INTERCOMPARISON AND EVALUATION OF GRIDDED PRECIPITATION PRODUCTS AND ENHANCED DROUGHT ANALYSIS IN THE SHABELLE BASIN, SOMALIA-ETHIOPIA.

PhD Dissertation

Abdinour Abshir HUSSEIN

Eskişehir 2025

INTERCOMPARISON AND EVALUATION OF GRIDDED PRECIPITATION PRODUCTS AND ENHANCED DROUGHT ANALYSIS IN THE SHABELLE BASIN, SOMALIA-ETHIOPIA

Abdinour Abshir HUSSEIN

PhD DISSERTATION

Department of Civil Engineering

Programme in Hydraulics

Supervisor: Prof. Dr. Ahmet BAYLAR

Eskişehir

Eskişehir Technical University

Institute of Graduate Programs

July 2025

FINAL APPROVAL FOR THESIS

This thesis titled INTERCOMPARISON AND EVALUATION OF GRIDDED PRECIPITATION PRODUCTS AND ENHANCED DROUGHT ANALYSIS IN THE SHABELLE BASIN, SOMALIA-ETHIOPIA has been prepared and submitted by Abdinour Abshir HUSSEIN in partial fulfillment of the requirements in "Eskişehir Technical University Directive on Graduate Education and Examination" for the Degree of PhD in Civil Engineering Department has been examined and approved on 02/07/2025.

Committee Members		**Title, Name and Surname**	**Signature**
Member	:	Prof. Dr. Ahmet BAYLAR	
Member	:	Prof. Dr. Mustafa TOMBUL	
Member	:	Prof. Dr. Serdar GÖNCÜ	
Member	:	Prof. Dr. Recep BAKIŞ	
Member	:	Doç. Dr. Yıldırım BAYAZIT	

Prof. Dr. Harun BÖCÜK
Director of the Institute of Graduate Programs

02/07/2025

SUPERVISOR APPROVAL

PhD student Abdinour Abshir HUSSEIN, whom I supervise, has completed this thesis titled INTERCOMPARISON AND EVALUATION OF GRIDDED PRECIPITATION PRODUCTS AND ENHANCED DROUGHT ANALYSIS IN THE SHABELLE BASIN, SOMALIA-ETHIOPIA. According to my inspections, the work is scientifically and ethically appropriate for the student to the thesis defense exam.

Supervisor

Prof. Dr. Ahmet BAYLAR

ABSTRACT

INTERCOMPARISON AND EVALUATION OF GRIDDED PRECIPITATION PRODUCTS AND ENHANCED DROUGHT ANALYSIS IN THE SHABELLE BASIN, SOMALIA-ETHIOPIA

Abdinour Abshir HUSSEIN

Department of Civil Engineering

Program in Hydraulics

Eskişehir Technical University, Institute of Graduate Programs, July 2025

Supervisor: Prof. Dr. Ahmet BAYLAR

This study compiled gauge-based precipitation observations (1998–2016) for the Shabelle River Basin in eastern Ethiopia and compared them with multiple satellite and reanalysis gridded precipitation products. Each dataset's performance was evaluated against ground measurements using diverse statistical metrics (correlations, error indices, efficiency scores) across daily, monthly, seasonal, and annual scales, as well as categorical indices for rainfall event detection. Results indicate that a satellite–rain gauge blended product (CHIRPS) and a modern reanalysis (ERA5) achieved the highest agreement with station observations, accurately capturing spatial rainfall gradients and bimodal seasonal cycles, whereas a gauge-only analysis (NOAA CPC Unified) consistently underestimated precipitation. Using the best-performing datasets, the Standardized Precipitation Index (SPI) and Standardized Precipitation–Evapotranspiration Index (SPEI) were calculated to conduct an enhanced drought analysis. These indices derived from CHIRPS and ERA5 closely mirrored ground-station drought histories, successfully capturing the timing and severity of major drought events from 1999 to 2016. Furthermore, multiple machine learning models were applied for station-level drought prediction; notably, a Naïve Bayes classifier outperformed more complex methods (random forest, k-nearest neighbors, linear regression) in predicting short-term SPI variations. The findings provide a comprehensive evaluation of precipitation data uncertainties in the Shabelle Basin and offer practical guidance on selecting optimal datasets for drought monitoring in data-sparse region.

Keywords: Gridded precipitation dataset, Shabelle basin, Flood, Drought.

ÖZET

SOMALİ-ETİYOPYA ŞABELLE HAVZASI'NDA KAFESLİ YAĞIŞ VERİ ÜRÜNLERİNİN KARŞILAŞTIRMALI DEĞERLENDİRMESİ VE GELİŞMİŞ KURAKLIK ANALİZİ

Abdinour Abshir HUSSEIN

İnşaat Mühendisliği Anabilim Dalı

Hidrolik Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz 2025

Danışman: Prof. Dr. Ahmet BAYLAR

Bu çalışmada, Doğu Etiyopya'daki Shabelle Nehri Havzası için 1998–2016 dönemine ait yağış gözlemleri derlenmiş ve çeşitli uydu ile yeniden analiz tabanlı kurgusal yağış ürünleriyle karşılaştırılmıştır. Her bir veri setinin performansı; günlük, aylık, mevsimsel ve yıllık ölçeklerde korelasyon, hata endeksleri ve verimlilik katsayıları gibi ölçütler ile yağış olaylarını tespit etme becerisine göre yer istasyonlarına karşı değerlendirilmiştir. Bulgular, uydu ve yağış ölçer verilerini harmanlayan CHIRPS ürünü ile ERA5 yeniden analiz veri setinin istasyon gözlemleriyle en yüksek uyumu sağladığını göstermektedir; bu iki veri seti, mekânsal yağış desenlerini ve çift modlu mevsimsel döngüleri doğru şekilde yansıtabilirken, yalnızca yağış istasyonlarından türetilen NOAA CPC Unified ürünü yağışı sistematik olarak eksik tahmin etmektedir. En iyi performans gösteren veri setleri kullanılarak Standartlaştırılmış Yağış İndeksi (SPI) ve Standartlaştırılmış Yağış-Evapotranspirasyon İndeksi (SPEI) hesaplanmış ve kapsamlı bir kuraklık analizi yürütülmüştür. CHIRPS ve ERA5 kaynaklı bu indisler, yer istasyonlarının kuraklık geçmişlerini yakından yansıtmakta ve 1999–2016 yılları arasında meydana gelen başlıca kuraklık olaylarının zamanlamasını ile şiddetini başarıyla yakalamaktadır. Ayrıca, istasyon ölçeğinde kuraklık tahmini için çeşitli makine öğrenmesi modelleri uygulanmıştır; özellikle Naive Bayes yöntemi, kısa vadeli SPI değişimlerini öngörmede daha karmaşık yöntemlere (rastgele orman, en yakın komşu, doğrusal regresyon) kıyasla üstün performans sergilemiştir. Bu çalışma, Shabelle Havzası'ndaki yağış verilerinin belirsizliklerine dair kapsamlı bir değerlendirme sunmakta ve veri kıtlığı yaşanan bölgelerde kuraklık izlemesi ile su kaynakları yönetimi için en uygun yağış veri setlerinin seçimine yönelik önemli bilgiler sağlamaktadır.

Anahtar Sözcükler: Izgara yağış ürünleri, Şebelle havzası, Taşkin, Kuraklık.

ACKNOWLEDGEMENTS

First, I thank Allah for giving me the strength and guidance to finish this work. I am truly grateful to my supervisor, Prof. Dr.Ahmet Baylar for his constant support and advice. His help and mentorship have been instrumental in every stage of this PhD work. Without his help, this research would not have been possible.

I would like to express my sincere gratitude to Prof. Dr. Mustafa Tombul, for his thoughtful encouragement and guidance during the course of this PhD academic journey and the thesis.

I would also like to extend my appreciation to the distinguished jury members — Prof. Dr. Serdar Göncü, Prof. Dr. Recep Bakış, and Doç. Dr. Yıldırım Bayazıt — for their valuable time, critical insights, and feedback, all of which contributed meaningfully to the academic rigor of this thesis.

Finally, I owe everything to my family and friends and am deeply grateful to them- especially my beloved mother and father for their unconditional prayer, love and endless support.

Abdinour Abshir HUSSEIN

02/07/2025

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with "scientific plagiarism detection program" used by Eskişehir Technical University, and that "it does not have any plagiarism" whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Abdinour Abshir HUSSEIN

CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

GLOSSARY OF SYMBOLS AND ABBREVIATIONS

AgERA5	:	Agrometeorological ERA5, a bias-corrected 9 km resolution version of ERA5 tailored for agriculture (also referred to as ERA5-Ag)
APHRODITE	:	Asian Precipitation – Highly-Resolved Observational Data Integration Towards Evaluation, a high-resolution gauge-based precipitation dataset for Asia
GO	:	Ground Observation
CFSR	:	Climate Forecast System Reanalysis, a global reanalysis of climate data (often shortened to CFS)
CHIRPS	:	Climate Hazards Group InfraRed Precipitation with Stations, a satellite–rain gauge blended precipitation dataset
CMORPH	:	CPC Morphing Technique, a NOAA satellite-based precipitation estimation method
CPC	:	Climate Prediction Center, often referring to the NOAA CPC Unified gauge-based global precipitation analysis
CPC UPP	:	Climate Prediction Center Unified Precipitation Project, a global daily precipitation dataset from NOAA's CPC (gauge-only
CRU TS	:	Climatic Research Unit Time Series, a global gridded climate dataset from the UK (University of East Anglia)
CSI	:	Critical Success Index, a metric for event detection accuracy (also known as the threat score)
DJF	:	December–January–February, the winter season (boreal winter) in seasonal analysis
ERA5	:	ECMWF Reanalysis v5, the fifth-generation global atmospheric reanalysis from the European Centre (ECMWF)

ERA5-Land	:	ERA5-Land, a land-surface downscaled version of ERA5 with ~9 km grid resolution
ERA5-Ag	:	ERA5 (Agricultural), a bias-adjusted version of ERA5 (AgERA5) optimized for agricultural applications and reduced bias
FAR	:	False Alarm Ratio, the fraction of predicted events that did not occur (false alarms).
FEWS	:	Famine Early Warning Systems, a network (often "FEWS NET") for monitoring food security and drought early warnings
F1-Score	:	F1 Score, the harmonic mean of precision and recall in classification (rain/no-rain or drought category accuracy)
GPCC	:	Global Precipitation Climatology Centre, a gauge-based global precipitation dataset (Deutscher Wetterdienst, Germany)
GPM	:	Global Precipitation Measurement, a joint NASA/JAXA satellite mission (successor to TRMM) for rainfall observation
HSS	:	Heidke Skill Score, a skill metric measuring forecasting accuracy relative to random chance
IMERG	:	Integrated Multi-Satellite Retrievals for GPM, the blended satellite precipitation product from the GPM mission
JJA	:	June–July–August, the main summer rainy season in the region (boreal summer)
KGE	:	Kling–Gupta Efficiency, a statistical measure of agreement (combining correlation, bias, and variability)
KNN	:	k-Nearest Neighbors, a non-parametric machine learning algorithm for regression/classification
Kappa	:	Cohen's Kappa, a statistic measuring agreement (here, between dataset predictions and observations) adjusted for chance

MAE	:	Mean Absolute Error, the average of absolute differences between estimated and observed values
MAM	:	March–April–May, the spring secondary rainy season in Ethiopia
MERRA-2	:	Modern-Era Retrospective Analysis for Research and Applications (version 2), a NASA global reanalysis dataset
NSE	:	Nash–Sutcliffe Efficiency, a normalized statistic indicating how well a time series model (or dataset) predicts observations (1.0 is perfect, 0 means no skill)
NB	:	Naïve Bayes, a probabilistic machine learning classifier based on Bayes' theorem (here used for drought index prediction)
PET	:	Potential Evapotranspiration, the climatic evaporative demand of the atmosphere (used in SPEI calculations)
POD	:	Probability of Detection, the fraction of actual events (e.g. rain days) that were correctly detected (hit rate)
PERSIANN-CDR	:	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record, a long-term satellite-based precipitation dataset (with monthly bias correction)
RF	:	Random Forest, an ensemble decision-tree machine learning model
R^2	:	Coefficient of Determination, the proportion of variance in observations explained by a model or dataset ($R^2 = 1$ indicates a perfect fit)
SEDI	:	Symmetric Extremal Dependence Index, a metric for extreme event detection skill that is less sensitive to event rarity
SON	:	September–October–November, the autumn transitional season
SPI	:	Standardized Precipitation Index, a meteorological drought index based on precipitation anomalies over a specified accumulation period

SPEI	:	Standardized Precipitation–Evapotranspiration Index, a drought index similar to SPI but incorporating precipitation minus evapotranspiration (a water balance approach)
TMPA	:	TRMM Multi-satellite Precipitation Analysis, the TRMM satellite's merged rainfall estimation product (e.g. 3B42/3B43)
TRMM	:	Tropical Rainfall Measuring Mission, a NASA satellite mission (1997–2015) that provided rainfall estimates (the predecessor to GPM)
UDel	:	University of Delaware Precipitation Dataset, a global gridded precipitation analysis derived from station data (University of Delaware)

1. INTRODUCTION

1.1. Background of the Study Area

Water availability in semi-arid regions hinges on accurate precipitation estimates, yet ground-based observations in remote basins are often sparse (Tadesse et al., 2022). The Shabelle River Basin of eastern Ethiopia – one of the country's largest basins – exemplifies this challenge. This basin's rainfall is delivered in a bimodal pattern due to the seasonal oscillation of the Inter-Tropical Convergence Zone (Tadesse et al., 2022).

However, its rain gauge network is limited in coverage and data continuity, falling below World Meteorological Organization recommendations for station density (Tadesse et al., 2022). Historical instability and resource constraints have further hindered consistent measurements and the result is significant uncertainty in quantifying spatial rainfall gradients, which tend to decrease from the high-elevation headwaters toward the lowlands region.

The Shabelle River Basin originates from the Bale Mountains in the Ethiopian highlands, situated at an elevation of about 4,230 meters above sea level. It has a catchment area of 297,000 km² (Thiemig et al., 2010). The basin is mostly located in Ethiopia, accounting for around 63.5% of its total area, while the remaining third is found in Somalia. In Ethiopia, many tributaries contribute to the flow of the river. Initially, the river runs southeast, then eastward, and then northeast to the East Hararge Zone of the Oromiya Region. The river then runs into the limestone tablelands in the basin's centre. River floods have severely degraded this landmass, but erosion is gradually decreasing near the Somali Region's beginning. At this point, the river turns southeast slightly upstream of Immi and flows into Somalia via Gode, Mustahil, and Ferfer (Thiemig et al., 2010)

The Shabelle River flows mostly in a south-easterly direction, passing through dry areas in eastern Ethiopia and cutting through expansive lowlands in southern Somalia. The river spans a total distance of about 2,526 km, with 1,290 km of its course being inside Ethiopia. However, it usually does not flow all the way to the Indian Ocean. Instead, it disperses and spreads out over a dry area in southern Somalia, where the sand absorbs it, nourishing a sensitive environment and replenishing nearby underground water sources. The Shabelle River only manages to merge with the Juba River and flow into the Indian Ocean during periods of very strong rainfall.

The basin has an average annual precipitation of 425 mm, with the mountainous regions receiving over 1,500 mm and the border areas receiving about 200 mm (Elmi Mohamed, 2013). The river also has a high saline content, even during periods of high flow which affects the agricultural activities in the surrounding areas. The fluctuating water levels and salinity of the river create challenges for farmers who rely on it for irrigation purposes.The Shabelle Basin is prone to recurrent droughts and floods because to its meteorological circumstances(Elmi Mohamed, 2013). These geographic and climatic contrasts make the Shabelle Basin prone to both flash floods and prolonged droughts, challenging water resource management and agricultural planning.

Accurate precipitation information is fundamental for water resource management, agriculture, and drought monitoring, especially in semi-arid regions. In the Horn of Africa, rainfall variability drives river flows, soil moisture, and drought incidence, directly impacting livelihoods. The Shabelle River Basin in eastern Ethiopia is a prime example: its predominantly rain-fed agricultural and pastoral communities are highly vulnerable to rainfall fluctuation. However, obtaining reliable precipitation data in this region is challenging due to the sparse distribution of rain gauges and gaps in long-term records.

Monitoring and understanding precipitation patterns in this basin is therefore critical. However, ground-based rain gauge networks in Ethiopia are sparse and unevenly distributed, especially in remote lowland areas like parts of Shabelle basin(Dinku, 2019). Many stations have short or incomplete records, and vast areas (including parts of the Shabelle Basin) are under-monitored (Dinku, 2019). This lack of dense observations limits our ability to capture the true spatio-temporal variability of rainfall (Bayissa et al., 2017). To bridge observational gaps, gridded precipitation datasets derived from satellites, reanalysis models, and gauge interpolation have become invaluable.

These products provide quasi-global coverage over multiple decades, enabling hydrological assessments in data-scarce basins. Yet, each data source has strengths and weaknesses (Van Dijk, et al., 2017). Satellite estimates can capture convective storm patterns but may suffer biases due to algorithm assumptions and orographic effects . In recent decades, global gridded precipitation datasets have emerged as a valuable alternative or complement to ground stations (Samim et al., 2024) These datasets provide continuous spatial coverage by combining information from satellite sensors, ground gauges, and/or atmospheric models (Samim et al., 2024).

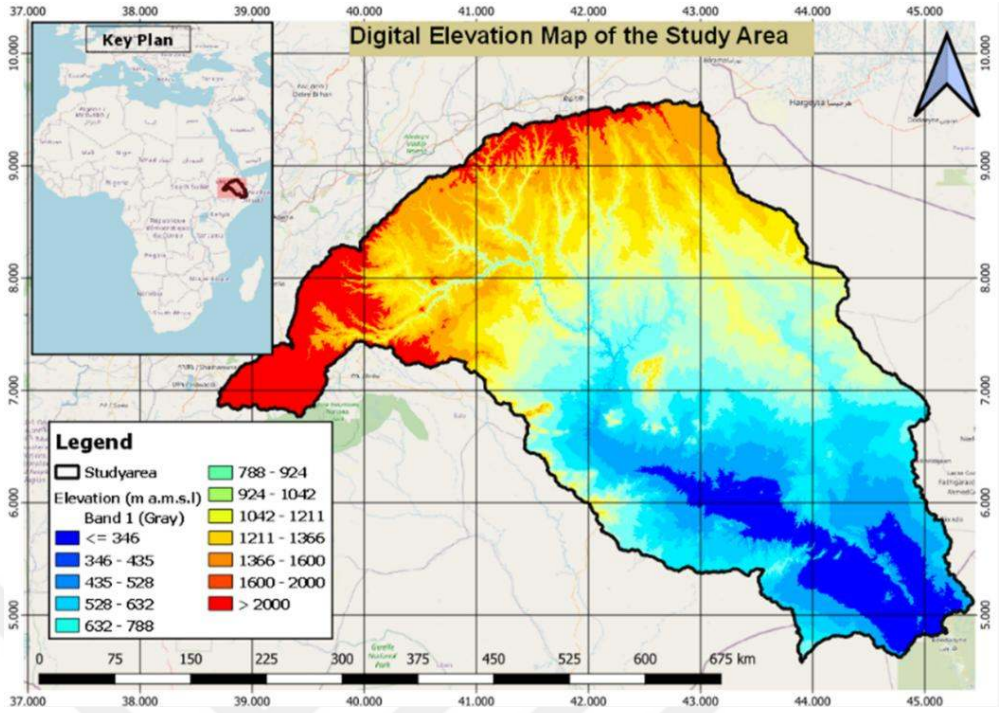

Figure 1.1. *Location of the study area (Shabelle Basin)*

Figure 1.1 show the location of the study area with Elevation ranges from the highlands of the Bale Mountains (above 3,200 m) to lowland plains near the Somalia border (below 500 m), leading to strong spatial gradients in climate (mean annual rainfall varies from over 1000 mm in highlands to under 300 mm in the dry lowlands).

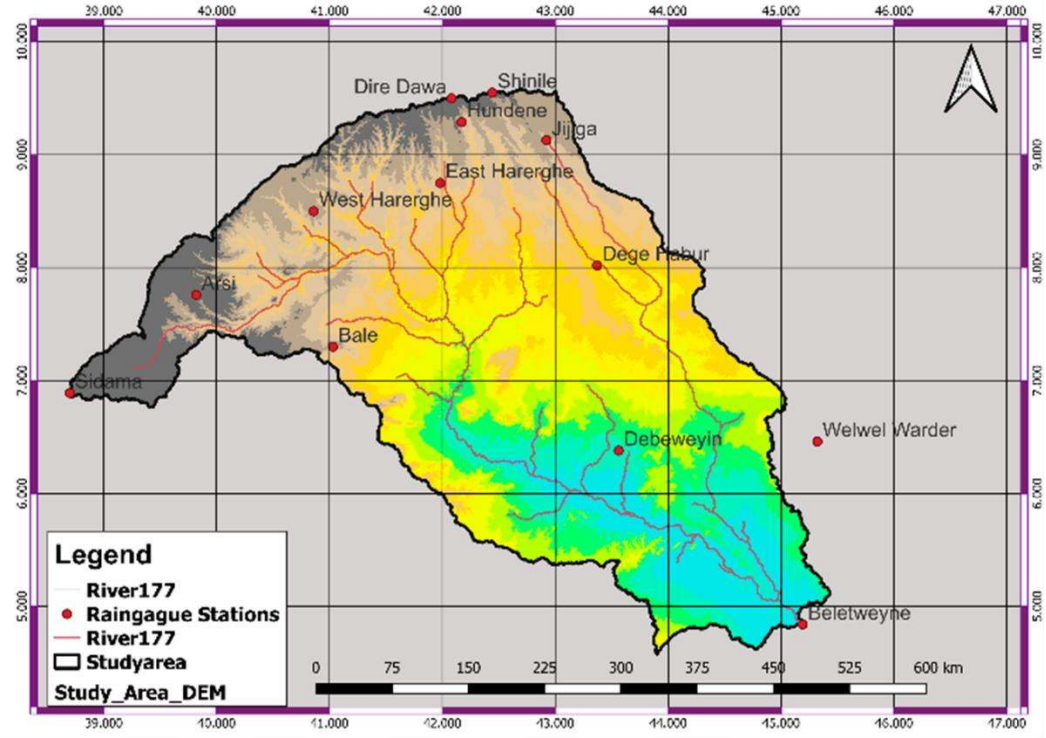

Figure 1.2. *Location of the upper stream rain gauge station on Shabelle River*

Locations of the available upper stream rain gauge station on Shabelle River that have fewer missing data are shown in Figure 1.2.

