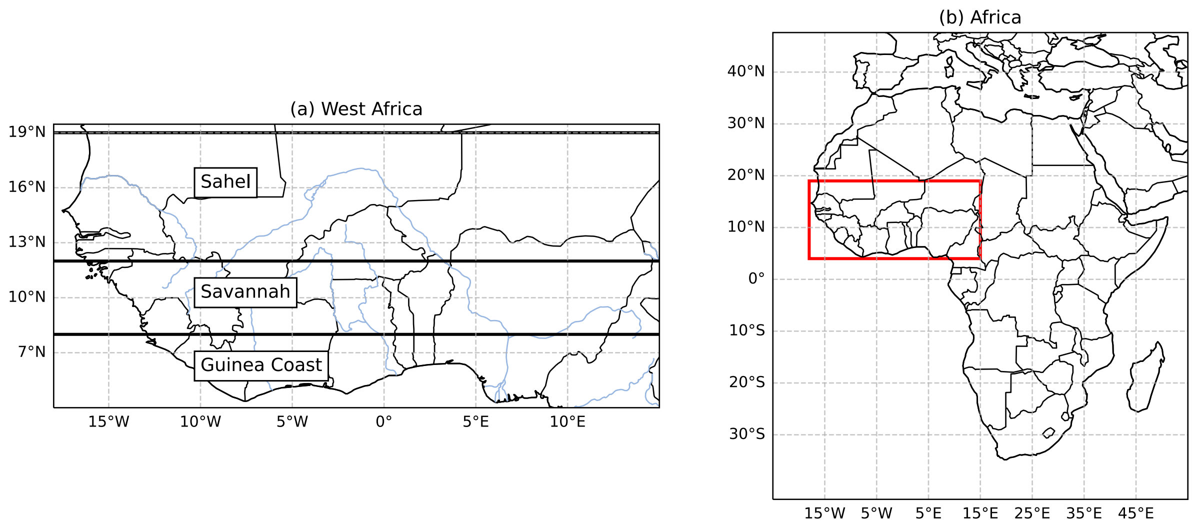

Figure 3.2. Map of West Africa, showing the demarcated climatic zones of West Africa.

3.2 Methods

3.2.1 Trend Analysis

The Mann-Kendall test (Mann, 1945; Kendall, 1948), which is often used to test for trends in hydro-climatological time series (Oguntunde *et al*., 2012; Oguntunde *et al*., 2017), is used to test for the presence of trends in this analysis. Before applying the Mann-Kendall (M-K) test, the lag-one autoregressive AR(1) was removed from the data using a pre-whitening technique described by Yue *et al*. (2002) as follows:

a. The data was detrended by first estimating the trend (T_t), which is assumed to be linear, and subtracting the estimated trend from the original data (X_t) using the equation:

$$X'_t = X_t - T_t = X_t - bt \tag{3.1}$$

b is the slope of the trend.

b. The lag-1 serial correlation coefficient (r_1) of the detrended data (X'_t) is computed.
c. The AR(1) component is removed from the X'_t using the formula:

$$Y'_t = X'_t - r_1 X'_{t-1} \tag{3.2}$$

Trend analysis was performed on the same 1990–2013 annual data for all zones to allow comparison between zones. The M-K test is applicable in cases where the data values x of a time series can be assumed to obey the model.

$$x = f(t) + \sum t \tag{3.3}$$

Where $f(t)$ is a continuous monotonic increasing or decreasing function of time, and the Residual $\sum t$ can be assumed to be from the same distribution with a zero mean. The Mann-Kendall test statistic S is given as:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \mathrm{sgn}(x_j - x_k) \tag{3.4}$$

where n is the length of the time series $x_i \ldots x_n$, and sgn(.) is a sign function, x_j and x_k are values in years j and k, respectively. The expected value of S equals zero (E[S] = 0) for series without trend and the variance is computed as:

$$\sigma^2(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p - 1)(2t_p + 5)\right] \tag{3.5}$$

Here q is the number of tied groups, and t_p is the number of data values in p^{th} group. The test statistic Z is then given as:

$$Z = \begin{cases} \frac{S-1}{\sqrt{\sigma^2(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\sigma^2(S)}} & \text{if } S < 0 \end{cases} \tag{3.6}$$

As a non-parametric test, no assumptions as to the underlying distribution of the data are necessary. The Z-statistic is then used to test the null hypothesis, H_o, which states that the data is randomly ordered in time, against the alternative hypothesis, H_1, which states an increasing or decreasing monotonic trend. To estimate the slope of an existing trend, Theil-Sen's nonparametric method (Theil, 1950; Sen, 1968) is used. First, the f(t) is equal to:

$$f(t) = Q_t + B \tag{3.7}$$

where Q is the slope and B is a constant. To get the slope estimate Q, first, slopes of all data value pairs were calculated as:

$$Q_i = \frac{x_j - x_k}{j - k} \tag{3.8}$$

where j > k. If there are n values xj in the time series, we get as many as N = n (n -1)/2 slope estimates Q_i. Sen's estimator of slope is the median of these N values of Q_i. The estimated slope is multiplied by the number of years to indicate the magnitude of change over the period.

3.2.2 Establishing Oceanic Relations with Ghana Thunderstorm Frequency

The GMet station data were aggregated into three zones based on Figure 3.1, and anomalies were computed using the annual thunderstorm frequency data. The anomalies were computed using Equation 3.9. The computed anomalies for the annual thunderstorm frequency data for each zone were first detrended and correlated with detrended annual global SST anomalies (SSTAs), and correlation maps were created. Both datasets were detrended by fitting a linear regression to each time series, and the fitted values from linear regression models were subtracted from the datasets to obtain the detrended time series.

A correlation map for each zone is made to show the portions of the global SSTs that are correlated with the thunderstorm frequencies in the study area. The thunderstorm and SST data were from 1990–2013.

$$\text{Anomaly} = X_i - \overline{X} \tag{3.9}$$

Where X_i is the annual thunderstorm frequency for the i^{th} year and $\overline{X}$ is the zone-specific 1990-2013 mean.

3.2.3 Machine Learning Models

Tree-Based Models

Three ML algorithms that are based on decision trees are used in this work. These include Random Forest (RF), LightGBM, and XGBoost. For the training of RF, a random subsection of the data is used to train each decision tree. Because RFs fit each tree to a different subset of the data, in the next step, the variance is reduced by averaging the prediction of the model. The randomness and ensemble nature of RF help to increase the model's robustness and generalisation.

Gradient-boosted decision trees like LightGBM and XGBoost are built sequentially, and each tree models the errors in the trees built thus far (de Burgh-Day & Leeuwenburg, 2023). In boosting, each new tree is grown using information from the previously grown trees. Boosting algorithms are slow learners. LightGBM is a type of gradient-boosting decision tree that is efficient, is designed to be fast, requires lower memory usage, and still achieves high accuracy (Ke *et al*., 2017). XGBoost has been used extensively in weather and climate modelling with good results (de Burgh-Day & Leeuwenburg, 2023). The difference between the tree growth of the LightGBM and XGBoost models is that LightGBM grows trees leaf-wise, while XGBoost grows trees level-wise. Although both are similar, they could give different results. This work uses both since the purpose is to get the best model for thunderstorm prediction.

3.2.4 Machine Learning Workflow for thunderstorm models

For developing a model that predicts the monthly thunderstorm frequencies over Ghana, the features were lagged by a year, and the ERA5 reanalysis data passed through a low-pass filter, and the ML models were trained and developed. All models were run using Python. The ML workflow is outlined below:

1. The thunderstorm (feature) data were processed and divided such that January 1990 (1989)–December 2007 (2006) was used for model training, January 2008 (2007)–

December 2010 (2009) to validate, and January 2011 (2010)–December 2013 (2012) to test the model.

2. For each of the ML models (LightGBM, XGBoost, and RF), Recursive Feature Elimination (RFE) with cross-validation was used to find the most relevant features from a list of over 60 features. The RFE was performed from 'sklearn.feature_selection' (Pedregosa, 2011). 5-fold cross-validation was applied: during the 5-fold split, the data was not shuffled.
3. To reduce the risk of collinearity, features that were well correlated were not used in the model. No two features with a Pearson correlation coefficient greater than 0.5 or less than -0.5 were used. Figure 3.3 is the correlation matrix of all the features used in training the final models.
4. The 'BayesianOptimization' module from the 'bayes_opt' library was used to fine-tune the hyperparameters by defining a function for each model and evaluating their performance using the negative mean squared error (MSE) on the validation set. The hyperparameters that were fine-tuned for LightGBM, XGBoost, and RF and the selected values are presented in Table 3.2. To reduce the risk of model overfitting in the boosted models (Hastie *et al*., 2001), regularization methods was applied to the boosted models. These are also added to Table 3.2.
5. Each model was then evaluated using the independent data set aside for testing the models.
6. SHapley Additive exPlanations (SHAP) was applied to visualise model behaviour.

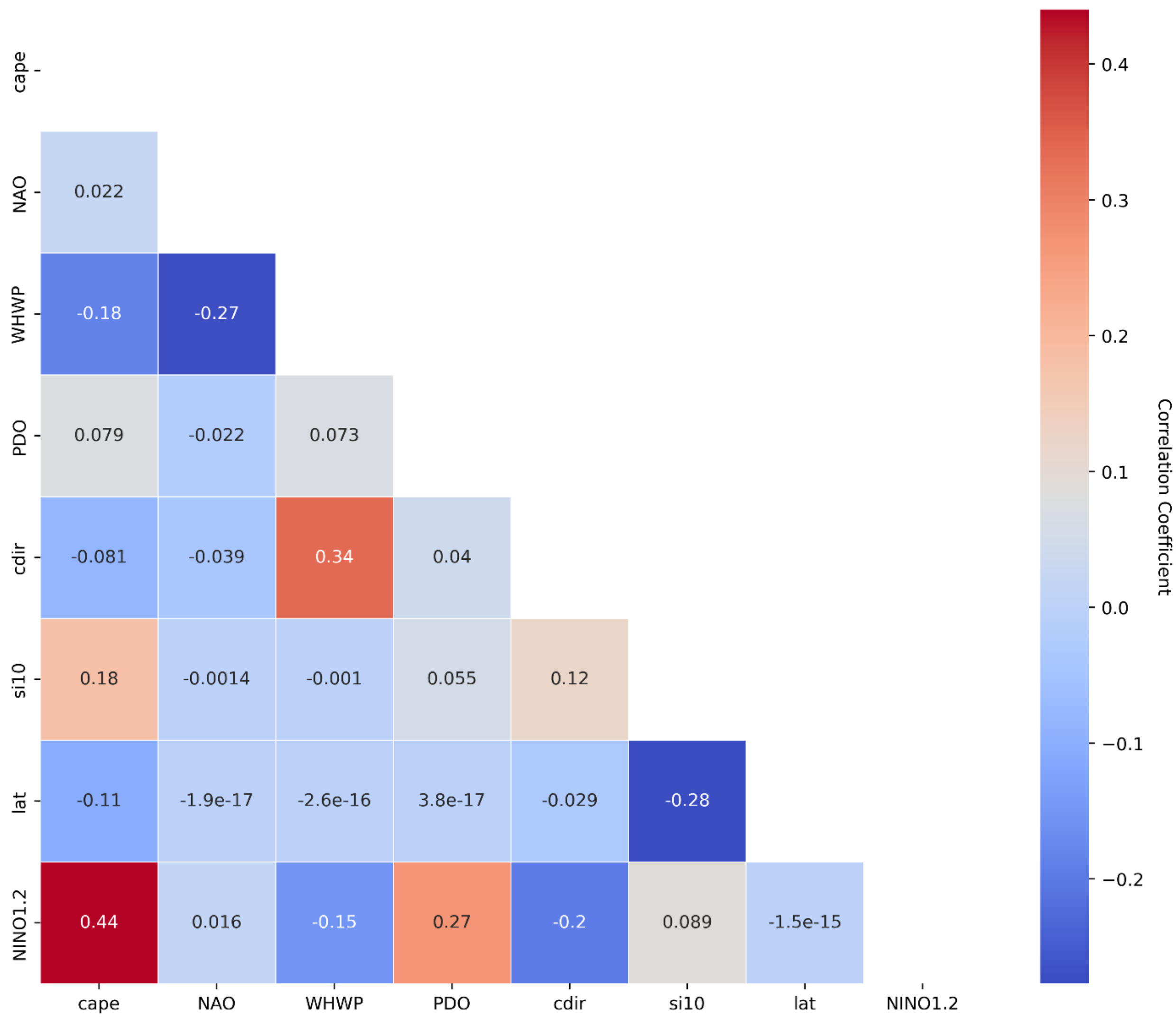

Figure 3.3: Correlation matrix of the selected features used in developing the ML models.

3.2.5 Machine Learning Workflow for precipitation models

The ML workflow for the precipitation models is as follows:

1. The precipitation data and features were preprocessed and manually divided into three subsets: the training set, the validation set, and the test set, as detailed in Table 3.4. The model was trained using the training set. Time series cross-validation was conducted with 'TimeSeriesSplit' from 'sklearn' on the training set to fine-tune the hyperparameters and ensure a reliable evaluation of the model. The validation set was also used to monitor the model's performance during training and to implement early stopping to prevent overfitting. Finally, the model's performance was evaluated using the independent test set.
2. The first three PC time series were used; these accounted for approximately 86% of the variance in the ERA5 reanalysis features. The following climate indices were added: Western Hemisphere warm pool (WHWP), Atlantic Meridional Mode (AMM), and El Niño/Southern Oscillation Regions 1+2 (Niño 1+2). The longitude and latitude were also used as features.
3. Bayesian optimisation was used to fine-tune the hyperparameters. The selected hyperparameter values are shown in supplementary Table 3.3. LightGBM and XGBoost are prone to overfitting if no regularisation method is applied (Hastie *et al*., 2009); therefore, the following hyperparameters are added to the model: L1 and L2 regularisations on weight and the minimum sum of instance weights, or the Hessian (min_child_weight).
4. Bayesian optimisation was conducted via the 'BayesianOptimization' module from the 'bayes_opt' library. A function was defined that trains each model with the

hyperparameters listed in supplementary Table 3.3 and evaluates their performance via the negative mean squared error (MSE) on the validation set. 100 iterations were performed during the optimisation.

5. Different LightGBM and XGBoost models were developed using the following objective functions: Tweedie and regression (mean squared error or L2 loss).

   a. Tweedie Objective: Suitable for continuous data with a mix of zero and positive values, such as precipitation data, bridging Gaussian and Poisson distributions.

   b. Mean Squared Error (regression) Objective: A standard regression objective that minimises the squared differences between observed and predicted values, aiming to predict the mean of the target variable.

6. An early stopping round of 10 was used during the model optimisation process.
7. Each model was then evaluated using the unseen data, and an ensemble model was developed and evaluated.
8. Monthly mean bias correction was applied using the training dataset's climatology to adjust the ensemble model outputs.

Table 3.2. The list of hyperparameters fine-tuned using Bayesian optimization for the thunderstorm models, showing the range the hyperparameters and the selected values for each model. Hyperparameters that were not used in a particular model are left empty in the column of that model.

Hyperparameter	Range	Selected value for LightGBM	Selected value for XGBoost	Selected value for RF
Number of Leaves	2 to 400	189		
Maximum depth	2 to 400	323	27	79
Learning Rate	0.001 to 0.6	0.30	0.02	
Number of Estimators	2 to 400	381	194	118
L1 Regularization Term (Alpha)	0 to 10	9.41	4.86	
L2 Regularization Term (Lambda)	0 to 10	3.15	0.23	
Minimum Sum of Instance Weight (Hessian) Needed in a Child	0.01 to 10	5.51	6.96	
Minimum Loss Reduction (gamma)	0 to 0.5		0.97	
Subsample Ratio of Training Instances	0.2 to 1		0.38	
Column Subsampling (colsample bytree)	0.2 to 1		0.66	
Minimum Number of Samples required to Split a Node (min_samples_split)	2 to 20			11
Minimum Number of Samples required to be a Leaf Node (min_samples_leaf)	2 to 10			8
Maximum Number of Features Considered for Splitting a Node (max_features)	0.1 to 1			0.39

Table 3.3. The list of hyperparameters fine-tuned using Bayesian optimization for the precipitation models, showing the range the hyperparameters and the selected values for each model for precipitation prediction. Hyperparameters that were not used in a particular model are left empty in the column of that model.

Hyperparameter	Range	LightGBM Regression	LightGBM Tweedie	XGBoost Regression	XGBoost Tweedie
Number of Leaves	2 to 1000	985	885		
Maximum depth	2 to 1000	591	752	12	16
Learning Rate	0.0001 to 0.5	0.11	0.27	0.37	0.10
Number of Estimators	2 to 2000	563	1166	316	314
L1 Regularization Term (Alpha)	0 to 20	0.94	0.00		
L2 Regularization Term (Lambda)	0 to 20	19.43	9.57		
Minimum Sum of Instance Weight (Hessian) Needed in a Child	0.001 to 20	5.43	0.001	3.12	11.49
Tweedie variance power	1.10 to 1.99		1.10		1.56
Subsample	0.5 to 1			0.53	0.99
Gamma	0 to 10			9.51	
Colsample bytree	0.5 to 1			0.69	0.96

Table 3.4: Division of features and target data into training, validation and test sizes.

Data Split	Features Data Periods	Target Data Periods
Training	6/1999 - 8/2016	6/2000 - 8/2017
Validation	9/2016 - 8/2019	9/2017 - 8/2020
Test	9/2019 - 8/2022	9/2020 - 8/2023

3.2.6 Model Evaluation

To evaluate the performance of the models, the following metrics were used: coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), model bias, index of agreement (d), and and Kling-Gupta efficiency (KGE). Below are the equations for computing these metrics. In all equations, P_i is the predicted i^th^ value, O_i is the observed i^th^ value. r is the Pearson correlation coefficient, α is the variability ratio, and β is the bias ratio.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(O_i - P_i)^2}{\sum_{i=1}^{m}(O_i - \bar{P})^2} \tag{3.10}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|O_i - P_i| \tag{3.11}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(O_i - P_i)^2 \tag{3.12}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(O_i - P_i)^2} \tag{3.13}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i) \tag{3.14}$$

$$d = 1 - \frac{\sum_{i=1}^{m}(O - P_i)^2}{\sum_{i=1}^{m}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \tag{3.15}$$

$$KGE = 1 - \sqrt{r^2 + \alpha^2 + \beta^2} \tag{3.16}$$

3.2.7 SHapley Additive exPlanation (SHAP)

SHapley Additive exPlanation (SHAP) is one of the approaches used to interpret ML models; it is particularly useful for complex models. SHAP was initially proposed by Shapley based on game theory (Shapley, 1953). SHAP helps in understanding the importance and contribution of a given feature to the model output. A Python package has been developed that computes the SHAP values of the features used in ML models like LightGBM, XGBoost, etc. (Lundberg and Lee, 2017). SHAP values and plots have practical advantages, especially for meteorology, in understanding how features drive and impact a desired target like rainfall or temperature.

Given $S \subseteq F$, where F is the set of all the features, to compute the effect of a feature (i) on the model prediction, a model $fS \cup \{i\}$ is trained with i present, and a second model fS is trained while withholding the i feature. The prediction from the second model is subtracted from the first model: $fS \cup \{i\}(xS \cup \{i\}) - fS(xS)$. Where xS are the input feature values in the subset S.

Shapley values are determined from:

$$\emptyset_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [fS \cup \{i\}(xS \cup \{i\}) - fS(xS)] \tag{3.17}$$

$\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the normalization term based on the number of choices of the subset.

3.2.8 Partial Dependence Plot (PDP)

Complex nonparametric models like LightGBM can be difficult to interpret. However, partial dependence plots (PDPs) can help show the relationship between features of interest and the target variable, independent of other features. PDP is used to study the relationship and interaction between the features used in this work and the target variables (monthly precipitation amount and thunderstorm frequencies).

For a set of predictors (features) $x = \{x_1, x_2, \ldots, x_p\}$ in a model with a prediction function of $\hat{f}(x)$, partitioned into predictors of interest, z_s, and its complement, $z_c = x \setminus z_s$, the partial dependence of the target response to z_s is given as:

$$f_s = E_{z_c}\left[\hat{f}(z_s, z_c)\right] = \int \hat{f}(z_s, z_c)\, p_c(z_c) dz_c \tag{3.18}$$

Where $p_c(z_c)$is the marginal probability density of the compliment features (z_c).

$p_c(z_c) = \int p(x) dz_s$. Equation (2) was estimated by calculating equation (3), the Monte Carlo method:

$$\overline{f_s}(z_s) = \frac{1}{n} \sum_1^n \hat{f}\left(z_s, z_{i,c}\right) \tag{3.19}$$

Where $z_{i,c}(i = 1,2, \ldots, n)$ are the values of the complimentary features (z_c) that are present in the training dataset.

3.2.9 Low Pass Filtering

Before using the one-year lagged ERA5 reanalysis data as features for the thunderstorm frequency models, they were passed through a low-pass filter with varying period lengths (pl) using a Fast Fourier Transform (FFT). Thus, the high-frequency part of the signal was subtracted, and only the low-frequency part of the signals was retained to predict the thunderstorms one year ahead. The different period lengths used include 2, 3, 4, 5, 6, and 12 months. A period length of 3 months gave the best results in all the models, as shown in Table 4.5. The results presented in the rest of this work are the results of a period length of 3 months.

Similarly, for the precipitation model, the ERA5 reanalysis was passed through a low-pass filter with a period length of 3 months, and then the filtered data underwent principal component analysis to reduce the dimensionality.

3.2.10 Empirical Orthogonal Function (EOF)

To reduce the dimensionality of the ERA features, Empirical Orthogonal Function (EOF) was performed on the low-pass filtered features. EOF analysis is widely used in climatology to extract dominant spatial and temporal patterns (Hannachi *et al*., 2007). The following features were used in the EOF analysis: 10 m u wind component (U10), 10 m v wind component (V10), surface pressure (SP), evaporation (E), clear-sky direct solar radiation at surface (CDIR), vertical integral of eastward cloud liquid water flux (P88.162), total totals index (TOTALX), mean sea-level pressure (MSL), k-index (KX), convective available potential energy (CAPE), and instantaneous northward turbulent surface stress (INSS). Convective inhibition (CIN) was excluded from this list because it had many missing values. The retained time series were used along with the climate indices as features to predict monthly precipitation over West Africa. The percentage variances of the first 5 principal components (PCs) were 57.8%, 21.0%, 7.2%, 3%, and 2% for PC1, PC2, PC3, PC4, and PC5, respectively. The data was then divided into training, evaluation, and test data.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Trends and Drivers of Thunderstorms

4.1.1 Thunderstorm Climatology

4.1.1.1 Annual Cycle of Thunderstorm Frequencies over Ghana

Monthly thunderstorm frequency data exhibit clear seasonal patterns. From the three main zones (Figure 3.1), the northern zone shows a close-to-monomodal distribution, whereas the coastal and middle zones experience a clearly bimodal rainfall pattern (Figure 4.1), peaking in May and November, except for the Tema synoptic station, which peaks in October. The 'little dry season' is seen in August (Figure 4.1a). During this period, morning drizzles result from cold winds that reduce the temperature (Cornforth *et al*., 2017). As a result of these cooler conditions, the atmospheric energy required for thunderstorm development is insufficient, leading to a reduced likelihood of thunderstorms during this period.

The middle zone has similar patterns compared to the coast. However, the influence of the 'little dry season' is reduced towards the north, as shown in Figure 4.1b. Furthermore, the middle zone has the second peak of thunderstorm frequencies in October. This second peak is higher than the first peak in May, unlike in the coastal zone. In the middle zone, what is mostly referred to as the second rainy season or "minor rainfall season" has more thunderstorms than the first rainy season. The northern zone stations have thunderstorm frequencies peaking in May and then experiencing only a small reduction in June, July, and August before peaking again in September, except for the Navrongo station, where there is a single peak in August. Thus, the 'little dry season' is weak or non-existent in this zone.

The coastal and middle zones have less than five thunderstorms in August, except for the Abetifi, Wenchi, and Kete Krachi stations (Figure 4.1b). In contrast, all the stations in the northern zone had more than five thunderstorms in August during the study period. In fact, the northernmost station in Ghana (Navrongo) has its thunderstorm peak in August (Figure 4.1c).

The intra-annual cycle of thunderstorms is explained by the movement of the intertropical convergence zone (ITCZ) in response to the apparent movement of the sun. The two main wind regimes that affect Ghana are the moist southwesterlies and dry northeasterlies. When a cold tongue (low sea surface temperature; SST) develops over the Gulf of Guinea by boreal mid-spring in response to the amplification of the St Helena anticyclone (Thiaw *et al*., 2017), moist southwesterlies prevail over Ghana. During the dry season, Ghana is predominantly affected by dry northeasterly winds (Fink *et al*., 2017).

As the ITCZ moves northward in May, much of Ghana is under moist air conditions and an active zone of thunderstorms. As the ITCZ moves further north, the active zone of thunderstorms also moves northward. When the ITCZ reaches its northernmost position around August, the northernmost station in Ghana, Navrongo, experiences its peak number of thunderstorms. As the ITCZ retreats south, the other northern stations have peak thunderstorms in September. However, over the middle zone, the second peak in response to the ITCZ retreat is in October. While the southernmost coastal zone experiences its second peak in November.

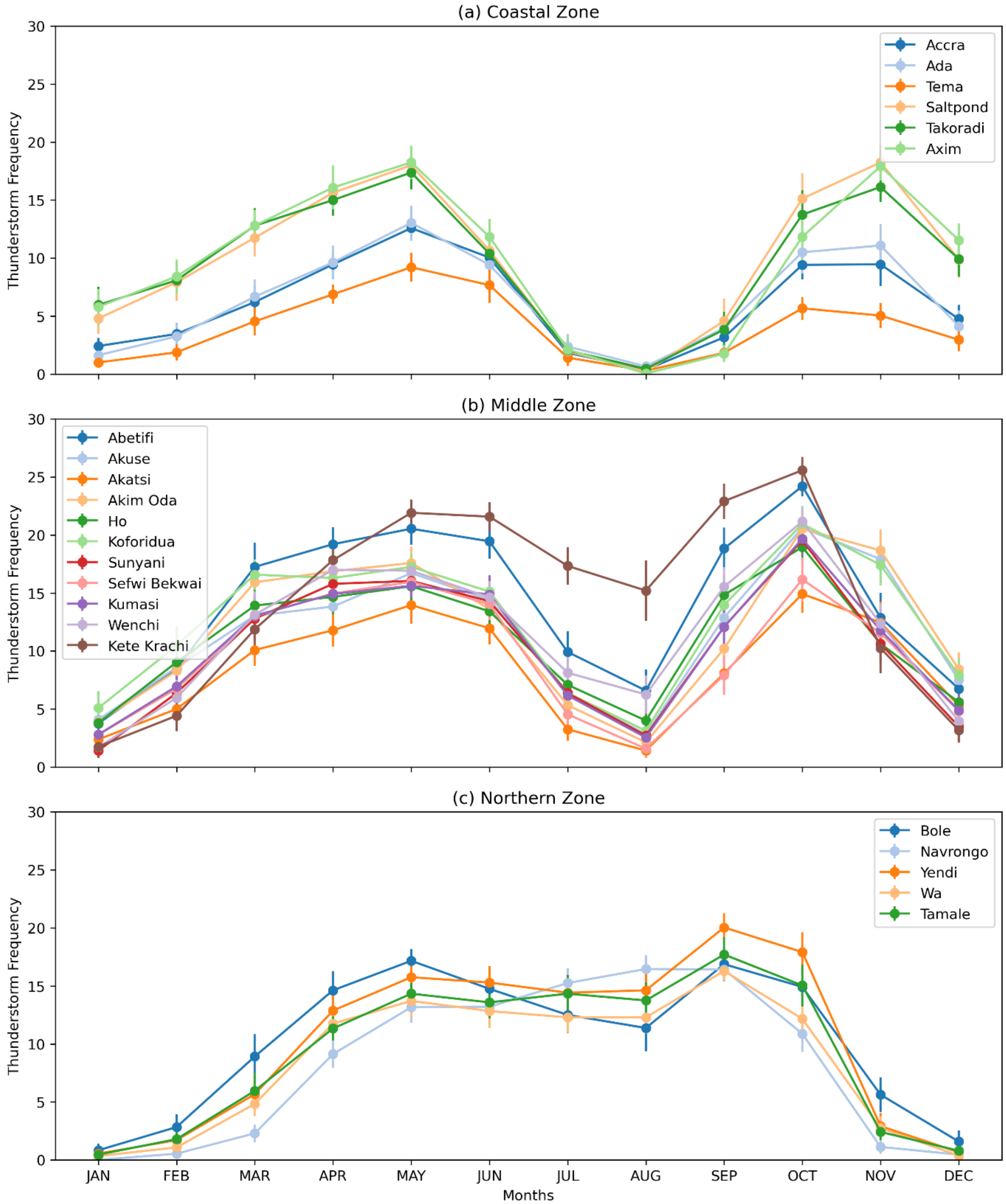

Figure 4.1. Mean annual cycle of monthly thunderstorm frequencies over Ghana from 1990-2013, depicting the various zones: (a) coast, (b) middle, and (c) north. Vertical bars show the 95th confidence interval.

4.1.1.2 Interannual Variability of Thunderstorm Frequencies over Ghana

In this section, annual mean thunderstorm data per zone are investigated. The interannual variability of thunderstorms in the coastal zone is the greatest among all zones (Table 4.1). In this zone, the Accra and Ada stations (Table 4.1) had the highest coefficient of variation (CV) of 25.3% and 23.3%, respectively. Akatsi and Ho recorded the highest variability in the middle zone, with CVs of 20.9% and 14.3%, respectively.

The Tema station (Figure 4.2c) recorded the lowest number of annual thunderstorms in the coastal zone and throughout the country. Generally, the eastern coastal stations had a lower number of annual thunderstorms compared to the western coastal stations (Figure 4.2). These eastern coastal stations are found to be in the Ghana-Benin dry zone (Vollmert *et al*., 2003), the so-called Dahomey gap. In the middle zone (Figure 4.3), all the stations recorded at least 100 thunderstorms per year during the study period, except the Akatsi and Sefwi Bekwai stations. The Kete Krachi station in the middle zone recorded the highest number of mean annual thunderstorms. In the northern zone (Figure 4.4), the Bole station recorded both the least and highest number of annual thunderstorms (Figure 4.4a).

Table 4.1: Interannual coefficient of variation of thunderstorms in the 22 synoptic stations in Ghana

Zone	Station	Coefficient of variation
Coast	Accra	23.3
	Ada	25.3
	Tema	20.8
	Saltpond	14.4
	Takoradi	10.2
	Axim	10.9
Middle	Abetifi	10.8
	Akuse	9.5
	Akatsi	20.9
	Akim Oda	10.4
	Ho	14.3
	Koforidua	8.4
	Sunyani	9.9
	Sefwi Bekwai	10
	Kumasi	10.3
	Wenchi	9.9
	Kete Krachi	12.7
North	Bole	15.1
	Navrongo	11.1
	Yendi	12.7
	Wa	14.4
	Tamale	10.2

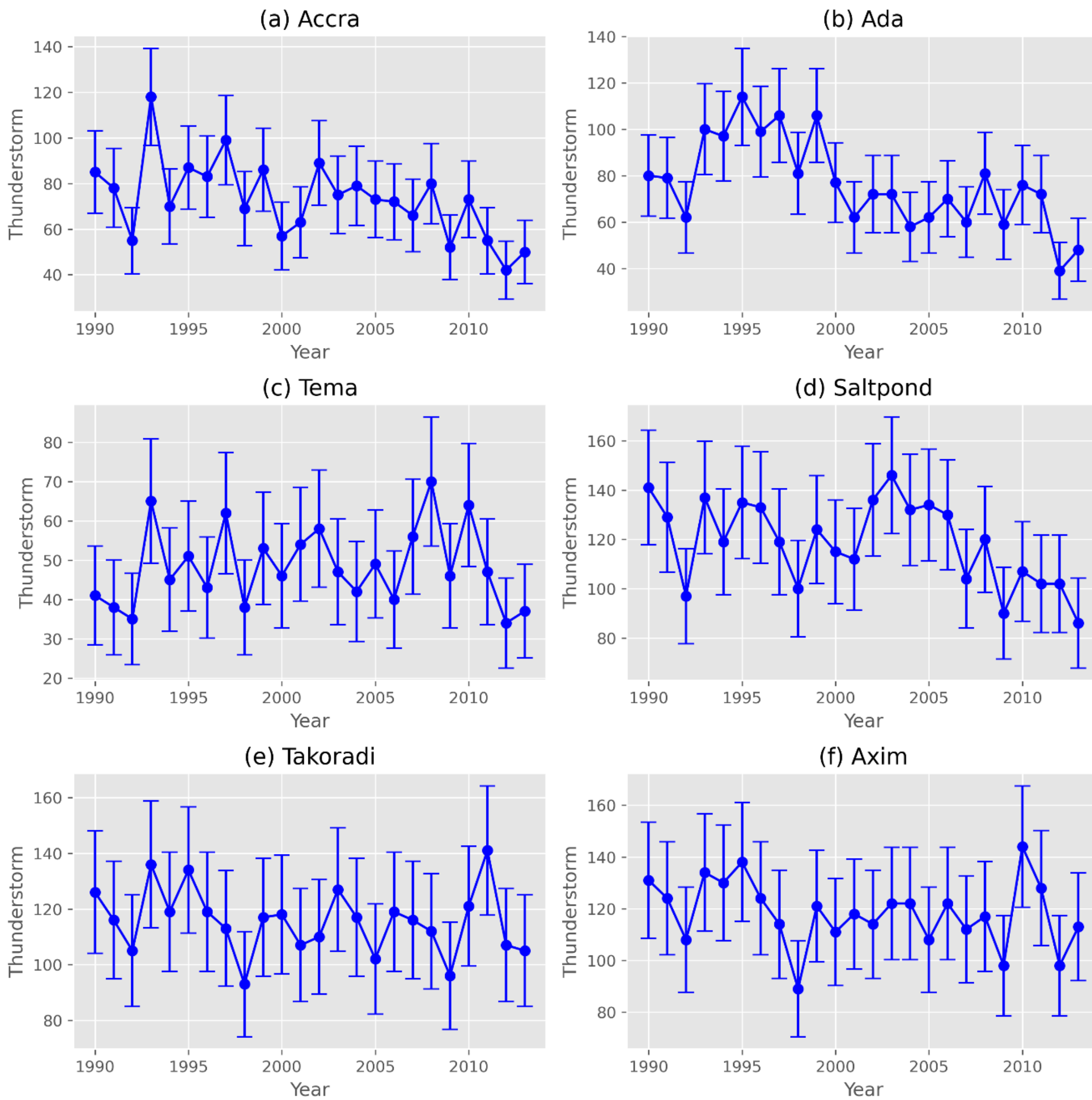

Figure 4.2. Interannual thunderstorm frequency variability over the coastal zone of Ghana.

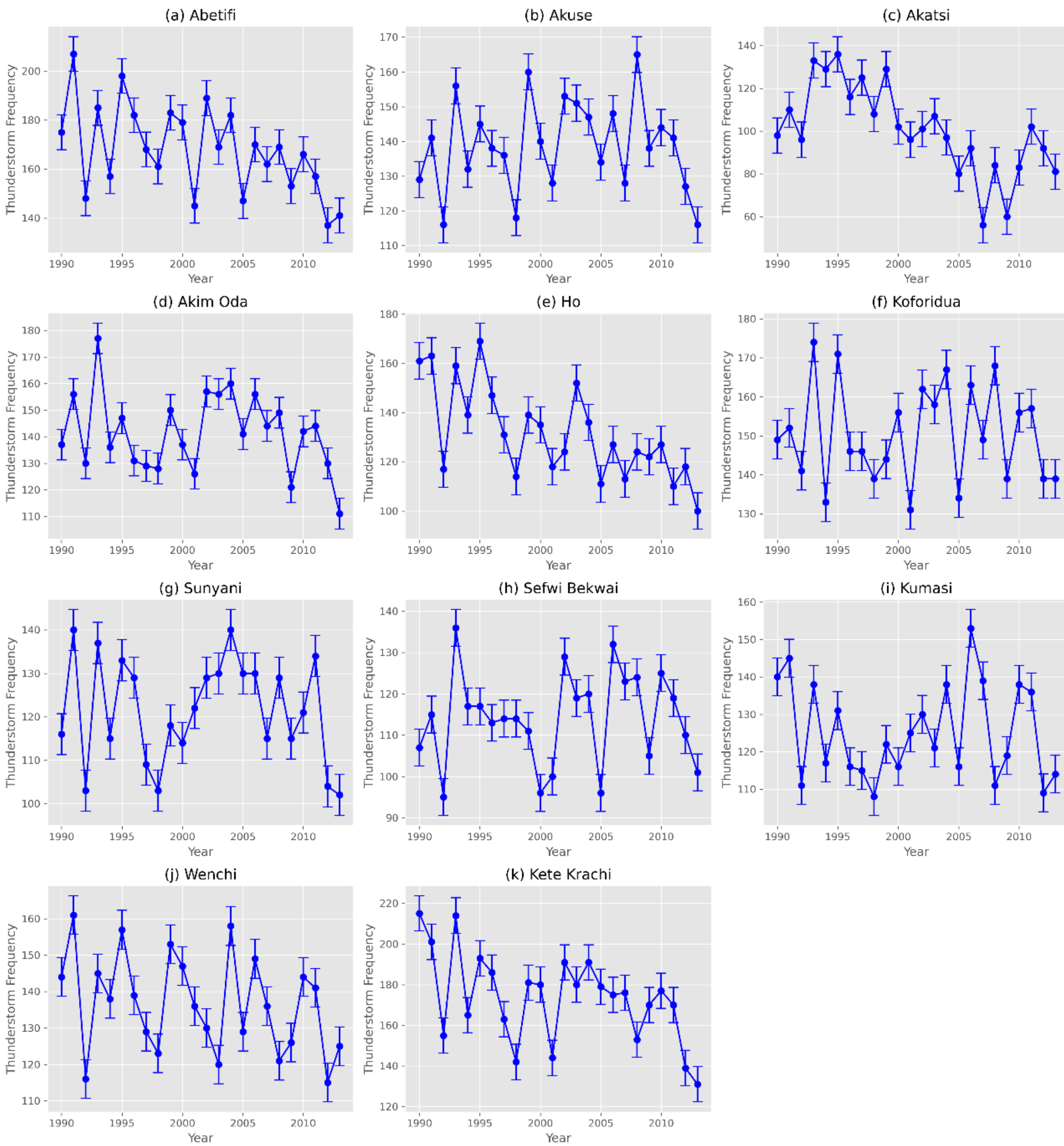

Figure 4.3. Interannual thunderstorm frequency variability over the middle zone of Ghana.

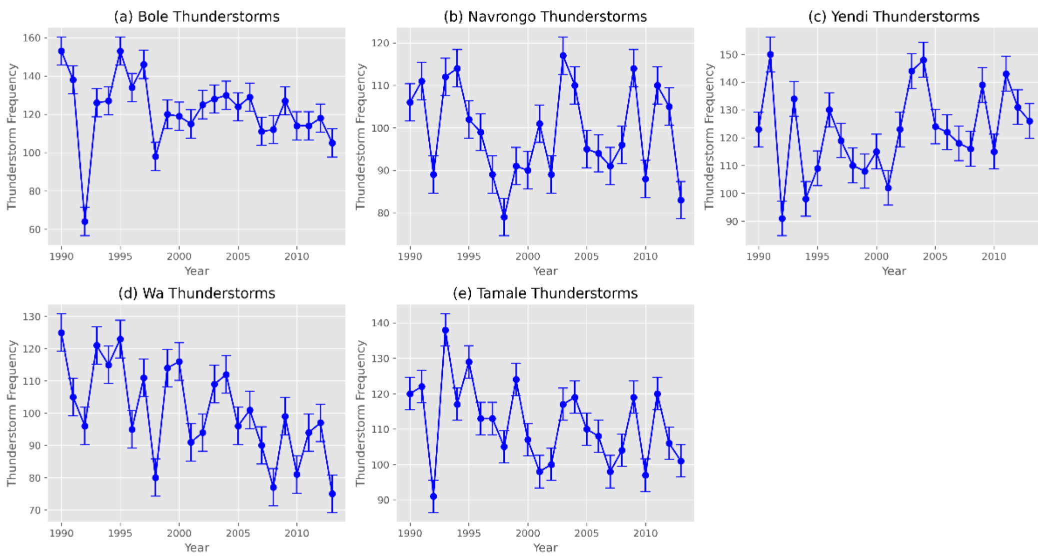

Figure 4.4. Interannual thunderstorm frequency variability over the northern zone of Ghana.

4.1.2 Thunderstorms Relation with Global SST

To investigate the relationship of the thunderstorm frequencies with global SST, the anomalies of the thunderstorm frequencies were correlated with the corresponding period of the SSTAs from 1990–2013, and the results are presented in Figure 4.5.

In parts of West Africa, rainfall variability and global SST have been shown to relate in the following ways: dry periods are associated with warmer SSTs in the south and equatorial Atlantic (Lamb, 1978; Vizy and Cook, 2001; Rowell, 2013), cooler Mediterranean (Jung *et al*., 2006), warmer SSTs in the Indian Ocean (Lu and Delworth, 2005), and ENSO in the Pacific Ocean (Janicot *et al*., 2001). However, it is important to note that these relationships may not necessarily apply to thunderstorm frequencies in all the zones under study. Firstly, the thunderstorm frequencies also include "dry thunder" events where the sound of thunder is heard or lightning is seen in the station but no rainfall is recorded. Secondly, although the majority of rainfall in West Africa is from convective rainfall, there is a good percentage of rainfall in the region that is not accompanied by thunderstorms.

According to Figure 4.5a, the correlation of the coastal zone's interannual thunderstorm frequency variation with interannual SSTAs indicates that the coastal zone's thunderstorms are negatively correlated with a small region of the Pacific around latitude 15^0S and longitude 120^0W. Over the Atlantic, the thunderstorm frequencies are positively correlated with Gulf of Guinea SST and equatorial and south Atlantic SST. Earlier research concluded that the SST influence on Guinea coast rainfall is less complex and that dry conditions are only driven by cooling in the equatorial and near-equatorial south Atlantic Ocean (Rowell, 2013). Similar observations are made for the thunderstorm frequencies on the coast of Ghana in this study.

Figure 4.5b shows that interannual thunderstorm frequencies in the middle zone correlated positively with the near-equatorial Atlantic and Gulf of Guinea. The middle zone negatively correlated with a region in the Indian Ocean around latitudes 0^0 to 30^0S and longitude 120^0E. Warming in the equatorial Atlantic and near-equatorial South Atlantic could increase the frequency of thunderstorms over the coastal and middle zones of Ghana.

Figure 4.5c illustrates that thunderstorm frequencies in the northern zone correlated negatively with parts of the Indian Ocean. A prominent feature of this correlation result is in the Pacific Ocean, in the region of latitudes 0^0 to 30^0 north, where the thunderstorm frequencies correlate positively with the SST. The signatures of these results are those of WHWP and ENSO. The zone also correlated positively with the Mediterranean. The negative correlation with the Indian Ocean is ascribed to the warming of the Indian Ocean enhancing subsidence and stability over West Africa (Martin *et al*., 2014).

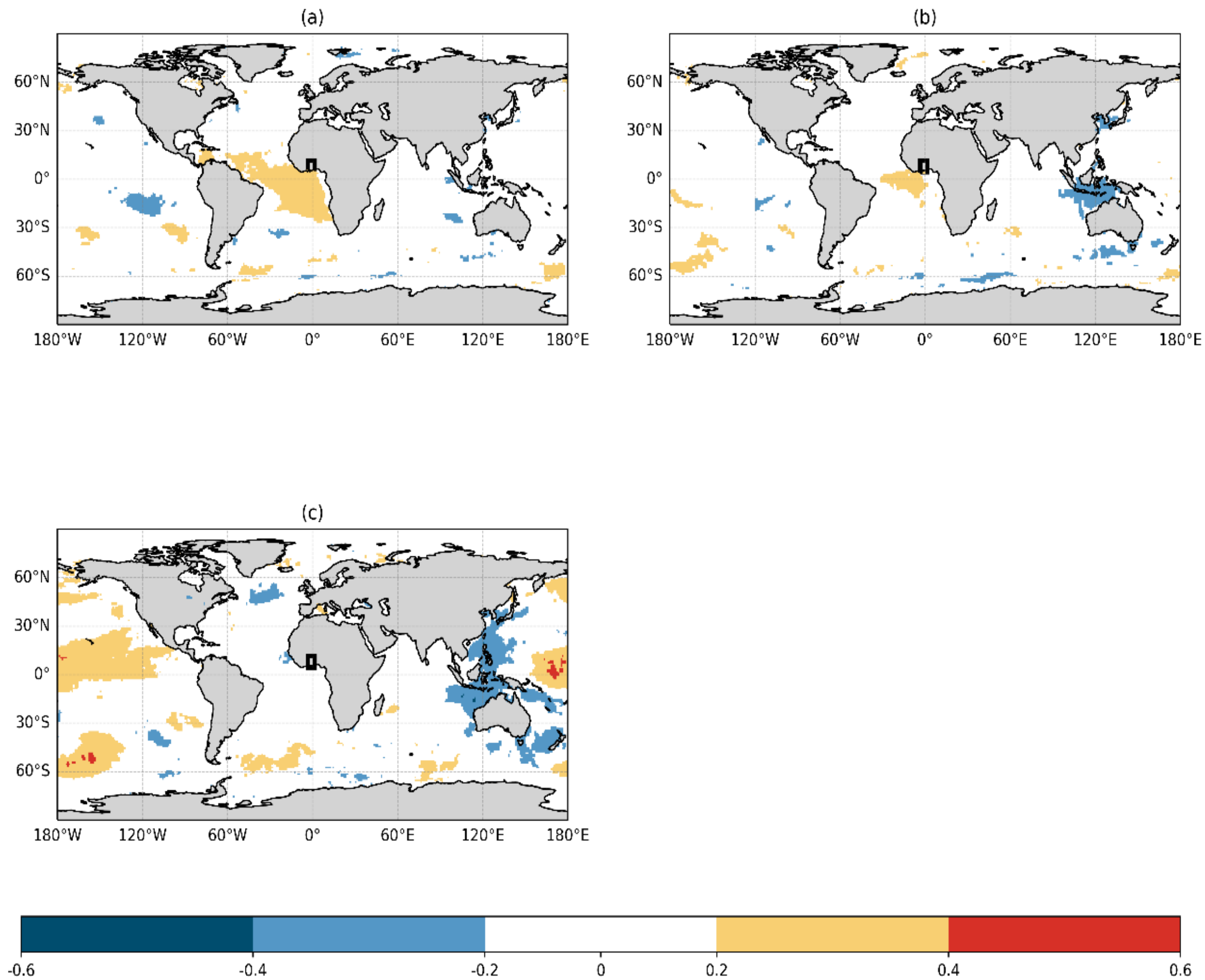

Figure 4.5. Significant (95%) correlation between coastal zone (a), middle zone (b), and northern zone (c) thunderstorm frequencies in Ghana and global SST anomalies. The black box over Africa is the approximate location of Ghana.

4.1.3 Impact of Teleconnections on Thunderstorm Variability

To elucidate further the relationship between thunderstorm frequency variation and climate indices, a machine learning regression analysis was done using the LightGBM regressor, with thunderstorm frequencies as the target and climate indices as the features. In this section, the analysis is done on station level.

To evaluate the overall performance of the regression models, the coefficient of determination (R^2) and index of agreement (d) for each station were computed (Table 4.2).

Stations in the northern zone had R^2 values ranging from 0.51 to 0.60, except for Tamale station, which had a value of 0.36. Thus, for most of the stations in this zone, five (5) climate indices explained at least 51% of the thunderstorm variability, and over the Tamale station, they explained 36%. The index of agreement (d) values also ranges from 0.71 to 0.87. The proportion of variance explained by these indices, however, has decreased in the other zones. The transition zone had an R^2 (d) of 0.20 (0.56) for Wenchi and 0.42 (0.77) for Kete Krachi. The two southernmost zones (coast and middle) had the lowest values of both R^2 and index of agreement. The middle zone had R^2 values ranging from 0.13 to 0.27, and the coastal zone values ranged from 0.17 to 0.32.

The middle zone recorded the lowest R^2 values in the LightGBM model. The lower values of R^2 and d over most stations in the middle and coastal zones suggest that other indices not used in this work, or other factors like vegetation, soil moisture, atmospheric circulation, etc. (Pielke, 2001; Taylor *et al*., 2010; Piper *et al*., 2019) other than the selected climate indices are the major drivers of thunderstorm variability over these areas.

To further investigate the contribution of individual indices to the model, the feature importance of the LightGBM model is analysed to quantify the impact of the climate indices in driving the variation of thunderstorm frequencies. Figure 4.6 is an example of the feature importance of the Navrongo station, indicating the contribution of each feature used in the LightGBM model over the station. This example shows that the AMO feature had the highest contribution to the thunderstorm frequency model over the Navrongo station, followed by the PDO, Niño 1+2, WHWP and AMM. These features, therefore, have the strongest impact on the thunderstorm frequency variability over the station of Navrongo during the study period.

Further information on all the features used in the final model for each station is detailed in Table 4.3. The feature importance does not show the direction of the influence of the feature on the thunderstorm variability. The most important feature for each station is highlighted in bold in Table 4.3. For easy comparison, the feature importance is scaled by dividing the individual feature importance by the sum of all the feature importance values for that station.

Figure 4.7 depicts the features that have been selected in at least three different zones, and their frequency of usage per zone. It shows that WHWP and Niño 1+2 have been identified as the most relevant features in all zones. In general, feature importance tends to decrease from north to south, reflecting decreasing persistence of the thunderstorm dynamics towards the coast. Two major exceptions are AMM in the northern zone, and PDO in the coastal zone.

The fact that the middle and the coastal zones exhibit a less consistent pattern in terms of TSA, PNA, NAO, and SOI (each of them being selected for less than 50% of the stations), and that in general more features have been selected here seems to be more related to lower model performance in general rather than to clear causal relationships: the model tries to “squeeze out” even random “information” from the data, a typical phenomenon of overtraining. Thus, these patterns are not interpreted.

Three of the northern stations (Figure 4.7) had their most important features from the Atlantic, these being AMO and AMM. The AMO and AMM have been found to be the most important features for many other stations in the other zones (Figure 4.7). The AMM is associated with the latitudinal movement of the ITCZ in Ghana; therefore, it is a strong driver of monthly thunderstorm variability in Ghana. The warm phase of AMO has been observed to strengthen the Sahara heat low (Martin and Thorncroft, 2014). When the heat low is strengthened, it will strengthen the inflow of moist south westerlies over Ghana, which in turn will lead to an increase in thunderstorm frequency over Ghana, especially over the northern zone. An intense AMO could lead to fewer thunderstorms over the coastal zone stations, as these stations will be outside the active zone of the ITCZ.

The ENSO indices have been found to influence thunderstorm variability in all zones of Ghana. West Africa is affected by El Niño, such that during El Niño events, there is enhanced northeast wind flow over the region (Camberlin *et al*., 2001). The northeasterlies are dry, which leads to dry conditions and hence lower thunderstorm frequencies. Other indices in the Pacific Ocean that have been found to drive the variability of monthly thunderstorms in Ghana are the PDO and WHWP.

PDO has been previously shown to negatively correlate with Sahel rainfall (Diatta and Fink, 2014). The frequency of the WHWP in the final models for all the zones is particularly outstanding. Variations in the WHWP have been found to affect the Walker circulation and tropical convection, where the eastward extension into the tropical Atlantic relates to convection in northwest Africa (Wang and Enfield, 2001). Indeed, it has been shown that this eastward extension of the WHWP into the tropical Atlantic allows the WHWP to supply moisture to the Atlantic ITCZ (Drumond *et al*., 2011) and hence could enhance thunderstorm activity. The

WHWP has also been shown to induce cooling in the Pacific that moves equatorward to form La Niña (Park *et al*., 2018). The proposed reason for the great influence of the WHWP on Ghana thunderstorm frequencies is that the WHWP relates to Ghana convection activities through both the tropical Atlantic and Pacific.

Figure 4.8 is a plot of the LightGBM model using only climate indices to predict the thunderstorm frequency over Navrongo. The climate indices were able to moderately predict the variability of the thunderstorms over the Navrongo station. The influence of the Atlantic Ocean and Pacific Ocean on the thunderstorm variability over Ghana is not only shown in the correlation maps but also in the LightGBM regression model.

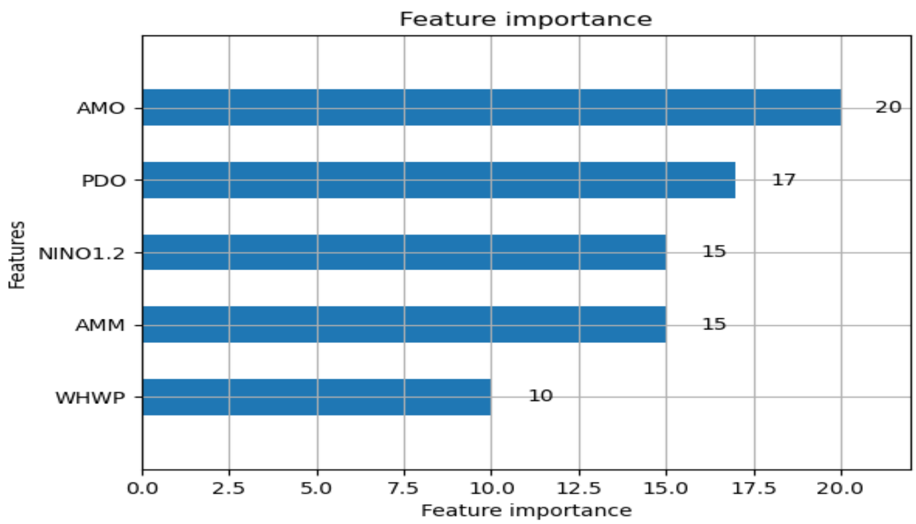

Figure 4.6. LightGBM feature importance plot for the Navrongo synoptic station in the northern zone.