(Dinku et al., 2018) offer potential for monitoring rainfall and drought in data-scarce regions like the Shabelle Basin. (Beck et al., 2019)However, their accuracy can vary with region and dataset methodology. (Dembélé & Zwart, 2016)Satellite estimates may struggle in complex terrain or very dry conditions (Samim et al., 2024), and (Beck et al., 2019; Hersbach et al., 2020a) reanalysis products (which assimilate various observations into weather models) can have biases in rainfall timing and amount (Samim et al., 2024). It is therefore essential to evaluate how well these products represent actual rainfall in our study area before they are used for decision-making or drought assessment.

Reanalysis datasets assimilate diverse observations to produce a coherent climate record, but their accuracy depends on the representation of regional climate dynamics and available inputs.(Dinku et al., 2018; Funk et al., 2015) Gauge-based gridded products offer fidelity where station density is high, but in regions like Shabelle with sparse gauges, they must extrapolate over large areas. (Gebremicael et al., 2019;. Therefore, it is crucial to evaluate the performance of various precipitation datasets against the ground truth in specific basins.Recurrent droughts have severe socio-economic impacts in the Horn of

Africa, and effective early warning relies on trustworthy rainfall data. (Nwayor & Robeson, 2024; Wagesho & Claire, 2016) Indices like the Standardized Precipitation Index (SPI) and Standardized Precipitation-Evapotranspiration Index (SPEI) distill precipitation (and evapotranspiration) anomalies into standardized drought metrics. These indices are widely used for drought frequency and severity analysis (C. Chen et al., 2020; Gao et al., 2020; Ramadhan et al., 2022; Wei et al., 2021) . However, different precipitation inputs can produce divergent drought signals. (Fenta et al., 2018; Gao et al., 2020; Ramadhan et al., 2022)For instance, a recent evaluation in China found that bias-corrected satellite products (such as GPM IMERG-Final) and gauge-based datasets (like GPCC) were the most reliable for drought monitoring via SPI, whereas near-real-time satellite estimates often lagged in detecting drought onset. (Maghsood et al., 2020; Navidi Nassaj et al., 2022)Similar studies in Iran have highlighted that some global datasets underestimate precipitation deficits, affecting drought index values. Thus, understanding dataset biases is key for robust drought analysis.

The primary work of this thesis focuses on the Spatio-temporal evaluation of multiple precipitation datasets in the Shabelle River Basin from 1998–2016, and on comparing their representation of meteorological droughts against ground observations. In doing so, it addresses: (1) which gridded datasets (satellite-based, reanalysis, and gauge-based) best capture observed rainfall patterns in the basin; (2) how well these products detect and characterize drought events at monthly to annual scales (through SPI and SPEI indices).

1.2. Problem Statement

Despite the proliferation of global precipitation products, there remains a knowledge gap on their performance in semi-arid East African basins such as the Shabelle (Kabite Wedajo et al., 2021). Past studies have evaluated satellite and reanalysis rainfall data in various regions of Africa and the globe (Ahmed et al., 2024; Bayissa et al., 2017; Bitew et al., 2012; Degefu et al., 2022; Dembélé & Zwart, 2016; Dinku et al., 2018; Gebremicael et al., 2019; Hussein & Baylar, 2023; Kabite Wedajo et al., 2021; Satgé et al., 2020; Tadesse et al., 2022; Taye et al., 2020; Ware et al., 2023),but none has comprehensively focused on the Shabelle Basin with its unique climatic setting in-terms of detailed spatio-temporal evaluation and their applications of different machine learning algorithm. The basin's sparse gauge network makes it unclear which gridded dataset can

be trusted to reflect local rainfall patterns. Additionally, while drought indices like the Standardized Precipitation Index (SPI) and Standardized Precipitation-Evapotranspiration Index (SPEI) are widely used for drought monitoring (Gumus, 2023; Nwayor and Robeson, 2024; Pei et al., 2020), it is uncertain how consistently different precipitation datasets reproduce drought signals when compared to ground observations. In summary, the core problem is the uncertainty in which data sources most accurately capture the spatial-temporal characteristics of rainfall in the Shabelle Basin, and consequently how reliable they are for drought analysis in this data-scarce region. İt is also important to identify which machine learning algorith works best for Shabelle basin which the study will figure out to find. Addressing this problem is vital for improving drought early warning and climate risk management in the basin.

1.3. Research Questions

To tackle the above problem, this study is guided by several key research questions (RQs):

- RQ1: Which gridded precipitation datasets (satellite-based and reanalysis) most accurately represent observed rainfall patterns in the Shabelle River Basin?
- RQ2: How do these datasets compare in detecting and characterizing meteorological droughts in the basin across various time scales (monthly, seasonal, annual)?
- RQ3: What are the spatial patterns of rainfall distribution (e.g. by elevation zones and rainfall regimes) in the Shabelle Basin as depicted by gauge data and by the gridded products?
- RQ4: By applying common drought indices (such as the SPI, SPEI at 3, 6, and 12 month scales), do the satellite and reanalysis products reliably reproduce drought occurrences and severity when compared to ground observations?
- RQ5: Which precipitation product (or combination of products) can be recommended for operational drought monitoring in data-sparse basins like the Shabelle, and what are the implications for water resource management in the region?

- RQ6: By using different machine learning and identification of which machine learning algorithm works best for the drought prediction?

1.4. Objectives of the Study

In line with the research questions, the specific objectives of this dissertation are:

- Objective 1: Compile and quality-control a dataset of in-situ rainfall observations from the available stations within the Shabelle Basin, and assemble a suite of publicly available satellite-based and reanalysis precipitation products covering the same region and period (1998–2016)
- Objective 2: Evaluate the performance of these gridded precipitation products against ground station data using continous statistical measures (e.g. KGE, NSE, MAE,R, CC, RMSE, Bias) and categorical metrics (e.g. POD, CSI, FAR, SEDI, SR, HSS, Accuracy, F1-Score, Kappa) at daily, monthly, seasonal, and annual time scales. This will quantify how well each product captures observed rainfall amounts, variability, and rain/no-rain occurrence.
- Objective 3: Characterize the basin's precipitation regime spatially and temporally using the ground observations and examine how well each gridded product reproduces these observed spatial patterns and seasonality. This includes assessing whether the products capture known gradients (e.g. wetter highlands vs drier lowlands) and seasonal cycle features of Shabelle Basin rainfall.
- Objective 4: Identify the top-performing precipitation products in terms of agreement with ground observations, and utilize these best products to conduct an in-depth drought analysis.This involves computing drought indices (SPI and SPEI at 3-, 6-, and 12-month accumulation periods) from the precipitation data to assess historical drought frequency, duration, and intensity over the past two decades in the Shabelle Basin.
- Objective 5: Analyze the results of the drought assessment to understand the strengths and limitations of each precipitation dataset for drought monitoring. Based on these insights, provide recommendations on which dataset(s) are most suitable for operational use in similar data-sparse

regions, and whether blending multiple data sources could improve drought early warning capabilities.

- Objective 6: Apply different machine learning algorithm and identify which model works best for drought prediction.

Through these objectives, the study aims not only to evaluate data accuracy but also to translate those findings into practical guidance for water resource management and drought preparedness in the Shabelle Basin and comparable environments.

1.5. Significance of the Study Area

This research is significant for several reasons. First, it provides a comprehensive evaluation of multiple precipitation estimation products in one of East Africa's critical river basins. By benchmarking a broad array of gridded datasets against local observations, the study will shed light on data uncertainties in the Shabelle Basin and improve confidence (or caution) in using such products for drought analysis. Second, the findings directly support drought monitoring and climate risk management. Reliable precipitation data are the cornerstone of effective drought indices; thus, identifying which dataset best detects drought conditions will enhance early warning systems in the region. This is especially valuable for humanitarian and agricultural agencies operating in Ethiopia's Somali region, where on-the-ground information is limited. Third, the study addresses the challenge of data scarcity by testing alternative data sources. It contributes to the broader scientific understanding of how satellite and reanalysis rainfall products perform in dry, topographically diverse regions. Previous evaluations in other countries have shown varied results – for example, some global products work well in certain areas while failing in others. This research will fill a geographical gap in the literature by adding detailed knowledge for the Shabelle Basin. Lastly, the methodology and insights from this work can be generalized to similar basins across Africa. The approach of combining spatio-temporal validation with drought index analysis is an innovative aspect that can guide future studies. Ultimately, the study's outcomes aim to inform operational decision-making, helping stakeholders choose appropriate precipitation data for modeling river flows, planning irrigation, and mitigating drought impacts in data-sparse regions.

1.6. Structure of the Thesis

This thesis is structured for five chapters. Chapter one which introduction where the background of the study area, problem statement, research questions, objectives and the significance of the study area are discussed. Chapter two reviews literature on gridded precipitation datasets, droughts in Ethiopia and drought indices that will be used for the study. Chapter 3 discusses methodology used for the study where chapter 4 presents the results and discussions of the findings in detail. Chapter 5 draws conclusions and the recommendations.

2. LITERATURE REVIEW

2.1. Overview of Gridded Precipitation Datasets and Their Use in Africa

A wide array of gridded precipitation datasets has been developed over the past few decades, each using different input data and algorithms (Beck, Van Dijk, et al., 2017; Gao et al., 2020; Navidi Nassaj et al., 2021, 2022; Samim et al., 2024; Satgé et al., 2020). Broadly, these products fall into three categories: satellite-based estimates, reanalysis-based datasets, and gauge-based interpolations (some products involve a blend of these). This study will review and evaluate number of gridded datasets from each category, namely CHIRPS, TRMM, PERSIAN CDR (satellite-related products); ERA5(and derivatives), MERRA-2, CFS (reanalysis products); and CPC Unified Precipitation (gauge-based). We highlight findings from recent validation studies around the world and in similar climates (East Africa and semi-arid regions) to provide context on each dataset's performance.

2.1.1. Gauge-based gridded datasets

Gauge-based gridded datasets are constructed by interpolating measurements from rain gauge networks onto a regular grid. They represent the "ground truth" more directly than satellite or reanalysis products, at least in areas where station density is high. However, their accuracy is entirely dependent on the distribution and quality of the underlying gauges (Navidi Nassaj et al., 2021; Wei et al., 2021). In regions with abundant stations, gauge analyses can be very reliable; in sparsely gauged regions, they may have large uncertainties in between station locations (Samim et al., 2024). The NOAA CPC Unified Precipitation Project (CPC UPP) is one such global dataset, and is included in this study. It provides daily precipitation on a 0.5° (~50 km) grid globally, generated by interpolating thousands of gauges (from national networks and synoptic reports) using algorithms that ensure global consistency (Beck et al., 2019; Gebremicael et al., 2019; Navidi Nassaj et al., 2022; Schneider et al., 2013; Sun et al., 2018; Wei et al., 2021c). The CPC analysis spans several decades (from ~1979 onward). Other notable gauge-based globals are the GPCC (Global Precipitation Climatology Centre) dataset and the CRU TS (Climatic Research Unit Time Series) product, as well as the University of Delaware (UDel) dataset – each using somewhat different input station archives and interpolation methods.

Several studies have benchmarked these gauge-based datasets, often finding them to be top-performers when and where gauge coverage is sufficient. For example, in Iran, using 96 stations as reference, GPCC was found to have the best agreement with observations in a majority of locations, outperforming CRU, CPC, and others (Navidi Nassaj et al., 2022). GPCC's strength was also evident in reproducing the probability distribution of precipitation closely to observed in many grids (Navidi Nassaj et al., 2022). A 2018 evaluation across Iran similarly concluded that GPCC was the most suitable dataset nationally, with UDel and CRU next, while a satellite product (PERSIANN-CDR) was weakest. In arid Balochistan (Pakistan), GPCC again emerged as the best among gauge-based products compared (versus APHRODITE and UDel) according to error metrics like mean absolute error (K. Ahmed et al., 2019; Samim et al., 2024). These findings indicate that when an area's climate network is reasonably dense (as in parts of Asia), the GPCC or similar gauge products leverage that data effectively.

However, in data-sparse regions such as Eastern Africa, the situation can differ. Fewer gauges mean that large areas have to be extrapolated. A study in the Eastern Nile Basin (which covers parts of Ethiopia and Sudan) found that GPCC still showed the best performance overall (by correlation and agreement metrics), but interestingly, the satellite product TRMM was the next best, while CRU and UDel performed relatively poorly (Abdelwares et al., 2020; Basheer & Elagib, 2019; Bayissa et al., 2017).

The poor performance of CRU/UDel in that case was attributed to their coarse native resolution and limited station inputs in the region, causing them to miss local variations. This underscores that gauge-based datasets are not inherently infallible – their quality hinges on data availability. For the Shabelle Basin during 1997–2016, the gauge network within Ethiopia was quite sparse (only a handful of stations with continuous data), so global gauge analyses like CPC or GPCC likely had to interpolate across large distances. In fact, CPC (gauge-only) and GPCC are known to underestimate precipitation in areas where very few stations exist. This literature review found an example in mainland China's drought analysis: GPCC performed best among gauge products, but still had notable regional biases; it was judged the most reliable largely because of China's relatively dense network (Wei et al., 2021c). In contrast, for Ethiopia/Somalia, the reliability of GPCC or CPC might be lower (Basheer & Elagib, 2019). This is one reason satellite-gauge blends like CHIRPS have become popular in Africa – they at least

introduce satellite estimates in gauge-void regions, which can be more realistic than a broad interpolation from a distant gauge.

CPC Unified specifically has been less frequently singled out in literature compared to GPCC and CRU. In (Wei et al., 2021c)'s ranking for China, GPCC was the top gauge-based product and CPC was slightly behind (they noted GPCC had higher accuracy in drought detection) . In Afghanistan's recent assessment, CPC and GPCC both showed high correlations (median $R > 0.6$ at monthly scale) with observations across many stations (Samim et al., 2024), indicating strong performance of gauge-based datasets even in an environment of data scarcity, provided some station data exist.

In summary, gauge-based gridded datasets are often the benchmark for evaluating other products when gauge density is adequate (Samim et al., 2024). They tend to capture the timing of rainy seasons correctly (since they rely on actual observations) but may misrepresent the magnitude in under-monitored areas. Our study uses the CPC Unified Precipitation dataset as a representative gauge analysis. This study will cross-reference its performance with that of GPCC (from literature) to contextualize results. Overall, the literature suggests that no single dataset is universally superior (Navidi Nassaj et al., 2022) ; therefore, combining the strengths of multiple data sources (e.g., satellite + gauge) or at least understanding their error characteristics is crucial for robust hydro-climatic studies in regions like Shabelle basin.

2.1.2. Reanalysis-based precipitation datasets

Reanalysis datasets offer another pathway to gridded precipitation, by integrating available observations into a weather forecasting model to reconstruct the climate state. Modern reanalyses provide complete spatial coverage and a range of meteorological variables at relatively high resolution (Hersbach et al., 2020a; Muñoz-Sabater et al., 2021; Saha et al., 2010; Wei et al., 2021) . Precipitation from reanalyses is a model output constrained indirectly by other assimilated fields (pressure, humidity, etc.), since rain gauges are typically not directly assimilated . Three reanalysis-based products are considered in this study: ERA5 (and its land-focused variants), MERRA-2, and CFS.

ERA5 is the fifth-generation global reanalysis from ECMWF, available from 1979-present at ~31 km grid spacing (0.25°) and hourly time steps and it represents a major improvement over the earlier ERA-Interim, with data assimilation of satellite-era observations and advanced model physics (J. S. Ahmed et al., 2024a; Hersbach et al.,

2020b). For land surface applications, ERA5-Land was produced as a replay of ERA5's land component at 0.1° (~9 km) resolution, providing finer detail of precipitation and soil moisture by downscaling the atmospheric forcing (ERA5-Land Hourly Data from 1950 to Present, n.d.; Muñoz-Sabater et al., 2021). Additionally, a specialized daily dataset known as AgERA5 (referred here as "ERA5-Ag-9km") provides agro-climatology variables including precipitation at ~9 km, bias-adjusted for agricultural studies (Agrometeorological Indicators from 1979 to Present Derived from Reanalysis, n.d.; Global Weather for Agriculture - AgERA5 - Datasets - "FAO Catalog," n.d.). These ERA5-derived products are highly relevant for our study region, as they potentially resolve orographic effects better (ERA5-Land) and may incorporate local bias corrections (ERA5-Ag) to improve rainfall estimates. However, global analyses are known to have systematic biases in precipitation. (Navidi Nassaj et al., 2022) found that ERA5 exhibited a consistent wet bias across all seasons and regions in Iran, tending to overestimate precipitation totals relative to gauges. Other evaluations have noted that ERA5 underestimates extreme daily rainfall (Dembélé & Zwart, 2016) and sometimes mis-timings peak rainy seasons in parts of Ethiopia (Gebremicael et al., 2019). These biases likely stem from model parameterizations of convection and the lack of direct gauge assimilation. By comparing ERA5 and its high-resolution derivatives with ground data, recent studies aim to assess whether finer resolution alleviates some bias. For example, (Cao et al., 2020) report that increased spatial resolution can reduce differences with point observations. Literature reviews indicate that ERA5's high resolution (and ERA5-Land's 9 km grid) often yields better spatial pattern matching but does not automatically solve biases in magnitude. Therefore, ERA5's strong representation of seasonal patterns will be scrutinized alongside any systematic offset in the Shabelle basin context.

MERRA-2 (Modern-Era Retrospective Analysis for Research and Applications v2) is NASA's latest reanalysis (1980–present) that includes an improved treatment of the water cycle, such as corrections for gauge-based precipitation in its land surface forcing. MERRA-2 has a resolution of ~0.5° latitude by 0.625° longitude and assimilates satellite precipitation indirectly by adjusting soil moisture when large precipitation biases are detected. Interestingly, MERRA-2 has shown robust performance in some intercomparisons. In a comprehensive study over China, MERRA-2 had the best performance among reanalysis datasets for both precipitation estimation and drought index calculation (SPI), outperforming even ERA5 in that context(Li et al., 2020; Pei et

al., 2020; Wei et al., 2021). It captured monthly rainfall variability reasonably well and contributed to more accurate drought monitoring than other reanalyses. In Iran, MERRA-2 also performed competitively; although gauge-based datasets led overall, MERRA-2 was noted for not exhibiting the strong wet biases that ERA5 did (Navidi Nassaj et al., 2021) In arid northwestern. These mixed findings imply that MERRA-2's bias can be region-specific – relatively wetter than observed in some cases, drier in others. Its performance in Ethiopia is not yet fully established in literature; given Ethiopia's complex highland climate, it will be insightful to see if MERRA-2's adjustments (like the Corrected Precipitation algorithm) yield closer agreement with observed totals. Prior work in Africa often did not include MERRA-2, focusing instead on satellite products, so our study helps fill that gap by evaluating MERRA-2 for the Shabelle Basin.

NCEP's Climate Forecast System Reanalysis (CFS) and its subsequent operational analysis (CFSv2) provide another source of model-based precipitation data. CFSR (available ~1979–2010 at ~0.3°) and CFSv2 (available from 2011 onward as part of a coupled forecast system) offer ~0.2° resolution precipitation outputs (Saha et al., 2010, 2014). This study refers to this collectively as "CFS" dataset for our 1997–2016 period. The CFS model includes a fully coupled atmosphere-ocean system and assimilates a variety of observations (though not local rain gauges directly). It produces precipitation as a forecast variable up to 6 hours ahead, which is then stitched into a continuous record.

In summary, reanalysis datasets provide complete spatio-temporal coverage and multi-decadal records, which are invaluable for places like the Shabelle Basin with limited data. Yet, literature emphasizes caution: reanalysis precipitation is not a direct observation and must be validated. Bias correction is often needed before use in water balance studies. This literature review suggests MERRA-2 and ERA5 family as among the most advanced reanalyses, with MERRA-2 sometimes excelling in arid conditions (Gelaro et al., 2017). ERA5's higher resolution (and derivatives) are promising for terrain-following detail, though biases (e.g. too wet in Iran) are documented (Navidi Nassaj et al., 2022). CFS provides an additional perspective of a coupled model's output, which some studies found surprisingly reliable. By comparing all three in the Shabelle Basin, this study will contribute to understanding how these reanalysis products perform in an East African setting that includes both highlands and arid lowlands like Shabelle basin.

2.1.3. Satellite-based and satellite-gauge plended products

Satellite precipitation estimates have become indispensable for hydrologic monitoring in regions with sparse gauges. They typically rely on satellites' infrared (IR) or microwave sensors to infer rainfall(Aghakouchak, Feldman, and Hsu, 2011; Bayissa et al., 2017; Belayneh et al., 2020; Dembélé & Zwart, 2016; Z. Wang et al., 2017; Wei et al., 2021) . A prominent example is the Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA 3B42) eproduct, available at 0.25° (~25 km) resolution from 1998 to 2019 (Huffman et al., 2007; Prakash et al., 2016; X. Wang et al., 2019). TRMM 3B42 merges passive microwave rainfall retrievals from satellites and calibrates IR estimates, with a monthly bias correction using rain gauge data . Studies have shown TRMM's performance to be regionally variable. For instance, in the Eastern Nile Basin of East Africa, TRMM was found to be the second-best dataset (after a gauge product) in capturing monthly rainfall patterns (Abdelwares et al., 2020) . Generally, TRMM tends to struggle in areas of complex terrain or where convective storm structures differ from the tropical ocean calibration baseline. Nevertheless, TRMM (and its successor, GPM IMERG) remain widely used and have been benchmarked in numerous global studies (Wei et al., 2021).

CHIRPS (Climate Hazards Group InfraRed Precipitation with Stations) is another key dataset, specifically designed for monitoring rainfall in Africa and other drought-prone regions. CHIRPS provides quasi-global coverage at 0.05° (~5 km) resolution, with data from 1981-present (J. S. Ahmed et al., 2024a; Dinku et al., 2018; Hussein & Baylar, 2023; Popovych & Dunaieva, 2021) . It blends satellite IR estimates with in-situ rain gauges, incorporating station data in a two-step bias correction process . This combination leverages satellites for spatial coverage and gauges for reducing bias. CHIRPS has been extensively validated over Africa (Dinku et al., 2018). Previous studies indicate that CHIRPS captures spatial and temporal rainfall patterns well in many parts of Ethiopia(J. S. Ahmed et al., 2024a; Hussein & Baylar, 2023). Notably, it has been used for drought monitoring applications given its long record and fine resolution, and it was specifically developed to support the US FEWS (Famine Early Warning Systems) by detecting rainfall deficits (Wei et al., 2021c). However, CHIRPS is not without limitations; it may underperform during extreme dry spells or in detecting isolated convective downpours . For example, in regions or seasons with very sparse station input, CHIRPS can inherit errors from the satellite IR estimates, leading to underestimation of intense storms or false

detection of light rain. Despite these caveats, CHIRPS generally ranks among the top-performing satellite-gauge products in East Africa. In semi-arid Afghanistan, CHIRPS showed relatively strong correlations with station data (median Pearson R > 0.4 in dry season) and was among the higher-scoring datasets countrywide , though its performance varied with topography (Samim et al., 2024).