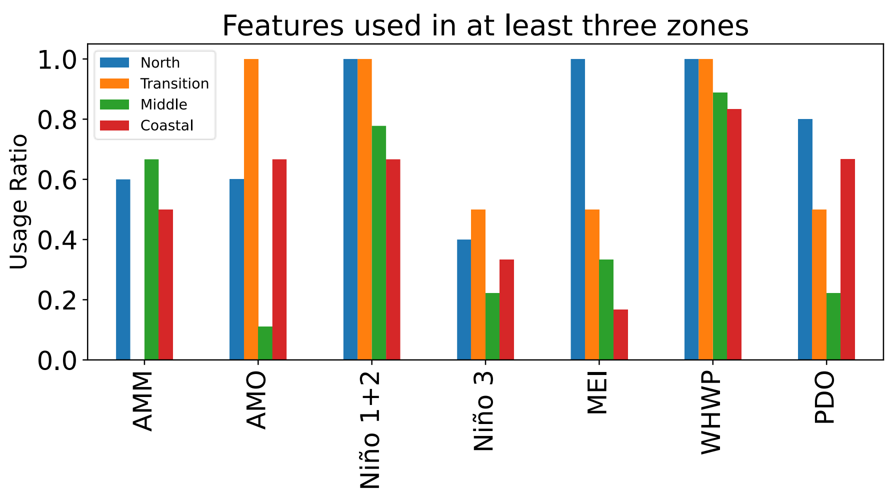

Figure 4.7. Climate indices that have been used at least in three zones and their frequency of having been selected by the regression models for the various zones in Ghana.

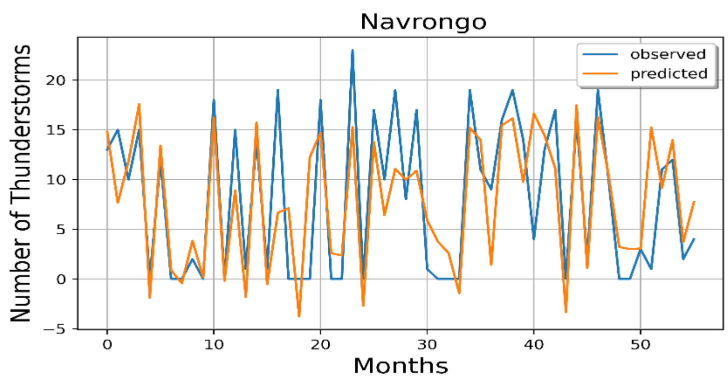

Figure 4.8. Time series of LightGBM model prediction of thunderstorm frequency and the observed thunderstorms over Navrongo synoptic station.

Table 4.2. LightGBM model evaluation parameters for all the stations used in the study. The highest metrics for each zone are given in bolden.

Zone	Station	R^2	d
Northern	Yendi	**0.60**	0.85
	Wa	0.51	0.83
	Tamale	0.36	0.71
	Navrongo	0.59	**0.87**
	Bole	0.58	0.86
Transition	Kete Krachi	**0.42**	**0.77**
	Wenchi	0.20	0.56
Middle	Sunyani	0.25	0.62
	Sefwi Bekwai	0.17	0.56
	Ho	0.22	0.63
	Kumasi	0.13	0.51
	Akuse	0.18	0.61
	Koforidua	0.24	0.63
	Akim Oda	0.23	**0.67**
	Akatsi	0.20	0.65
	Abetifi	**0.27**	0.60
Coastal	Accra	0.18	0.63
	Ada	0.19	0.51
	Takoradi	**0.32**	0.65
	Tema	0.27	0.62
	Axim	0.22	**0.67**
	Saltpond	0.17	0.59

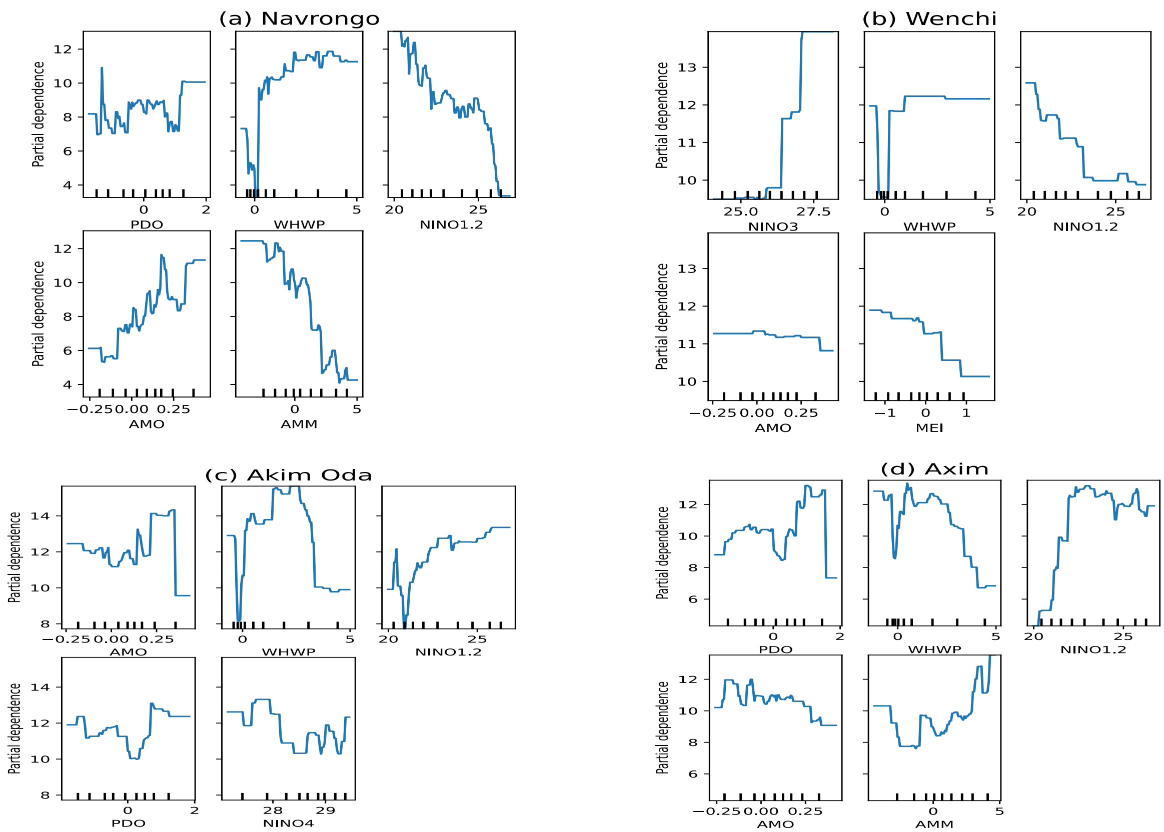

Figure 4.9. Partial dependence plot of Navrongo station in the northern zone (a), Wenchi in the transition zone (b), Akim Oda in the middle zone (c), and Axim in the coastal zone (d).

The partial dependence plot of the LightGBM regression model of the selected features on selected stations in each zone is presented in Figure 4.9. These plots help further explain the effect of the indices on thunderstorms in Ghana. For instance, as the WHWP becomes warmer, the number of thunderstorms over the northern zone increases (Figure 4.9a). However, the transition zone (Figure 4.9b) experiences increasing thunderstorms as the WHWP is above 0^0C but flattens when the WHWP gets to about 2^0C. Further south, WHWP about 2^0C and above, shows the middle and coastal zones experiencing decreases in thunderstorms (Figure 4.9c,d). These observed patterns support our argument that the WHWP supplies moisture to the Atlantic ITCZ (Drumond *et al*., 2011) and thus influences Ghana's thunderstorms. When the WHWP region begins to warm, it supplies just enough moisture, making the ITCZ to be just above the coast and middle zones, causing increases in thunderstorms. However, as it intensifies, the ITCZ is moved further north, favouring the transition and northern zones and leaving the middle and coastal zones out of the active zone of the ITCZ. Similar observations are made for the AMO, where very high values of AMO appear to favour the northern zone only.

Another interesting result from the partial dependence plot is the Niño1+2 effect on the thunderstorms. Increasing values of Niño1+2 relate to lower thunderstorms in the north and transition zones, but oppositely in the middle and coastal zones. It must be pointed out that increasing MEI coincided with decreasing thunderstorms in all the zones. The partial dependence plots of the features have shown that high MEI leads to low thunderstorm frequencies in Ghana, clearly showing the influence of ENSO on thunderstorms in Ghana. However, warmer WHWP seems to favour only thunderstorm occurrence in the northern stations and to inhibit the occurrence of thunderstorms in the middle and coastal zones.

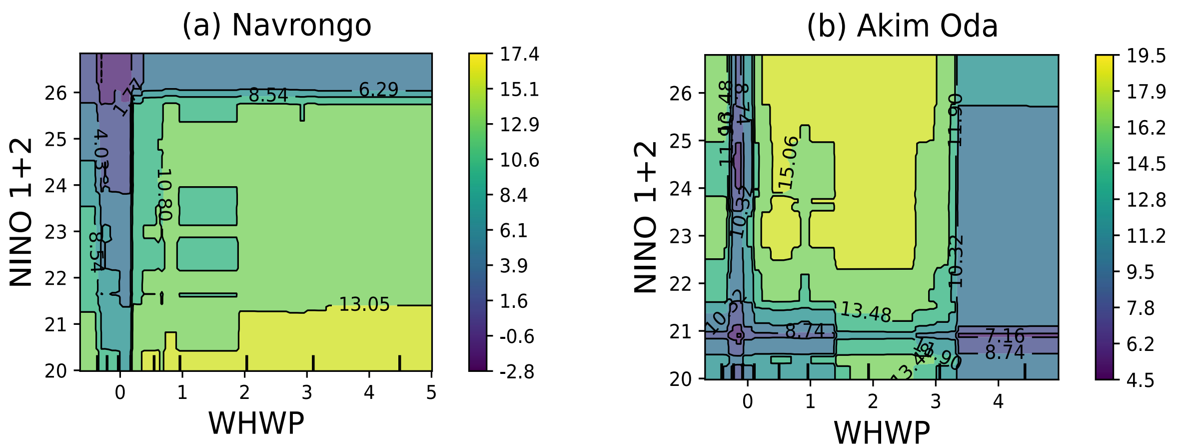

Figure 4.10. Two-way partial dependence plot of the two most important features, showing the interaction between the features at a northern station (a) and a southern station (b).

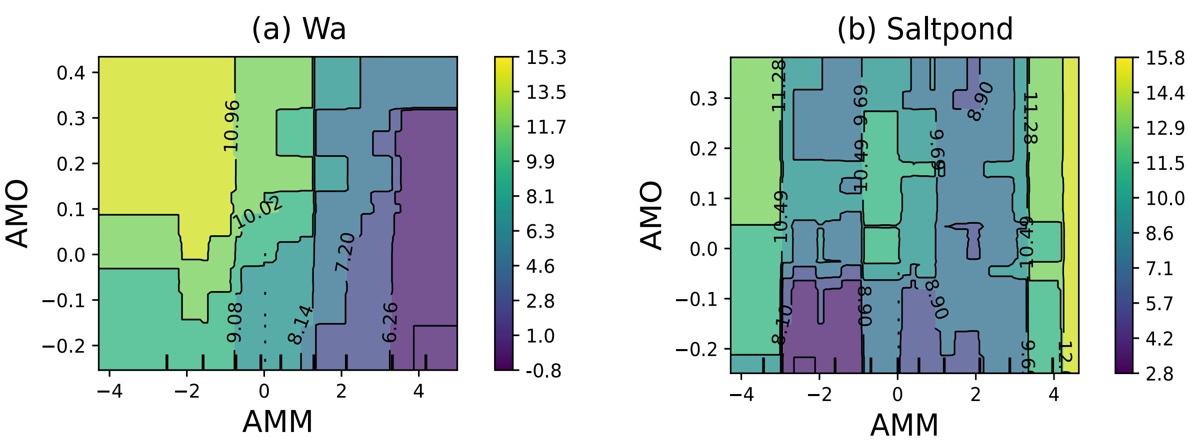

Figure 4.11. Two-way partial dependence plot of the most frequent features that reside in the Atlantic Ocean, showing the interaction between the features at a northern station (a) and a southern station (b).

The two-way partial dependence plot shows how the predicted thunderstorms change as the two features change while keeping other features constant. To show the interaction between the two most frequent features (WHWP and Niño1+2), two-way partial dependence plots are shown for a station in the north of Ghana and another for the south. The predicted number of thunderstorms is higher over the northern station (Figure 4.10a) when the WHWP is warmer and the Niño 1+2 cool, with the lowest thunderstorms predicted when the WHWP is cooler and the Niño 1+2 warmer. The plot also shows the interaction between the two features (WHWP and Niño1+2). When WHWP is low, variations in Niño1+2 have little influence on the thunderstorm variation compared to when WHWP is higher. Similarly, when the Niño1+2 values are extremely high, variations in WHWP have little to no influence on thunderstorm variation. However, over the southern station (Figure 4.10b), the highest number of thunderstorms occur when the Niño 1+2 is warmer and the WHWP has mean values. This appeared to be the trend for most of the stations in southern Ghana that had both WHWP and Niño1+2 features in their final model.

The interaction between the two most frequent features that reside in the Atlantic Ocean (AMM and AMO) shows that over the northern station (Figure 4.11a) higher thunderstorms occur when the AMO is in a positive phase and the AMM is in a negative phase. However, over the southern stations (Figure 4.11b), the relationship is more complex. The influence of AMO variation on the predicted thunderstorms is lower when AMM is high than when AMM is low (Figure 4.11b). Thus, when AMM is high, the influence of AMO on thunderstorms in southern Ghana is less. This implies that the interaction between the AMO and thunderstorms in the southern stations is modulated by the AMM.

Table 4.3. Selected climate indices and their relative LightGBM regression features importance (normalized by the sum of all feature importance values per station) of the various synoptic stations in Ghana. The most important feature for each station is boldened on the table.

Zone	Station	AMM	AMO	Niño 1+2	Niño 4	Niño 3	IOD	NAO	PNA	TSA	SOI	MEI	AO	WHWP	PDO
Northern	**Yendi**	0.23	0.19	**0.24**										0.23	0.11
	Wa	**0.24**	**0.24**	0.16										**0.24**	0.14
	Tamale			**0.30**		0.28						0.16	0.15	0.12	
	Navrongo	0.16	**0.26**	0.17										0.17	0.24
	Bole			**0.23**		0.21						0.19		0.20	0.17
Transition	**Kete Krachi**		**0.25**	0.21	0.17									0.18	0.19
	Wenchi		0.06	**0.29**		0.20						0.27		0.18	
Middle	**Sunyani**	0.17		0.21				0.16				0.15		**0.31**	
	Sefwi Bekwai			**0.26**		0.20						0.18	0.17	0.20	
	Ho	0.16		0.22	0.19			0.19						**0.24**	
	Kumasi			**0.26**		0.22					0.13	0.18		0.21	
	Akuse	0.19		0.19				0.15					0.18	**0.29**	
	Koforidua	0.11			0.16					0.22			0.24	**0.27**	
	Akim Oda		0.20	**0.21**	0.20									0.20	0.19

Zone	Station	AMM	AMO	Niño 1+2	Niño 4	Niño 3	IOD	NAO	PNA	TSA	SOI	MEI	AO	WHWP	PDO
Middle	**Akatsi**	0.20							**0.23**	0.19	0.15				**0.23**
	Abetifi	0.20		**0.21**	0.19			0.18						**0.21**	
Coast	**Accra**			0.20		0.21				0.13		0.20		**0.27**	
	Ada			0.07				0.20		0.23				**0.33**	0.17
	Takoradi		0.17			0.21	0.18		0.21	**0.23**					
	Tema	0.22	0.20								0.13			**0.23**	0.21
	Axim	0.21	0.17	0.21										0.18	**0.23**
	Saltpond	0.21	0.17	0.20										**0.24**	0.18

4.1.4 Drivers of Thunderstorm Variability

Model Evaluation

Using recursive feature elimination (RFE), the following features are found to be the most impactful drivers of thunderstorms: convective available potential energy (CAPE), convective inhibition (CIN), top-of-atmosphere incident solar radiation (TISR), 10 m wind speed (SI10), and longitude (Lon). Although climate indices have been used to predict thunderstorms in previous studies (Sátori *et al*., 2009; Pinto, 2015), they have not been found to be the most impactful features in this study.

All data on the overall performance of the regression model refer to the test data set. The model had R^2, MAE, and model bias values of 0.79, 2.26, and 0.24, respectively. Thus, the features used in the model explained 79% of the monthly thunderstorm variability. The results of the Q-Q plot (Figure 4.12) show that the model captured the observed quantiles well.

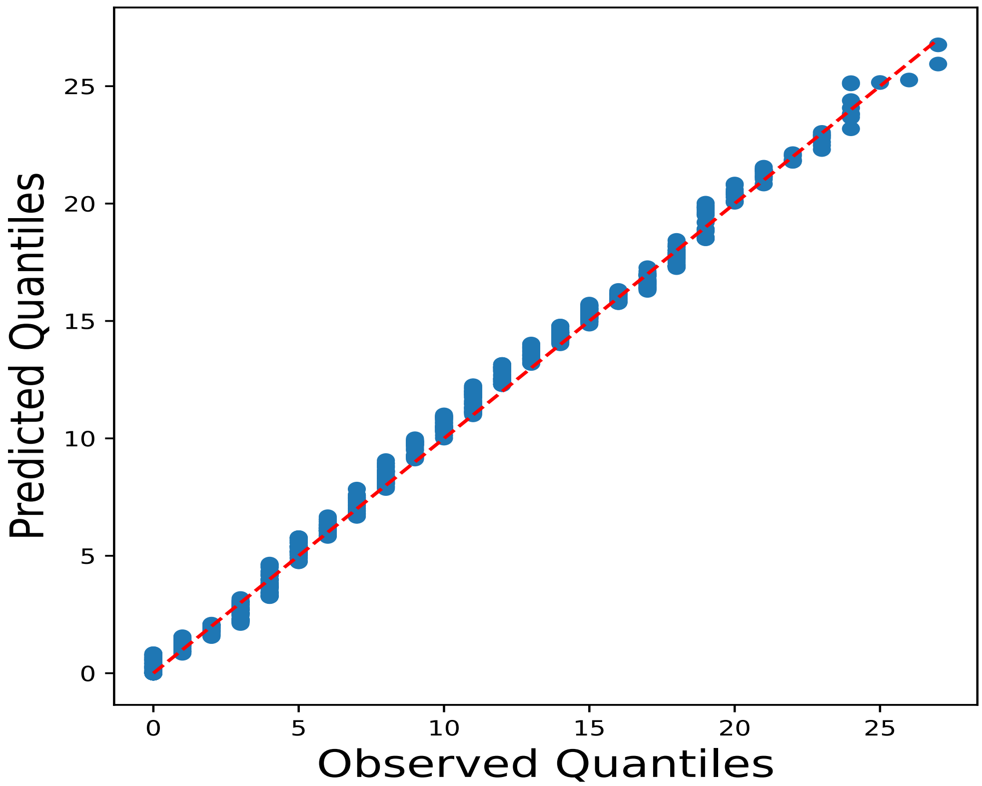

Figure 4.12. Q-Q plot showing the distribution of the predicted and observed monthly frequency data of the various stations.

4.1.5 Partial dependence plots (PDP) and SHapley Additive exPlanations (SHAP) plots

The selected features from the recursive feature elimination (RFE) were further investigated to understand the contribution of individual features to the model. The partial dependence and SHapley Additive exPlanations (SHAP) plots of the LightGBM regression model are presented in

Figures 4.13 and 4.14. These plots help further explain the effect of the selected features on thunderstorms in Ghana.

4.1.5.1 SHAP plots

To understand how the selected features impact the model output and to compare the behaviour of the features, SHAP plots are presented in Figure 4.13. The features are sorted according to relevance, starting from the top. CAPE contributes the most to the output of the model, in particular such that lower values of CAPE impact the model output negatively (Figure 4.13). TISR is the second most important feature, being positively correlated with thunderstorm variability. The third most important driver of thunderstorms among the selected features is CIN. Higher CIN values lead to a lower number of thunderstorms, while lower CIN values lead to a higher number of thunderstorms. Similarly, the next important feature, SI10, is negatively related to predicted thunderstorm frequencies. Longitude is the fifth driver of thunderstorms. Generally, a lower number of thunderstorms are predicted over the west compared to the east. On average, the importance of the selected features on the model output is in this order: CAPE > TISR > CIN > SI10 > lon.

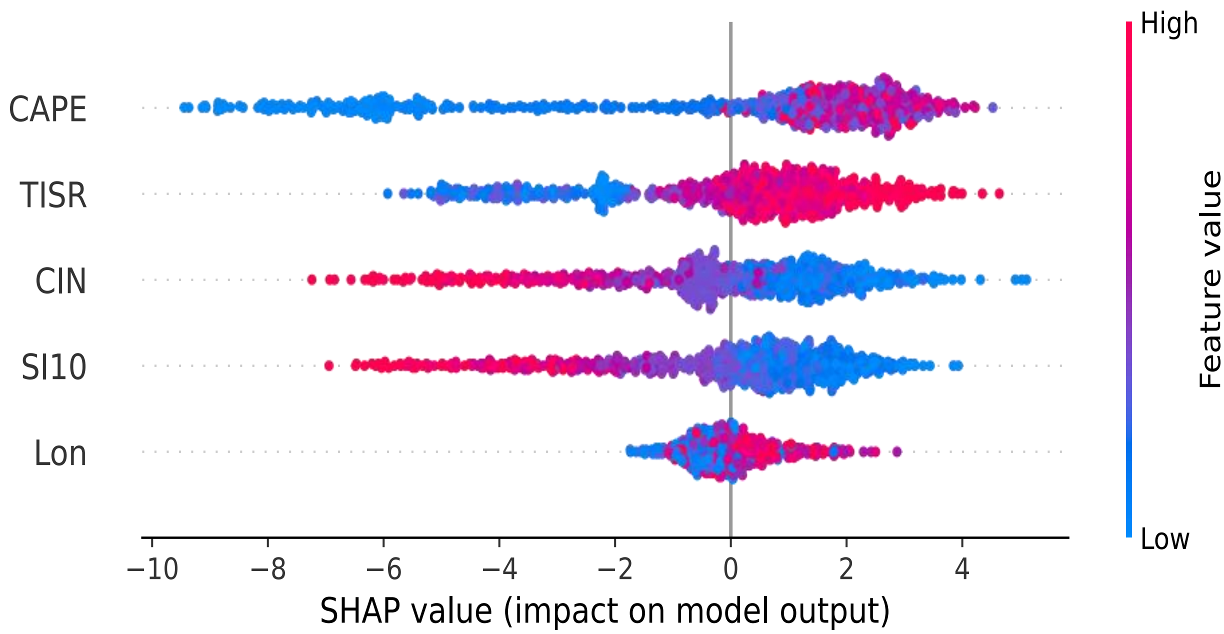

Figure 4.13. SHAP plots showing how the features impact the model output. At the y-axis features are sorted by relative importance.

4.1.5.2 Partial dependence plots

Increasing the value of the convective available potential energy (CAPE) generally increases the number of thunderstorms (Figure 4.14). Initially increasing the values of CAPE leads to a sharp increase in the number of thunderstorms, and the effect levels off at CAPE values greater than 500 J/kg. Top-of-atmosphere incident solar radiation (TISR) is positively and nearly linearly related to thunderstorm frequencies in Ghana, and convective inhibition (CIN) and 10 m wind speed (SI10) are negatively related to thunderstorm frequencies. For the longitude feature, the highest number of thunderstorms is predicted between 1 degree west and longitude 0.

The behaviour of the ERA5 atmospheric features in driving thunderstorms in Ghana meets the current understanding of how these features affect convection. For instance, CAPE represents atmospheric instability, while CIN represents the ability of the lower troposphere to prevent convection and hence thunderstorms from occurring if the CIN values are high (Chen *et al*., 2020). High values of TISR increase the energy available for convection. The behaviour of SI10 is attributed to the wind patterns in Ghana. During the dry harmattan season, high wind speed prevails over parts of Ghana, especially over the northern stations. These strong and dry winds will suppress the occurrence of thunderstorms compared to the weak south-westerly winds during the rainy season (Parker and Diop-Kane, 2017).

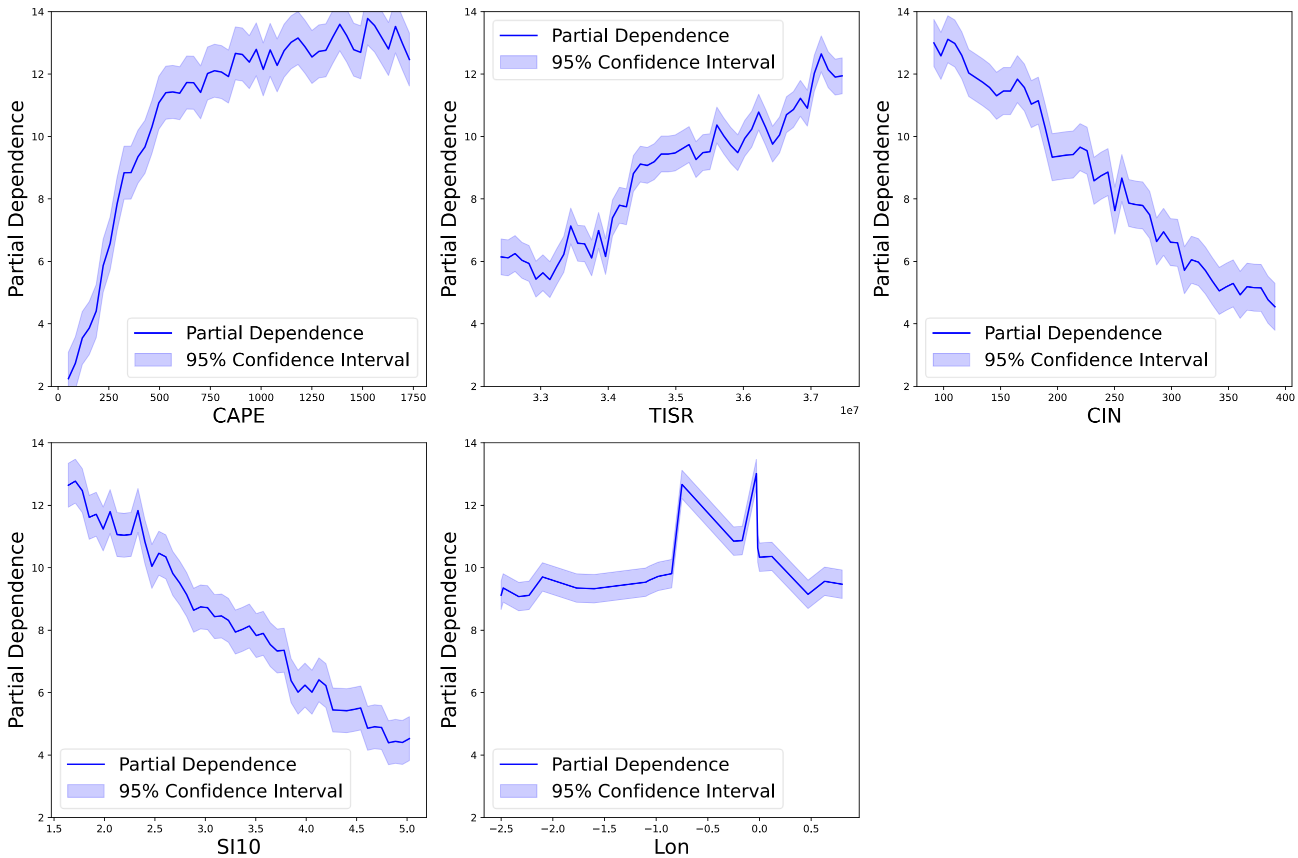

Figure 4.14. Partial dependence plot of LightGBM model, with 95th confidence interval. Showing the monthly thunderstorm dependence on the features.

4.1.6 Trend Analysis

In this section, the results of the trend analysis are presented. Identifying the trends of thunderstorms and the identified climate drivers of thunderstorms will provide useful information into whether the trend of thunderstorms can be attributed to shifts in these key climate variables. Both monthly thunderstorm data and features were converted to annual values. The yearly data were aggregated across all stations to create a single dataset for each variable. Each aggregated dataset was then used to perform the trend analysis.

The trend values and confidence intervals are presented in Table 4.4. The trend analysis of thunderstorm frequency over Ghana has revealed a statistically significant downward trend. The thunderstorm trend value is -0.73 thunderstorms per year. The downward trend of thunderstorm frequencies over Ghana is consistent with earlier studies that have reported a downward trend of rainfall in Ghana (Owusu *et al*., 2008; Nkrumah *et al*., 2014). Both downward trends in thunderstorms and rainfall have many implications for Ghana's agriculture, which is mainly rain-fed (Abbam *et al*., 2018). This will also affect hydropower generation due to reduced water levels (Kabo-Bah, 2016).

The trend results of the identified drivers of thunderstorms are as follows: CAPE recorded a significant trend of -3.92 J/kg per year. CIN also recorded a significant trend value of 1.44 J/kg per year. SI10 and TISR recorded no statistically significant trend.

The downward trend of CAPE has earlier been reported to occur in the tropics (Taszarek *et al*., 2021). This downward trend of CAPE, the primary driver of thunderstorm frequency in Ghana, agrees with the observed downward trend of thunderstorms. The LightGBM model revealed a positive effect of CAPE on thunderstorms (Figure 4.13, Figure 4.14). This is consistent with synchronous trends of CAPE and thunderstorm frequencies. CIN, the third most important driver of thunderstorms, recorded a significant upward trend. Since CIN is negatively related to thunderstorm frequencies (Figure 4.13, Figure 4.14), the increasing trend of CIN could reduce the number of thunderstorms, and hence the downward trend of thunderstorms over Ghana. The second and fourth most important drivers did not have a significant trend. This indicates that the trend of CAPE and CIN are the most important drivers in contributing to the reduction in thunderstorms interannually.

Table 4.4. Trend analysis results of Ghana annual thunderstorm frequencies from 1990-2013. Included in the table are the lower and upper confidence intervals (CI).

Variable	**Trend per year**	**Lower CI**	**Upper CI**
Thunderstorm frequency	-0.73*	-1.23	-0.12
CAPE	-3.92* J/kg	-7.46	-0.36
TISR	99 J/m^2	-533	613
CIN	1.44* J/kg	0.92	1.97
SI10	0.00 m/s	0.00	0.00

* Trend is significant at $\alpha = 0.05$.

4.2 Thunderstorm Prediction Model

4.2.1 Model Performance and Evaluation

The best features selected from the RFE approach for the models to predict thunderstorms one year ahead are Convective Available Potential Energy (CAPE), clear-sky direct solar radiation at the surface (CDIR), 10 m wind speed (SI10), latitude (lat), El Niño/Southern Oscillation (ENSO) region 1+2 (Niño 1+2), Pacific Decadal Oscillation (PDO), North Atlantic Oscillation (NAO), and Western Hemisphere Warm Pool (WHWP).

Although atmospheric stability indices are well known to be associated with thunderstorms and some have been used in developing models to predict diurnal thunderstorm occurrence (e.g., Ukkonen & Mäkelä, 2019), only CAPE was found to be among the eight most relevant features in this study.

Table 4.6 represents the performance metrics of the Climatology, LightGBM, XGBoost, RF, and Ensemble models. It shows the Ensemble model to be the best performing model in predicting monthly thunderstorm frequencies over Ghana, capturing 76% of the variance of monthly thunderstorms in Ghana. The Climatology model captured the least variance of 70%; it also recorded the least performances in all the other metrics as well. LightGBM, XGBoost, and RF captured 74%, 75%, and 75% of the variance, respectively. All the other models show an improvement over the Climatology model.

Figure 4.15 shows a quantile-quantile (Q-Q) plot comparing the performance of the three ML models and the Ensemble model, i.e., the quantiles of the predicted values plotted against those of the observations. The multiple values per quantile in the Q-Q plot are due to repeated values in both the observed and predicted thunderstorm counts per month, as these data are discrete and from 22 different stations over three years, causing clustering at specific quantiles. The LightGBM model predicts better across all quantiles compared to the other models. The other models predict better in the lower and middle quantile ranges than in the upper quantile. The XGBoost particularly underpredicts in the upper quantile more than the other models.

To further examine the performance of the models across all the stations, line plots of the observed and predicted thunderstorm frequencies are made and presented in Figures 4.16-4.18.

The model performance over the coastal zone is presented in Figure 4.16. The results show that all the models captured the number of thunderstorms over this zone well. However, the models underestimated the number of thunderstorms in some months (Figure 4.16a,b,c,e,f). All models captured well the thunderstorms in the middle and northern zones (Figures 4.17, 4.18). However, in these zones as well, some of the extremely low and high thunderstorm frequencies are not well captured by the models.

Table 4.5: Metric table showing the performance of the models for the different period lengths used in developing the thunderstorm prediction model.

Period length (months)	LightGBM (R^2)	XGBoost (R^2)	Random Forest (R^2)	LightGBM (MAE)	XGBoost (MAE)	Random Forest (MAE)
2	0.68	0.73	0.73	2.93	2.63	2.63
3	**0.74**	**0.75**	**0.75**	**2.63**	**2.55**	**2.59**
4	0.72	0.68	0.73	2.71	2.85	2.65
5	0.71	0.73	0.74	2.74	2.71	2.64
6	0.70	0.71	0.71	2.75	2.66	2.75
12	0.68	0.70	0.71	2.86	2.75	2.82

Table 4.6: Evaluation table for the regression models. The predicted values were evaluated against the test set (thunderstorm frequency per month). Bold numbers highlight the best performance each.

Metrics	Climatology	LightGBM	XGBoost	RF	Ensemble
MAE	2.65	2.62	2.55	2.59	**2.49**
RMSE	3.49	3.42	3.37	3.36	**3.27**
MSE	12.16	11.72	11.35	11.27	**10.70**
R^2	0.70	0.74	0.75	0.75	**0.76**
Bias	1.65	0.12	-0.33	0.06	**-0.05**

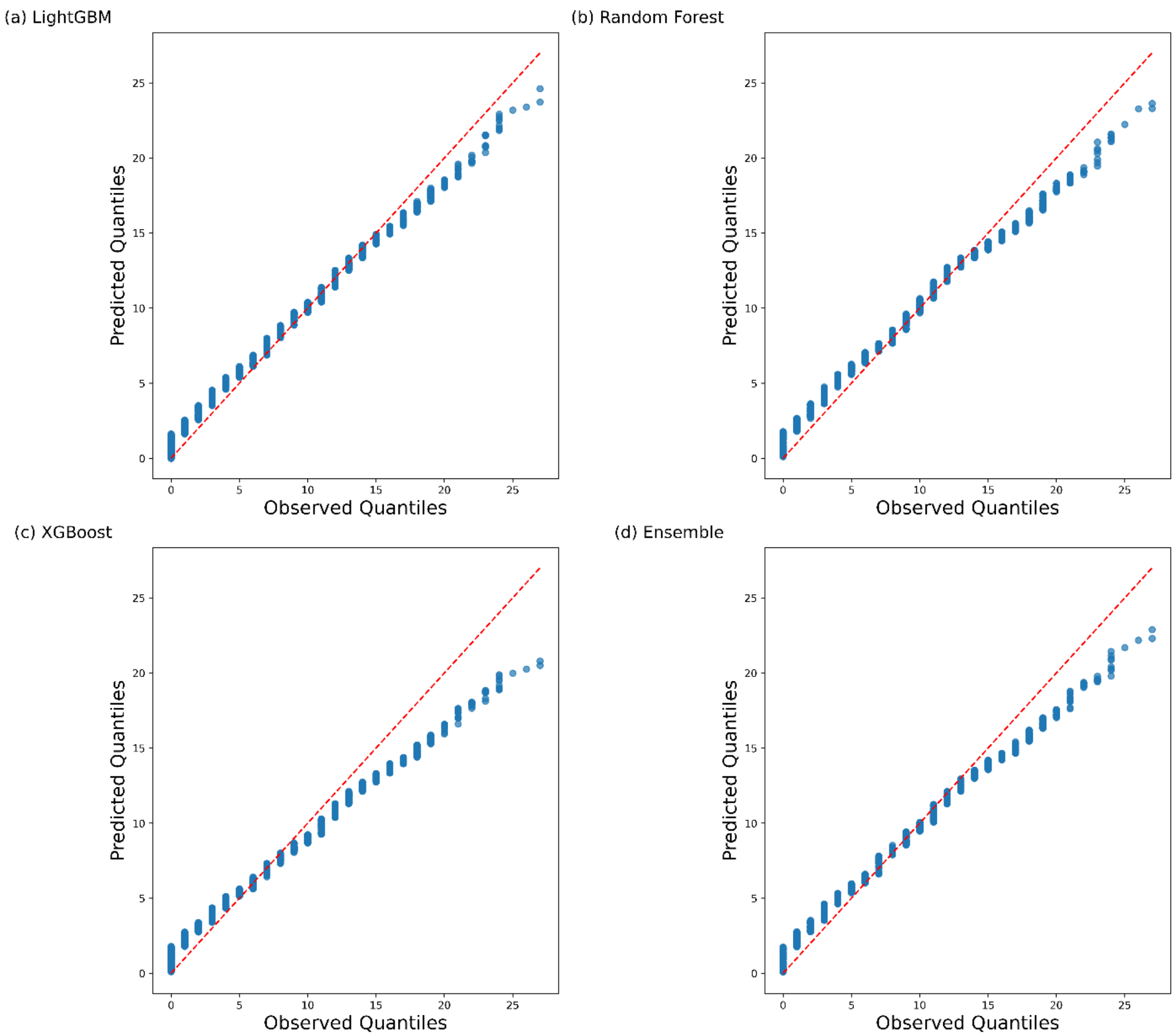

Figure 4.15: Q-Q plot of the observed and predicted thunderstorms for (a) LightGBM, (b) Random Forest, (c) XGBoost, and (d) Ensemble models.

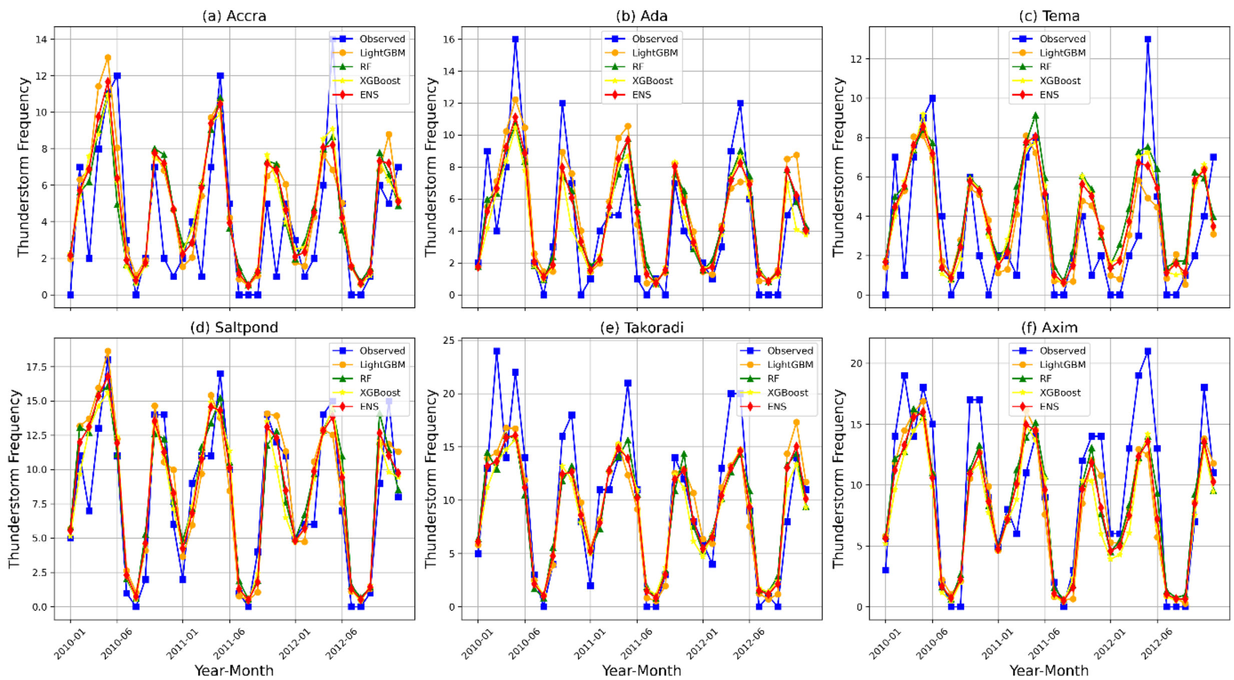

Figure 4.16. Model evaluation plots, showing the performance of the various models over the Coast stations.

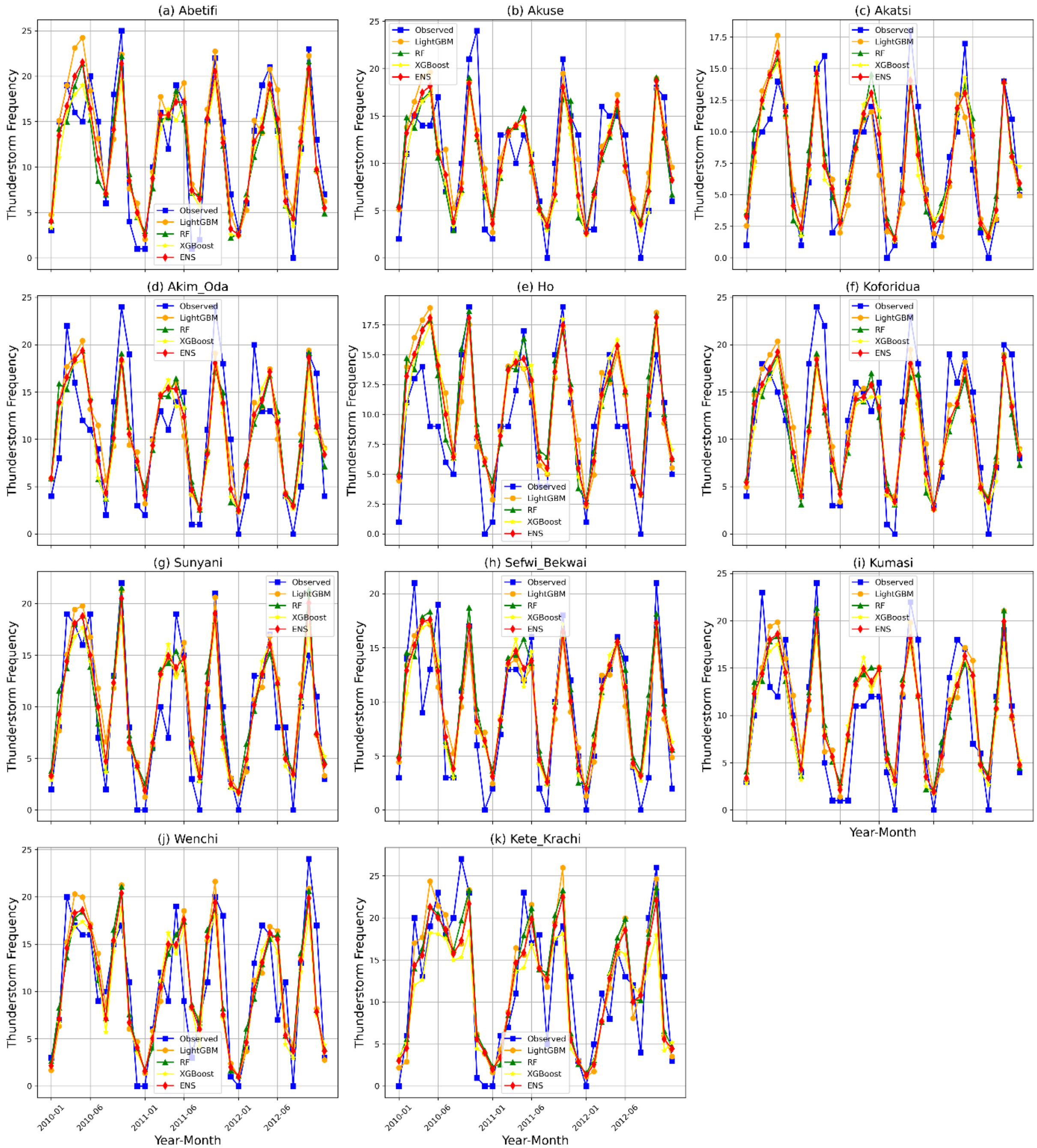

Figure 4.17. Model evaluation plots, showing the performance of the various models over the Middle stations.

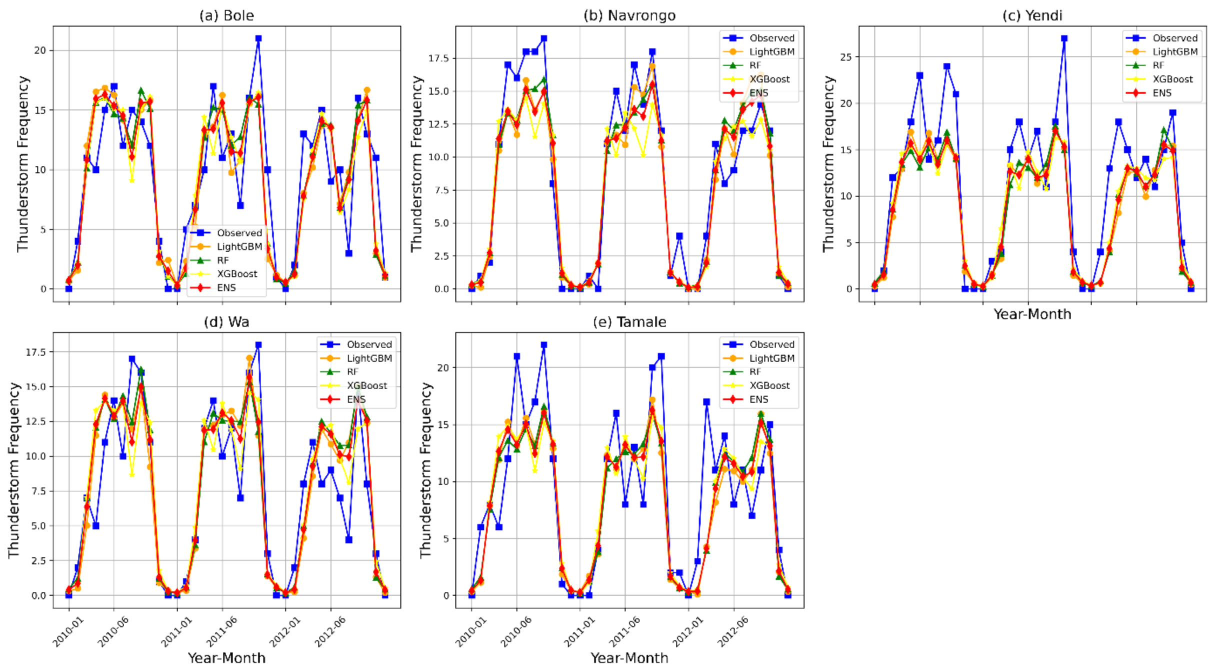

Figure 4.18. Model evaluation plots, showing the performance of the various models over the Northern stations.

4.2.2 Model Behaviour Analysis

The Shapley additive explanations (SHAP) plot presented in Figure 4.19 shows the contribution of the selected eight features to the prediction of the RF model. The most important features are in the upper part of the plot. The order of feature importance in the model is CAPE > CDIR > SI10 > WHWP > lat > Niño 1+2 > NAO > PDO.

The SHAP values show that high amounts of CAPE and clear-sky direct solar radiation at the surface positively impact the model, and lower values (blue dots) of both CAPE and CDIR impact the model negatively. The impact of CAPE and CDIR on the model output meets common expectations. High CAPE values lead to instability and convection. When the solar radiation at the surface (CDIR) is high, it provides enough energy for convection, and a high occurrence of thunderstorms is expected. The 10 m wind speed impact on the model output is also very plausible. Strong winds mostly occur over Ghana during the harmattan (dry) season, especially over the northern stations. Therefore, the strong and dry harmattan winds will suppress thunderstorm formation in Ghana as shown in the SHAP plot. Although the mean south-westerly winds during the rainy season are very weak, very strong gales can occur from thunderstorms (Parker & Diop-Kane, 2017).

WHWP is the climate index that contributes the most to the model output. Generally, lower WHWP and NAO values negatively impact the model output, and higher values impact the model output positively. The inverse is true for Niño 1+2. PDO had the least impact on the model of all the selected eight features. The high concentration of both high and low values of PDO around the centre of the SHAP plot indicates this feature has a weak influence on the model output.

All the developed ML and the Ensemble models outperformed the Climatology model. The Climatology model has a very high model bias and captures the least thunderstorm variability. The year-ahead forecasts of monthly thunderstorm frequency with good evaluation metrics of the ML and Ensemble models will help hydropower managers optimise water storage and release strategies based on anticipated thunderstorm-induced rainfall, leading to more efficient power generation and resource management. It can also help farmers plan and mitigate the risks of floods caused by thunderstorms. This is crucial given the uncertainty of convection in a changing climate (Taszarek *et al*., 2021). The Ghana Meteorological Agency could play a central role by providing seasonal forecasts of thunderstorm frequencies based on the developed models in this work.

From the SHAP results, the behaviour of the ERA5 reanalysis features is less complex compared to the climate indices. The complex relationship of the climate indices and thunderstorms in Ghana is discussed. When WHWP is strengthened and extends eastward into the tropical Atlantic, it supplies moisture to the Atlantic intertropical convergence zone (ITCZ) (Drumond *et al*., 2011). Therefore, in this phase, the WHWP helps to push the ITCZ northward, favouring moisture supply to Ghana and an increase in convection. Therefore, the SHAP plot has captured well the behaviour of WHWP on Ghana thunderstorms. Indirectly, the WHWP has been reported to induce cooling in the Pacific Ocean, and this cooling can move equatorward to form La Niña conditions (Park *et al*., 2018). La Niña conditions enhance rainfall over West Africa.

The SHAP plot values of Niño 1+2 are also consistent with current knowledge. Generally, ENSO events lead to enhanced northeasterlies over West Africa and thus low moisture influx and weaker monsoons (Janicot *et al*., 2001). Therefore, high values of Niño 1+2 will lead to a lower number of thunderstorms, as shown by the SHAP plot. In contrast, the effects of NAO and PDO in the SHAP plot are surprising. Positive NAO generally is accompanied by the intensification of the Azores high-pressure system, which in turn leads to the intensification of the dry harmattan winds (Fiedler *et al*., 2015). Therefore, high values of NAO are expected to suppress thunderstorm occurrences over Ghana, contrary to what the SHAP plot depicts. PDO has also been shown to negatively correlate with West Africa Sahel rainfall (Diatta & Fink, 2014), contrary to the SHAP plot in this work.

For the latitude feature, generally, higher latitudes impact the model output negatively, and lower values impact the model positively. Thus, areas within low-latitude areas generally have higher thunderstorms predicted than areas within higher latitudes in Ghana.

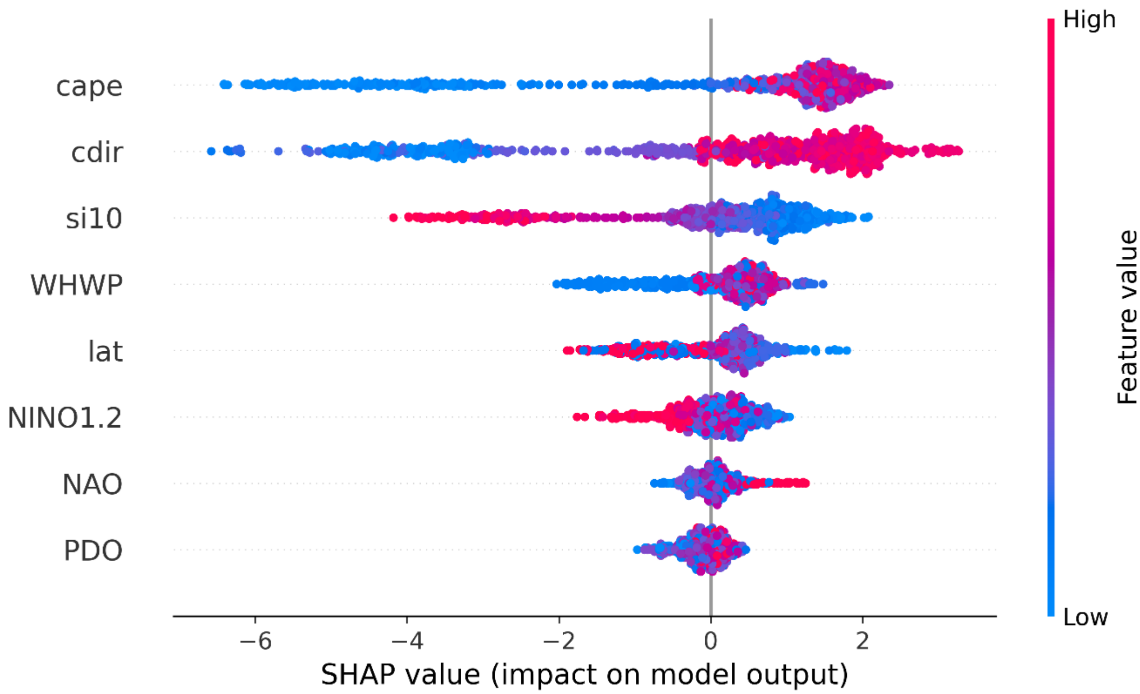

Figure 4.19: Shapely additive explanations (SHAP) plot of the final features used in training the Random Forest model over Ghana.