PERSIANN-CDR (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record) is a long-term satellite precipitation dataset spanning 1983–present at 0.25° resolution (Ashouri et al., 2015) . It is based on IR satellite imagery calibrated with rain gauge data via neural network techniques. PERSIANN-CDR offers daily estimates suitable for climatic analysis. In practice, its performance has been mixed. Evaluations across various regions consistently rank PERSIANN-CDR slightly lower than CHIRPS or gauge-based datasets in accuracy . For instance, a country-wide study in Iran found PERSIANN-CDR to have the weakest performance among the datasets tested (which included GPCC, CRU, and UDel gauge products). Similarly, (Wei et al., 2021c) examining drought monitoring in China, noted that near-real-time PERSIANN products had poorer skill in detecting monthly precipitation than bias-corrected satellite products. One recurring issue is significant underestimation of precipitation by PERSIANN in many cases. In Turkey, for example, a recent study observed that the original PERSIANN algorithm drastically underestimated precipitation totals at regional scale, especially during wet seasons (Uysal & Şorman, 2021). It is worth noting that many satellite products are now blends of satellite and gauges (CHIRPS is one, and even TRMM 3B42 V7 uses monthly gauge adjustments). Such blending generally improves bias and rainfall frequency detection.

Table 2.1. *Classification of different precipitation datasets used for the study*

Dataset	Full Name	Data Source(s)	Spatial Resolution	Temporal Resolution	Spatial Coverage	Temporal Coverage	Reference
CHIRPS	Climate Hazards Group InfraRed Precipitation with Stations	USGS/CHG	0.05° (~5.5 km)	Daily	50°S–50°N	1981–Present	(Funk et al., 2015b)
GROUND	Ground-based Meteorological Data	G	Point stations	Daily	Ethiopia	1997–2016	Historical Precipitation Data for Ethiopia (1997–2016).
CFS	Climate Forecast System Reanalysis	NOAA/NCEP	0.5° (~55 km)	Hourly	Global	1979–NRT	(Saha et al., 2010)
CPC-UPP	CPC Unified Gauge-Based Analysis	NOAA/CPC	0.5° (~55 km)	Daily	Global	1948–Present	(Schneider et al., 2013)
ERA5	ECMWF Reanalysis 5	ECMWF	0.25° (~31 km)	Hourly	Global	1950–Present	(Hersbach et al., 2020b)
TRMM-3B42	Tropical Rainfall Measuring Mission Multi-Satellite Precipitation Analysis	NASA/JAXA	0.25° (~27 km)	3-hourly	50°S–50°N	1998–2019	(Huffman et al., 2007)

Table 2.1. *(Continued) Classification of different precipitation datasets used for the study*

ERA5-Land	ECMWF Reanalysis 5 Land	ECMWF	0.1° (~11 km)	Hourly	Global	1950–Present	(*ERA5-Land Hourly Data from 1950 to Present*, n.d.; Muñoz-Sabater et al., 2021)
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, V2.	NASA/GMAO	0.5° (~55 km)	Hourly	Global	1980–Present	(Gelaro et al., 2017; *MERRA-2*, n.d.)
PERSIANN-CDR	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks – Climate Data Record	NOAA/CPC	0.25° (~27 km)	Daily	60°S–60°N	1983–2022	(Ashouri et al., 2015)
AgERA5	ECMWF Reanalysis 5 for Agriculture	ECMWF	0.1° (~9 km)	Daily	Global	1979–Present	(Agrometeorological Indicators from 1979 to Present Derived

Beyond these categories, some blended datasets combine multiple sources (satellite + gauge, or satellite + reanalysis + gauge) to leverage their complementary strengths

(Dinku et al., 2018). Notable examples are CHIRPS (which blends satellite estimates with station data) and the GPCC Monitoring Product (which blends gauges with some model data) (Basheer & Elagib, 2019). The PERSIANN-CDR dataset calibrates satellite estimates with monthly gauge-based climatology (GPCP) (Ashouri et al., 2015; Ceferino-Hernández et al., 2024; Sadeghi et al., 2021; Uysal & Şorman, 2021). The goal of blending is often to improve bias and obtain better absolute accuracy. Indeed, in East Africa, blended products like CHIRPS have become popular for applications such as agricultural drought monitoring(Hussein & Baylar, 2023; Tadesse et al., 2022; Taye et al., 2020; Ware et al., 2023b) . However, blending cannot fully compensate for sparse ground data – if very few stations exist in a basin, even a blended dataset's quality will suffer, as is the case in parts of the Shabelle Basin.

In conclusion, the literature reveals a wide array of precipitation datasets available for hydrological studies, each with distinct strengths and weaknesses. Table 2-1 summarizes the key characteristics of the data sets selected for this study. Previous evaluations in various regions (Asia, Africa, Middle East) highlight that no single dataset is universally "best"; performance depends on climatic regime, terrain, and the presence of ground data for calibration (K. Ahmed et al., 2019; Beck, Vergopolan, et al., 2017; Katsanos et al., 2016; Navidi Nassaj et al., 2021, 2022; Satgé et al., 2020). This motivates the study's comparative approach in the Shabelle Basin. By reviewing these studies, this study anticipates that gauge-informed products (like CHIRPS or CPC) may excel in overall bias, satellites (like TRMM or CMORPH) may capture event timing well, and reanalyses (like ERA5 or MERRA-2) may provide consistency over space – but their true accuracy in Shabelle remains to be tested.

2.2. Performance of Gridded Products in Drought Monitoring

Quantifying drought requires condensing complex hydro-meteorological data into simple indices that reflect the severity of dry conditions. Over the years, many drought indices have been developed, but two of the most widely used in meteorology are the Standardized Precipitation Index (SPI) and the Standardized Precipitation-Evapotranspiration Index (SPEI). These indices are used for drought analysis and monitoring. These indices translate precipitation (and temperature/evaporation for SPEI) into standardized anomalies, facilitating comparison of drought severity across space and time. If a gridded dataset accurately captures rainfall, it should also reproduce observed

drought index variations. Conversely, biases or errors in precipitation will affect drought identification (e.g., a wet-biased dataset might underestimate drought frequency).

Recent studies have directly evaluated precipitation datasets for drought analysis (Degefu et al., 2022; Gumus, 2023; Li et al., 2020; Navidi Nassaj et al., 2021; i et al., 2020; Philip et al., 2018; Taye et al., 2020; Tesfamariam et al., 2019; Vicente-Serrano et al., 2010; Wei et al., 2021). They found that the ranking of datasets for drought monitoring closely mirrored their performance in rainfall estimation. In other words, datasets that better matched observed precipitation (e.g., GPCC, MERRA-2, bias-corrected satellites) also produced SPI time series more consistent with the benchmark. Notably, (Wei et al., 2021c)reported that MERRA-2 and GPCC enabled the most accurate drought detection (based on SPI), while some real-time satellite products (with biases) showed inferior SPI behavior (Wei et al., 2021c). This suggests that bias correction is crucial for drought indices – even if two products both "see" a rainfall deficit, the magnitude matters for the standardized index value.

In regions of Africa, similar comparisons are emerging (Taye et al., 2020). For instance, (Navidi Nassaj et al., 2022)included a drought perspective by comparing how different datasets capture dry spells in Iran; they concluded that no single dataset was best in all aspects, reinforcing that one must be cautious in generalizing results. A study in Iran focusing on drought monitoring with gauge-based gridded data (GPCC, CRU, and CPC) found that all three could reproduce broad drought patterns, but differences in precipitation amount led to discrepancies in drought intensity classification. Gauge-only products tended to miss some extreme drought events in poorly gauged areas. In Ethiopia, although detailed literature is sparser, there have been efforts to use CHIRPS and other satellite products for computing SPI for early warning purposes. These efforts generally show that CHIRPS-based SPI correlates well with ground-based SPI in areas near stations but may diverge in remote areas or at finer spatial scales.

Drought analysis in data-scarce regions has increasingly leveraged the above datasets. The Standardized Precipitation Index (SPI), which normalizes precipitation over a chosen accumulation period (e.g. 3, 6, 12 months) to a probability distribution, is a preferred tool for meteorological drought because it requires only precipitation data. Many studies have compared SPI time series computed from different precipitation products to gauge-based SPI as a benchmark. In Iran, for instance, (K. Ahmed et al., 2019)evaluated SPI-12 from three global gridded datasets (GPCC, CRU, and a gauge-

based Asian dataset) and concluded that while all three tracked major droughts, the magnitude and duration of mild droughts varied, with GPCC being most reliable overall. In Pakistan's arid climate, (K. Ahmed et al., 2019)similarly found that global products could identify prolonged drought periods but sometimes disagreed on short-term dryness due to differences in drizzle and light rain detection.

The Standardized Precipitation-Evapotranspiration Index (SPEI) extends SPI by incorporating atmospheric demand via potential evapotranspiration (PET) (Gumus, 2023; Nwayor & Robeson, 2024; Ojha et al., 2021; Pei et al., 2020). SPEI is sensitive to temperature changes and has been used to assess drought under global warming scenarios. For example, (Abbasi et al., 2019) introduced a combined precipitation–temperature decile index for Lake Urmia, Iran, to account for warming effects on drought severity. They found that including temperature data revealed an increasing risk of meteorological drought under climate change that precipitation alone (SPI) might underestimate. In East Africa, rising temperatures and periodic extreme El Niño events have prompted use of SPEI to gauge not just rainfall deficits but also enhanced evaporation during heatwaves. (Haile et al., 2020; Omondi & Lin, 2023; Polong et al., 2019) A study of East African drylands (Kalisa et al., 2020) using SPEI identified significant shortening of rainy seasons and longer drought spells post-1980s. These findings underscore why our study incorporates SPEI: in the Shabelle basin, high evaporation rates in hot years could intensify drought impacts beyond what SPI indicates.

When multiple products are used for drought monitoring, researchers often perform ensemble or comparative analyses. Study by (Wei et al., 2021c), for example, compared 17 precipitation products for SPI-based drought monitoring in China. They evaluated detection capability for historical drought events and found that products with higher monthly accuracy also ranked higher in drought index performance. Notably, bias-adjusted GPM (IMERG-F) and the GPCC gauge product had the best agreement with the reference drought index, while near-real-time satellite products and certain reanalyses (ERA-Interim) lagged. This suggests that the quality of precipitation data directly affects drought index reliability, a logical but important confirmation. In another study, (S. Chen et al., 2022) examined how well satellite products (including TRMM and CHIRPS) could monitor drought across China. They reported that all datasets generally captured the broad 1999–2011 drought in North China, but in smaller scale droughts, the differences were larger – with gauge-informed datasets giving more accurate local intensity.

Over Africa, (Bayissa et al., 2017) demonstrated an application of satellite rainfall for drought monitoring in the Upper Blue Nile (Ethiopia). They evaluated CHIRPS and other estimates for computing SPI-3 and SPI-6, finding that CHIRPS-based SPI correlated well with gauge-based SPI (correlation >0.8 at multiple stations) and effectively captured the timing of meteorological drought onsets. This gives confidence that similar satellite data can be used in the Shabelle basin. Meanwhile, (Dahri et al., 2021) assessed a host of gridded products in the high-altitude Indus basin for drought monitoring. They concluded that high-resolution gauge-corrected products (like APHRODITE for that region) provided the most reliable drought indices, but even reanalyses could be used after bias adjustment to fill gaps.

From these studies, a common approach emerges to use multiple datasets to cross-validate drought signals, and apply bias corrections or ensemble averaging to mitigate individual dataset errors. For this study, these lessons inform the methodology – this study will evaluate each dataset's baseline precipitation accuracy and then interpret SPI/SPEI differences in that context. By doing so, it can identify which data source (or combination) is most suitable for operational drought monitoring in the Shabelle basin's data-scarce environment. The literature strongly indicates that no single dataset is perfect; thus, a blended or carefully chosen set of products often yields the best insight.

In summary, the literature suggests that no single precipitation product is universally "best." Each has strengths and weaknesses, and using them in combination (e.g., for ensemble or integrated analysis) can provide more robust insights. This dissertation's approach is grounded in these findings: we evaluate a representative suite of datasets (satellite, reanalysis, gauge) in the specific context of the Shabelle Basin, then apply the data to drought index calculations with an understanding of each dataset's uncertainties. This dual use (evaluation + application) follows recommendations by researchers who advocate that dataset validation should be an integral part of climate impact analyses like drought assessment . With the background of existing knowledge established, we proceed to detail our methodology for dataset evaluation and drought analysis.

2.3. Drought Prediction with SPI and Machine Learning

East Africa is highly vulnerable to recurring droughts that impact agriculture, water resources, and livelihoods. In particular, the Shabelle River Basin in Ethiopia and

neighboring Somalia frequently experiences severe drought conditions, necessitating improved early warning and prediction systems. A widely used metric for quantifying drought severity is the Standardized Precipitation Index (SPI), which translates rainfall anomalies into standardized drought/wetness categories. Since 2018, there has been a surge in research coupling SPI analysis with machine learning (ML) algorithms to enhance drought prediction and management. Studies in East Africa – for example, in the Borana zone of southern Ethiopia and the Wami sub-basin of Tanzania – have begun to explore ML-based SPI forecasting, though this field is still emerging in the region (Piraei et al., 2024). To contextualize these efforts, this review examines how four common ML models – Random Forest (RF), K-Nearest Neighbors (KNN), Linear Regression, and Naïve Bayes – have been applied in SPI-based drought prediction and hydrological studies from 2018 to 2025. We compare findings from East Africa with broader results in similar climates, highlighting each model's performance, advantages, and limitations as reported in peer-reviewed literature.

2.3.1. Standardized Precipitation Index (SPI) in drought studies

SPI is a probabilistic index that measures precipitation deviation from the long-term normal, normalized to a standard distribution (Piraei et al., 2024). An SPI value is computed over a chosen time scale (e.g.3, 6, 12 months) and indicates wet or dry conditions; for example, SPI ≤ -1 signifies moderate to extreme drought, while SPI $\geq +1$ indicates excess rainfall. Because SPI can be calculated at multiple time scales, it is effective for characterizing short-term meteorological droughts (e.g. 3-month SPI) as well as long-term hydrological droughts (e.g. 12- or 24-month SPI) (Piraei et al., 2024). In East Africa, SPI has been the primary index for drought monitoring. Mulualem et al. (2024) analyzed SPI-based drought patterns across Ethiopia's major river basins (including the Shabelle) for 1981–2018, confirming SPI's utility in capturing regional drought variability (Mulualem et al., 2024; Worku, 2024a). SPI is also favored because it requires only precipitation data, which is often more readily available than evapotranspiration or streamflow data in data-sparse regions. However, SPI on its own is a retrospective indicator; to predict future SPI (and thus anticipate drought), researchers have increasingly integrated SPI time series with data-driven ML models (Piraei et al., 2024). This combination leverages historical climate patterns to forecast upcoming SPI values, thereby providing advance warning of drought or recovery.

2.3.2. Machine learning approaches for drought prediction

Machine learning algorithms can learn complex, non-linear relationships between climate inputs and drought indices, potentially improving prediction accuracy over traditional statistical models. In drought studies, ML models are typically trained to forecast SPI at a future time (e.g. SPI at month $t+1$) using prior climate observations (SPI or rainfall at months $t, t-1, \ldots$)(Piraei et al., 2024). Many studies adopt a comparative approach: they train multiple ML models on the same data and evaluate performance using metrics like the coefficient of determination (R^2), root-mean-square error (RMSE), and Nash–Sutcliffe efficiency (NSE). Below, we review findings for four ML algorithms – RF, KNN, linear regression, and Naïve Bayes – emphasizing applications in East Africa or analogous environments.

In the context of East Africa and the Shabelle Basin, the literature suggests that importing methodologies that succeeded in similar environments can be fruitful. For example, the performance of RF and SVR in semi-arid Iran or northern Africa (Piraei et al., 2024) is encouraging for analogous climates in Ethiopia and Somalia. This result aligns with global trends where support vector and ensemble methods often excel in drought index forecasting. The study underscores the importance of evaluating multiple models to identify the optimal one for a given region. Likewise, the Tanzanian study (Lalika et al. 2024) and others show a growing regional adoption of ML for drought early warning, focusing on algorithm comparisons and customization to local data.

In summary, the period 2018–2025 has seen significant advancements in applying ML techniques to SPI-based drought prediction. Random Forest has emerged as a leading algorithm due to its strong predictive accuracy and resilience in various climate settings (Piraei et al., 2024; Poudel et al., 2024). K-Nearest Neighbors and linear regression, while conceptually and computationally simpler, generally serve as baseline models and are often outperformed by more flexible methods (Citakoglu & Coşkun, 2022; El Ibrahimi & Baali, 2018) . Naïve Bayes is notably under-represented in recent drought literature, likely due to its inherent assumptions not aligning with the data characteristics, and thus remains an avenue for potential exploration rather than a proven tool. Many studies favor hybrid and ensemble approaches, reflecting the multifaceted nature of drought phenomena. The East African context, including the Shabelle Basin, stands to benefit from these developments. Early research in the region confirms that ML models can improve drought

forecasts, with the choice of model affecting accuracy and lead time. As climate data availability in East Africa improves (through remote sensing and regional climate services) and as computational capacity grows, we expect more local studies will validate the findings from other regions. Going forward, an integrated approach – combining SPI with machine learning, and even blending multiple ML models – appears to be the most promising path for effective drought early warning systems in East Africa. This approach will enable stakeholders in Ethiopia and neighboring countries to better anticipate droughts and implement mitigation strategies, thereby enhancing climate resilience.

3. RESEARCH METHODS AND METHODOLOGY

3.1. Data Acquisition

3.1.1. Study area

The study was conducted in the Shabelle River Basin, a transboundary basin spanning parts of the southeastern Ethiopian highlands and southern Somalia. This region is characterized by bimodal rainfall regimes driven by the intertropical convergence zone, with a primary wet season in March–May and a secondary wet season in late summer to autumn (Tadesse et al., 2022). Rainfall is highly variable in space and time, ranging from over 1400 mm annually in the upper highlands to below 300 mm in the downstream arid zones (Tadesse et al., 2022). Given the sparse distribution of rain gauges in the basin and the pronounced spatial gradients in rainfall, satellite and reanalysis products provide an important data source for comprehensive precipitation analysis (Ayehu et al., 2018).

3.1.2. Observed rainfall data

Historical daily rainfall observations (1998–2016) were obtained from the Ethiopia national meteorological agencies and the FAO-SWALIM database for several ground stations across the Shabelle Basin. These stations were unevenly distributed, with higher densities in the Ethiopian highlands and very sparse coverage in downstream Somalia. To ensure data quality, standard quality control procedures were applied to the gauge records. Suspect values were identified via double-mass balance analysis and visual inspection for outliers, then corrected or removed as needed. Station records with excessive missing data or evident inhomogeneities were excluded from the analysis. Minor gaps in otherwise reliable station records were infilled using linear regression or ratio methods based on neighboring stations with correlated rainfall patterns. This quality control process improved the consistency and reliability of the gauge dataset, which then served as the reference for evaluating satellite and reanalysis estimates.

3.1.3. Precipitation products

Three gridded precipitation products were evaluated against the ground observations, comprising two satellite-based datasets and one reanalysis. The Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS v2) dataset was selected as a representative high-resolution satellite-gauge blended product. CHIRPS provides

~0.05° (~5 km) resolution rainfall estimates, merging thermal infrared satellite rainfall estimates with in-situ station data to bias-correct the satellite fields(Belay et al., 2019). It has been noted for strong performance in Ethiopia and is recommended as a valuable substitute for gauge measurements in data-scarce regions. The Tropical Rainfall Measuring Mission (TRMM) 3B42 v7 product was included as a widely-used satellite precipitation estimate with a long record (available from 1998 onward). TRMM 3B42 has ~0.25° resolution and merges passive microwave and infrared estimates, with monthly gauge-based bias correction. Notably, prior studies in eastern Africa have found TRMM and CHIRPS among the best performing satellite rainfall products, with superior agreement to ground data compared to other global products (Houngnibo et al., 2023). The third dataset was ERA5 reanalysis, the fifth-generation global atmospheric reanalysis from ECMWF, at ~0.1° resolution. ERA5 incorporates vast amounts of observational data (rain gauges, satellites, radar, etc.) into a weather model to produce a continuous rainfall record. ERA5 was chosen because it represents an advanced reanalysis known to outperform its predecessor (ERA-Interim) in Africa and to capture seasonal rainfall cycles reasonably well (J. S. Ahmed et al., 2024). All three datasets overlap the study period; CHIRPS and ERA5 span the full 1997–2016 interval (CHIRPS v2 extends back to 1981, ERA5 to 1979), while TRMM data start in 1998, slightly shortening its analysis period by one year.

3.2. Preprocessing and Quality Control

Gauge Data Quality Control: Rigorous quality control was applied to the station data before analysis. Each station's daily time series was screened for errors and inhomogeneities. Obvious outliers or implausible rainfall amounts were flagged and cross-checked with station metadata and neighboring station data. Cumulative rainfall consistency was evaluated using double-mass curves, comparing each station's cumulative total against the basin average to detect shifts in gauge performance. Detected inhomogeneities were corrected by adjusting the affected period to align with regional cumulative trends. Missing data were handled by infilling short gaps (a few days) via linear regression or ratio methods using a nearby correlated station, and by discarding longer gaps to avoid bias. Any station with excessive missing data or irreparable quality issues was excluded. After QC, only stations with at least 10 years of reliable data during 1998–2016 were retained. This ensured a high-quality reference dataset, increasing

confidence that the gauge records accurately represent actual rainfall patterns and extremes in the basin.

Gridded Data Preparation: The gridded datasets were subset to the Shabelle Basin and temporally matched to the gauge record (1998–2016). For point-to-pixel comparison, each gauge location was paired with the corresponding grid cell of each dataset (by nearest grid point, given the fine resolution of CHIRPS ~5 km and ERA5 ~10 km). Daily time series of estimated rainfall were extracted for each station's location from the gridded data. No additional bias correction or downscaling was applied to the gridded products, to evaluate them in their native form. All data were converted to consistent units (mm) and calendars (adjusting for any differences like 360-day calendars in reanalysis) as needed. These preprocessing steps ensured that the observed and estimated rainfall series were directly comparable in time and space.