4.3 Precipitation model for seasonal forecasting

4.3.1 Model performance and evaluation

The performance of the various models in predicting precipitation over West Africa is presented in Table 4.7. The models captured the variability of precipitation ranging from 77% to 84%.

To show the spatial distribution of biases and errors in the model, the model bias and MAE values were computed for each grid cell using the test and predicted data and displayed in Figure 4.20. The model underpredicts over the Guinea Coasts of Ivory Coast and Ghana and over the Gulf of Guinea region just below Nigeria (Figure 4.20a). The model overpredicts in most parts of eastern and northern Nigeria and the Guinean Dome. Generally, over most parts of West Africa, the bias values are not high. The model error values are highest over the Gulf of Guinea (Figure 4.20b). This could be partly because the rainfall values over those areas are high.

Figure 4.21 shows how the bias-corrected ensemble model simulated the mean monthly precipitation over West Africa one year ahead, capturing areas with high precipitation and other important features of West African precipitation, like the Ghana-Benin dry zone (Vollmert *et al*., 2003).

Figure 4.22 shows the model's performance at nine randomly selected sites, three sites each in the Guinea Coast, Savanah, and Sahel. The model is able to capture the variability of the monthly rainfall amount in all nine randomly selected points. The model appears to perform better in the selected Savanah and Sahel points than on the Guinea Coast.

Table 4.7: Evaluation table for the LightGBM and XGBoost regression models of different objective functions. The predicted values were evaluated against the test set (precipitation per month). Bold numbers highlight the best performance each.

Metric	XGBoost-Regression	XGBoost-Tweedie	LightGBM-Tweedie	LightGBM-Regression	Ensemble	Bias corrected Ensemble
MAE	36.50	28.87	32.24	31.10	29.04	**26.44**
RMSE	58.74	53.92	57.48	53.87	52.40	**49.38**
KGE	0.83	0.88	0.87	0.86	0.87	**0.89**
R^2	0.77	0.80	0.78	0.80	0.82	**0.84**
Model bias	9.33	-1.03	3.81	2.11	**0.54**	0.91

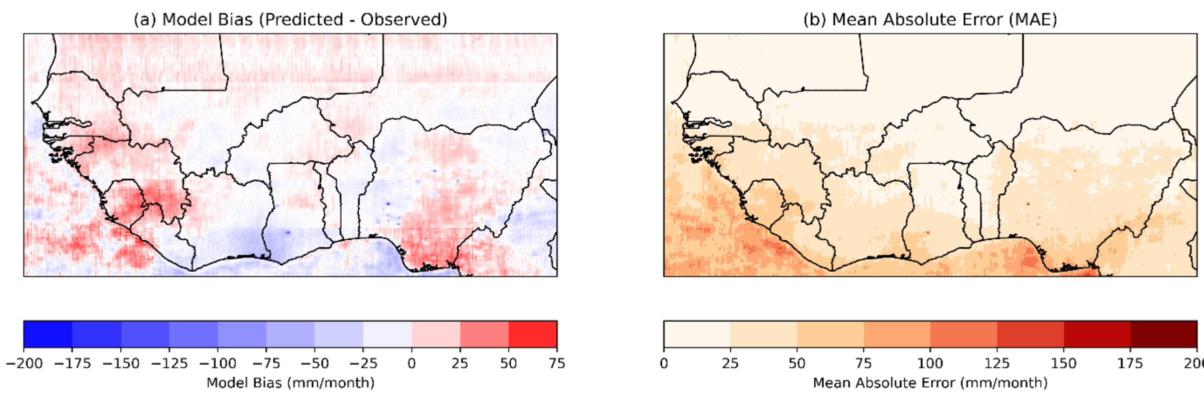

Figure 4.20: The spatial evaluation of the bias-corrected ensemble model bias (a) and mean absolute error (b). These were computed using the test and predicted data.

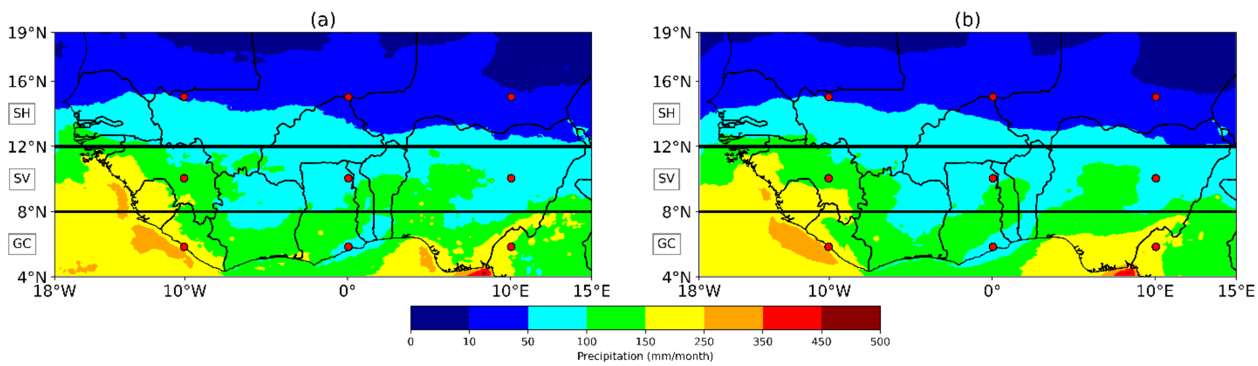

Figure 4.21: Mean test (observed) (a) and mean predicted (b) monthly precipitation over West Africa using the bias-corrected ensemble model for the test period of 9/2020–8/2023. The red dots indicate the selected sites for the timeseries plot of the observation and prediction in Figure 4.22. The black solid line shows the division of the various zones from south to north: Guinea Coast (GC), Savanah (SV), and Sahel (SH).

Figure 4.23 is the latitude-time plot to help understand the model's ability to capture the monthly precipitation amount change in time and latitude. The model captures the "monsoon jump" from June on the Guinea Coast to August further north; the "monsoon jump" in the predicted plot (Figure 4.23b) is almost as distinct as what is seen in the observation (Figure 4.23a). The model also captures the little dry season in July and August on the Guinea Coast, but less distinct in July (Figure 4.23b). However, there are some differences between the observation and the predicted precipitation amount. For instance, the model slightly underpredicts the amount of precipitation over the Guinea Coast in June and slightly overpredicts the amount in July over the Guinea Coast. Therefore, the little dry season is not as distinct in the prediction (Figure 4.23b) as it is in the observation (Figure 4.23a).

LightGBM and XGBoost regression models were developed in this work to predict monthly precipitation over West Africa one year ahead. The models were developed with IMERG precipitation as the target and ERA5 reanalysis data, climate indices, latitude, and longitude as features.

The ML models performed considerably well in space and time over West Africa for the study period, capturing key spatial and temporal features such as the 'monsoon jump', 'little dry season', and Ghana–Benin dry zone. The models captured 77% to 84% of the monthly precipitation variance, indicating strong predictive ability. Olaniyan *et al*. (2018) evaluated the European Center for Medium-range Weather Forecast's Sub-seasonal to Seasonal (ECMWF-S2S), with a 46-day lead time. They observed that ECMWF-S2S performed best in Savannah, with a synchronisation of 75%; the lowest performance occurred in the Gulf of Guinea, with a synchronisation slightly above 50%. Although the ECMWF-S2S and current ML models operate at different lead times, both exhibit lower performance along the Guinea Coast. The ML has shown improved accuracy in the Savannah and Sahel regions. These patterns suggest that precipitation predictability varies spatially across West Africa, with models generally performing better in drier northern zones than in humid coastal areas.

The performance of S2S models over West Africa has been observed to decrease significantly with increasing lead time. The C3S seasonal forecast models, evaluated by Gebrechorkos *et al*. (2022), demonstrate good skill at a lead time of one month, with ECMWF emerging as the most skilful model. However, the skill declines rapidly beyond a two-month lead time. Similarly, Siegmund *et al*. (2015) concluded that CFSv2 precipitation forecasts in West Africa for up to six months exhibit low to moderate skill. In contrast, the strong predictive performance of the current ML models at a one-year lead time suggests their potential as valuable forecasting tools for the region.

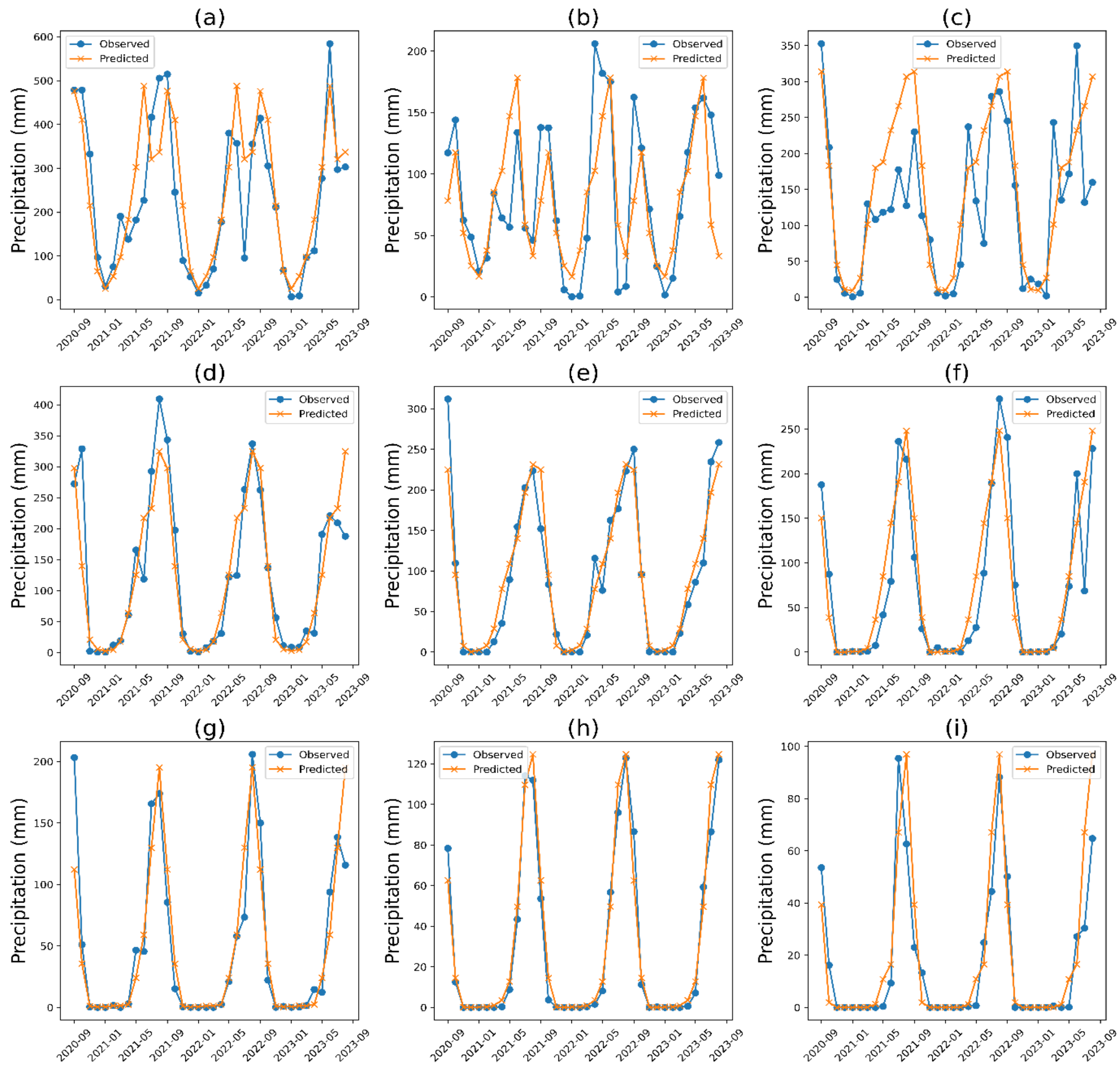

Figure 4.22: Plots of timeseries of observed and predicted precipitation at three points at the Guinea Coast (a-c), three points in the Savanah (d-f), and three points in Sahel (g-i).

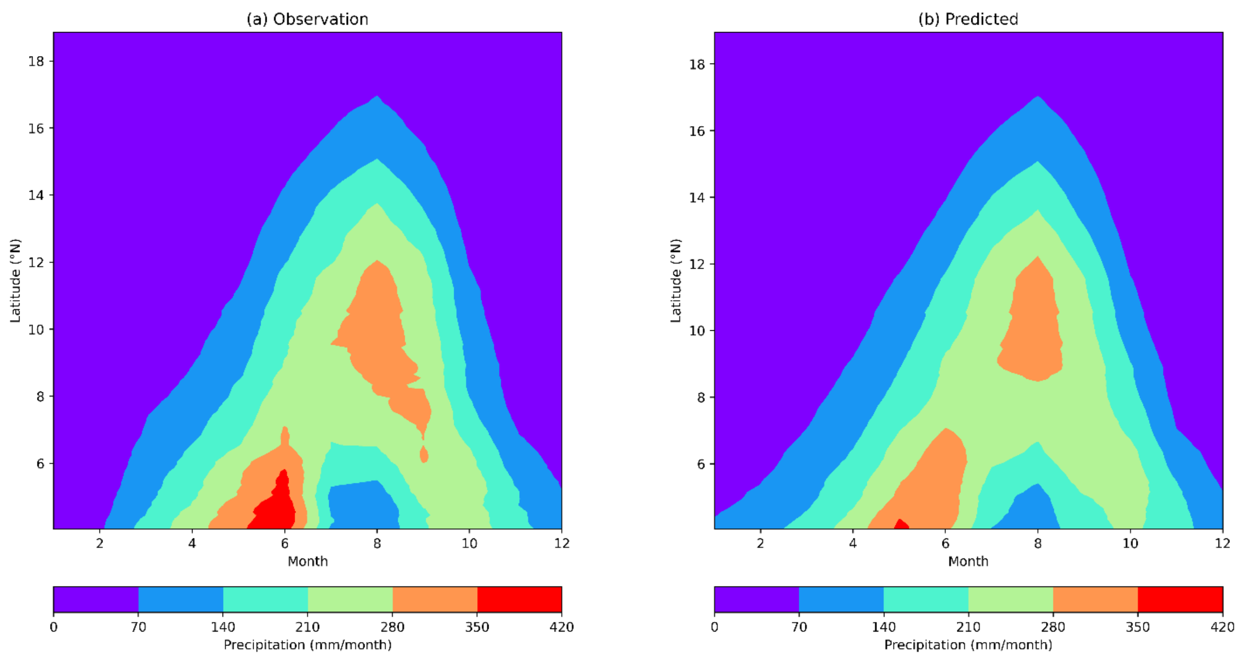

Figure 4.23: Latitude-time plot of the test (observed) (a) and predicted (b) monthly precipitation over West Africa.

Table 4.8: The selected latitudes and longitudes used for SHAP and PDP analyses. The longitude ranges were kept the same in all zones, for the precipitation models.

Zone	Latitude range	Longitude range
Guinea Coast	4 – 6 degrees north	2 degrees west – 2 degrees east
Savanah	8 – 10 degrees north	2 degrees west – 2 degrees east
Sahel	12 – 14 degrees north	2 degrees west – 2 degrees east

4.3.2 Model Behaviour Analysis

To analyse the model behaviour and how various features drive the monthly precipitation variability, SHAP and PDP plots were employed. To help understand how the features behave in the various zones (Guinea Coast, Savanah, and Sahel), these zones were analysed separately; however, using the same model. Thus, by varying the latitudes, how the features behave differently in the various zones in predicting monthly precipitation is shown. The selected subregions are shown in Table 4.8. To understand and compare the behaviour of the models in predicting the mean precipitation, extreme low and high precipitation, SHAPs and PDPs are made using LightGBM with the following objective functions: regression (for the precipitation amount) and quantile with alpha of 0.95 (for extreme high precipitation) and alpha of 0.05 (for extreme low precipitation). All these were done over the selected areas as shown in Table 4.8.

To help in the comparison, all the longitudes, model hyperparameters, and the number and type of features were all kept the same. The only changes are the latitude and the values of the features and target depending on the location.

4.3.2.1 Features behaviour during precipitation amount prediction

The SHAP plots of the LightGBM with regression as the objective function are presented in Figure 4.24, showing the feature values and the direction of the impact.

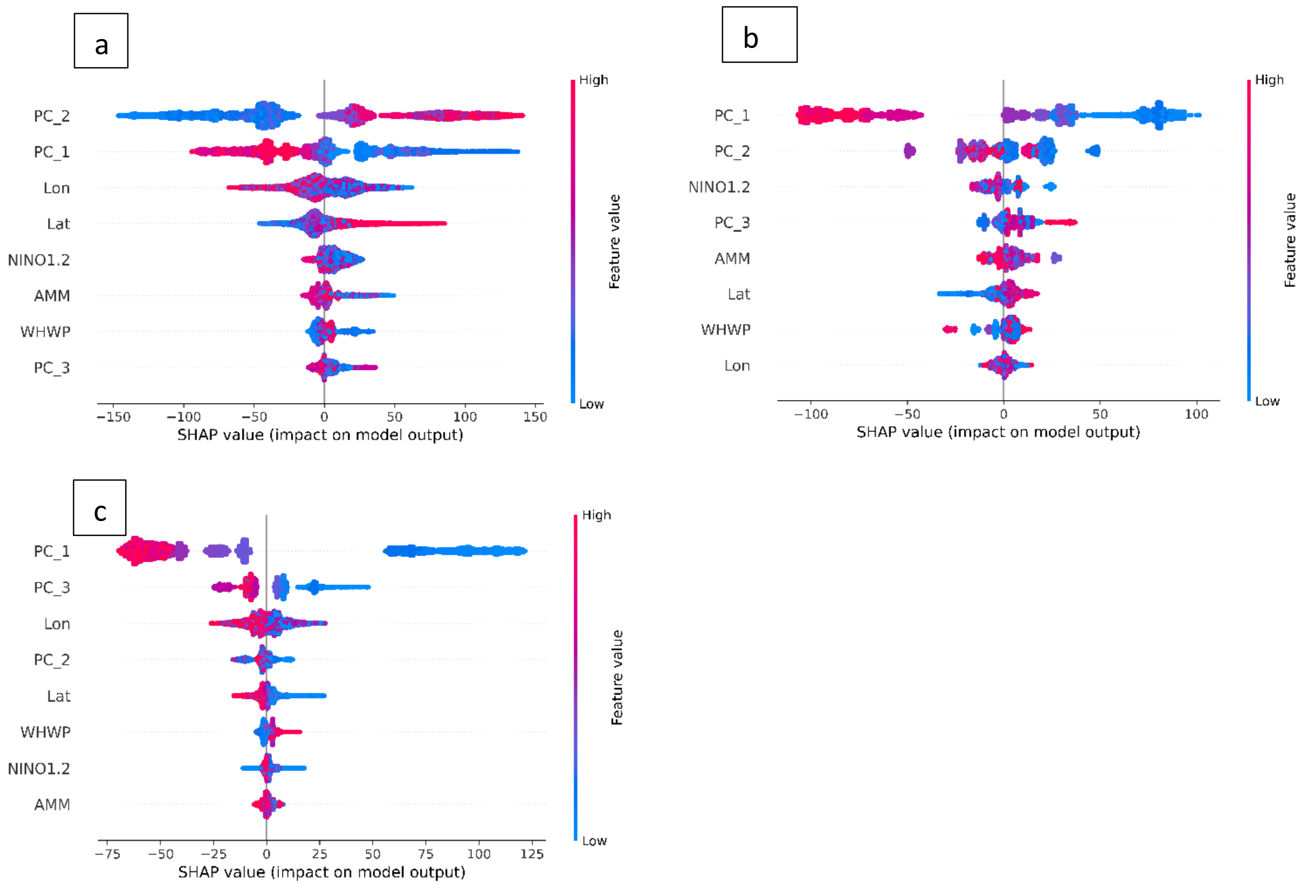

Figure 4.24: SHAP plots showing how the features impact the LightGBM (with regression as the objective function) model output in the Guinea Coast (a), Savanah (b), and Sahel (c). At the y-axis, features are sorted by relative importance, differing between the three zones.

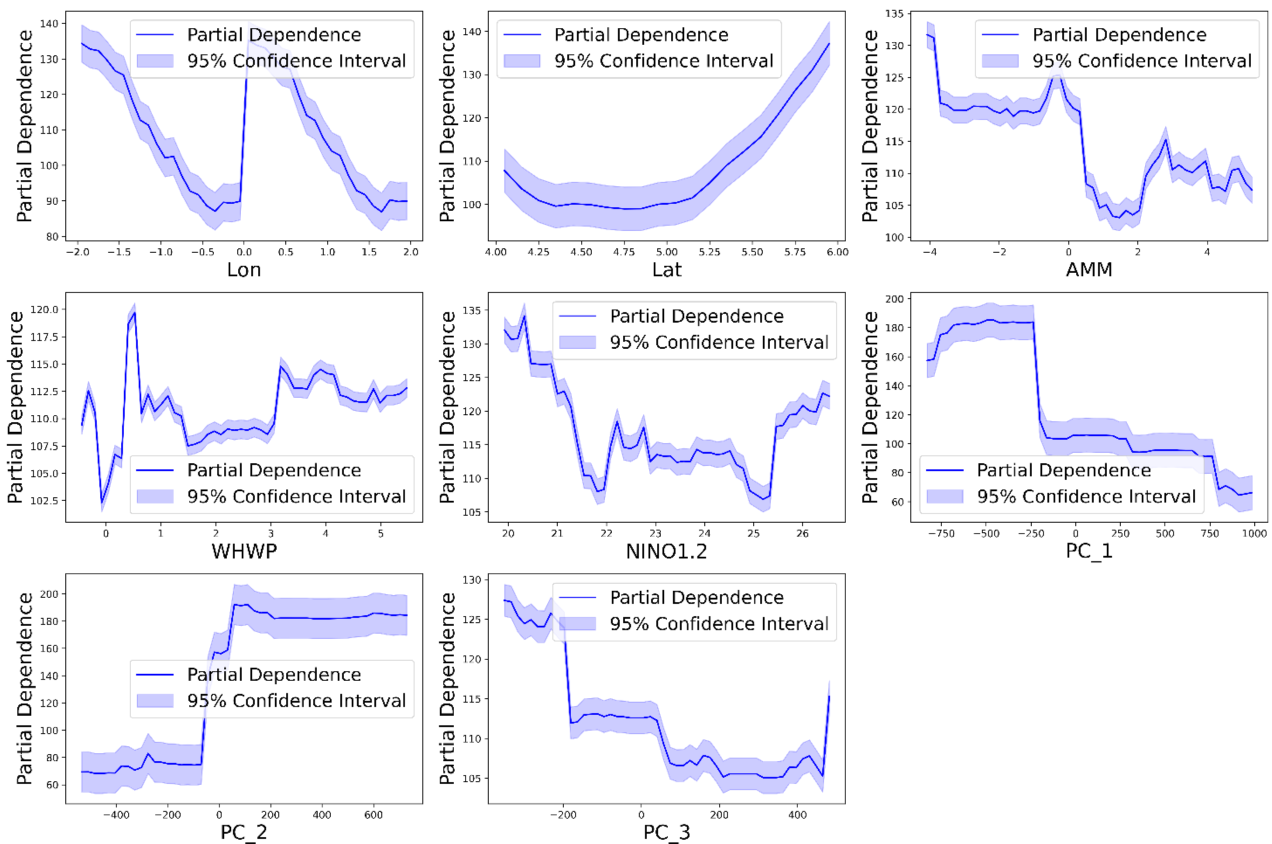

Figure 4.25: Partial dependence plots showing how the features impact the LightGBM (with regression as the objective function) model output in the Guinea Coast.

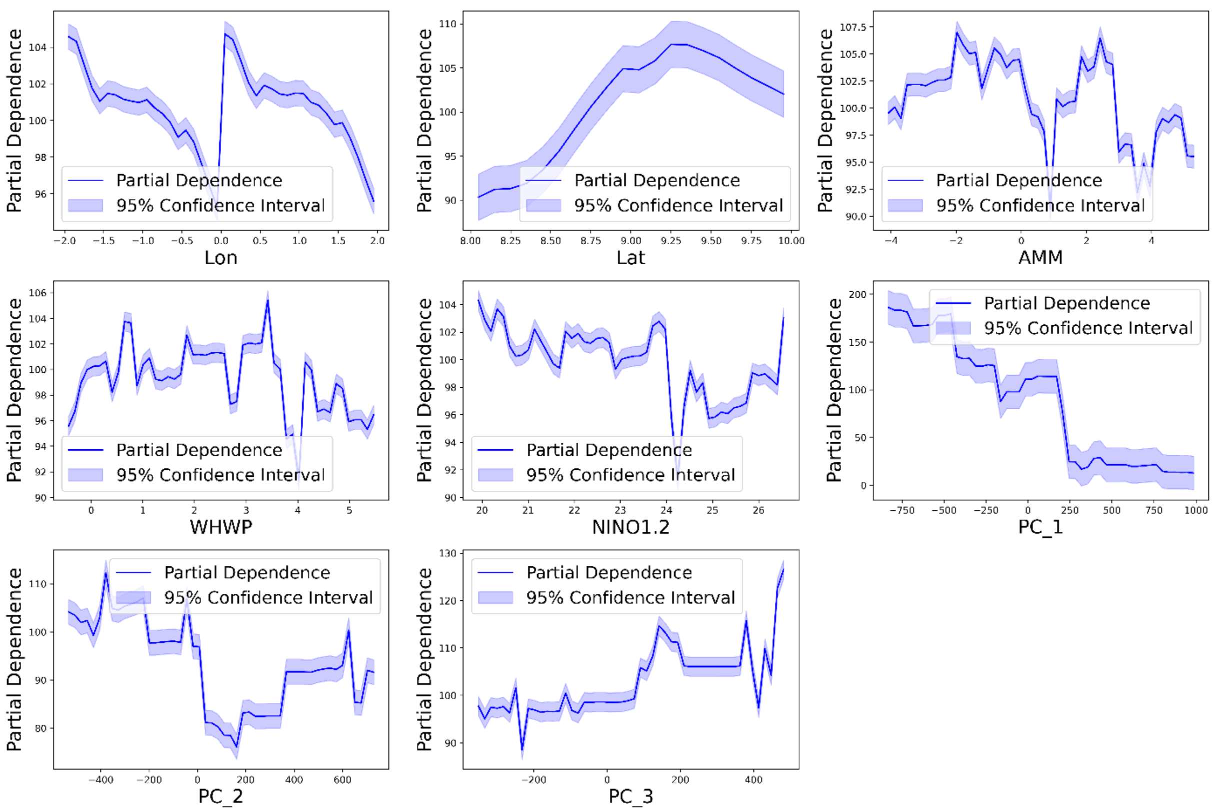

Figure 4.26: Partial dependence plots showing how the features impact the LightGBM (with regression as the objective function) model output in the Savanah.

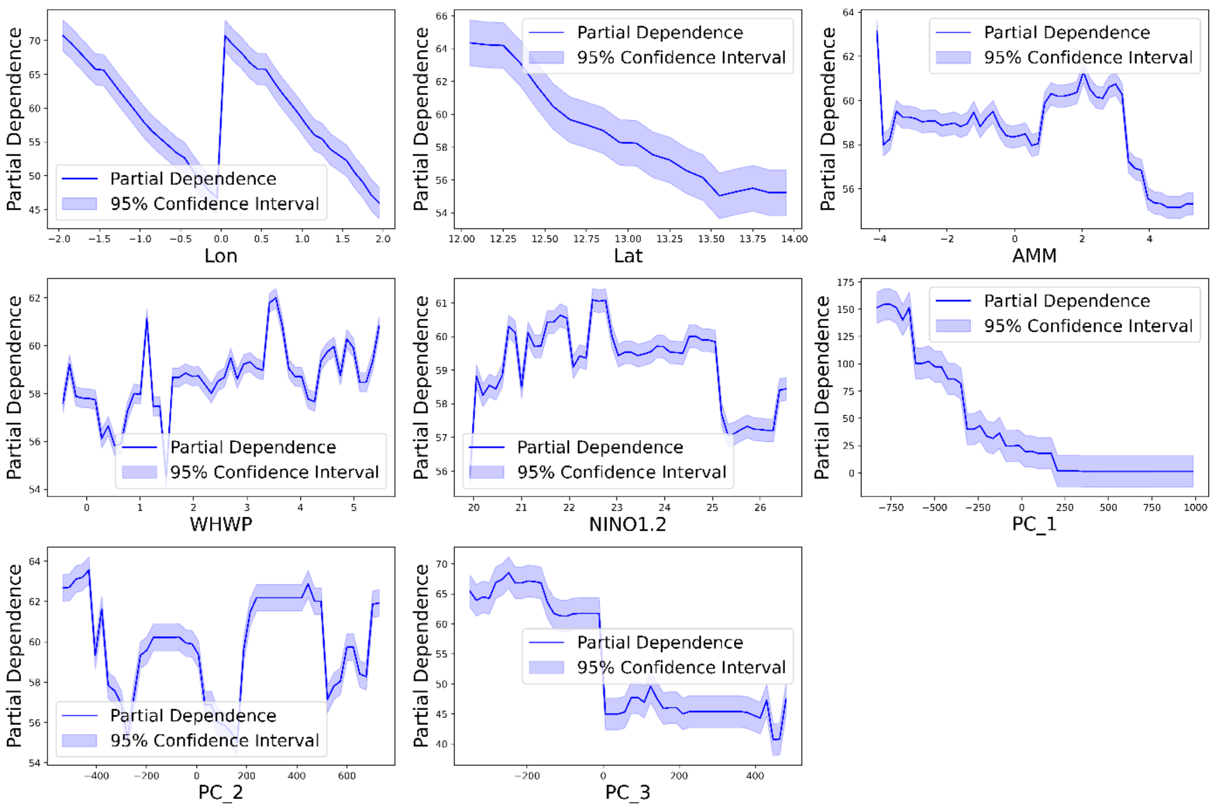

Figure 4.27: Partial dependence plots showing how the features impact the LightGBM (with regression as the objective function) model output in the Sahel.

For each zone, the order of feature importance in driving the monthly precipitation is: Guinea Coast: PC2 > PC1 > longitude > latitude > Niño 1+2 > AMM > WHWP > PC3 (Figure 4.24a). Savanah: PC1 > PC2 > Niño 1+2 > PC3 > AMM > latitude > WHWP > longitude (Figure 4.24b). Sahel: PC1 > PC3 > longitude > PC2 > latitude > WHWP > Niño 1+2 > AMM (Figure 4.24c). PC1 is the most important driver of precipitation over Savanah and Sahel (Figure 4.24b, c) and the second most important driver over the Guinea Coast (Figure 4.24a). In all the zones, high values of PC1 impact the model outputs negatively. Thus, high values of PC1 lead to a lower amount of precipitation predicted over West Africa. The impact of PC2 on the model output over the Guinea Coast (GC) is such that high values of PC2 impact model output positively. However, the impact of high values of PC2 on model output over the Savanah is negative; its impact over the Sahel is not very clear. The behaviour of PC3 over GC is such that low and medium values of PC3 impact the model output positively, and high values impact the model output negatively. High values of PC3 impact the model outputs over Savanah positively, but negatively over the Sahel.

For geographic coordinates, for the selected regions over West Africa, generally, a low amount of precipitation is predicted over the east compared to the west. A high amount of precipitation is predicted over higher latitudes than over lower latitudes in the GC and Savanah. Over the Sahel, higher precipitation is predicted over lower latitudes than higher latitudes.

The impacts of the climate indices on the model output in the various zones are as follows: Guinea Coast (GC) (Figure 4.24): Low and medium values of Niño 1+2 impact the output positively. High values of AMM impact the model output negatively, and low and medium values impact the model output positively. High values of WHWP impact the model output negatively, and low values impact the model output both positively and negatively. Savanah (Figure 4.24): Generally, high values of both AMM and Niño 1+2 impact the model output negatively. The impact of WHWP on model output over Savanah is such that very low and high values impact the model negatively. Sahel (Figure 4.24): Higher values of WHWP impact model output in Sahel positively. The impact of Niño 1+2 is not very clear, as low values impact the model both positively and negatively. High values of AMM appear to impact model output both positively and negatively.

The LightGBM regression model was developed to predict monthly precipitation over West Africa one year ahead by lagging the features by a year. Efforts were put in place to enhance model interpretability. The model was developed using GPM-IMERG precipitation as the target and ERA5 reanalysis data, climate indices, latitude, and longitude as features.

The model's evaluation has shown it performed considerably well in space and time over West Africa during the study period, capturing spatial and temporal features of the study area like the 'monsoon jump', 'little dry season', and the Ghana-Benin dry zone. The good performance of the model has implications for national meteorological seasonal forecasts. Much of the study area's agriculture is rainfed (Abbam *et al*., 2018); having a model that can help predict precipitation amounts with good performance one year ahead will help in agricultural planning and productivity in West Africa. Countries within the study area that generate electricity through hydroelectric power stations could also benefit from this model, as it could help in water management and planning. Indeed, other surface water resource managers could use the model in their planning.

Partial dependence plots (PDPs) with 95th confidence intervals are also made and presented to help understand how variations in the features drive precipitation. Over the selected area in the GC (Figure 4.25), longitude shows a non-linear relationship with monthly precipitation. As shown in the SHAP plot in the GC (Figure 4.24), higher amounts of precipitation are predicted over higher latitudes (Figure 4.25). The climate indices all have a non-linear relationship with precipitation. Generally, high precipitation values are predicted when AMM, WHWP, and Niño 1+2 have lower values than when they have higher values. High values of PC1 and PC3 contribute negatively to precipitation over the GC; however, the high values of PC2 contribute positively to precipitation.

Over the Savanah (Figure 4.26), the lowest amount of precipitation is predicted at longitude zero, while the highest amount of precipitation is predicted just east of longitude zero. Similar to the GC, higher precipitation values are predicted over higher latitudes than over lower latitudes in the Savanah zone. All the climate indices have a non-linear relationship with precipitation in the Savanah zone. For instance, initially increasing the values of AMM leads to an increase in the precipitation amount predicted, reaching a maximum around an AMM value of -2. Further increasing the AMM values beyond -2 leads to a decrease in the precipitation values, reaching a minimum when the AMM value is around 1. Also, increasing the AMM value from 1 to about 2 leads to an increase in precipitation; further increasing AMM values from 2 to 4 leads to a reduction in precipitation values. Both WHWP and Niño 1+2 also have a complex relationship with precipitation. Generally, initially increasing WHWP values lead to an increase in precipitation, peaking at a WHWP value of about 3.5, where further increases lead to reductions in precipitation values. Niño 1+2 values ranging from 21-24 lead to high precipitation prediction, and generally, values beyond 24 lead to a reduction in precipitation amount. High values of PC1 contribute negatively to precipitation in the Savanah zone, and low values of PC3 contribute negatively to precipitation amount. PC2 had a non-linear relationship with precipitation in the Savanah zone, where increasing PC2 values from -400 to about 200 leads to a decrease in precipitation, and further increasing PC2 values beyond 200 leads to an increase in precipitation over the Savanah.

Since the longitude values are the same over all the zones, the behaviour of longitude in the Sahel is similar to the GC and Savanah. The partial dependence of precipitation on latitude is such that high precipitation is predicted in lower latitudes than in high latitudes in the Sahel (Figure 4.27).

The impact of the climate indices is also non-linear in this zone. Generally, increasing AMM values from -4 to about 3 leads to increasing values of precipitation. However, further increasing AMM values beyond 3 leads to a decrease in precipitation. Generally, increasing WHWP values lead to higher values of precipitation in the Sahel. Initially increasing Niño 1+2 values from 20 to about 23 leads to an increase in precipitation; however, further increases beyond 23 lead to low precipitation. PC1 and PC2 have negative contributions to precipitation in the Sahel. PC3 has a non-linear relationship with precipitation in the Sahel (Figure 4.27).

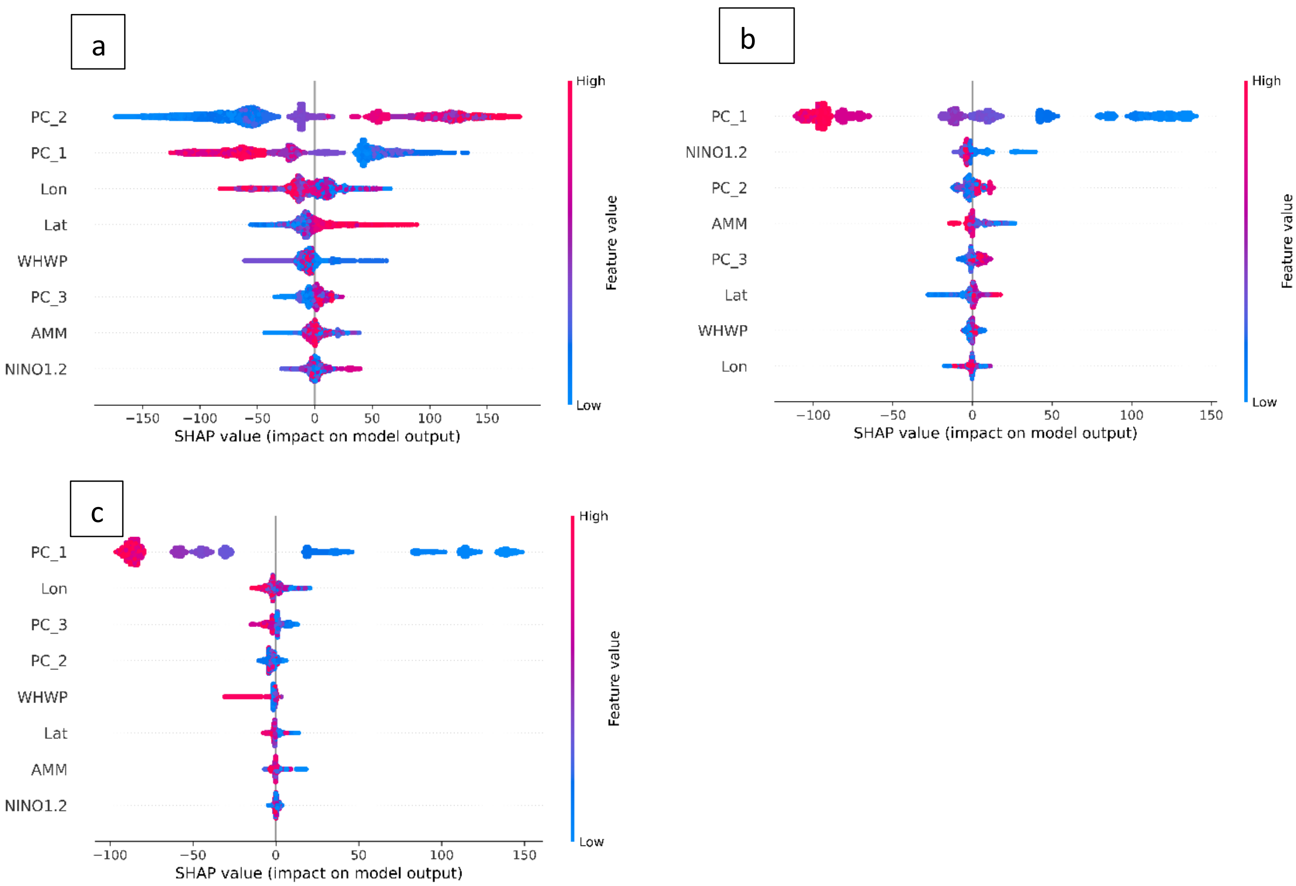

Figure 4.28: SHAP plots showing how the features impact the quantile at an alpha of 0.95 model output in the Guinea Coast (a), Savanah (b), and Sahel (c). At the y-axis, features are sorted by relative importance, differing between the three zones.

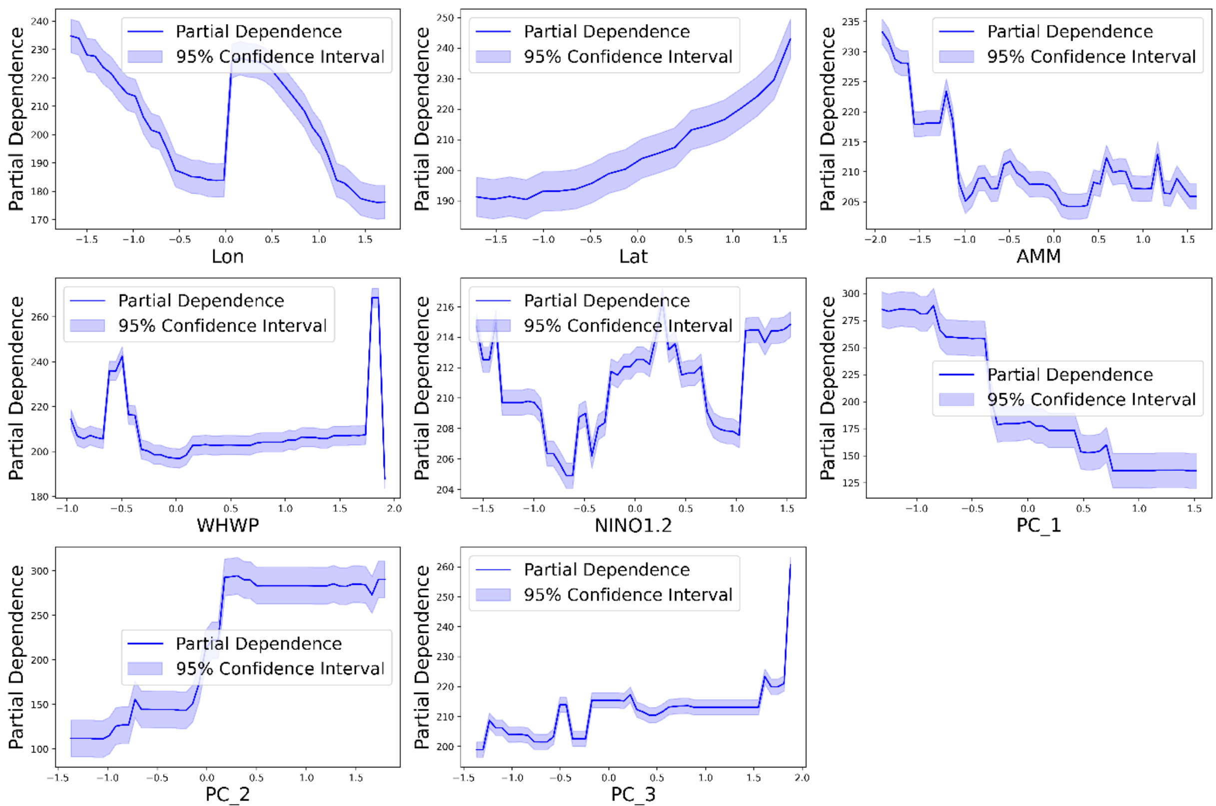

Figure 4.29: Partial dependence plots showing how the features impact the quantile at an alpha of 0.95 model output in the Guinea Coast.

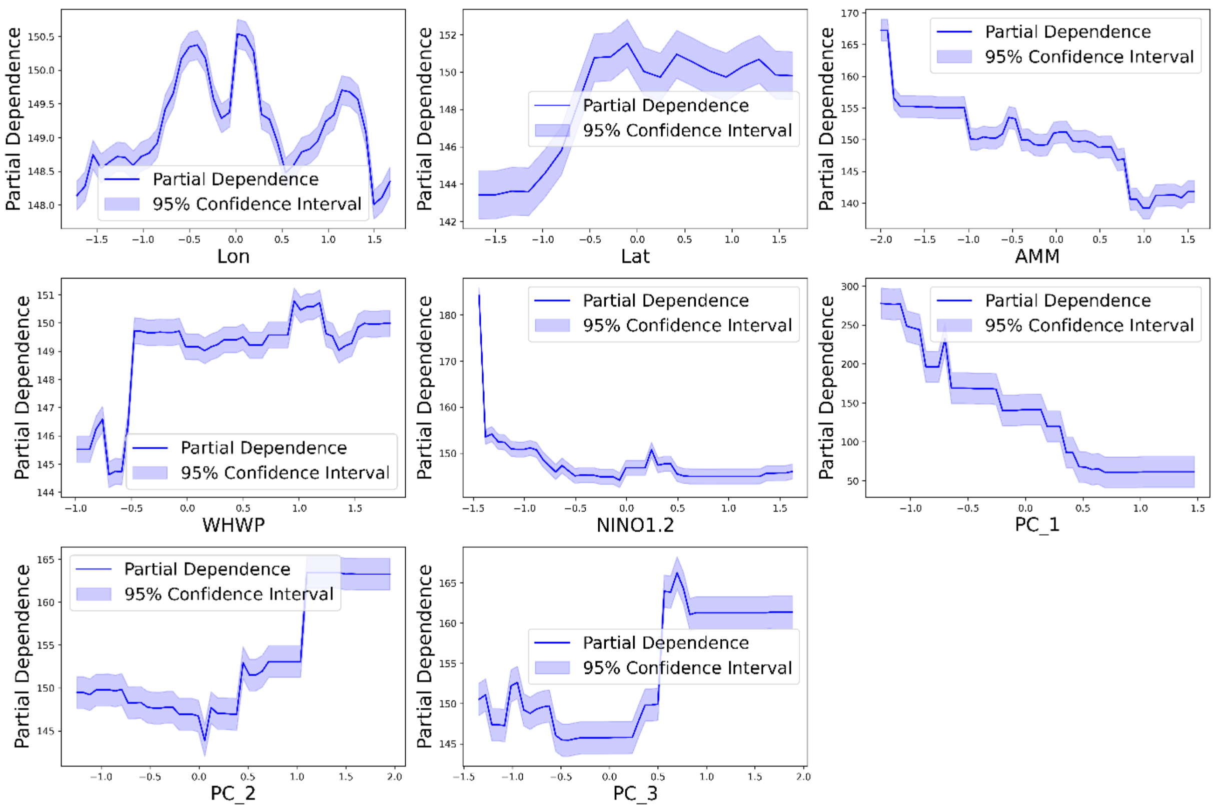

Figure 4.30: Partial dependence plots showing how the features impact the quantile at an alpha of 0.95 model output in the Savanah.

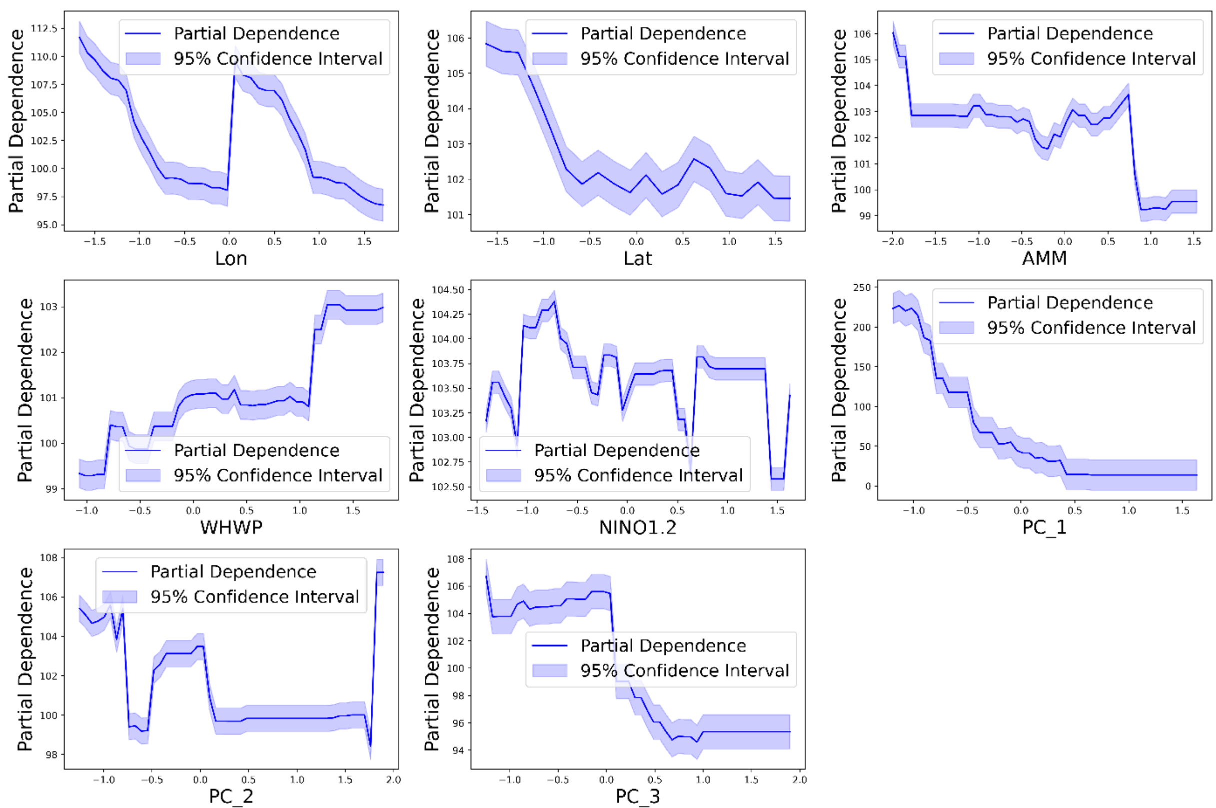

Figure 4.31: Partial dependence plots showing how the features impact the quantile at an alpha of 0.95 model output in the Sahel.

4.3.2.2 Features behaviour during extremely high precipitation prediction

To understand how the features impact and drive extreme precipitation prediction over the various zones, SHAP and PDPs are made using the LightGBM model with an objective of quantile and an alpha of 0.95. Therefore, the behaviour of the features in predicting the extremely high precipitation over the various zones is presented in this section.

The SHAP results (Figure 4.28) show the order by which features drive the extreme high precipitation. In the GC (Figure 4.28a), the order is as follows: PC2, PC1, longitude, latitude, WHWP, PC3, AMM, and Niño 1+2. High values of PC2, latitude, PC3, and Niño 1+2 contribute positively to extreme precipitation. Thus, high values of these features lead to increases in extreme precipitation values. However, higher values of PC1 and longitude lead to a lower amount of extreme precipitation. The impacts of WHWP and AMM are more complex. Medium and high values of WHWP lead to low precipitation prediction, and low values lead to high precipitation. High values of AMM concentrate around the zero-output value; however, medium values impact the model output positively, and low values impact the output negatively.

In the Savanah zone (Figure 4.28b), the order of features driving the extreme precipitation is as follows: PC1, Niño 1+2, PC2, AMM, PC3, latitude, WHWP, and longitude. The following have a negative impact on the precipitation: PC1, Niño 1+2, AMM, and longitude. Thus, high values of these features lead to a low amount of extreme precipitation prediction. PC2, PC3, and latitude had a positive impact on precipitation in the Savanah zone. The impact of WHWP in the Savanah zone is not very clear. However, it appears high values of WHWP lead to extreme precipitation occurrence.

In the Sahel zone (Figure 4.28c), the order by which the selected features drive precipitation is PC1, longitude, PC3, PC2, WHWP, latitude, AMM, and Niño 1+2. Higher values of the following features lead to a low amount of precipitation: PC1, longitude, PC3, WHWP, latitude, and AMM. How the rest of the features drive precipitation in this zone is not very clear.

Generally, in the different zones, the order and direction of how the features drive extreme precipitation are different.

To further understand how variations in one feature while keeping the rest constant affect the extreme precipitation values in the various zones, PDPs are made and analysed. In the GC (Figure 4.29), the highest values of precipitation are predicted around 1.50 degrees west; however, the precipitation values reduce as one moves toward the east, getting to a low around longitude zero. Moving further east from latitude zero leads to a sharp increase in precipitation to around 0.50 degrees east. However, from 0.50 east to around 1.50 east, there is a reduction in the amount of precipitation in the GC. The partial dependence of extreme precipitation on latitude is such that higher precipitations are predicted over higher latitudes than over lower latitudes (Figure 4.29). All the climate indices have a non-linear relationship with the extreme precipitation in the GC zone. However, generally, a high amount of precipitation is predicted when AMM has lower values than when AMM has higher values (Figure 4.29). Also, low values of WHWP lead to higher precipitation than medium and high values, except around a WHWP value of 1.7, where there is a spike in precipitation (Figure 4.29). The partial dependence of extreme precipitation on Niño 1+2 is as follows: High values of precipitation are predicted when Niño 1+2 has low values; as the values of Niño 1+2 increase to about -0.5, the predicted precipitation reduces. Further increases in Niño 1+2 values from -0.5 to about 0.2 lead to an increase in predicted precipitation, peaking at a Niño 1+2 value of about 0.2. Further increasing Niño 1+2 values from 0.2 to 1.0 leads to a reduction in predicted precipitation. However, further increases in Niño 1+2 values generally lead to an increase in precipitation values in the GC. The partial dependence of extreme precipitation on the principal components (PCs) is as follows: High values of PC1 generally lead to low precipitation being predicted compared to low values of PC1. However, high values of PC2 and PC3 generally lead to the prediction of high precipitation values (Figure 4.29).