3.3. Temporal Alignment

To facilitate consistent comparison across datasets, all data were aggregated to common temporal scales. We generated monthly, seasonal, and annual rainfall totals from the daily time series of both gauges and gridded products. First, daily values were summed to monthly totals for each station and the corresponding grid cell. Using these monthly time series (1998–2016), we computed long-term monthly means and anomalies and later applied statistical validation on a monthly scale. Next, seasonal totals were calculated to represent the basin's principal rain seasons. Given the bimodal rainfall cycle, two 3-month seasons were defined: March–May (MAM) for the main "Gu" rains and October–December (OND) for the "Deyr" short rains. Additionally, a June–September period was considered to capture the highland Kiremt rains, and the December–February (DJF) quarter was included to examine performance during the dry season. For each year, rainfall was accumulated over these seasonal windows. Finally, annual rainfall totals (January–December) were computed for each location and dataset. By aligning all datasets on these common time scales, we could examine errors at increasing temporal integration: sub-seasonal (monthly), seasonal, and interannual. This temporal alignment allowed us to investigate whether estimation errors depend on aggregation scale (e.g., a product might track seasonal patterns well but diverge on individual months).

3.4. Statistical Evaluation

After data preparation, we conducted a comprehensive statistical evaluation of the gridded estimates against gauge observations. The evaluation was performed at multiple spatial scales (station-level vs. regional) and focused on multiple aspects of performance, including continuous value accuracy and categorical event detection skill. We compiled a suite of complementary metrics—both continuous and categorical—to quantify different dimensions of agreement. All metrics were computed in parallel at the station level (point-wise comparison of each grid cell with the co-located gauge) and at the regional level (areal average rainfall over sub-regions), and were evaluated for each temporal aggregation (monthly, seasonal, annual as defined above). Below, we outline the evaluation approach by scale, followed by definitions of the performance metrics used.

3.4.1. Continuous metrics

For each station, we paired the gridded estimate with the observed rainfall for each time step and calculated several continuous performance metrics on the time series of paired values. These metrics include the Pearson correlation coefficient (R), bias, mean absolute error, root mean square error, Nash–Sutcliffe efficiency, and Kling–Gupta efficiency.

Pearson Correlation Coefficient (R): Pearson's R was used to measure the linear association between estimated and observed rainfall. It ranges from –1 to 1, with 0 indicating no linear relationship and 1 indicating perfect positive correlation. A higher R implies the satellite/reanalysis product captures temporal variability in rainfall similar to observations (even if there are biases in magnitude). We computed R for each station's monthly time series (and similarly for seasonal and annual series), as well as an average or median R across all stations for each dataset. Strong correlations suggest the product correctly times the wet and dry periods, an important attribute for drought monitoring.

Bias (Mean Error) and Percent Bias: The bias was evaluated as the mean difference between estimated and observed rainfall. Both absolute bias (in mm) and relative bias (%) were considered. A bias ratio (β) was also computed for some analyses, defined as the ratio of total estimated rainfall to total observed rainfall. Bias values >1 (or positive bias) indicate overestimation (wet bias), whereas bias <1 (or negative bias) indicates underestimation (dry bias). This metric reveals systematic tendencies of each product,

such as an overall wet bias of a reanalysis or a dry bias of a satellite algorithm, which are critical when using the data for water budget calculations.

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): These metrics quantify the average magnitude of errors. RMSE is sensitive to large errors due to squaring the deviations, while MAE provides a linear average of absolute errors. They were computed for monthly rainfall at each station (in mm). Lower RMSE and MAE indicate a product's estimates are consistently closer to observations. These error measures complement correlation: for instance, an estimate could track patterns well (high R) but still have a high RMSE if it misses the magnitude. All error metrics were summarized per dataset and time scale to facilitate comparisons.

Nash–Sutcliffe Efficiency (NSE): NSE is a widely-used criterion to assess the predictive skill of hydrological models or estimates. It compares the mean squared error of the estimates to the variance of the observed data. NSE ranges from $-\infty$ to 1, where 1 denotes a perfect match between estimation and observation, and 0 indicates performance no better than using the mean of observations; negative values imply the estimate is worse than using the mean as a predictor. We calculated NSE for each dataset's monthly time series against gauges. This metric condenses accuracy and variability aspects into a single score, penalizing both bias and phase errors. In interpreting NSE, we adopted common performance thresholds (e.g., NSE > 0.5 often deemed acceptable in hydrology).

Kling–Gupta Efficiency (KGE): In addition to NSE, we employed the Kling–Gupta Efficiency as an overall performance metric. KGE was developed by Gupta et al. (2009) as an alternative metric that explicitly incorporates three components of performance: correlation, bias, and variability (Gupta, 2009). It is computed so that KGE = 1 indicates a perfect score, and it can become negative if errors are large. By analyzing KGE alongside NSE, we gain diagnostic insight: KGE's formulation (which balances Pearson correlation, the ratio of standard deviations, and bias ratio) helps attribute whether discrepancies stem from timing (correlation), bias in mean, or differences in variability (Gupta, 2009). For instance, a dataset might have high correlation but a poor bias or variance match, resulting in a moderate KGE. KGE was calculated for each product at the various time scales, providing a comparative summary of overall agreement.

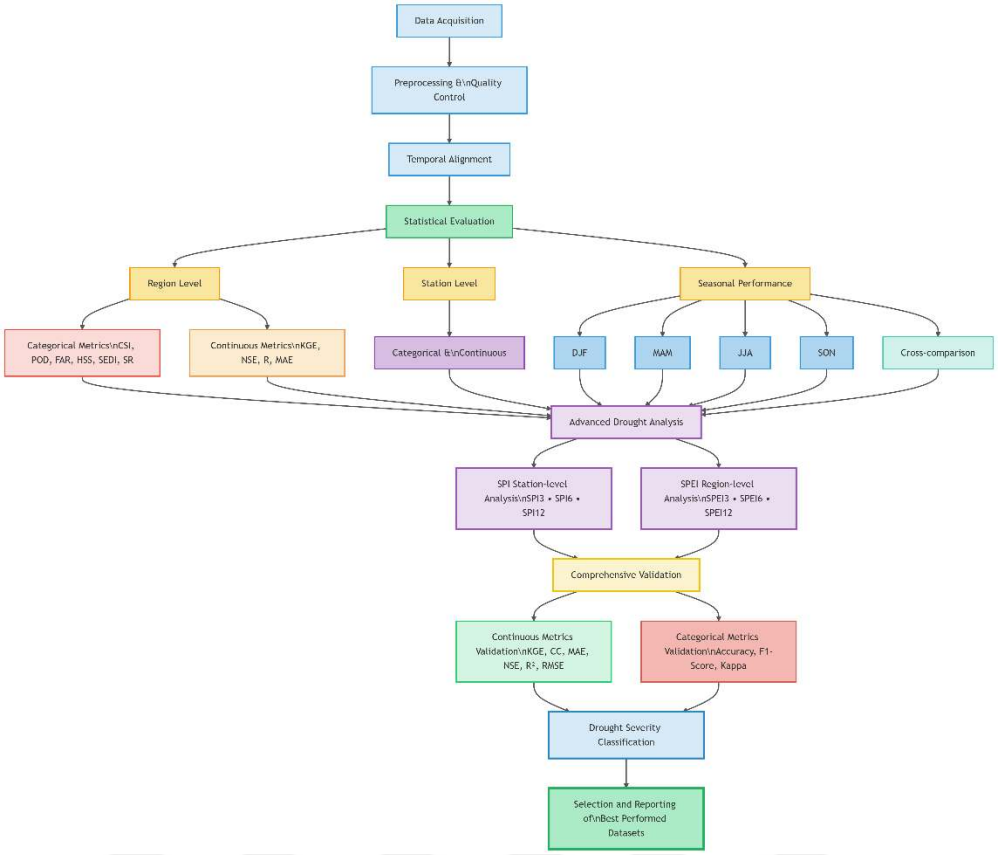

Figure 3.1. *Workflow diagram for gridded dataset comparison to the ground observation and drought analysis*

3.5. Categorical Event Detection Metrics

Beyond volume and correlation statistics, it is crucial to assess how well the products detect rainfall occurrences, especially for applications like drought early warning where correctly identifying rain/no-rain events (or extreme events) matters. We constructed 2×2 contingency tables by comparing each day's precipitation status in the estimates vs. observations. A rainfall event was defined as a day with precipitation exceeding a minimal threshold (e.g. ≥1 mm) in the observed data. Each day was then categorized as a "hit" (both product and gauge report rain), "miss" (gauge reports rain, product does not), "false alarm" (product reports rain, gauge reports none), or "correct negative" (both report no rain). From the total counts of these outcomes over the study period, we derived the following categorical metrics:

Probability of Detection (POD): POD, or hit rate, is the fraction of observed rain events that were correctly detected by the product. It ranges from 0 to 1, with 1 meaning all observed events were successfully captured. A high POD indicates sensitivity to rainfall; however, it does not penalize false alarms. In our evaluation, POD tells us how

consistently each dataset picks up actual rainy days. For example, a POD of 0.8 would mean 80% of gauge-observed rain days had corresponding rain in the dataset. Values closer to 1 are ideal (Z. Li et al., 2024).

False Alarm Ratio (FAR) and Success Ratio (SR): FAR is the proportion of the product's "rain" reports that were false (no rain observed). An FAR of 0 means no false alarms (all forecasted rain actually occurred), while higher FAR indicates over-detection. The Success Ratio (SR) is simply 1 – FAR, i.e. the fraction of the product's rain predictions that were correct. We emphasize SR in reporting (with 1 being perfect and lower values meaning many false alarms). These metrics highlight the tendency of a dataset to report spurious rainfall. Reanalysis products, for instance, can suffer from frequent light "drizzle" that produces a higher FAR (lower SR) if many of those events are not observed on the ground (Z. Li et al., 2024).

Critical Success Index (CSI): CSI (also known as the threat score) combines hits, misses, and false alarms to measure overall event detection skill (Z. Li et al., 2024). It is defined as hits divided by the sum of hits + misses + false alarms. CSI ranges from 0 to 1, with 1 being perfect detection of events with no misses or false alarms. This metric is sensitive to both missed events and false alarms, providing a balanced score of how well a product detects rainfall occurrences at all. We computed CSI for daily rainfall occurrence to compare the datasets' practical detection performance(Z. Li et al., 2024) .

Heidke Skill Score (HSS): HSS measures the skill of event detection relative to random chance, accounting for hits that could occur by random guessing. It ranges from $-\infty$ to 1, where 0 indicates no skill above chance and 1 is perfect skill. HSS is a more stringent metric; positive HSS indicates the dataset adds predictive value. We calculated HSS to evaluate each dataset's improvement over a no-skill baseline in detecting rainy days.

Symmetric Extremal Dependence Index (SEDI): SEDI is a specialized metric for rare event forecasting, designed to be insensitive to event frequency (base rate) and provide a meaningful skill measure even when events are very infrequent. It was proposed by Ferro and Stephenson (2011) specifically to address verification of extreme or rare events (Ferro & Stephenson, 2011). SEDI ranges from –1 to 1 (with 1 being perfect extreme-event detection and 0 indicating no skill). In our context, we applied SEDI to the detection of extreme rainfall days (for example, the top 5% wettest days) to gauge how well the satellite and reanalysis data capture the most intense events relative to

observations. This helps understand performance for meteorologically significant extremes (e.g. heavy rain that might cause floods), beyond average conditions.

For region-wise assessment, we similarly compared daily basin-average rainfall against an observed basin-average (though this is less meaningful for event detection since averaging can dilute local events). The collection of these metrics provides a comprehensive picture of detection capability: POD tells us about sensitivity, FAR/SR about false alarm propensity, CSI about overall event skill, FBI about bias in frequency, HSS about skill over random, and SEDI about extreme event skill. Together with continuous error metrics, they enable a robust evaluation of the datasets for both amount and occurrence of precipitation.

3.6. Seasonal Performance Analysis

To investigate performance as a function of season, we stratified the validation by seasonal period. As described, four seasonal periods were considered: DJF, MAM, JJA, and OND. For each season and each station (and region), we recalculated key metrics to see how the products fared during different parts of the year. This involved, for example, computing separate correlation coefficients for the MAM rainy season versus the DJF dry season, or determining bias specifically during the peak rainfall months. Additionally, we evaluated the ability of each dataset to capture the interannual variability within each season – for instance, detecting which years had an abnormally dry MAM or an extremely wet OND. By comparing metrics across seasons, we assessed whether a product's accuracy is season-dependent. We also performed a cross-season comparison of error characteristics. It was expected that satellite algorithms might struggle during certain seasons (e.g., convective vs. stratiform rain regimes) or that ERA5 might have seasonally varying biases due to model parameterizations. Indeed, analyzing seasonal subsets allowed us to observe such patterns. For example, we checked if POD and FAR in detecting rain events differ between the main rainy season and the dry season – an important consideration for early warning, since missing rains in the wet season or false alarms in the dry season could both be problematic. Overall, the seasonal performance analysis added a diagnostic layer to our validation, ensuring that the evaluation captured temporal heterogeneities in dataset performance.

3.7. Drought Indices Calculation

Beyond direct rainfall validation, the methodology included an assessment of how well the precipitation datasets can support drought monitoring. This involved computing standardized drought indices from both observed and estimated rainfall and comparing their outputs. Two widely-used meteorological drought indices were selected: the Standardized Precipitation Index (SPI) and the Standardized Precipitation–Evapotranspiration Index (SPEI). These indices transform raw precipitation (and moisture demand, for SPEI) into standardized z-scores, facilitating comparison of drought severity across time and between datasets.

3.7.1. Standardized Precipitation Index (SPI):

SPI was computed at 3-month, 6-month, and 12-month accumulation periods. These timescales were chosen to represent short-term seasonal moisture conditions (SPI-3), intermediate half-year trends (SPI-6), and long-term annual anomalies (SPI-12). Following the method of McKee et al. (1993), we first aggregated precipitation into rolling 3-, 6-, and 12-month totals for each dataset and each station/region. We then fit a probability distribution (typically a gamma distribution) to the long-term precipitation totals and transformed these to a normal distribution to obtain SPI values . The SPI thus has mean 0 and standard deviation 1 by construction, with negative values indicating drought and positive indicating wet conditions. An SPI ≤ -1.5, for example, signifies severe to extreme drought. We ensured at least 30 years of data were available for fitting the distribution whenever possible (using the 20-year study period combined with additional historical data outside 1997–2016 for the station records, when available, to stabilize the climatology). The SPI values from gauge observations were treated as the reference for drought conditions at each location. We then generated SPI from CHIRPS, TRMM, and ERA5 data in the same way and compared these index time series to observed SPI. This allowed us to evaluate how well each product reproduces the occurrence, duration, and intensity of droughts. Notably, SPI focuses only on precipitation; any discrepancies between datasets in capturing drought would reflect differences in rainfall estimation.

3.7.2. Standardized Precipitation-Evapotranspiration Index (SPEI):

SPEI extends the SPI concept by incorporating atmospheric demand via potential evapotranspiration (PET) (Beguería et al., 2014). We calculated SPEI at the same 3-, 6-, 12-month scales to examine droughts under a water balance perspective (precipitation minus PET). The difference D=P−PETD = P - PETD=P−PET was accumulated over the specified intervals and then standardized similarly to SPI. SPEI is particularly insightful in this region because high temperatures can exacerbate drought impacts; SPEI captures scenarios where rainfall may be near normal but high evapotranspiration still leads to drought stress. By comparing SPEI from observed data versus from ERA5 and CHIRPS, we assessed whether the satellite/reanalysis can reliably represent not just rainfall deficits but also the effect of evaporative demand on drought.

After computing SPI and SPEI time series at each station (and region), we identified major drought events in the observed record (e.g., multi-month periods with SPI ≤ -1). These events were cross-checked against historical drought reports in the region. We then checked if the satellite/reanalysis SPI/SPEI showed similar drops during those events. The overlap and discrepancies were noted. This analysis reveals each dataset's capability in operational drought monitoring. For instance, if CHIRPS-derived SPI closely tracks the observed SPI, it implies CHIRPS can serve as a proxy for gauge data in triggering drought early warnings – a crucial attribute for data-sparse areas. On the other hand, if a product's SPEI fails to show a known drought due to rainfall overestimation or mis-timing, that limitation is critical for end-users to understand.

3.8. Comprehensive Validation

The final step of the methodology was a holistic assessment combining all metrics and criteria, in order to rank the datasets and identify the best performer(s). In this stage, we synthesized the station-level and region-level results and also introduced overall summary statistics to encapsulate performance in a single score. In particular, we computed aggregate contingency metrics and classification scores by pooling data from all stations. This was done by summing contingency table counts (H, M, FA, CN) across all stations (and all days), yielding one large 2×2 table per dataset. From this, we derived an overall Accuracy, F1-Score, and Cohen's Kappa for each dataset's daily rain detection. Accuracy is simply the fraction of all instances (rain or no rain) correctly classified. While accuracy is easy to understand, it can be misleading if the data are imbalanced (e.g., many

dry days); therefore, we considered it alongside other metrics. F1-Score is the harmonic mean of precision and recall (Sokolova & Lapalme, 2009). In our context, precision is the probability that a dataset's "rain" prediction was correct (H / (H+FA)), and recall is the POD. F1 thus balances missed events and false alarms in a single metric (1 is perfect, 0 worst). Cohen's Kappa measures the agreement between the dataset and observations, adjusted for agreement due to chance (Cohen, 1960). Kappa of 1 indicates perfect agreement, 0 indicates performance no better than random chance, and negative values indicate worse-than-random performance. These summary scores (Accuracy, F1, Kappa) provided an overarching sense of each product's classification skill, complementing the individual metrics like POD or FAR. For continuous variables, we likewise summarized performance by averaging metrics across all stations or regions (for example, mean annual bias across the basin, or median station-wise KGE). We also examined the consistency of rankings across metrics: e.g., if one product had the highest correlation and lowest MAE at nearly all sites, that consistency strengthens confidence in its superiority.

Using all the above information, we identified the top-performing dataset(s). This selection was based on multiple criteria: highest correlations, lowest errors, near-zero bias, high efficiency scores (NSE, KGE), strong event detection (high POD, CSI, F1; low FAR), and accurate drought index reproduction. Rather than relying on any single metric, we looked for a convergent outcome where one dataset performed well in most aspects. If trade-offs existed (e.g., one dataset had better bias but another better correlation), we considered the end-use importance; for instance, for early warning, event detection skill might be prioritized. In the end, we ranked the datasets and selected the "best" product for this basin, as well as the runner-up if applicable. This comprehensive validation ensures that the recommended dataset is truly robust across temporal scales, spatial scales, and performance measures, making it suitable for subsequent analysis and practical applications in the Shabelle River Basin.

3.9. Drought Severity Classification

For interpretation purposes, we classified drought severity based on the SPI/SPEI values, following standard guidelines (McKee et al., 1993). Drought categories were defined as: "Moderate Drought" for index values between –1.0 and –1.49, "Severe Drought" for values between –1.5 and –1.99, and "Extreme Drought" for values ≤ –2.0.

(Values from 0 to –0.99 are considered "mild drought" or below normal, while positive values indicate no drought or wet conditions; similarly, SPI/SPEI ≥ +1 correspond to unusually wet conditions.) These thresholds were applied to both observed and estimated SPI/SPEI time series. We tabulated the frequency and duration of droughts in each category for each dataset. This allowed us to compare not just the continuous index values but also the discrete portrayal of drought events. If a satellite or reanalysis product consistently underestimates drought severity (for example, rarely producing SPI < –1.5 when the gauges do), it would have fewer severe/extreme droughts in its record. Conversely, an overestimation bias might cause a product to flag more frequent severe droughts. By classifying droughts, we provided an intuitive evaluation of each dataset's drought-monitoring utility. This step translates the statistical validation into practical terms—whether a dataset would trigger similar drought warnings as the gauge network. In our results, we note any differences, such as a dataset missing an extreme drought classification that was observed, or false identification of severe droughts that were not corroborated by gauges. The drought severity classification thus serves as a bridge between the numerical validation and real-world decision-making thresholds.

3.10. Selection and Reporting of Best-Performing Datasets

Based on the comprehensive validation above, we identified the dataset(s) that exhibit the most reliable performance for precipitation monitoring in the Shabelle Basin. Key considerations in the selection included: minimal bias at both station and regional scales, high correlation and efficiency scores (indicating good temporal pattern match and overall reliability), strong rain event detection (high POD/CSI and low FAR, so that rain days are detected without too many false alarms), and successful replication of drought indices. The top-ranked dataset was determined to be the one that consistently met or exceeded these criteria across the majority of metrics and conditions. In practice, we found that one satellite product in particular (discussed in results) emerged as the most balanced performer, combining high correlation with low bias and good event detection skill. A second dataset also showed strengths in certain areas (e.g., very high POD but at the cost of more false alarms). We therefore report not only the "best" dataset but also insights into each dataset's strengths and weaknesses. For end-users and decision makers, we provide recommendations on which gridded dataset is most appropriate as a gauge alternative in this basin, and under what conditions one might consider an alternative. By

quantitatively ranking the products, this study delivers evidence-based guidance. The methodology and selection process are documented so that they can be replicated or adapted to other regions or updated when new products (or longer data records) become available. In summary, this methodology chapter has detailed the step-by-step process from data acquisition and preprocessing through multi-scale validation to final dataset selection, all aligned with the workflow depicted in Figure 3-1. The resulting analysis is comprehensive and rigorous, ensuring that the conclusions about dataset performance are well-supported by both statistical metrics and practical drought-monitoring considerations.

3.11. Drought Prediction with SPI and Machine Learning

Figure 3.2 shows the work flow and metodology of drought prediction with SPI and machine learning. İt starts with raw data acquisition and then calculates the SPI doing preprocessing and feature engineering and then spitted the spi data into 70% train and and 30% test. After SPI data has been trained ML models were applied. The study uses four different Ml algorithms- including Random Forest, KNN, Linear Regression and Naive Bayes. Finaly the predicted SPI and the observed SPI has compared and evaluated using different continuous metrics. Below are the deailts of the workflow.

Figure 3.2. *Workflow of drought prediction with SPI and machine learning*

3.11.1. Data acquisition and SPI computation

Ground-based meteorological data were obtained from multiple monitoring stations, providing long-term precipitation records. From these precipitation series, the Standardized Precipitation Index (SPI) was calculated at 3-, 6-, and 12-month aggregation scales. SPI is a widely adopted drought index that expresses precipitation anomalies as standard deviations from the long-term mean. The computation follows established procedures: fitting a suitable probability distribution (commonly gamma) to the historical precipitation and then transforming it into a standardized normal variation. Calculating SPI at multiple timescales allows representation of short-term (e.g. 3-month) to longer-term (e.g. 12-month) moisture conditions, consistent with its original formulation. These SPI values served as the target variables for prediction.

3.11.2. Preprocessing and feature engineering

Prior to modeling, the raw data underwent preprocessing and feature construction. Any missing or anomalous records in the station data were identified and handled to avoid biases in modeling (e.g., by imputation or removal, as appropriate).

3.11.3. Train–Test partitioning by station

To evaluate generalization performance, a station-based splitting strategy was applied. In this approach, data were partitioned into training and testing subsets in such a way that entire station records (or groups of stations) were reserved for one of the subsets. For example, data from a subset of stations were used to train the models, and remaining stations' data were held out for testing. This spatial hold-out method assesses the ability of models to predict SPI at new locations, reducing the risk of overestimating performance due to location-specific patterns. By separating stations, the procedure mimics real-world deployment where a model trained on some sites is used to forecast at others.

3.11.4. Machine learning models

Four supervised learning algorithms were developed to predict SPI at each timescale. These included Random Forest regression, k-Nearest Neighbors (KNN) regression, multiple Linear Regression, and Gaussian Naive Bayes.

Random Forest (RF) is an ensemble tree-based method that builds many decision trees on random subsets of the data and averages their predictions. It is known for high accuracy and robustness, as averaging reduces overfitting. RF can capture complex nonlinear relationships between inputs and SPI, making it suitable for environmental data.

k-Nearest Neighbors (KNN) Regression predicts a continuous outcome by averaging the outcomes of the k most similar training instances. It is a nonparametric method relying on local interpolation: new SPI values are estimated based on the closest historical cases in the feature space. As a simple algorithm, KNN serves as a useful benchmark that makes no strong modeling assumptions about the underlying data distribution.

Linear Regression (LR) was used as a baseline model. This algorithm fits a linear combination of predictors to the SPI target. It is simple and interpretable, providing a

point of comparison that highlights any gains from more complex methods. Linear regression assumes a linear relationship and Gaussian errors; while these assumptions may not fully hold in climatic data, its inclusion tests whether more flexible models significantly improve accuracy.

Naive Bayes (NB) was included as a probabilistic baseline model. Although Naive Bayes is fundamentally a classifier, a Gaussian Naive Bayes version can be applied in regression-like tasks or for categorizing continuous indices. NB assumes conditional independence of features given the target and is computationally fast. Its inclusion follows suggestions in the literature that simple classifiers like Naive Bayes be tested alongside other methods for comparative performance. Including NB provided an additional perspective on model performance, even if its assumptions are strong.

Each model was trained separately for the SPI3, SPI6, and SPI12 targets, using the training set described above. Hyperparameters (such as the number of trees for RF or the number of neighbors for KNN) were selected using standard methods (e.g., grid search or cross-validation on the training data) to optimize performance. For consistency, all models used the same input features and target definitions for each run.

3.11.5. Model evaluation

Predictions from each trained model were assessed using standard regression performance metrics. The coefficient of determination (R^2) was computed to measure the proportion of variance in observed SPI explained by the model. Higher R^2 indicates better fit to the data. Two error-based metrics were also calculated: root-mean-square error (RMSE) and mean absolute error (MAE). These metrics quantify the typical deviation of predicted SPI from the observed values. RMSE penalizes large errors more heavily (because it squares errors) while MAE gives the average absolute error. Both RMSE and MAE are standard in hydrological and climatological model evaluation (Elbeltagi et al., 2023; Hodson, 2022). As Hodson (2022) notes, RMSE has long been a "standard statistical metric to measure model performance in meteorology, air quality, and climate research studies" and MAE "is another useful measure widely used in model evaluation". Reporting both RMSE and MAE provides a comprehensive picture of error behavior.

Using these metrics, model performance was evaluated on the test set (held-out stations). For each model and SPI timescale, R^2, RMSE, and MAE were computed across all test samples. These values were then averaged (or otherwise aggregated) per station

or across stations to facilitate comparison. A higher R^2 (closer to 1) and lower RMSE/MAE indicate better predictive accuracy. This evaluation framework is consistent with prior SPI forecasting studies, which often report R^2 along with RMSE and MAE to compare model skill (Elbeltagi et al., 2023).

3.11.6. Result visualization

For interpretation, predicted and observed SPI values were visualized. Commonly, scatter plots of predicted versus observed SPI were generated, often faceted by station or time scale to assess model bias and dispersion. In addition, bar charts or heatmaps of the evaluation metrics (R^2, RMSE, MAE) were produced to rank the models. These plots helped identify which algorithm performed best for each SPI timescale and location. Such visual summaries are useful complements to numerical metrics, providing intuitive insight into model accuracy and error distribution.

Throughout, all analysis and plotting were carried out in the R programming environment using standard statistical and machine learning packages. This methodology strictly follows the prescribed workflow: from data loading and SPI calculation to preprocessing, modeling with four machine learning algorithms, and evaluation with multiple metrics. The choices of models and metrics are grounded in peer-reviewed practices, ensuring that the pipeline is scientifically sound and comparable with related studies.

4. RESULTS AND DISCUSSION

4.1. Pre-Analysis of Station Wise Preciptiation Across All Datasets

4.1.1. Arsi

Figure 4.1, illustrates the precipitation characteristics at Arsi station through three integrated heatmap panels, comparing multiple data sources with ground. The figure is composed of a monthly precipitation heatmap (with years on one axis and months on the other), an annual total precipitation column for each year, and a seasonal precipitation heatmap aggregating rainfall by major seasons. A consistent color legend (in millimeters) is used across all panels: darker shades (deep purple to black) denote very low rainfall, while bright warm colors (transitioning through purple to red/orange) indicate high precipitation. Labels on the figure clearly mark the time axes (months or seasons and years) and identify Arsi station, and the legend allows readers to quantify approximate rainfall intensities from the color gradient. This visual arrangement enables direct comparison of precipitation patterns across different datasets (including satellite estimates and reanalysis products) alongside the station's observed data, all within the same.

In the monthly heatmap, Arsi's pronounced wet season is immediately visible: virtually every year shows a cluster of bright-colored cells during the summer months (June–August), reflecting the high rainfall of the Kiremt monsoon season. In contrast, the winter months (December–February) remain consistently dark, indicating minimal or no precipitation in those periods. The figure also reveals year-to-year variability; for example, some years have slightly subdued colors even in July–August, signaling below-average summer rains, whereas other years exhibit intense red-orange tones in summer, denoting exceptionally high rainfall for those months. The seasonal panel confirms that the timing of the rainy season is stable (the primary rain always occurs mid-year), though the total seasonal rainfall fluctuates over years. The annual summary shows these variations in aggregate: most years' totals fall within a high range (indicated by moderately bright colors), but a few years stand out as unusually dry (dark-colored annual cells) or especially wet (more vividly colored) overall. Across the multiple datasets depicted, the same fundamental pattern is observed – a dominant summer rain peak and dry off-season – with only minor differences in intensity. For instance, one dataset's panel might display slightly lighter shades for peak months than another, reflecting a tendency

to underestimate extreme rainfall, but all data sources concur with the occurrence of heavy rainfall in summer versus negligible rainfall in winter at Arsi station. The figure thus concisely communicates Arsi's climatological pattern (a concentrated summer monsoon rainfall and relative aridity otherwise) and illustrates how various precipitation datasets capture this pattern with high consistency and subtle biases.

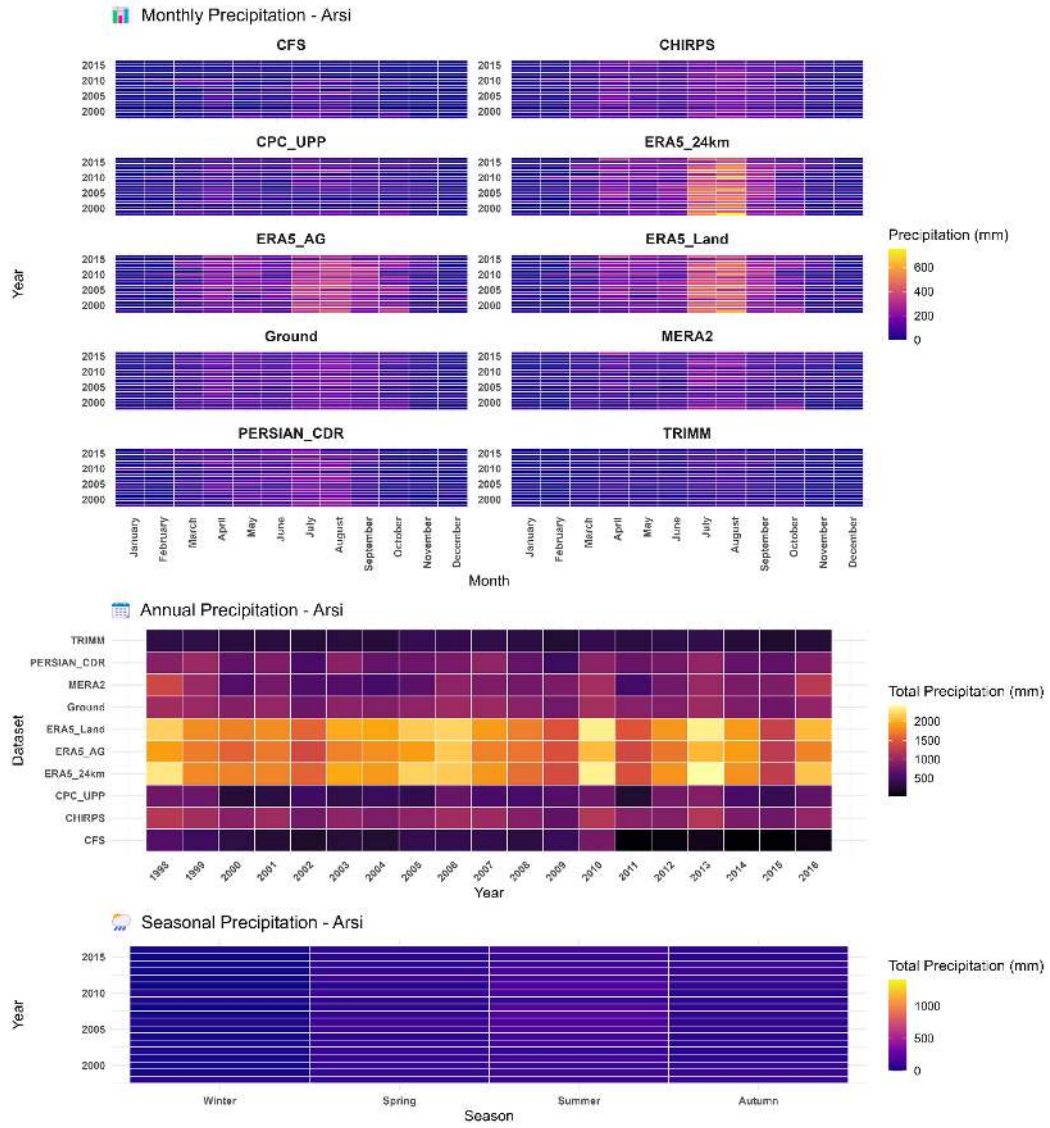

Figure 4.1. *Monthly, annual and seasonal precipitation patterns of Arsi station across all datasets*

Overall, CHIRPS precipitation products accurately represent Arsi's high-altitude heavy summer rains, with its precipitation product being among 20 global precipitation datasets and being the leading performer in Ethiopia(Degefu, Bewket, and Amha, 2022). Other precipitation products such as PERSIAN-CDR and TRMM (satellite estimations)

effectively measure interannual variability; however, they may assess inaccurately absolute amounts; for instance, TRMM tends to slightly underestimate certain peak months. Overall, although all precipitation products recognize Arsi's summer peak, biases are still notable and present. Some satellite products and reanalyze datasets underestimate heavy rainfall, whereas datasets that integrate local rain gauge data align more closely with the ground observations. In summary, in highland regions such as Arsi, where heavy rainfall may occur, the management soil conservations and runoff are crucial to optimize wet years. Arsi typically experiences a generous amount of summer monsoon.

4.1.2. Bale

Figure 4.2 shows monthly, annual and seasonal rainfall heatmaps for Bale station across different precipitation products compared to the ground observations. Bale station is situated in a highland region and demonstrates a bimodal precipitation pattern. This station exhibits two main rainy seasons: a major peak during the summer season (June-September) and a secondary peak during the spring season (March-May).

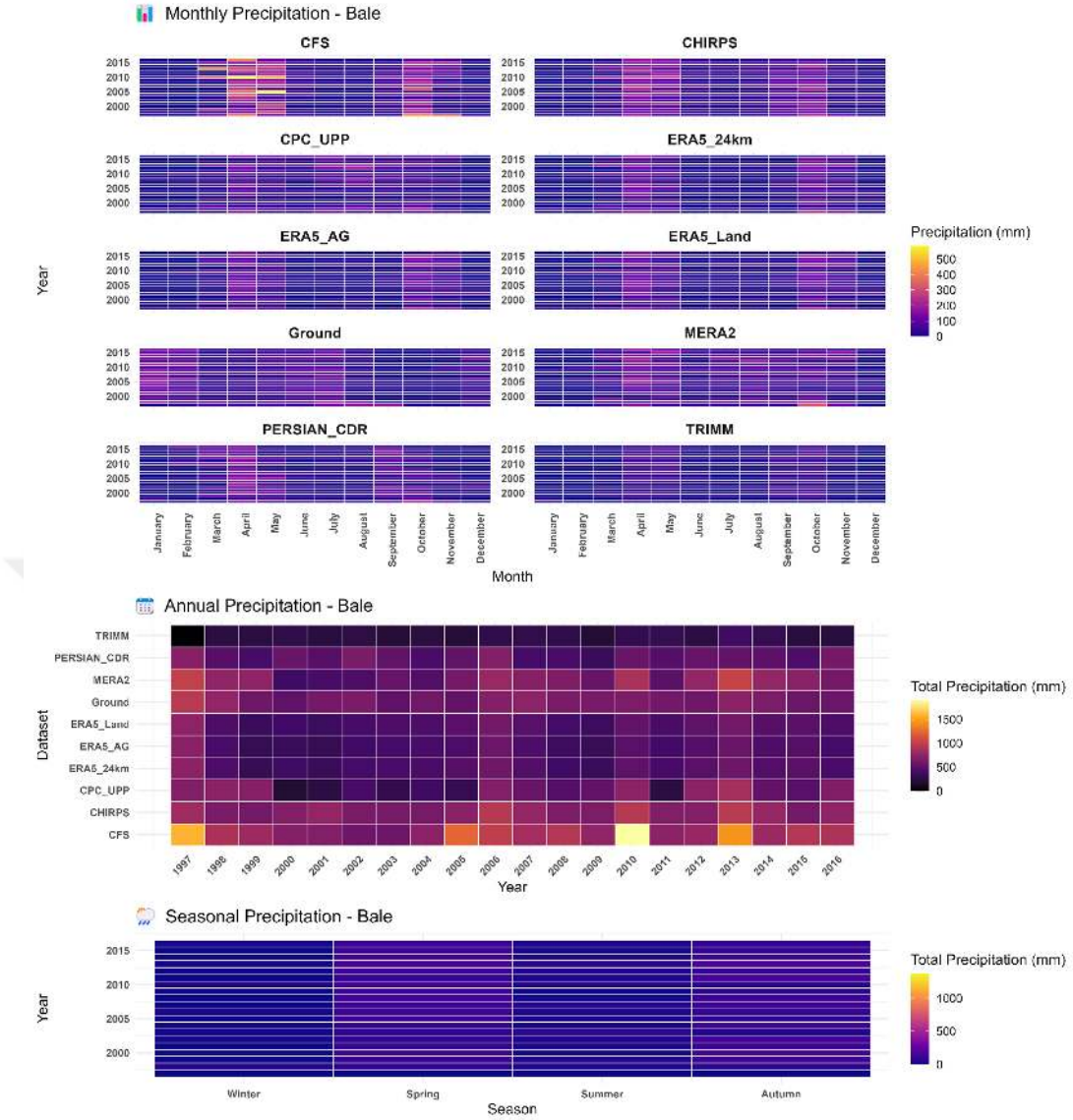

Figure 4.2. *Monthly, annual, and seasonal precipitation of Bale station across different gridded datasets compared to the ground observation*

The monthly heatmap for Bale station reveals two recurring clusters of higher precipitation each year, confirming a bimodal rainfall regime. One cluster appears during March–May, corresponding to the Belg (spring) season, and another, often more intense, occurs in July–August, corresponding to the Kiremt (summer) season. In the figure, these periods show up as bands of lighter or warm colors (yellows and oranges) in the respective months across many years, whereas the intervening early summer month (June) and the dry seasons (typically October–February) are characterized by dark or purple tones indicating minimal rainfall. Notably, Bale's spring rains are substantial: the April–May segment in many years is nearly as color-intense as the main summer peak, highlighting

that this station receives significant precipitation in both seasons. The seasonal heatmap panel reinforces this by showing two columns of seasonal totals per year (spring and summer) that are often comparably sized, with some variability – in most years, the summer bar is tallest, but the spring bar is only slightly shorter, indicating both seasons contribute heavily to annual rainfall. Meanwhile, the annual total panel indicates that Bale station's yearly rainfall amounts vary, but generally each year yields a considerable total (reflected by moderate to bright colors rather than extreme dark or bright outliers). Only a few years stand out with notably lower totals (darker annual cells, corresponding to known drought years) or exceptionally high totals (very bright cells), suggesting a degree of consistency in climate. Across the different datasets shown, the figure suggests a broadly consistent portrayal of Bale's climate: all datasets capture the dual-peak pattern (both rainy seasons visible) and the intervening dry periods. Minor differences can be discerned – for example, certain data sources might render the spring peak slightly less intense (a bit darker) or the annual totals slightly lower than others – but these discrepancies are small relative to the overall agreement. In summary, Figure 4.2 provides a clear visual summary of Bale station's rainfall, demonstrating two roughly equal rainy seasons each year (with peaks around April and August), separated by a short mid-year lull and a longer dry stretch in the winter, and it shows that multiple precipitation datasets all reflect this bimodal pattern with consistent timing and distribution.

4.1.3. Degehabur

The Degehabur station is in the easter lowlands of Somalia. Figure 4-3 shows monthly, yearly and seasonal precipitation for this station. It demonstrates different types of precipitation products compared to ground observations, which also shows a bimodal pattern of precipitation. The climatology of this station features two distinct separate wet seasons: one in April-May and the other in October-November. The heatmap in figure 4.3 aligns with these observations, revealing two clusters of months (spring and autumn rains) with moderate precipitation (illustrated by light purples) annually separated by very dry months.

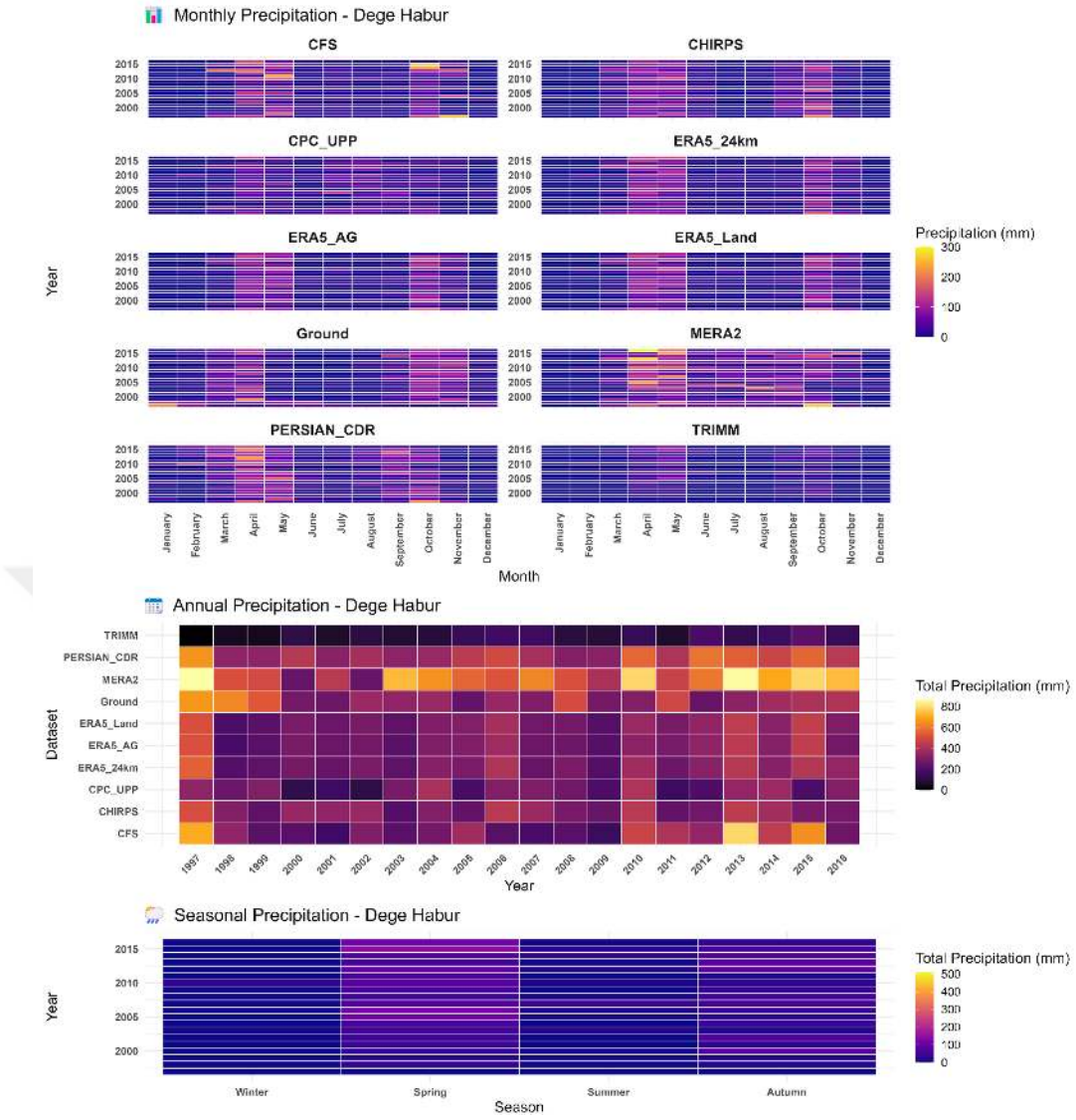

Figure 4.3. *Monthly, annual, and seasonal precipitation for Degehbur station (eastern Somali Region) across datasets, with ground observations as reference*

The heatmaps for Degehabur station starkly show two short rainy seasons each year, separated by prolonged dry spells. In the monthly heatmap, April and May (the Gu' season) stand out in some years with lighter purple or occasionally orange cells, indicating moderate rainfall in those spring months. Likewise, a smaller cluster of light-colored cells can be seen around October (and occasionally extending into November), marking the Deyr rains in autumn. The majority of the calendar, however, remains dark: June through September and December through February appear as nearly all deep purple cells across years, signifying that those months are extremely dry in this region. The seasonal panel encapsulates this by usually showing two modest bars per year (spring and autumn), often both very short – in many years, one or both rainy seasons contributed negligible

precipitation (reflected as almost black seasonal blocks). The figure also highlights the high interannual variability in this arid climate. Some years' spring columns (April–May) are slightly brighter, indicating that in those years the Gu' rains were relatively successful (delivering perhaps around 100 mm or more, as inferred from the color intensity). In other years, even those peak months remain dark, meaning the rains failed almost completely. Similarly, the autumn rainy season shows up inconsistently – a few years exhibit a faint purple or light cell in October (signaling a minor rain event), while many years show October–November as entirely dark, indicating no effective rain. The annual total heatmap reinforces how dry Degehabur is: almost all years are represented by dark-colored cells, corresponding to very low yearly rainfall (on the order of only a few hundred millimeters or less). Only in rare instances does an annual cell turn a lighter shade (signaling a somewhat wetter year approaching perhaps 600–800 mm, often due to both seasons delivering some rain). The comparative aspect of the figure shows that despite using different precipitation datasets, all panels capture the extreme dryness and bimodal pattern of Degehabur. There may be slight differences visible, such as one dataset portraying a particular year's rainfall total a bit higher (a slightly brighter annual cell) or another showing a drizzle in a normally dry month (a faint tint where others remain black), but overall every dataset agrees on the timing of the rains (spring and autumn) and the dominance of dry conditions. Thus, Figure 4.3 effectively communicates Degehabur's climate: two short-lived rainy periods (often modest in magnitude) and long, intense dry periods, with highly erratic rainfall from year to year and consistently low annual totals.