The partial dependence of extreme precipitation on the selected features in the Savanah zone is presented in Figure 4.30. Generally, moving from the west towards longitude zero leads to an increase in predicted precipitation, and moving from longitude zero towards the east leads to a reduction in predicted precipitation (Figure 4.30). In the selected area of the Savanah zone, a higher amount of precipitation is predicted in higher latitudes than in lower latitudes. The partial dependence of extreme precipitation on the climate indices is as follows: low values of AMM and Niño 1+2 lead to prediction of high precipitation, while high values of AMM and Niño 1+2 lead to prediction of low precipitation (Figure 4.30). Over the Savanah, increasing the values of WHWP leads to high values of precipitation. The partial dependence on the PCs is as follows: increasing PC1 values leads to a reduction in the precipitation values. However, both PC2 and PC3 have similar impacts on extreme precipitation in the Savanah zone. Increasing the values of PC2 and PC3 leads to high values of precipitation (Figure 4.30).

Figure 4.31 shows the partial dependence of extreme precipitation on the features in the Sahel zone. Generally, the amount of precipitation decreases from west to about longitude 0. Further increasing from longitude 0 leads to a sharp increase in precipitation values, where further increasing the longitude values leads to a reduction in precipitation (Figure 4.31). The partial dependence of precipitation on latitude in the Sahel zone is such that higher precipitation values are predicted in lower latitudes than in higher latitudes. The partial dependence of extreme precipitation on AMM is such that increasing AMM values lead to a low amount of precipitation. Increasing the values of WHWP leads to an increase in precipitation in the Sahel zone. The partial dependence of precipitation on Niño 1+2 in the Sahel zone is non-linear. Generally, increasing Niño 1+2 values lead to low precipitation values (Figure 4.31). Increasing the values of PC1 and PC3 leads to a decrease in the predicted precipitation in the Sahel zone. The partial dependence of extreme precipitation on PC2 is non-linear and complex. However, initially increasing values of PC2 lead to a reduction of precipitation. Further increasing PC2 values from -0.5 to 0 leads to an increase in predicted precipitation. Increasing the values of PC2 beyond 0 leads to a sharp decrease in predicted precipitation and flattens out around the PC2 value of 0.1. Extreme values of PC2 around 1.8 lead to a sharp increase in the predicted precipitation (Figure 4.31).

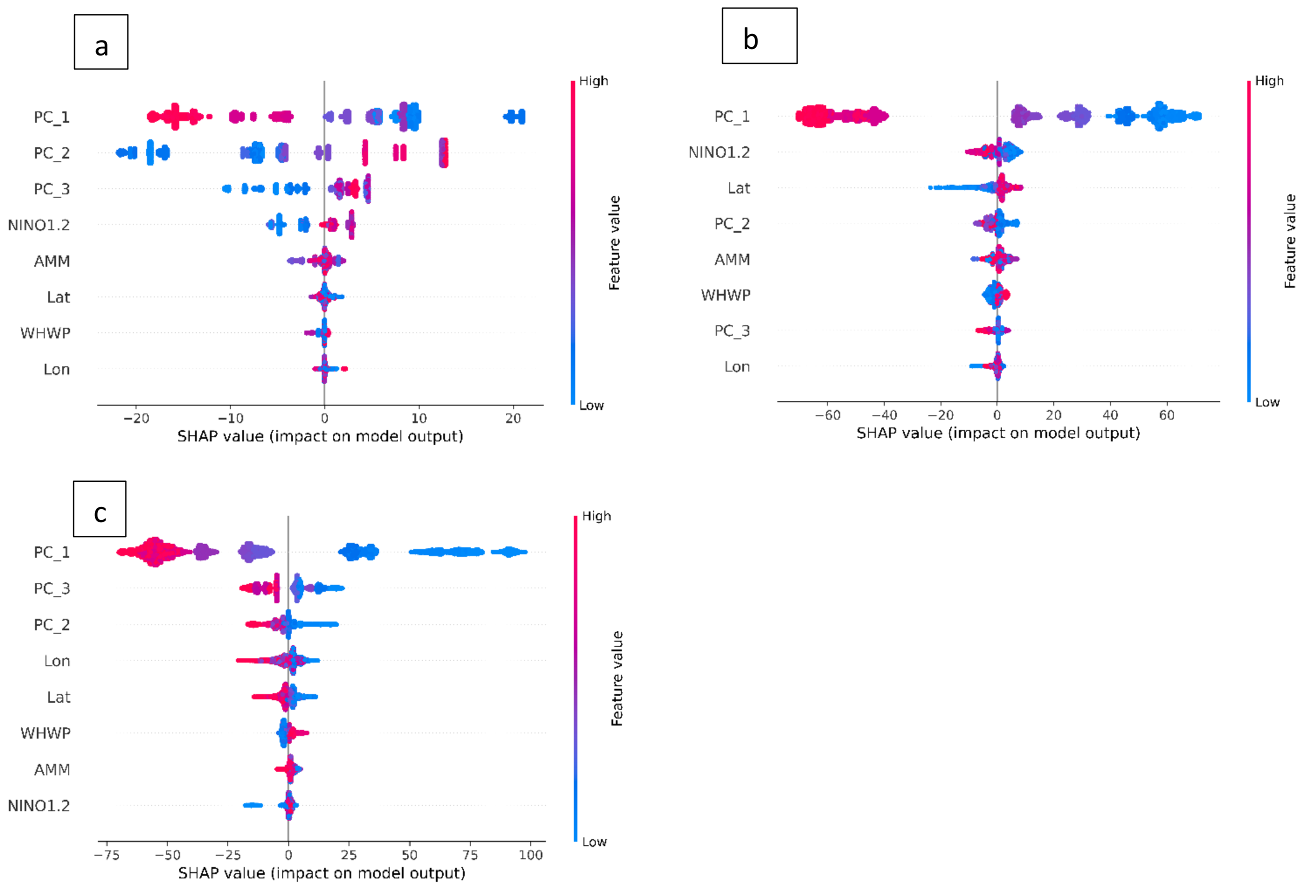

Figure 4.32: SHAP plots showing how the features impact the quantile at an alpha of 0.05 model output in the Guinea Coast (a), Savanah (b), and Sahel (c). At the y-axis, features are sorted by relative importance, differing between the three zones.

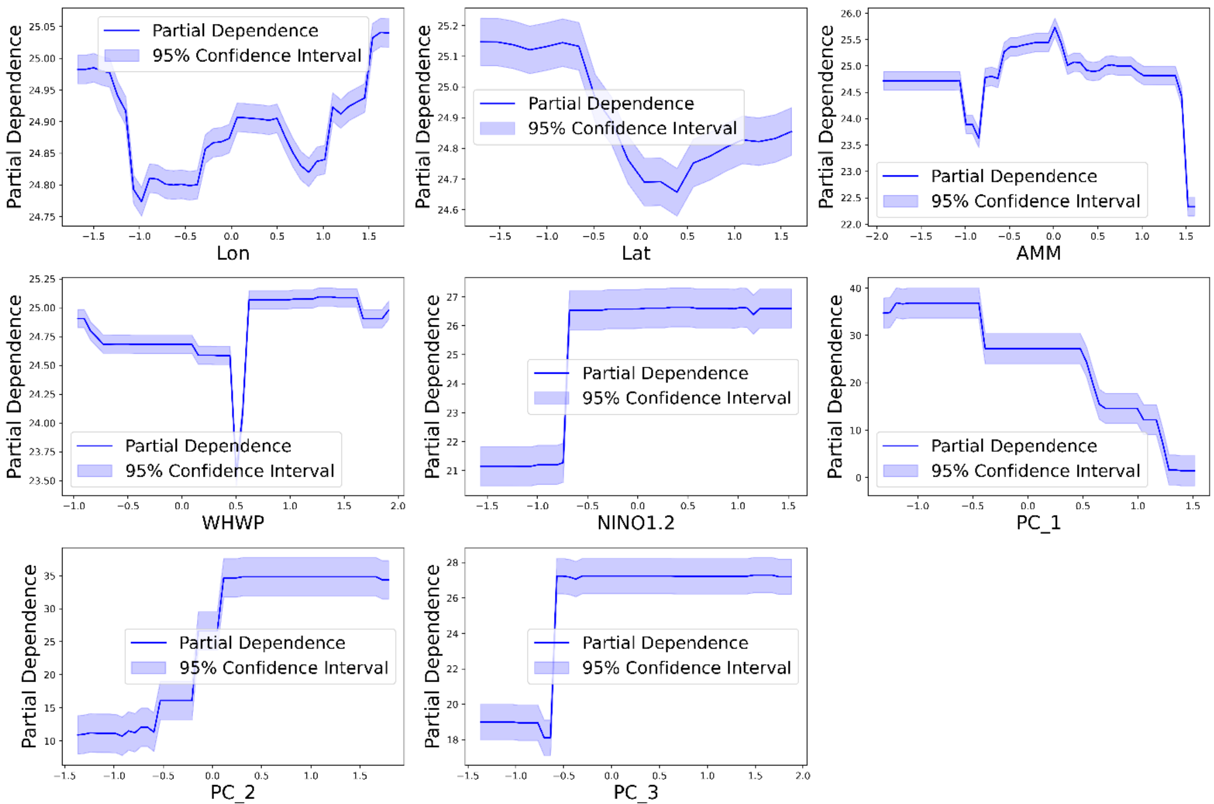

Figure 4.33: Partial dependence plots showing how the features impact the quantile at an alpha of 0.05 model output in the Guinea Coast.

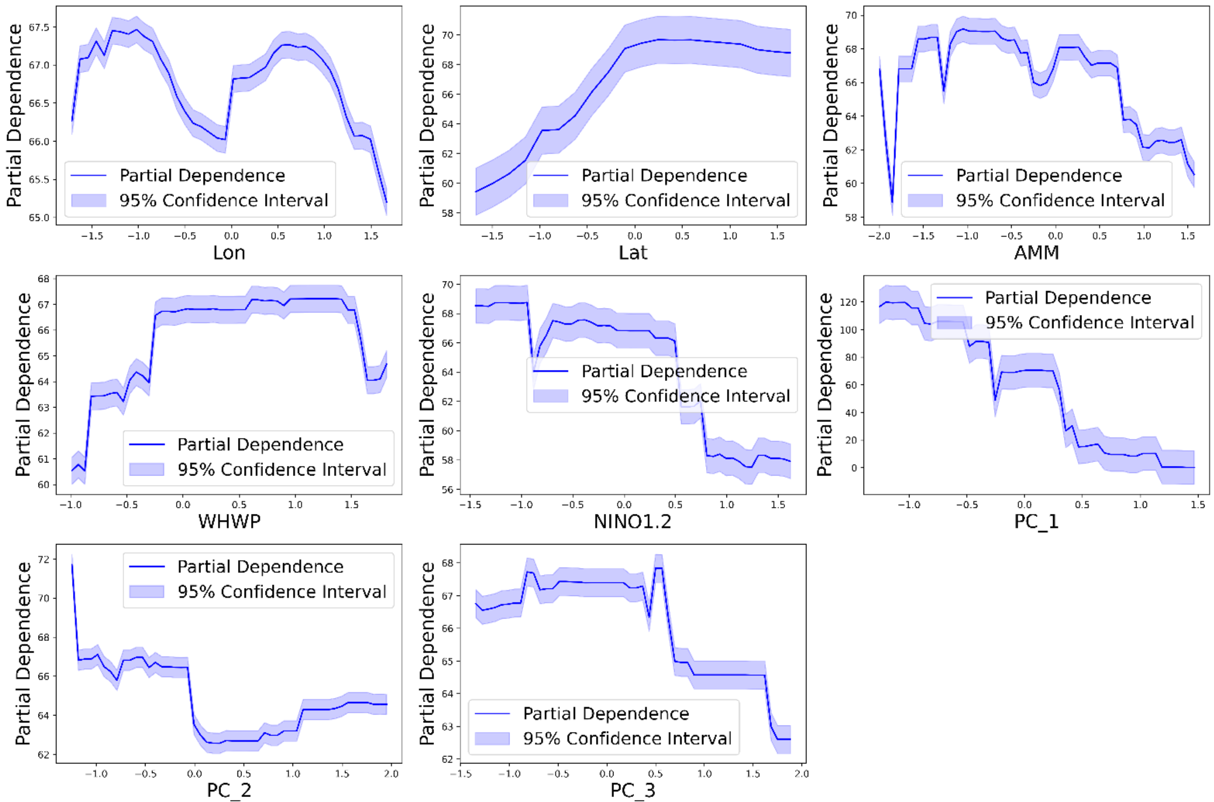

Figure 4.34: Partial dependence plots showing how the features impact the quantile at an alpha of 0.05 model output in the Savanah.

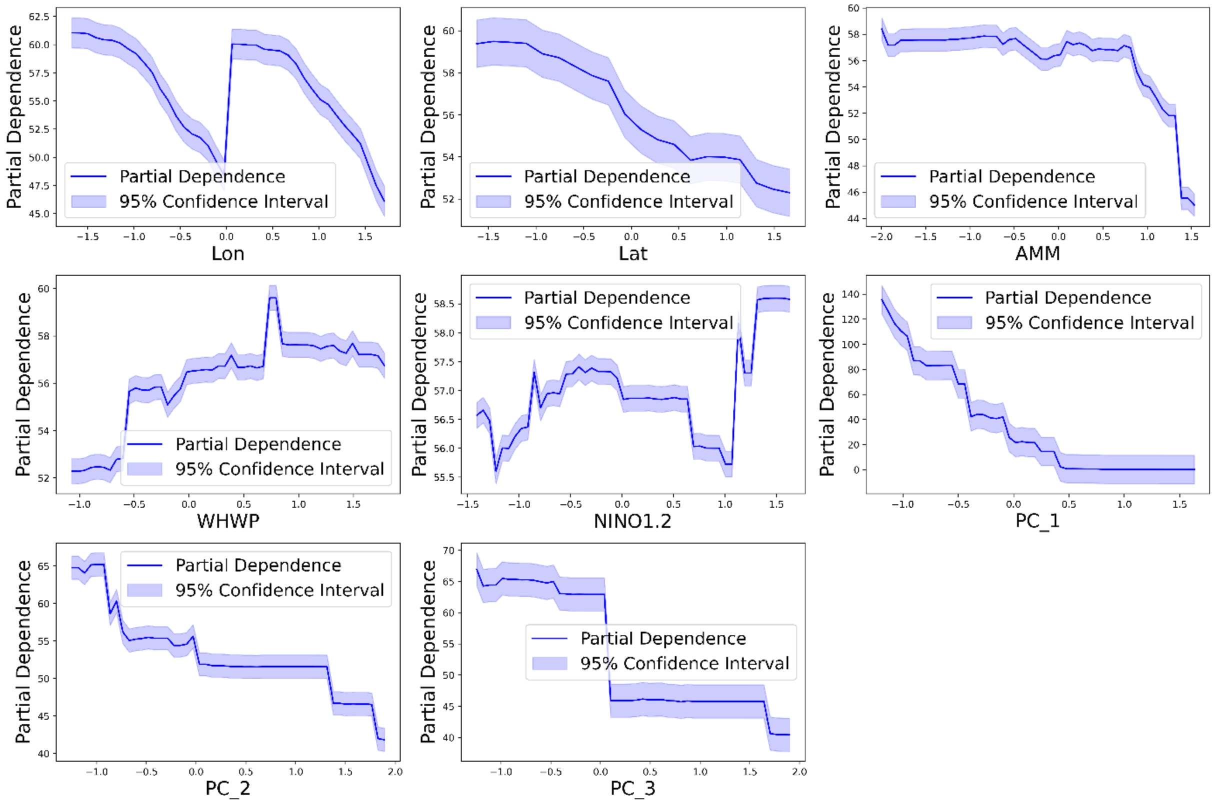

Figure 4.35: Partial dependence plots showing how the features impact the quantile at an alpha of 0.05 model output in the Sahel.

4.3.2.3 Features behaviour during extremely low precipitation prediction

To understand how the selected features behave and drive the extremely low precipitation in the various zones of West Africa, SHAPs and PDPs are made using the LightGBM quantile with an alpha of 0.05 over the selected areas of the various zones.

Figure 4.32 presents the SHAP results of GC (Figure 4.32a), Savanah (Figure 4.32b), and Sahel (Figure 4.32c). The order of importance by which the selected features drive extremely low precipitation in the various zones is as follows: GC (Figure 4.32a): PC1, PC2, PC3, Niño 1+2, AMM, latitude, WHWP, and longitude. Savanah (Figure 4.32b): PC1, Niño 1+2, latitude, PC2, AMM, WHWP, PC3, and longitude. Sahel (Figure 4.32c): PC1, PC3, PC1, longitude, latitude, WHWP, AMM, and Niño 1+2.

In the GC zone (Figure 4.32a), high values of PC1 contribute negatively to extremely low precipitation output prediction. However, the high values of the following features contribute positively to the precipitation output: PC2, PC3, and Niño 1+2. The impact of the other features is not clear in this zone. In the Savanah zone (Figure 4.32b), high values of PC1, Niño 1+2, PC2, and PC3 impact the precipitation output negatively. However, high values of latitude, WHWP, and longitude affect the model output positively. The remaining features have no clear pattern by which they impact the model output in the Savanah zone. In the Sahel zone (Figure 4.32c), high values of PC1, PC2, PC3, longitude, latitude, and AMM impact the extremely low precipitation output negatively. However, high values of WHWP and Niño 1+2 impact the output positively.

The partial dependence of extremely low precipitation on the selected features in the GC (Figure 4.33), Savanah (Figure 4.34), and Sahel (Figure 4.35) are presented. In the GC (Figure 4.33), initially increasing longitude values lead to a decrease in predicted precipitation; however, moving further from west to east leads to an increase in predicted precipitation in the GC. The partial dependence of extremely low precipitation on latitude in the GC is such that increasing values of latitude lead to lower precipitation. Increasing the values of AMM generally leads to an initial increase in predicted precipitation, peaking at an AMM value of 0. Further increasing the AMM values leads to a decrease in precipitation. Changing the values of WHWP has little impact on the predicted precipitation output; however, increasing values leads to an increase in precipitation (Figure 4.33). Partial dependence of precipitation on PC3 is such that increasing the values of PC2 generally leads to an increase in the predicted precipitation. The partial dependence of extremely low precipitation on the PC1 is such that increasing the values of PC1 leads to a reduction in precipitation. However, increasing the values of PC2 and PC3 leads to an increase in predicted precipitation (Figure 4.33).

Over the Savanah (Figure 4.34), the partial dependence of the extremely low precipitation on longitude is such that lower latitudes generally lead to higher predicted precipitation than higher longitudes. Lower latitude values lead to lower precipitation, and higher latitude leads to higher predicted precipitation (Figure 4.34). Increasing the values of AMM leads to an increase in predicted precipitation; however, further increasing AMM values leads to a decrease in precipitation in the Savanah zone (Figure 4.34). Generally increasing values of WHWP lead to increases in predicted precipitation in the Savanah, while increasing Niño 1+2 values lead to decreases in precipitation values (Figure 4.34). The partial dependence of the precipitation on PC1, PC2, and PC3 is such that increasing the values of the PCs leads to a decrease in predicted precipitation in the Savanah zone (Figure 4.34).

Over the Sahel zone (Figure 4.35), the partial dependence of extremely low precipitation on longitude is such that initially increasing longitude values lead to a decrease in the predicted precipitation until longitude 0. Where further increases in the longitude values lead to a sharp increase in predicted precipitation, however, further increasing the longitude values leads to a decrease in predicted precipitation (Figure 4.35). Increasing the values of AMM leads to a decrease in the predicted precipitation in the Sahel zone. The values of the predicted precipitation in the Sahel zone increase when AMM and Niño 1+2 values are increased. The partial dependence of precipitation on PC1, PC2, and PC3 in the Sahel is such that increasing the values of these features leads to a reduction in the predicted precipitation.

4.3.2.4 Discussion of climate drivers of precipitation over West Africa

To understand how precipitation in the various zones over West Africa (Guinea Coast, Savanah, and Sahel) is driven differently by the features, the model was applied to selected areas in each of the zones. Understanding model behaviour fosters model trust and model improvement. Model behaviour understanding could help in understanding how various drivers interact in driving precipitation over West Africa. To also understand how the features drive extreme precipitation variability over West Africa, SHAPs and PDPs are developed using LightGBM with quantile as the objective function, with alpha values of 0.05 and 0.95 for extremely low and extremely high precipitation, respectively.

The feature importance of the various zones is different. That is, although the number and list of features are the same, those that appear to be most important are different for different zones. This implies that each zone has different environmental dynamics. Even with these environmental dynamics, the model developed performed well in all the zones. Similarly, the environmental dynamics that drive extremely low and extremely high precipitation in the different zones of West Africa are different.

Low to medium values of Niño 1+2 lead to an increase in the values of monthly precipitation predicted over GC, Savanah, and Sahel. Generally, El Niño events have been reported to lead to dryness over West Africa because of enhanced northeast trade winds during El Niño events (Camberlin *et al*., 2001). Therefore, the behaviour of Niño 1+2 over GC, Savanah and Sahel from the SHAP and partial dependence results, where low Niño 1+2 values lead to high precipitation values and high Niño 1+2 values lead to low precipitation, is expected.

A previous study has shown that the WHWP supplies moisture to the Atlantic intertropical convergence zone (ITCZ) when it extends eastward into the tropical Atlantic (Drumond *et al*., 2011). This could explain the relationship of WHWP with monthly precipitation as observed in this work. The PDP and SHAP results have shown that higher values of WHWP generally lead to lower values of precipitation predicted in the GC (Figure 4.25). But in the Sahel zones, generally, higher values of WHWP lead to higher values of predicted precipitation (Figure 4.27). Since intensification of WHWP has been shown to supply moisture to the Atlantic ITCZ, when the WHWP is low, it might just be supplying enough moisture to move the ITCZ to dangle over the GC, thus leading to a higher amount of precipitation. However, as the WHWP intensifies, the moisture supply is high enough to move the ITCZ further north, favouring the Sahel zone to the disadvantage of the GC. This could explain why higher WHWP values appear to increase precipitation over the Sahel zone but reduce precipitation over the GC. The impact of WHWP over the Savanah is weak.

A positive AMM phase is characterised by a warmer than usual SST in the tropical North Atlantic and a cooler than usual tropical South Atlantic, leading to a stronger equatorial wind flow from the Southern Hemisphere into the Northern Hemisphere (Diatta and Fink, 2014). This wind flow influences the latitudinal positioning of the ITCZ. AMM has a strong influence on Sahel rainfall through the ITCZ (Lamb, 1978; Hastenrath and Greischar, 1993; Giannini *et al*., 2003; Foltz *et al*., 2012). Diatta and Fink (2014) reported that AMM correlated positively with Sahel rainfall and negatively with GC; however, only the correlation value with Sahel was statistically significant. Their classification of GC included the Savanah region (of this work) as part of GC. The partial dependence and SHAP results of this work show that high values of AMM impact GC precipitation negatively. Also, in the Savanah and Sahel, the highest amount of precipitation occurs when AMM has medium values.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

5.1.1 Trends and Drivers of Thunderstorm Variability

Drivers and trends of thunderstorm dynamics in Ghana were studied using monthly thunderstorm frequency data from 22 synoptic stations operated by the Ghana Meteorological Agency (GMet). Drivers were identified using LightGBM and climate indices and ERA reanalysis data as features. Trend analysis was done on the thunderstorm data and the identified drivers. The LightGBM model explained 79% of the observed thunderstorm variance as a function of CAPE, TISR, CIN, SI10, and longitude. Among these, CAPE was identified as the primary driver. Whereas the model identified monotonic or even linear relationships between these drivers and thunderstorm frequency, longitude was exceptional insofar as it indicated a maximum between 1 degree west and longitude 0.

The trend analysis indicated a significant decrease in thunderstorm frequency all over Ghana from 1990-2013. That trend agrees with earlier studies that reported decreasing rainfall over Ghana. The negative trend of thunderstorms could greatly affect the agricultural sector in Ghana, which is mainly rain-fed. The trend analysis results of the driving climate variables, especially CAPE and CIN, indicate that changes in these climatic factors are driving the observed downward trend of thunderstorms over Ghana.

The current thunderstorm model is limited by the short data period and limiting trends and climate interactions post-2013. Since the current work is limited to Ghana, future research can extend over West Africa for a broader understanding.

5.1.2 Predicting Thunderstorms One Year Ahead

ML models have been developed and evaluated to predict thunderstorm frequencies over Ghana one year ahead, using mainly ERA5 and climate indices as features. The ERA5 atmospheric features were passed through a low-pass filter with a period length of three (3). The features were lagged by a year, so that thunderstorm frequencies were predicted one year in advance. The study also explored the model behaviour to understand how features drive thunderstorms.

The following ML models were developed: LightGBM, XGBoost, RF, and Ensemble. They were then compared with a simple climatological model. All the machine learning models captured well the variance of thunderstorm frequencies in Ghana, with R^2 values ranging from 0.74 to 0.76. The two best models based on the evaluation metrics were the RF and Ensemble models, with coefficients of determination of 0.75 each. Therefore, thunderstorm frequencies over Ghana can be predicted one year in advance using the developed models.

5.1.3 Drivers and Prediction of Precipitation One Year Ahead in West Africa

Finally, monthly precipitation over West Africa is predicted one year ahead using LightGBM and XGBoost and evaluated. The GPM-IMERG final run version 7 precipitation is used as the target variable, spanning from June 2000 to August 2023. The features are ERA5 reanalysis data (June 1999 to August 2022), climate indices (June 1999 to August 2022), and geographic coordinates (longitude and latitude). The ERA5 data were passed through a low-pass filter, and then principal component analysis was done on the low-pass filtered ERA5 data to reduce the dimensionality. To understand model behaviour for trust, accountability, model improvements, and understanding drivers of monthly precipitation over West Africa, PDP and SHAP plots were used.

The LightGBM and XGBoost models captured 77%–84% of the variance in monthly precipitation with lower computational demands. This will allow for the use of the model by meteorological and hydrological agencies or individuals in West Africa, even for those that do not have high-performance computers. The models' predictions captured the “monsoon jump,” “little dry season,” and the dry Ghana-Benin zone over West Africa.

This model provides West African meteorological agencies with a reliable tool for sub-seasonal to seasonal forecasts. This will support agricultural planning by predicting rainfall critical for rain-fed farming, enhance reservoir management and flood risk mitigation for hydropower-dependent countries, and aids disaster risk preparedness by identifying high precipitation areas and months in advance.