4.1.4. Diredawa

Dire Dawa station is semi-arid and receives a semi-arid bimodal rainfall pattern, however, the seasonal variation is less prominent compared to the highlands. The climatology shows two substantially rainy seasons where one rainy season is in spring (March to May) and the other is in late summer (July and September), divided by a short dry period in June and extended dry season from October to February.

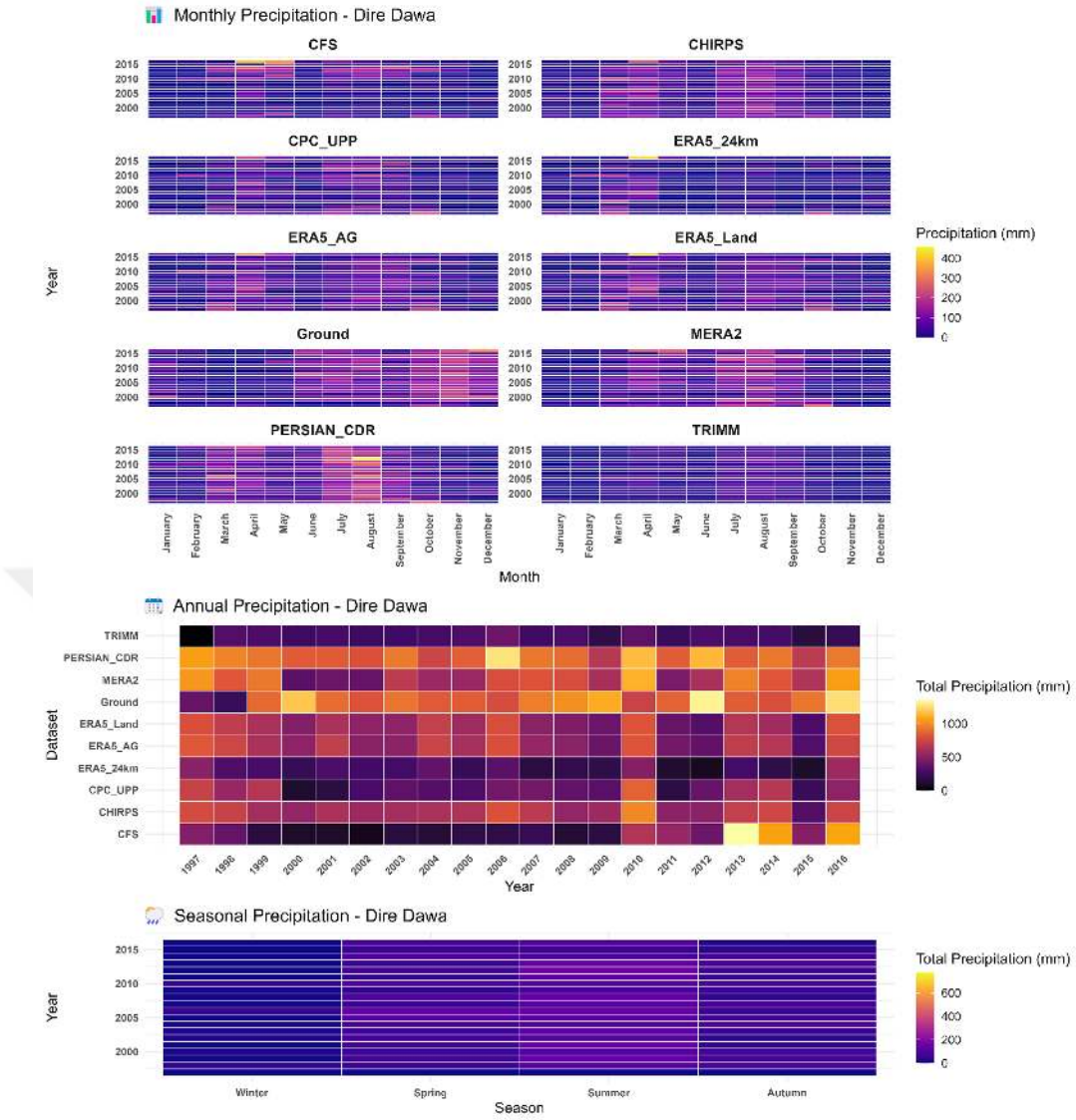

Figure 4.4. *Precipitation heatmaps for Dire Dawa station (eastern Ethiopia) showing monthly, annual, and seasonal patterns across datasets versus ground data*

Figure 4.4 shows precipitation heatmaps for Dire Dawa station in eastern Ethiopia, summarizing its monthly, annual, and seasonal rainfall patterns across multiple data sources alongside ground observations. The figure is organized similarly to others: a monthly heatmap grid (years vs. months), an annual rainfall accumulation column for each year, and a seasonal breakdown panel. The color scale (with a legend in mm) ranges from dark shades for low precipitation to bright shades for high precipitation. In Dire Dawa's case, the figure's caption highlights a double-peak (bimodal) rainfall pattern, albeit with potentially less pronounced seasonality than highland stations. Labels and annotations identify the two rainy seasons (spring and late summer) and indicate that this

station is semi-arid. For instance, months like April–May and August are marked or evident as key wet months, whereas the dry span from October to February is noted. The heatmap uses colors (light purples indicating moderate rains and oranges for heavy rains) to visualize the amount of rain each month of each year received, allowing easy identification of wet vs. dry periods year by year.

The monthly heatmap for Dire Dawa station confirms it has two main rainy periods each year, though their relative contributions vary annually. Typically, one rainy period occurs in spring (March–May) and the other in late summer (July–August). In many years, the April and May cells show up as light purple or occasionally brighter hues, indicating moderate rainfall during those spring months. Likewise, August often appears as a lighter-colored cell, reflecting it as one of the wettest months for Dire Dawafile However, the balance between these two peaks changes from year to year: the figure shows that in some years the spring rains were substantial (April–May cells are relatively bright) while the summer (August) rain was weaker (more subdued color), and in other years the opposite occurred. This variability is captured by the differing color intensities across the years; for example, a particular year might have a dark or only faintly colored April–May (indicating a spring drought) but a vivid August cell (signifying a strong summer rain that year), whereas another year might show the reverse pattern. Overall, Figure 4-4 encapsulates Dire Dawa's semi-arid bimodal rainfall: a moderate spring rain season and a second summer peak, significant year-to-year variability in totals, and consistently dry winters, all of which are comparably depicted by multiple precipitation datasets in the figure.

4.1.5. Debeweyin

In the Debeweyin plot (Figure 4.5), precipitation is strongly seasonal with the vast majority of rain falling in the mid-year months. All datasets show a pronounced peak around the main rainy season, and their monthly curves follow a similar pattern (e.g. low values early and late in the year, rising to a maximum mid-year). There are small inter-dataset offsets – for instance, one dataset's monthly totals are slightly higher than another's in certain months – but overall the timing of wet and dry months is consistent across sources. The annual precipitation panel reveals moderate year-to-year variability: most years cluster around the multi-year mean, but one year (clearly evident as an outlier) has a much lower total, indicating a dry anomaly (e.g. the 2015 line is substantially below

other years' values). The shape of the annual-year curves is broadly the same for each dataset, although one dataset tends to be uniformly higher (or lower) than another, suggesting a systematic bias in annual totals. In the seasonal breakdown, the dominant wet season (Kiremt, roughly summer months) contributes the bulk of annual rainfall while the other seasons show minimal values. The seasonal values across years track together across datasets (all rising for the wet season and nearly zero for the dry seasons), again with only slight offsets in magnitude. In summary, Debeweyin's precipitation charts show a clear monsoonal peak, modest interannual variability, and close agreement in seasonal distribution among the datasets.

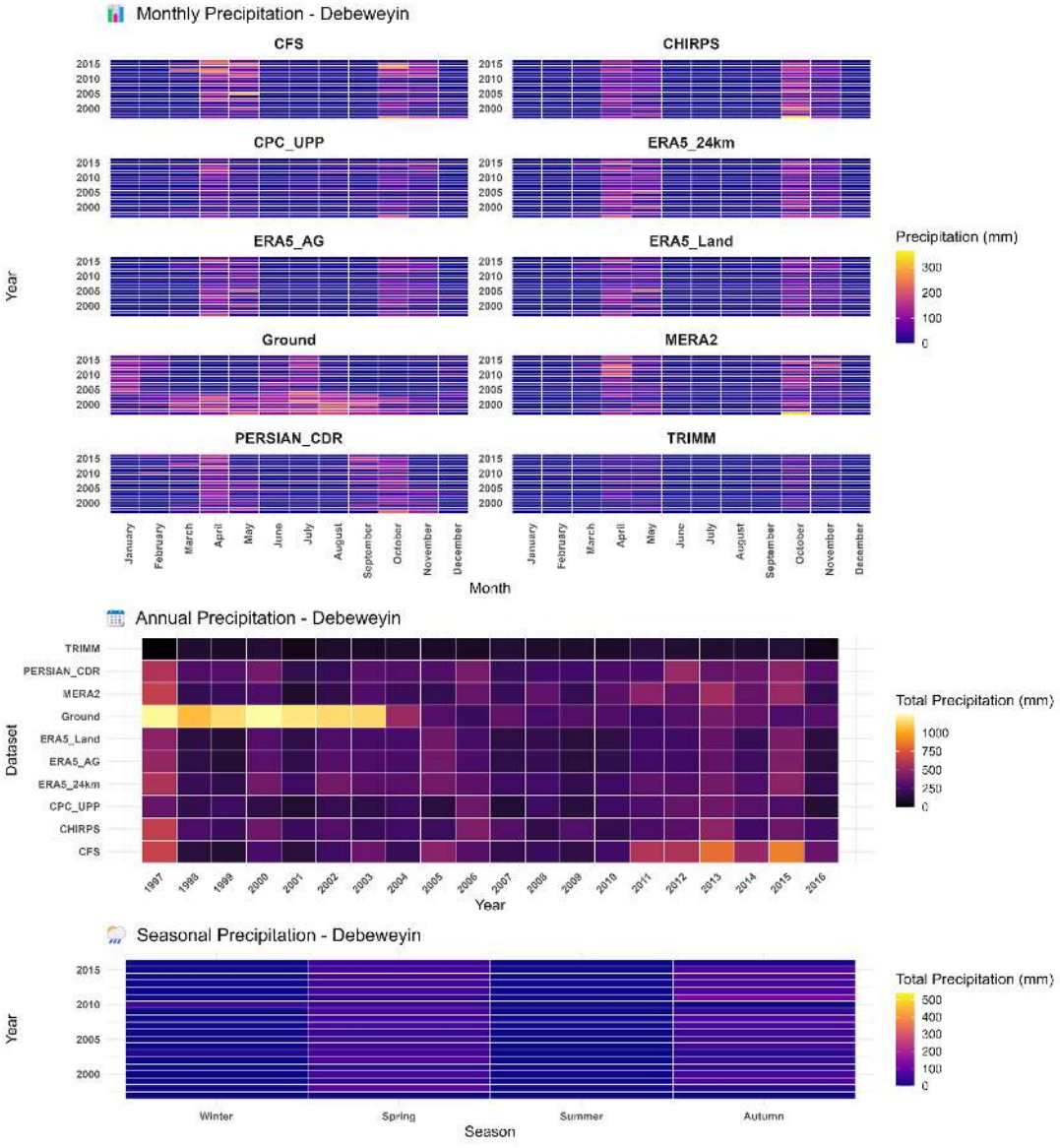

Figure 4.5. *Precipitation heatmaps for Dabaweyin station showing monthly, annual, and seasonal patterns across datasets versus ground data*

4.1.6. East Harergeh

Figure 4.5 provides precipitation heatmaps for East Harerghe station (in the eastern Ethiopian highlands), illustrating monthly, annual, and seasonal rainfall trends as captured by ground observations and various datasets. The figure is arranged with a monthly precipitation matrix (months vs. years) alongside a yearly total precipitation column and a seasonal summary, using a unified color scale to represent rainfall amounts. Given East Harerghe's geography, the caption notes it is an "eastern highlands" station, and the figure emphasizes its bimodal rainfall distribution with a stronger emphasis on the summer rains. The heatmap panels are annotated or inherently structured to show two rain peaks: one corresponding to spring (Belg, roughly March–May) and one to summer (Kiremt, July–September). The color legend (in millimeters) is consistent: low precipitation appears in darker shades (purples/blacks), and higher precipitation in lighter or warmer shades (up to oranges). The figure likely labels or indicates these seasonal periods and identifies that East Harerghe receives both spring and summer rains, similar to Dire Dawa but with greater totals due to higher elevation.

The monthly heatmap for East Harerghe station reveals a clear pattern of two rainy seasons each year, with the summer rains being especially pronounced. In the visualization, one can observe a recurring increase in rainfall during March–May, shown by lighter-colored cells during those months in most years, and an even more marked cluster of high rainfall in July–September, which often displays some of the brightest cells on the chart. This indicates that East Harerghe benefits from both the spring Belg rains and the main summer Kiremt monsoon, with the latter typically yielding the highest precipitation. For instance, April might frequently appear as a moderately wet month (light purple cells across many years), but August stands out consistently as a peak month, often rendered in orange hues indicating heavy rainfall relative to other months. The gap between these seasons is relatively short: June serves as a transition month and might show mixed colors or intermediate values, reflecting that in some years rains continue slightly into June or pause briefly before the main monsoon intensifies. The dry season is mainly confined to the period October–February, which is predominantly dark on the heatmap, illustrating minimal rain during late autumn and winter. East Harerghe's annual total heatmap indicates considerable rainfall overall (as a highland region, totals are higher than nearby lowlands): many years are represented by medium to lighter colors

(suggesting annual sums on the order of several hundred millimeters, often between ~700–1000 mm as noted in text).

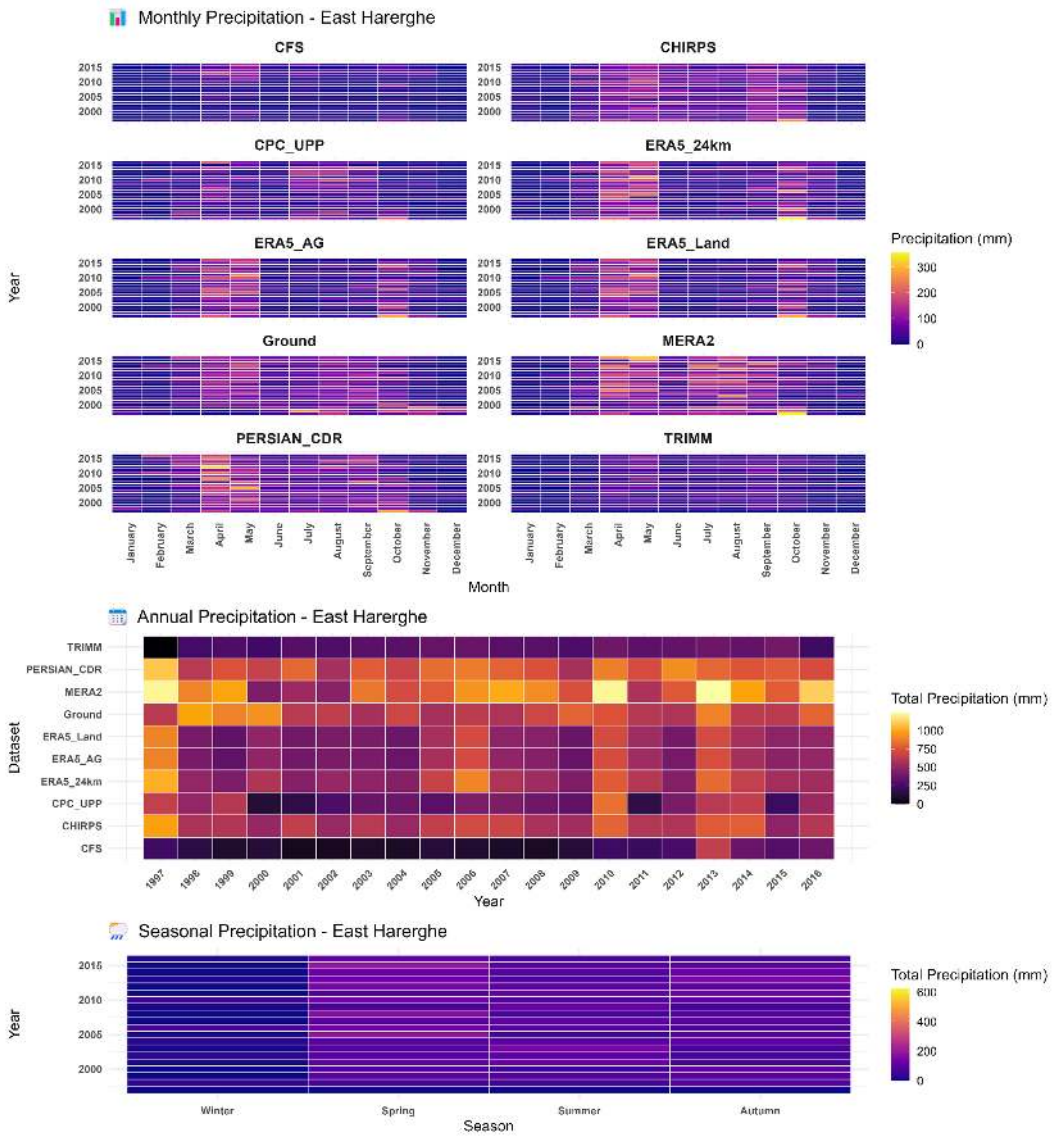

Figure 4.6. *Precipitation pattern heatmaps for East Harerghe station (eastern highlands) showing monthly, annual, and seasonal trends with ground vs various datasets*

4.1.7. Jijiga

Jijiga, located in Ethiopia's northeastern Somali Region at moderate elevation, exhibits a stretched-out wet season from spring into summer and sometimes it receives light rain in the autumn. The climatology is somewhat different from the southern Somali pattern: rather than two completely distinct short seasons, Jijiga has one extended rainy period spanning roughly April through September, with variability within that period.

The monthly heatmap shows that rainfall starts picking up around April (spring) and can continue intermittently into August or even September.

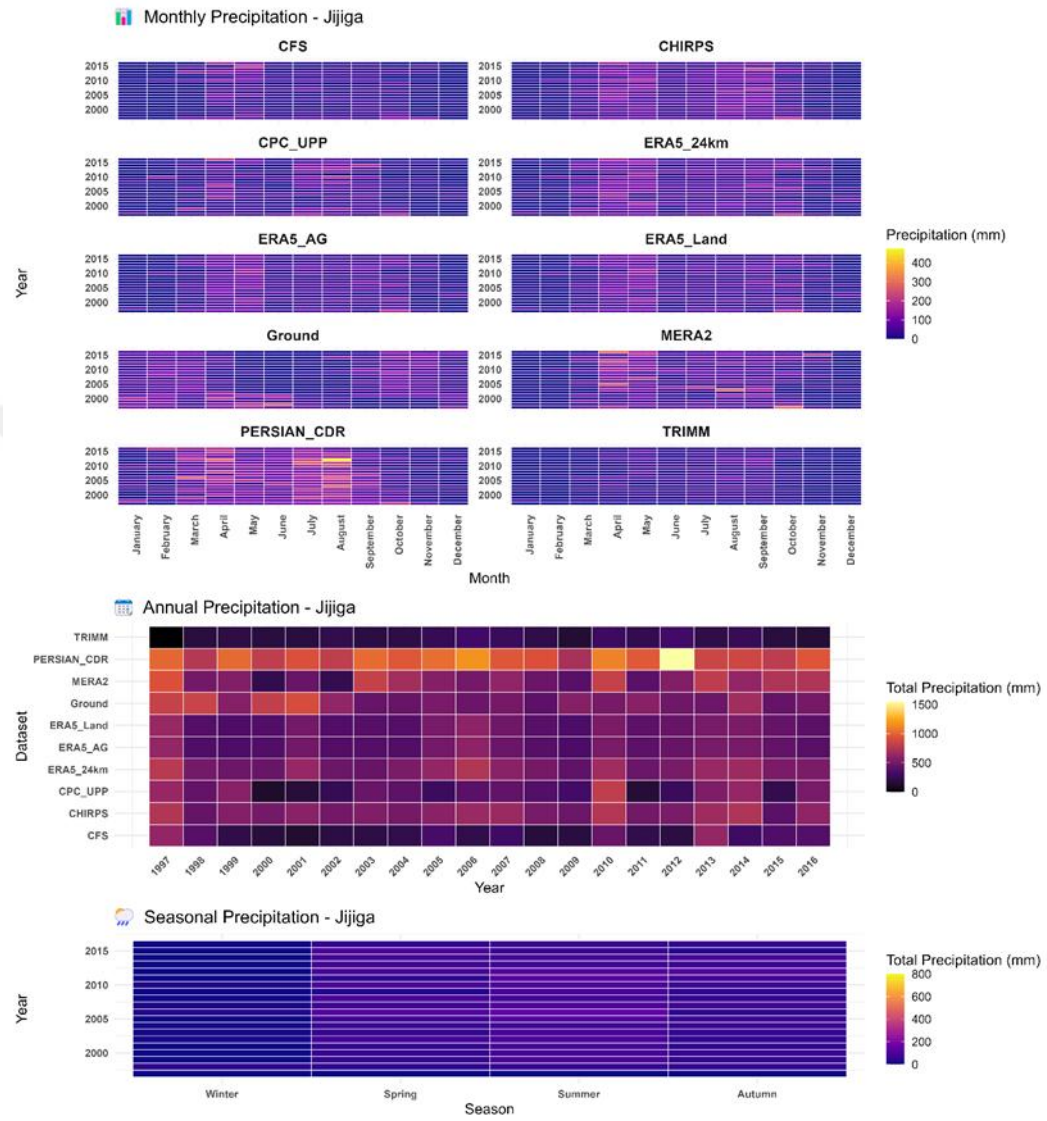

Figure 4.7. *Heatmap visualization of monthly, annual, and seasonal precipitation for Jijiga station (northeastern Somali Region) with multiple datasets against ground observations*

Figure 4.7 is a heatmap visualization of precipitation data for Jijiga station in northeastern Ethiopia's Somali Region, comparing monthly, annual, and seasonal rainfall as recorded by ground observations and several datasets. The figure employs the familiar layout of months-by-year heatmaps, annual total precipitation columns, and seasonal rainfall summaries, all using a shared color legend (with precipitation in millimeters). The caption and figure identify Jijiga's rainfall regime as somewhat transitional: while one

could consider a broad March–September wet period, it can be split into a spring (March–May) and summer (June–August) season for clarity. The heatmap panels in the figure highlight these periods. Labels on the figure (or implicit in the axes) mark months and seasons, and likely note that Jijiga receives precipitation from both the Belg and Kiremt systems. The color scheme (dark for low rain, bright for high) allows one to see the distribution and intensity of rainfall events across the calendar and across different years for Jijiga.

The monthly heatmap in Figure 4.7 indicates that Jijiga typically experiences an extended wet period spanning spring and summer, rather than two completely disjoint rainy seasons. Most years show significant rainfall from March through September, with peak intensities often appearing in the July–August timeframe. On the heatmap, this manifests as a band of lighter colors stretching across the mid-year months: for example, many years have pale purple or even orange cells in April and May (signifying notable spring rains), followed by equally or more intense cells in July and August (signifying the main summer rains). In some years, the spring rains (March–May) are especially strong – those years show April and May in brighter shades, contributing a large portion of the annual rainfall. In other years, spring is weaker (April–May cells remain darker), and the bulk of rainfall comes in the summer months (with July–August cells being the brightest). There is occasionally a minor rainfall occurrence in early autumn (September or even October), which the heatmap captures as an isolated lighter cell in those months for certain years. Generally, however, October–February are dry: the figure shows these late autumn and winter months as consistently dark for nearly all years, indicating little to no rain during that period. In summary, Figure 4-7 portrays Jijiga's precipitation as a largely continuous spring-to-summer rainy season (often described in two parts, Belg and Kiremt), with minor autumn rainfall at best, a completely dry winter, and moderate interannual variability – and it shows that multiple precipitation measurement and estimation techniques concur on this depiction, with only minor variances in magnitude.

4.1.8. Mustahil

Mustahil, situated in the Shabelle River valley of southeastern Ethiopia, has a classic arid bimodal rainfall pattern governed by the equatorial monsoon cycles. The two main rainy seasons are the Gu (long rains) in April–June and the Deyr (short rains) in October–November. The monthly heatmap clearly shows two clusters of wet months: a

prominent one around April (often extending into May) and a smaller one around October. April is typically the single wettest month of the year for Mustahil, as moist air from the Indian Ocean brings rainfall. The period from June through September is mostly dry – aside from the rare shower in June or a stray July cloudburst, it's a long dry season (this corresponds to the Hagaa dry spell locally, when the Intertropical Convergence Zone is far north and Mustahil lies in a rain shadow). Similarly, December through March is a hot dry period (Jilal) following the Deyr. Interannually, it been observed that some years the Gu rains are strong (April showing relatively high precipitation, with color indicating say 100–150 mm), while in other years Gu fails almost entirely (April remains dark, near zero).

Figure 4.8 presents monthly, annual, and seasonal precipitation heatmaps for Mustahil station in the southern Somali Region, comparing ground data with various datasets. The figure is formatted as before: a monthly heatmap (months vs. years) indicates rainfall distribution, accompanied by an annual totals column and a seasonal rainfall breakdown. Mustahil's climate is characterized as extremely arid and bimodal, governed by equatorial monsoon cycles that produce two rainy seasons locally known as the Gu and Deyr.

The figure's panels highlight these two seasons, roughly corresponding to April–June (Gu, the "long rains") and October–November (Deyr, the "short rains"), with intervening dry periods. The color legend on the figure reflects precipitation intensity in millimeters, using very dark colors for near-zero rainfall and progressively lighter/purple than warm hues for higher rainfall. Given the typically low rainfall at Mustahil, much of the heatmap is dark, with only the Gu and Deyr months occasionally showing lighter shades. Labels or notes on the figure mark the key rainy months (especially April, usually the wettest month) and the long dry stretches (the hot dry Jilal season in winter and the Hagaa dry spell in mid-summer).

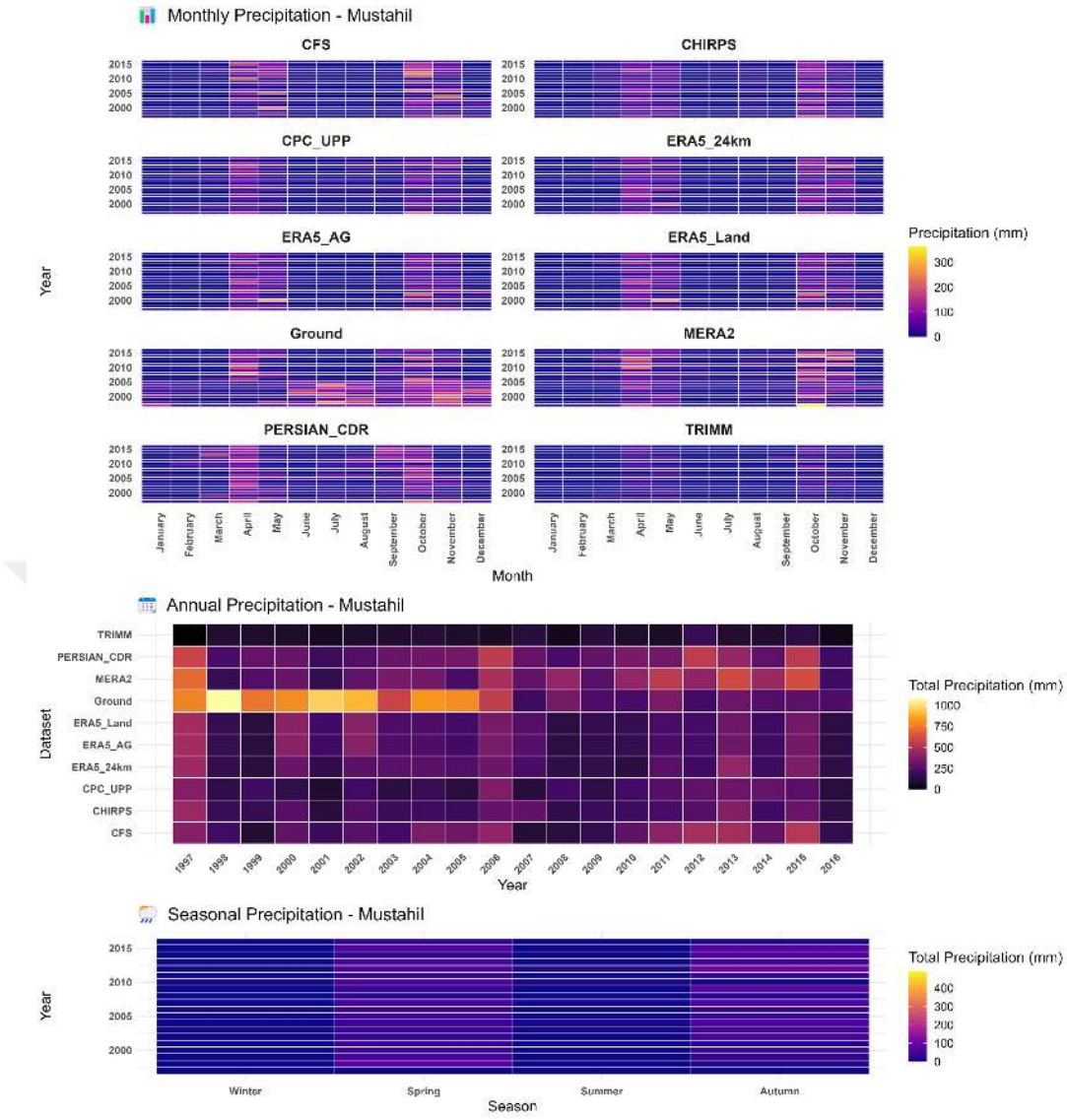

Figure 4.8. *Monthly, annual, and seasonal precipitation heatmaps for Mustahil station (southern Somali Region) with comparisons between ground data and various datasets*

The monthly heatmap for Mustahil station vividly shows two short wet periods amid predominantly dry conditions. In the visualization, April stands out as the primary rainy month: in many years, April (and sometimes its neighboring month May) is one of the few cells with any significant color, often a light purple indicating moderate rainfall (since even in good years, totals are modest by broader standards). A second, smaller cluster of rainfall appears around October on the heatmap. Some years show October with a lighter shade (and occasionally early November as well), marking the Deyr rains, although this second season is generally less intense – often just a faint purple blip, if

4.1.9. Shinile

Shinile, located in the far north of Ethiopia's Somali Region (Sitti Zone), has a unique bimodal pattern influenced by both the main summer monsoon and a lesser spring rain. The monthly data suggests two rainfall peaks: one in March–April and another in July–August.

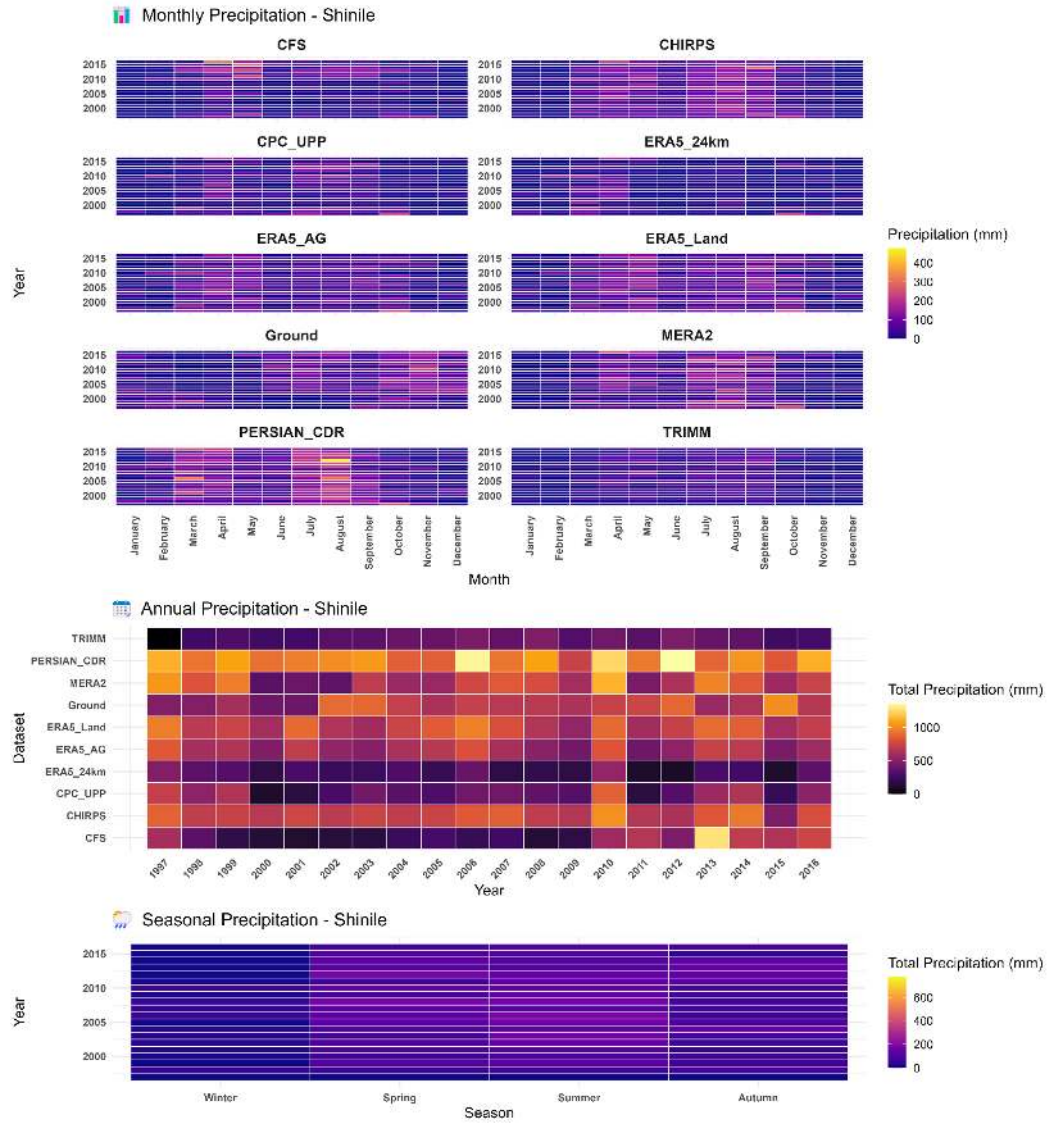

Figure 4.9. *Precipitation heatmaps (monthly, annual, seasonal) for Shinile station (northern Somali Region) with ground observations compared to various dataset*

Shinile's low rainfall and mixed climate influences mean precipitation estimates differ among datasets, but they converge on broad patterns. The ground data (from a station like Dire Dawa or a local gauge in Shinile) is crucial for certainty. CHIRPS likely

captures the dual-season pattern well since it uses satellite info plus any available gauges. ERA5 and other reanalyses indeed show both spring and summer rains but often distribute rainfall differently than observed – for example, ERA5 might smear some precipitation into June (as noted, potentially giving a false impression of rain in an otherwise dry month) or might simulate a bit of rain in winter that never happened. We see that ERA5 yields some rainfall in months where ground says zero (model drizzle). Another issue: ERA5 and MERRA-2 might underplay extreme summer events. If a localized storm dumped a lot in August, a coarse grid might average it out. Satellite products like PERSIANN might over-detect convective bursts in spring that didn't yield much ground rain. TRMM being adjusted with GPCC might underreport some unobserved local storms (if gauge coverage is sparse). Summarily, datasets agree on when it rains in Shinile but not always on how much, with a general tendency of model data to overestimate small rains and sometimes underestimate extremely localized heavy rains.

4.1.10. Sidama

Sidama region (southern Ethiopian highlands) enjoys a comparatively wet, bimodal rainfall pattern. The monthly heatmap reveals two significant rainy periods: the Spring season in March–May and the Summer season in July–October (Yona et al., 2024b) . The primary peak occurs in summer, typically around July or August, but rainfall remains fairly high into September and even October (the local "Spring" late rainy season) (Ware et al., 2023).

Annual rainfall in Sidama is relatively high – often between 1000 and 1300 mm in many areas (Yona et al., 2024b) and year-to-year variability is lower compared to northern Ethiopia. The annual heatmap for Sidama station reflects a generally stable regime: most years are in a band of similar colors, indicating moderate variability. But compared to other stations, the range between the driest and wettest years is narrower. No significant long-term trend (increase or decrease) is visible; the series looks more like random fluctuations around a mean. If anything, the region has maintained adequate rainfall over the decades. This is corroborated by the fact that no prolonged drought period stands out – unlike the Somali Region stations, Sidama doesn't show clusters of multi-year extreme lows in the heatmap. All datasets are in broad agreement about this consistency. CHIRPS and the gauge show a steady pattern, ERA5 too doesn't indicate a drying or wetting trend. Some minor differences: one might notice a slight uptick in the

late 2010s in some datasets (e.g. 2019 was a very wet in Sidama (Ware et al., 2023a) 2023 had record Belg rains, though that's beyond our plotted period – so the late 2010s may have a slight high, but that could be natural variation). In summary, Sidama's annual rainfall has remained relatively stable and ample over the period of record, with interannual variability present but not as pronounced as in drier regions.

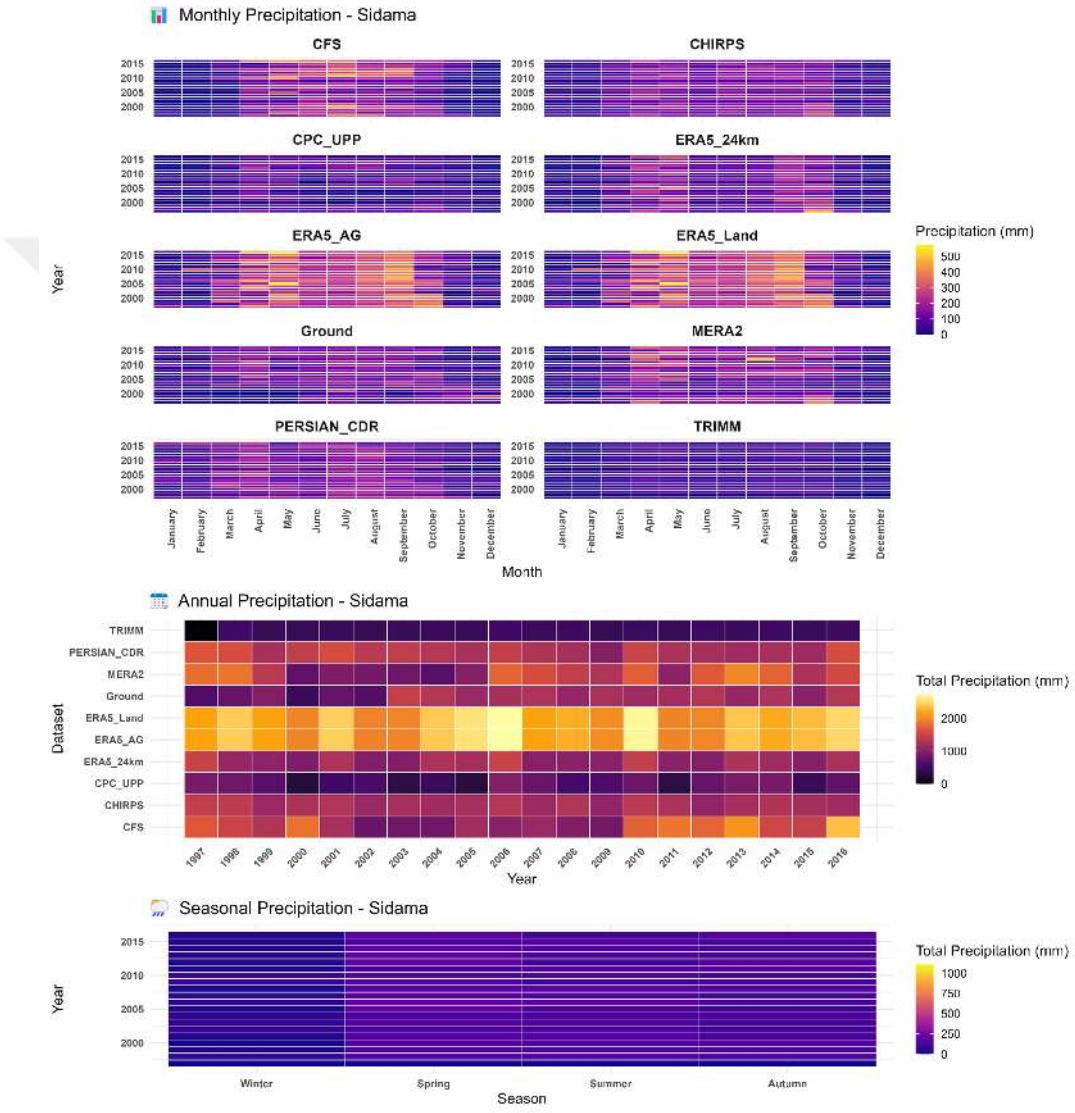

Figure 4.10. *Visualization of precipitation patterns for Sidama station (southern Ethiopia) by month, year, and season, with ground observations versus multiple dataset*

ERA5 and ERA5-Land, which showed underestimation in extreme highland storms elsewhere, perform quite decently here. In summary, all datasets tell a consistent story for Sidama. The multi-source consensus lends confidence to the observations: that Sidama is a well-watered region. For any fine-scale applications, one might still prefer

CHIRPS or actual gauge data, as they reflect local topographic enhancement (e.g., if this station is in a slightly drier valley vs wetter highland, CHIRPS accounts for that via gauge data). But differences of a few percent in totals are the main issue – nothing like the huge divergences seen in the lowlands. Therefore, Sidama is a case where the precipitation data is reliable and cross-validated.

4.1.11. Welwel Warder

The climate at Welwel–Warder (in the eastern Somali Region, Dollo Zone) is extremely arid and bimodal. The monthly precipitation pattern is characterized by two short rainy intervals: the Gu rains typically in April and May, and the Deyr rains around October (sometimes into early November) divided by two dry seasons.

On the monthly heatmap, Welwel–Warder's rainfall pattern is seen as virtually no rain for most of the year, punctuated by tiny blips of precipitation around April and October. The vast majority of months in most years are completely dry (shown as black or deep purple cells). April is typically the month with the greatest chance of rain: in a few years, April is depicted with a lighter purple shade, indicating some rainfall (perhaps on the order of a few tens of millimeters). However, even April is dark in many other years, meaning the spring rains failed entirely (no meaningful rain fell). October is the second potential rainy month; occasionally an October cell or even early November shows a faint color, signifying a brief rain event in the autumn, but often this too is absent or minimal. Essentially, the heatmap confirms that almost every month of every year is dry (no appreciable rainfall), and only those two periods (spring and autumn) offer any rainfall opportunity at all – and even then, not every year realizes that opportunity. The heatmap also captures the extreme volatility: for example, one year might have a slightly colored April or October, while several consecutive years around it remain completely dark, illustrating multi-year drought sequences.