5.2 Recommendations

Below are the recommendations derived from this work:

1. Promote climate-resilient agricultural practices to adapt to declining thunderstorms.
2. National meteorological agencies should adopt the developed ML models to improve long-lead-time forecasting of thunderstorms and precipitation.
3. Encourage interdisciplinary collaboration between data scientists, meteorologists, and policymakers to operationalise predictive models at local and regional levels.
4. Since the feature analysis revealed that the different zones have different environmental dynamics, future works could also consider developing zone-specific models, which may likely improve the model performance.
5. Ministries of Agriculture, Energy, and Environment should integrate the forecast outputs into early warning systems for drought and flood risk management.

5.3 Research Contributions

The contributions of the research are:

1. Developed ML-based forecasting models for thunderstorm frequencies over Ghana and precipitation over West Africa one year ahead.
2. Enhanced model trust and interpretability in climate science using SHAP and PDP explainability tools. The results show that different environmental dynamics drive precipitation in the various zones in West Africa.
3. Established a machine learning-based spatial forecasting framework for precipitation over West Africa, capable of simulating key meteorological features across diverse zones.
4. The research offers a scalable, low-cost forecasting tool suitable for operational use in resource-limited meteorological services in West Africa.

REFERENCES

Abbam, T., Johnson, F.A., Dash, J. & Padmadas, S.S. (2018) ‘Spatiotemporal variations in rainfall and temperature in Ghana over the twentieth century, 1900–2014’, Earth and Space Science, 5(4), 120-132.

Ahijevych, D., Pinto, J. O., Williams, J. K., & Steiner, M. (2016). Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. Weather and Forecasting, 31(2), 581–599.

Azad, M., Kalam, A., Islam, A. R. M., Rahman, M., & Ayen, K. (2021). Development of novel hybrid machine learning models for monthly thunderstorm frequency prediction over Bangladesh. Natural Hazards, 108(1), 1109-1135.

Baker, M. B., Christian, H. J., & Latham, J. (1995). A computational study of the relationships linking lightning frequency and other thundercloud parameters. Quarterly Journal of the Royal Meteorological Society, 121(527), 1525-1548.

Baker, M. B., Blyth, A. M., Christian, H. J., Latham, J., Miller, K. A., & Gadian, A. M. (1999). Relationships between lightning activity and various thundercloud parameters: Satellite and modelling studies. Atmospheric research, 51(3-4), 221-236.

Balogun, R.A., Ajayi, V.O. & Balogun, I.A. (2022) ‘Local Impacts and Surface Characteristics of Cold Pools Wind Gust Observed from Thunderstorm’, Atmospheric and Climate Sciences, 12(4), 679-697.

Balogun R. A. (2024a) Diurnal Circle of the West and Central Monsoon: Profiles of Wind and Moisture Flux during JJAS Season. *JEAR*, 7 (1), 220-244.

Balogun R. A. (2024b) Diurnal Circle of the West and Central African Monsoon: Profiles of Wind and Moisture Flux during MAM Season. *JEAR*, 7 (1), 245-266.

Bochenek, B., & Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—Applications and perspectives. Atmosphere, 13(2), 180.

Boluwade, A. (2020). Remote sensed-based rainfall estimations over the East and West Africa regions for disaster risk management. ISPRS journal of photogrammetry and remote sensing, 167, 305-320.

Brooks, H. E., Anderson, A. R., Riemann, K., Ebbers, I., & Flachs, H. (2007). Climatological aspects of convective parameters from the NCAR/NCEP reanalysis. Atmospheric Research, 83(2), 294–305. https://doi.org/10.1016/j.atmosres.2005.08.005

Burpee, R.W., 1972. The origin and structure of easterly waves in the lower troposphere of North Africa. Journal of the Atmospheric Sciences, 29(1), 77-90.

Camberlin, P., Janicot, S., & Poccard, I. (2001). Seasonality and atmospheric dynamics of the teleconnection between African rainfall and tropical sea-surface temperature: Atlantic vs. ENSO. International Journal of Climatology: A Journal of the Royal Meteorological Society, 21(8), 973-1005.

Campozano, L., Trachte, K., Célleri, R., Samaniego, E., Bendix, J., Albuja, C., & Mejia, J. F. (2018). Climatology and Teleconnections of Mesoscale Convective Systems in an Andean Basin in Southern Ecuador: The Case of the Paute Basin. Advances in Meteorology, 2018, 1–13. https://doi.org/10.1155/2018/4259191

Chen, J., Dai, A., Zhang, Y. & Rasmussen, K.L. (2020) 'Changes in convective available potential energy and convective inhibition under global warming', Journal of Climate, 33(6), 2025-2050.

Chen, R., Zhang, W., & Wang, X. (2020). Machine learning in tropical cyclone forecast modeling: A review. Atmosphere, 11(7), 676.

Christian, H. J., Blakeslee, R. J., Boccippio, D. J., Boeck, W. L., Buechler, D. E., Driscoll, K. T., ... & Stewart, M. F. (2003). Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. Journal of Geophysical Research: Atmospheres, 108(D1), ACL-4.

Cook, K. H. (1999). Generation of the African easterly jet and its role in determining West African precipitation. Journal of climate, 12(5), 1165-1184.

Cornforth, R., Mumba, Z. & Parker, D.J. (2017) 'Synoptic Systems', in Parker, D.J. & Diop-Kane, M. (eds.) Meteorology of Tropical West Africa: The Forecasters' Handbook. Chichester: John Wiley & Sons, pp. 40-84.

Cotton, W. R., R. L. George, P. J. Wetzel, and R. L. McAnelly, 1983: A long-lived mesoscale convective complex. Part I: The mountain-generated component. Mon. Wea. Rev., 111, 1893–1918, .

de Burgh-Day, C. O., & Leeuwenburg, T. (2023). Machine learning for numerical weather and climate modelling: A review. Geoscientific Model Development, 16(22), 6433–6477.

Diatta, S., & Fink, A. H. (2014). Statistical relationship between remote climate indices and West African monsoon variability. Int. J. Climatol, 34(12), 3348-3367.

Dirmeyer, P. A., Cash, B. A., Kinter, J. L., Jung, T., Marx, L., Satoh, M., *et al*. (2011). Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization. Climate Dynamics, 39(1-2), 399–418. https://doi.org/10.1007/s00382-011-1127-9.

Doswell, C. A., III, 1984: A kinematic analysis of frontogenesis associated with a nondivergent vortex. J. Atmos. Sci., 41, 1242–1248, .

Drumond, A., Nieto, R., & Gimeno, L. (2011). On the contribution of the Tropical Western Hemisphere Warm Pool source of moisture to the Northern Hemisphere precipitation through a Lagrangian approach. Journal of Geophysical Research: Atmospheres, 116(D21). https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD015397

Evan AT, Flamant C, Lavaysse C, Kocha C, Saci A (2015) Water vapour over the Sahara desert and drought in the Sahel. J Clim 28:108–123.

Fiedler, S., Kaplan, M. L., & Knippertz, P. (2015). The importance of Harmattan surges for the emission of North African dust aerosol. Geophysical Research Letters, 42(21), 9495–9504. https://doi.org/10.1002/2015GL065925

Fink, A.H., Engel, T., Ermert, V., van der Linden, R., Schneidewind, M., Redl, R., Afiesimama, E., Thiaw, W.M., Yorke, C. & Evans, M. (2017) 'Mean Climate and Seasonal Cycle', in Parker, D.J. & Diop-Kane, M. (eds.) Meteorology of Tropical West Africa: The Forecasters' Handbook. Chichester: John Wiley & Sons, pp. 37-95.

Foltz, G. R., McPhaden, M. J., & Lumpkin, R. (2012). A strong Atlantic meridional mode event in 2009: The role of mixed layer dynamics. Journal of Climate, 25(1), 363-380.

Fontaine, B. and Janicot, S., 1992. Wind-field coherence and its variations over West Africa. Journal of Climate, 5(5), 512-524.

Foresti, L., Sideris, I. V., Nerini, D., Beusch, L., & Germann, U. (2019). Using a 10-year radar archive for nowcasting precipitation growth and decay: A probabilistic machine learning approach. Weather and Forecasting, 34(5), 1547-1569.

Fulks, J. R., 1951: The instability line. Compendium of Meteorology, T. F. Malone, Ed., Amer. Meteor. Soc., 647–652.

Gebrechorkos, S. H., Pan, M., Beck, H. E., & Sheffield, J. (2022). Performance of state-of-the-art C3S European seasonal climate forecast models for mean and extreme precipitation over Africa. Water Resources Research, 58(3), e2021WR031480.

Geron, A. (2017). Hands-On Machine Learning with Scikit-Learn & Tensorflow O'Reilly Media, Inc. O'Reilly Media, Inc, 1005, 564.

Ghavidel, Y., Baghbanan, P. & Farajzadeh, M. (2016) 'A statistical survey on linking between thunderstorm frequency and teleconnection patterns in Iran', Arabian Journal of Geosciences, 9, 1-12.

Giannini, A., Saravanan, R., & Chang, P. (2003). Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. Science, 302(5647), 1027-1030.

Hamilton RA, Archbold JW, Douglas CKM. 1945. Meteorology of Nigeria and adjacent territory. Q. J. R. Meteorol. Soc. 71: 231–264. doi: 10.1002/qj.49707130905.

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. International journal of climatology, 27(9), 1119-1152.

Harats, N., Ziv, B., Yair, Y., Kotroni, V. & Dayan, U. (2010) 'Lightning and rain dynamic indices as predictors for flash floods events in the Mediterranean', Advances in Geosciences, 23, 57-64.

Harel, M., & Price, C. (2020). Thunderstorm trends over Africa. Journal of Climate, 33(7), 2741-2755.

Hastenrath, S., & Greischar, L. (1993). Circulation mechanisms related to northeast Brazil rainfall anomalies. Journal of Geophysical Research: Atmospheres, 98(D3), 5093-5102.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. New York: Springer.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803

Houze, R. A., Jr., 1993: Cloud Dynamics. Academic Press, 573 pp.

Ilesanmi, O. O., 1971: An empirical formulation of an ITD rainfall model for the tropics: A case study of Nigeria. J. Appl. Meteor., 10, 882–891.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Janicot, S., Trzaska, S., & Poccard, I. (2001). Summer Sahel-ENSO teleconnection and decadal time scale SST variations. Climate Dynamics, 18(3–4), 303–320. https://doi.org/10.1007/s003820100172.

Jung, T., Ferranti, L., & Tompkins, A. M. (2006). Response to the summer of 2003 Mediterranean SST anomalies over Europe and Africa. Journal of Climate, 19(20), 5439-5454.

Kabo-Bah, A. T., Diji, C. J., Nokoe, K., Mulugetta, Y., Obeng-Ofori, D., & Akpoti, K. (2016). Multiyear rainfall and temperature trends in the Volta river basin and their potential impact on hydropower generation in Ghana. *Climate*, *4*(4), 49.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W. & Ma, W. *et al*. (2017) 'Lightgbm: A highly efficient gradient boosting decision tree', Advances in Neural Information Processing Systems, 30.

Kendall, M.G. (1948) Rank correlation methods. London: Charles Griffin.

Knist, S., Goergen, K., & Simmer, C. (2018). Evaluation and projected changes of precipitation statistics in convection-permitting WRF climate simulations over Central Europe. Climate Dynamics, 1–17. https://doi.org/10.1007/s00382-018-4147-x.

Kyselý, J., Rulfová, Z., Farda, A., & Hanel, M. (2015). Convective and stratiform precipitation characteristics in an ensemble of regional climate model simulations. Climate Dynamics, 46(1-2), 227–243. https://doi.org/10.1007/s00382-015-2580-7.

Laing, A. G., Carbone, R. E., & Levizzani, V. (2011). Cycles and propagation of deep convection over equatorial Africa. Monthly Weather Review, 139(9), 2832-2853.

Lamb, P. J. (1978). Large-scale Tropical Atlantic surface circulation patterns associated with Subsaharan weather anomalies. Tellus A: Dynamic Meteorology and Oceanography, 30(3), 240. https://doi.org/10.3402/tellusa.v30i3.10338

Lavaysse C, Flamant C, Janicot S (2010a) Regional-scale convection patterns during strong and weak phases of the Saharan heat low. Atmos Sci Lett 11(4):255–264.

Lavaysse C, Flamant C, Janicot S, Knippertz P (2010b) Links between African easterly waves, midlatitude circulation and intraseasonal pulsations of the West African heat low. Q J R Meteorol Soc 136(S1):141–158

Lavaysse C, Flamant C, Janicot S, Parker DJ, Lafore JP, Sultan B, Pelon J (2009) Seasonal evolution of the West African heat low: a climatological perspective. Clim Dyn 33(2–3):313–330

Lavaysse, C., Flamant, C., Evan, A., Janicot, S., & Gaetani, M. (2016). Recent climatological trend of the Saharan heat low and its impact on the West African climate. Climate Dynamics, 47(11), 3479-3498.

Lélé, M. I., & Lamb, P. J. (2010). Variability of the intertropical front (ITF) and rainfall over the West African Sudan–Sahel zone. Journal of Climate, 23(14), 3984-4004.

LeMone, M. A., & Zipser, E. J. (1980). Cumulonimbus vertical velocity events in GATE. Part I: Diameter, intensity and mass flux. Journal of the Atmospheric Sciences, 37(11), 2444-2457.

Liu, N., Liu, C., Chen, B., & Zipser, E. (2020). What are the favorable large-scale environments for the highest-flash-rate thunderstorms on Earth?. Journal of the Atmospheric Sciences, 77(5), 1583-1612.

Lu, J., & Delworth, T. L. (2005). Oceanic forcing of the late 20th century Sahel drought. Geophysical Research Letters, 32(22), 2005GL023316. https://doi.org/10.1029/2005GL023316

Lucas, C., Zipser, E. J., & Lemone, M. A. (1994). Vertical velocity in oceanic convection off tropical Australia. Journal of Atmospheric sciences, 51(21), 3183-3193.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Mann, H.B. (1945) 'Nonparametric tests against trend', Econometrica: Journal of the Econometric Society, 13(3), 245-259.

Maranan, M., Fink, A. H., & Knippertz, P. (2018). Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment. Quarterly Journal of the Royal Meteorological Society, 144(714), 1628–1648. https://doi.org/10.1002/qj.3345.

Maranan, M., Fink, A. H., Knippertz, P., Amekudzi, L. K., Atiah, W. A., & Stengel, M. (2020). A process-based validation of GPM IMERG and its sources using a mesoscale rain gauge network in the West African forest zone. Journal of Hydrometeorology, 21(4), 729-749.

Martin, E. R., Thorncroft, C., & Booth, B. B. (2014). The multidecadal Atlantic SST—Sahel rainfall teleconnection in CMIP5 simulations. Journal of Climate, 27(2), 784-806.

Miller, R., 1972: Notes on analysis and severe-storm forecasting procedures of the Air Force Global Weather Central. Air Weather Service Tech. Rep. 200, 190 pp.

Mostajabi, A., Finney, D. L., Rubinstein, M., & Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. Npj Climate and Atmospheric Science, 2(1), 1-15.

Mul, M., Obuobie, E., Appoh, R., Kankam-Yeboah, K., Bekoe-Obeng, E., Amisigo, B. *et al*. (2015) Water resources assessment of the Volta River Basin. Vol. 166. Colombo: International Water Management Institute (IWMI).

Newell, R.E. and Kidson, J.W., 1984. African mean wind changes between Sahelian wet and dry periods. Journal of Climatology, 4(1), 27-33.

Niang, I., Ruppel, O.C., Abdrabo, M.A., Essel, A., Lennard, C., Padgham, J. & Urquhart, P. (2014) 'Africa', in Barros, V.R., Field, C.B., Dokken, D.J., Mastrandrea, M.D., Mach, K.J., Bilir, T.E., *et al*. (eds.) Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, pp. 1199-1265.

Nkrumah, F., Klutse, N.A.B., Adukpo, D.C., Owusu, K. & Quagraine, K.A. (2014) 'Rainfall variability over Ghana: model versus rain gauge observation', Theoretical and Applied Climatology, 118(1-2), 27-41.

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. Monthly weather review, 110(7), 699-706.

Nwachukwu, P. N., Satge, F., Yacoubi, S. E., Pinel, S., & Bonnet, M. P. (2020). From TRMM to GPM: how reliable are satellite-based precipitation data across Nigeria?. Remote Sensing, 12(23), 3964.

Oguntunde, P. G., Abiodun, B. J., Olukunle, O. J., & Olufayo, A. A. (2012). Trends and variability in pan evaporation and other climatic variables at Ibadan, Nigeria, 1973–2008. *Meteorological Applications*, *19*(4), 464-472.

Oguntunde, P. G., Lischeid, G., Abiodun, B. J., & Dietrich, O. (2017). Analysis of long-term dry and wet conditions over Nigeria. *International Journal of Climatology*, *37*(9), 3577-3586.

Olaniyan, E., Adefisan, E. A., Oni, F., Afiesimama, E., Balogun, A. A., & Lawal, K. A. (2018). Evaluation of the ECMWF sub-seasonal to seasonal precipitation forecasts during the peak of West Africa monsoon in Nigeria. Frontiers in Environmental Science, 6, 4.

Owusu, K., Waylen, P. & Qiu, Y. (2008) 'Changing rainfall inputs in the Volta basin: implications for water sharing in Ghana', GeoJournal, 71, 201-210.

Park, J.-H., Kug, J.-S., Li, T., & Behera, S. K. (2018). Predicting El Niño beyond 1-year lead: Effect of the Western Hemisphere warm pool. Scientific Reports, 8(1), 14957.

Parker DJ, Burton RR, Diongue-Niang A, Ellis RJ, Felton M, Taylor CM, Thorncroft CD, Bessemoulin P, Tompkins AM (2005) The diurnal cycle of the West African monsoon circulation. Q J R Meteorol Soc 131(611):2839–2860

Parker, D. J., & Diop-Kane, M. (2017). Meteorology of Tropical West Africa: The Forecasters' Handbook. John Wiley & Sons.

Pedregosa, F. *et al*. (2011) 'Scikit-learn: Machine learning in Python', Journal of Machine Learning Research, 12, 2825-2830.

Peyrillé P, Lafore JP (2007) An idealized two-dimensional framework to study the West African monsoon. Part II: large-scale advection and the diurnal cycle. J Atmos Sci 64(8):2783–2803. Pielke Sr, R. A. (2001). Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. Reviews of Geophysics, 39(2), 151-177.

Pinto, O. Jr (2015) 'Thunderstorm climatology of Brazil: ENSO and Tropical Atlantic connections', International Journal of Climatology, 35(6), 871-878.

Piper, D. & Kunz, M. (2017) 'Spatiotemporal variability of lightning activity in Europe and the relation to the North Atlantic Oscillation teleconnection pattern', Natural Hazards and Earth System Sciences, 17(8), 1319-1336.

Piper, D. A., Kunz, M., Allen, J. T., & Mohr, S. (2019). Investigation of the temporal variability of thunderstorms in central and western Europe and the relation to large-scale flow and teleconnection patterns. Quarterly Journal of the Royal Meteorological Society, 145(725), 3644–3666. https://doi.org/10.1002/qj.3647

Price, C. (2009). Thunderstorms, lightning and climate change. In Lightning: principles, instruments and applications (pp. 521-535). Springer, Dordrecht.

Price, C., 2008: Lightning sensors for observing, tracking and nowcasting severe weather, Sensors, 8, 157–170.

Riemann-Campe, K., Fraedrich, K. & Lunkeit, F. (2009) ‘Global climatology of convective available potential energy (CAPE) and convective inhibition (CIN) in ERA-40 reanalysis’, Atmospheric Research, 93(1-3), 534-545.

Rockwood, A. A., and R. A. Maddox, 1988: Mesoscale and syn-optic interactions leading to intense convection: The case of 7 June 1982. Wea. Forecasting, 3, 51–68.

Roehrig R, Bouniol D, Guichard F, *et al*. 2013. The present and future of the West African monsoon: a process-oriented assessment of CMIP5 simulations along the AMMA transect. J. Climate 26(17): 6471–6505. doi:10.1175/JCLI-D-12-00505.1.

Rowell, D. P. (2013). Simulating SST teleconnections to Africa: What is the state of the art? Journal of Climate, 26(15), 5397–5418.

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. Applied Sciences, 8(12), 2570.

Sanogo, S., Peyrillé, P., Roehrig, R., Guichard, F., & Ouedraogo, O. (2022). Extreme precipitating events in satellite and rain gauge products over the Sahel. Journal of Climate, 35(6), 1915-1938.

Sátori, G., Williams, E. & Lemperger, I. (2009) ‘Variability of global lightning activity on the ENSO time scale’, Atmospheric Research, 91(2-4), 500-507.

Schoenberg Ferrier, B., J. Simpson, and W.-K. Tao, 1996: Factors responsible for precipitation efficiencies in midlatitude and tropical squall simulations. Mon. Wea. Rev., 124, 2100–2125,.

Schumacher, C., and R. A. Houze Jr., 2003: Stratiform rain in the tropics as seen by the TRMM precipitation radar. J. Climate, 16, 1739–1756.

Sen, P.K. (1968) ‘Estimates of the regression coefficient based on Kendall's tau’, Journal of the American Statistical Association, 63(324), pp. 1379-1389.

Shabib Aftab, M. A., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. *Int J Adv Comput Sci Appl*, *9*(5), 143-150.

Shapley, L. S. (1953). A value for n-person games.

Sherwood, S. C., Bony, S., & Dufresne, J. L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37-42.

Shi, M., Zhang, W., Fan, P., Chen, Q., Liu, Z., Li, Q., & Liu, X. (2022). Modelling deep convective activity using lightning clusters and machine learning. International Journal of Climatology, 42(2), 952-973.

Siegmund, J., Bliefernicht, J., Laux, P., & Kunstmann, H. (2015). Toward a seasonal precipitation prediction system for West Africa: Performance of CFSv2 and high-resolution dynamical downscaling. Journal of Geophysical Research: Atmospheres, 120(15), 7316-7339.

Skofronick-Jackson, G., Petersen, W. A., Berg, W., Kidd, C., Stocker, E. F., Kirschbaum, D. B., ... & Wilheit, T. (2017). The Global Precipitation Measurement (GPM) mission for science and society. Bulletin of the American Meteorological Society, 98(8), 1679-1695.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., ... & Kalchbrenner, N. (2020). Metnet: A neural weather model for precipitation forecasting. arXiv preprint arXiv:2003.12140.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., ... & Miller, H. (2007). IPCC fourth assessment report (AR4). Climate change, 374.

Stith, J. L., J. A. Haggerty, A. Heymsfield, and C. A. Grainger, 2004: Microphysical characteristics of tropical updrafts in clean conditions. J. Appl. Meteor., 43, 779–794.

Takemi, T., 2007: A sensitivity of squall-line intensity to environmental static stability under various shear and moisture conditions. Atmos. Res., 84, 374–389,.

Tapia, A., Smith, J. A., & Dixon, M. (1998). Estimation of convective rainfall from lightning observations. Journal of Applied Meteorology and Climatology, 37(11), 1497-1509.

Taszarek, M., Allen, J. T., Marchio, M., & Brooks, H. E. (2021). Global climatology and trends in convective environments from ERA5 and rawinsonde data. Npj Climate and Atmospheric Science, 4(1), 1–11. https://doi.org/10.1038/s41612-021-00190-x.

Taylor, C. M., Harris, P. P., & Parker, D. J. (2010). Impact of soil moisture on the development of a Sahelian mesoscale convective system: A case-study from the AMMA special observing period. Quarterly Journal of the Royal Meteorological Society, 136(S1), 456-470.

Theil, H. (1950) ‘A rank-invariant method of linear and polynomial regression analysis. I, II, III’, Nederl. Akad. Wetensch., Proc., 53, pp. 386-392.

Thiaw, W.M., Janicot, S., Bekele, E., Cook, K.H., Fontaine, B., Mekonnen, A., Ndiaye, O., Tamo, P.H.K. & Vizy, E.K. (2017) ‘Subseasonal Forecasting’, in Parker, D.J. & Diop-Kane, M. (eds.) Meteorology of Tropical West Africa: The Forecasters’ Handbook. Chichester: John Wiley & Sons, pp. 255-277.

Tinmaker, M.I.R., Aslam, M.Y. & Chate, D.M. (2017) ‘Association of rainfall and stability index with lightning parameter over the Indo-Gangetic Plains’, American Journal of Climate Change, 6(3), 443-455.

Ukkonen, P., & Mäkelä, A. (2019). Evaluation of Machine Learning Classifiers for Predicting Deep Convection. Journal of Advances in Modeling Earth Systems, 11(6), 1784–1802. https://doi.org/10.1029/2018MS001561.

Uman, M. A. (2001). The lightning discharge. Courier Corporation.

Virts, K. S., Wallace, J. M., Hutchins, M. L., & Holzworth, R. H. (2013). Highlights of a new ground-based, hourly global lightning climatology. Bulletin of the American Meteorological Society, 94(9), 1381-1391.

Vizy, E. K., & Cook, K. H. (2001). Mechanisms by which Gulf of Guinea and eastern North Atlantic sea surface temperature anomalies can influence African rainfall. Journal of Climate, 14(5), 795–821.

Vollmert, P., Fink, A. H., & Besler, H. (2003). Ghana Dry Zone” und “Dahomey Gap”: Ursachen für eine Niederschlagsanomalie im tropischen Westafrika. Erde, 134(4), 375-393.

Weisman, M. L., and J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. Mon. Wea. Rev., 110, 504–520, .

Williams, E.R., and G. Satori, 2004: Lightning, thermodynamic and hydrological comparisons of the two tropical continental chimneys, J. Atmos. Sol. Terr. Phys., 66, 1213–1231.

Wu, M. L. C., Reale, O., Schubert, S. D., Suarez, M. J., Koster, R. D., & Pegion, P. J. (2009). African easterly jet: Structure and maintenance. Journal of Climate, 22(17), 4459-4480.

Yue, S., Pilon, P., Phinney, B. & Cavadias, G. (2002) 'The influence of autocorrelation on the ability to detect trend in hydrological series', Hydrological Processes, 16(9), 1807-1829.

Zainudin, S., Jasim, D. S., & Bakar, A. A. (2016). Comparative analysis of data mining techniques for Malaysian rainfall prediction. Int. J. Adv. Sci. Eng. Inf. Technol, 6(6), 1148-1153.

Zhang, S., Lu, L., Yu, J., & Zhou, H. (2016, July). Short-term water level prediction using different artificial intelligent models. In 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (pp. 1-6). IEEE.

Zipser, E. J., and K. R. Lutz, 1994: The vertical profile of radar reflectivity of convective cells: A strong indicator of storm intensity and lightning probability? Mon. Wea. Rev., 122, 1751–1759.

Ziv, B., Saaroni, H., Yair, Y., Ganot, M., Baharad, A. & Isaschari, D. (2009) 'Atmospheric factors governing winter thunderstorms in the coastal region of the eastern Mediterranean', Theoretical and Applied Climatology, 95(3), pp. 301-310.