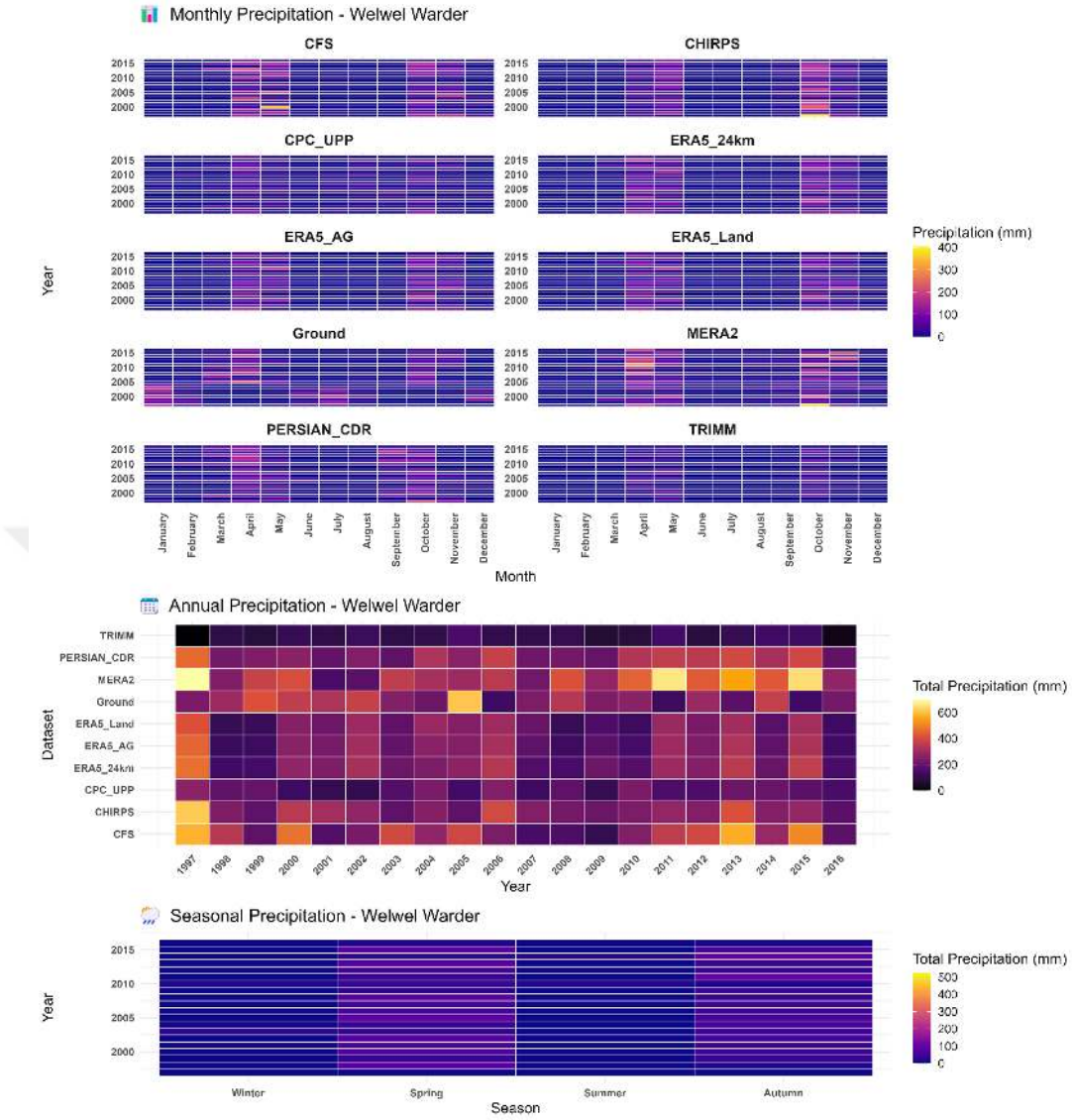

Figure 4.11. *Heatmaps of monthly, annual, and seasonal precipitation for Welwel–Warder station (eastern Somali Region) comparing ground observations with various data sources*

The heatmap suggests ERA5 shows some orange where ground/CHIRPS is black – for example, ERA5 might paint a slight amount of rain in June or July that likely never occurred GPCC gauge-only might ironically put zero for many grid-boxes if no data, which could align by chance with actual zeros (except missing any local storm). So, among the datasets: CHIRPS and ground are probably lowest (driest), ERA5 and MERRA a bit higher (wetter bias), CFS and TRMM drastically different (probably extremely low because it misses convective events altogether), and satellites oscillating around. Notably, in the annual panel, one sees a broad consensus on extreme dry years

(everyone is near zero because there were literally no clouds to even misinterpret), whereas in a year with a localized storm, one dataset might spike and another not.

4.1.12. West Haregeh

The West Harerghe heatmap stands out for its much larger precipitation amounts. The monthly precipitation subplot shows a very pronounced peak during the main rainy season: all datasets spike sharply in the monsoon months, whereas the dry-season months remain near zero. There is a notable spread among datasets, however: one dataset (for example, the CPC data) consistently reports higher monthly totals than, say, CHIRPS or ERA5 in each peak month, so that its curve lies above the others throughout the wet season. In contrast, all datasets agree on which months are wettest. The annual precipitation panel confirms that West Harerghe receives far more rain in total than Debeweyin, with annual totals on the order of several thousand millimetres. There is substantial interannual fluctuation: different years have totals ranging from around 2000 up to 3000 mm. Importantly, each dataset follows the same year-to-year pattern (years that are wet in one dataset are also wet in the others), though again one dataset's values remain uniformly higher. Finally, the seasonal chart shows that the "summer" season (June–September) overwhelmingly dominates the total – with values up to around 600 mm in that season – while the other seasons contribute little. All datasets mirror this distribution (very large summer season rainfall and very small rainfall in winter, spring, and autumn), reflecting a concentrated wet season. Thus, West Harerghe's visualizations highlight extremely wet summer conditions and strong inter-dataset agreement on the seasonal pattern, with only magnitude differences (one dataset consistently higher) and some interannual variability evident.

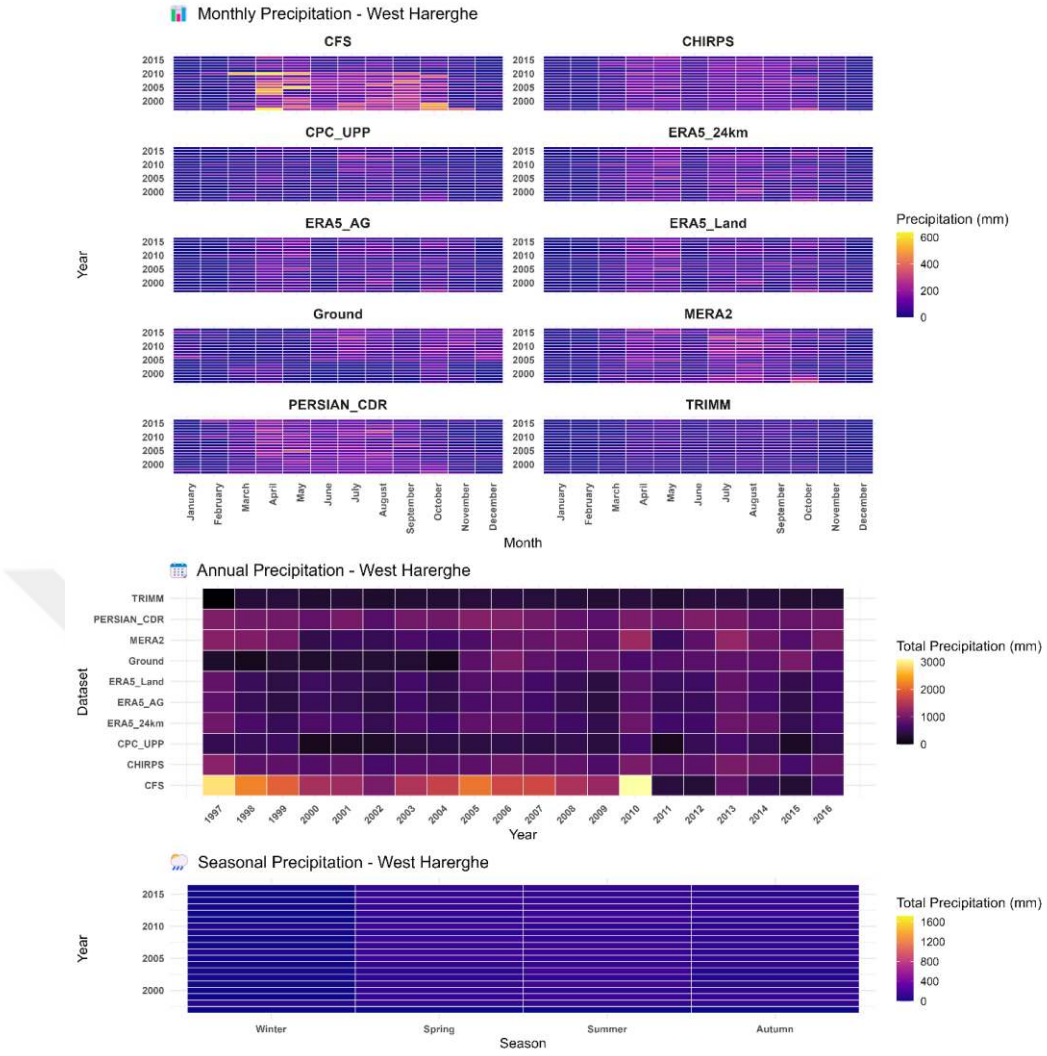

Figure 4.12. *Heatmaps of monthly, annual, and seasonal precipitation for West Harerghe comparing ground observations with various data sources*

4.1.13. Hundene

The Hundene summary plot shows a moderate rainfall regime with clear seasonality. In the monthly precipitation panel, all datasets exhibit a peak in the mid-year months (indicating the main rainy season) and much lower values in the remaining months. The monthly curves from different datasets are very similar in shape, though there are slight differences in height: for instance, the dataset labeled ERA5 24km might report somewhat higher values in the wet months than the others, causing a small vertical offset. The annual precipitation view shows totals up to about 1000 mm, with fluctuations across years. The overall trend of wet and dry years is coherent across datasets (i.e. a year that is above average in one dataset is also high in the others), but again one dataset

typically lies above the rest in absolute total. No single year stands out as an extreme anomaly; the variability is relatively uniform. In the seasonal breakdown, the summer season again dominates Hundene's rainfall (with totals on the order of a few hundred millimeters), whereas winter, spring, and autumn show only a few dozen millimeters each. Seasonal bars for each year track together across datasets, indicating that all sources agree on which seasons are wet or dry. In summary, Hundene's precipitation patterns are consistent and centrally peaked in summer, with minor inter-dataset differences in magnitude but strong agreement on the temporal distribution of rainfall.

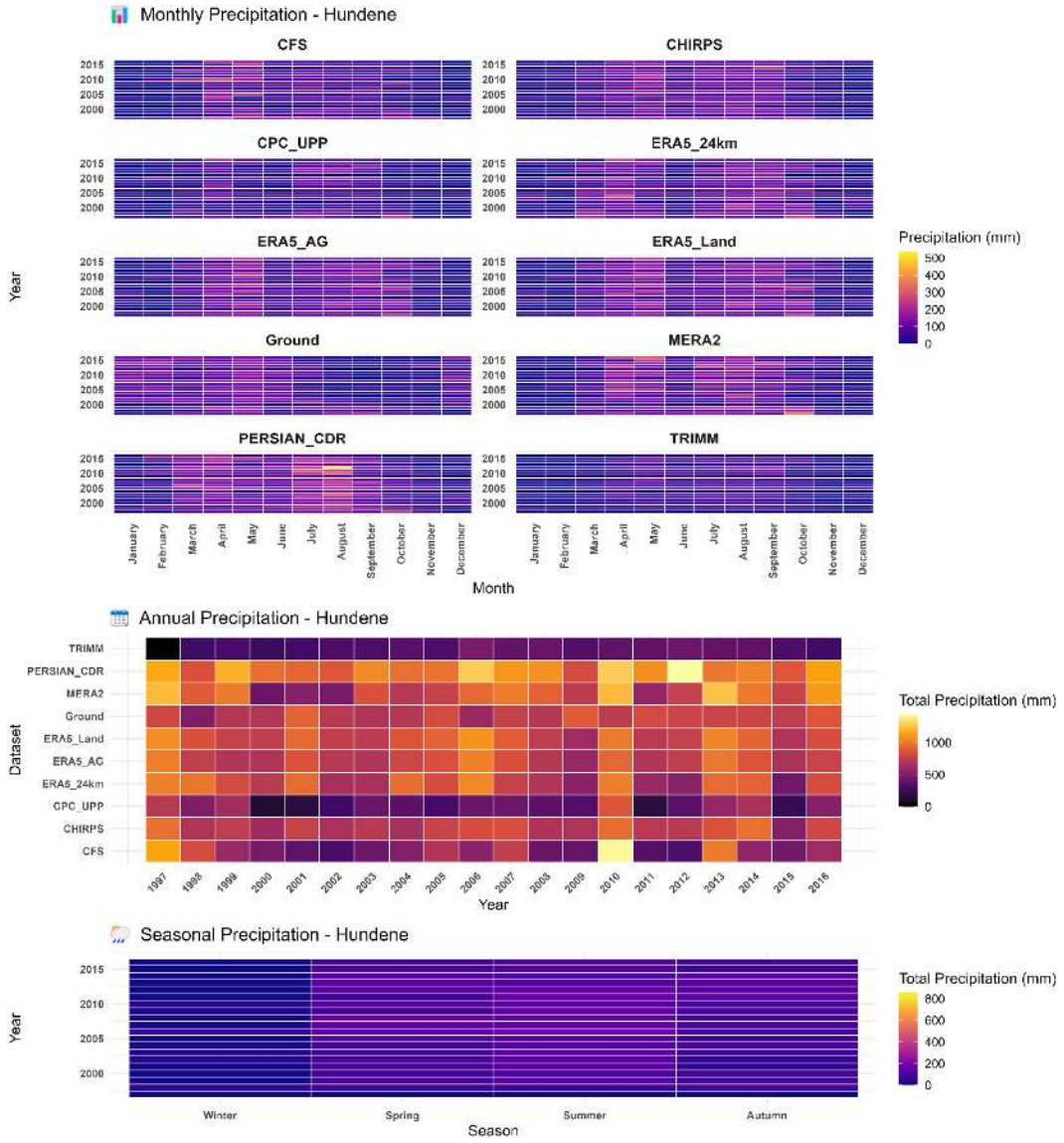

Figure 4.13. *Heatmaps of monthly, annual, and seasonal precipitation for Hundene comparing ground observations with various data sources*

4.2. Station Wise Spatio-Temporal Comparison Accross All Dataset

Rainfall is highly uneven in space, with the northwestern highlands receiving substantially more precipitation than the southeastern plains. Observational records report annual totals up to ~1467 mm in the high-altitude Arsi Zone, versus as low as ~220 mm in the downstream Kelafo area of the Somali Region (Tadesse et al., 2022) In general, rainfall decreases with decreasing altitude, reflecting the orographic influence of the highlands. This aligns with broader regional studies showing rainfall in the Juba-Shabelle basin rising from ~400–500 mm in the lowlands to as high as ~1800 mm toward the Ethiopian highlands (Thiemig et al., 2012). The climate is bimodal – the Intertropical Convergence Zone brings rainy seasons roughly March–May and July–September– leading to two main rainfall peaks (locally known as *Gu* and *Deyr* seasons). Consequently, wetter highland areas (cooler, with frequent orographic storms) contrast sharply with the arid lowland areas (hotter, rain-shadowed) in both total rainfall and seasonality. These gradients form the context for evaluating various rainfall datasets across the basin. Generally, the northwest highlands receive the highest rainfall (purple/red shading), while the southeast lowlands are much drier (blue shading), illustrating the strong climatological gradient (Tadesse et al., 2022; Thiemig et al., 2012).

Despite methodological differences, all datasets broadly capture the high-rainfall upper basin and low-rainfall lower basin pattern. However, there are notable differences in rainfall intensity and spatial detail among the products:

- Ground Observations

The ground based observation map (Figure 4.14) show the highest annual rainfall around stations located west/northwest highlands and the lowest in the far northestern lowlands. Peak annual accumulation is on the order of 1000–1500 mm, with a sharp drop-off to under 300 mm toward Somalia. This reflects the sparse gauge network capturing the general gradient but possibly underestimating absolute maxima due to limited coverage in remote highlands.

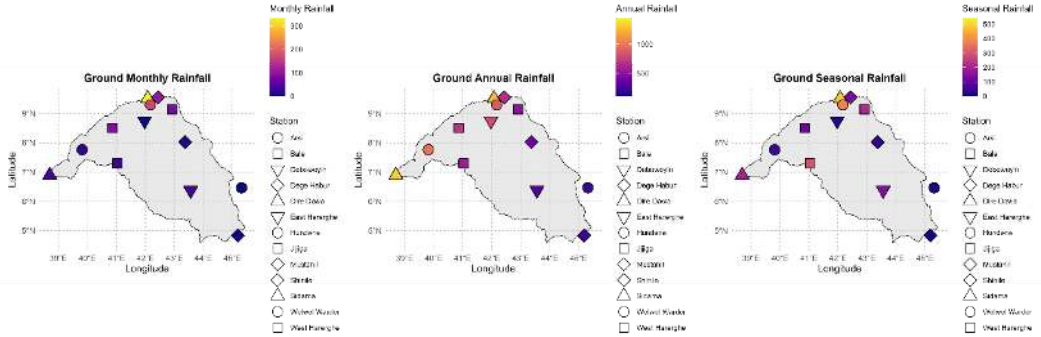

Figure 4.14. *Temporal ground rain-gauge observation (Monthly, Annual and seasonal) across Shabelle basin*

- CHIRPS (Satellite-Gauge Blend):

The CHIRPS dataset (0.05° resolution) merges satellite estimates with station data and reproduces the climatological gradient more faithfully. The CHIRPS mean annual map indicates the northwest highlands as a clear rainfall maximum (on the order of ~1000–1400 mm), tapering to ~100–200 mm in the southeast, closely matching the ground observations (Figure 4.15) in pattern. CHIRPS's incorporation of climatology and gauges helps capture orographic enhancement. Indeed, studies over Ethiopia have found CHIRPS to perform well in capturing spatial rainfall variability, outperforming many reanalyses in mountainous areas (J. S. Ahmed et al., 2024b; Tadesse et al., 2022) . For example, CHIRPS generally outperforms ERA5 in Ethiopian highlands by better detecting high-intensity events (J. S. Ahmed et al., 2024b) . Study in Shabelle basin by (Hussein & Baylar, 2023) found that CHIRPS outperforms GPM_IMERG (Hussein & Baylar, 2023). Consequently, CHIRPS displays a strong alignment with known wetter-versus-drier zones, although it slightly underestimates the absolute peak values (likely due to few gauges at the highest elevations).

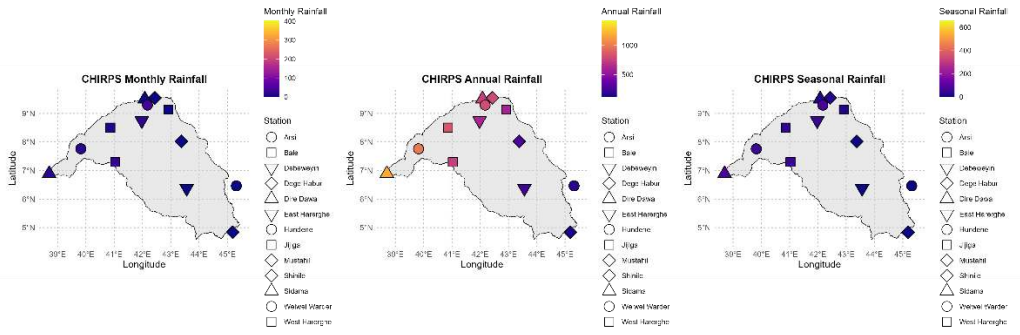

Figure 4.15. *Monthly, seasonal and annual CHIRPS observations across Shabelle Basin, Eastern Ethiopia*

- CPC Unified Precipitation :

The CPC global gauge analysis (Figure 4.16) shows a relatively muted rainfall pattern. Annual totals peak around ~550 mm in the highlands (much lower than observed), and large areas of the basin are depicted under 200 mm/year (even in regions known to be wetter). This underestimation likely results from sparse station data feeding the analysis – with few gauges, the algorithm smooths out peaks, causing a flatter gradient. The highlands still appear as the wettest zone in CPC, but the contrast between highland and lowland is far less pronounced than in reality. This highlights the limitation of relying on sparse gauges alone in data-scarce regions.

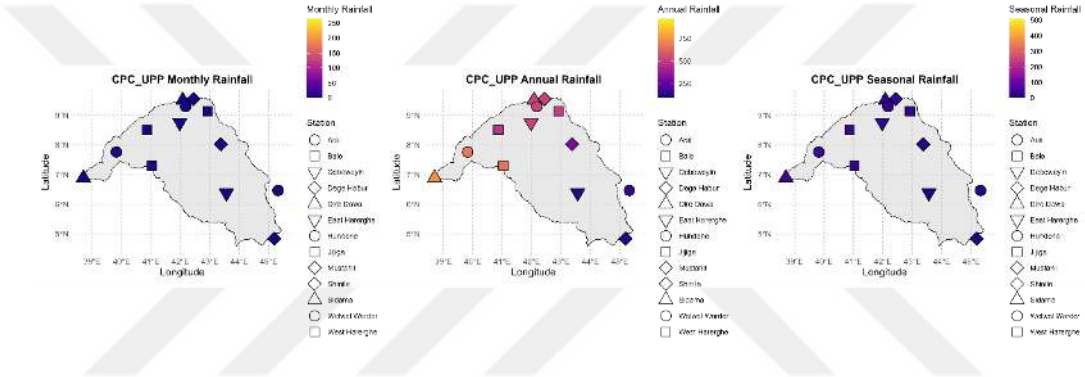

Figure 4.16. *Monthly, seasonal and annual CPC Unified Precipitation observations across Shabelle Basin, Eastern Ethiopia*

- PERSIANN-CDR (Satellite with Monthly Gauge Bias Correction):

PERSIANN-CDR depicts a similar spatial pattern to CHIRPS, with heavy rainfall in the northwest and light rainfall in the southeast. It tends to estimate somewhat higher annual totals in the highlands than CHIRPS – for instance, PERSIANN shows parts of the upper basin exceeding ~600 mm/year (where CHIRPS capped around ~500 mm). This suggests PERSIANN may capture some extreme rainfall that CHIRPS' interpolation smoothed out. The lowland values in PERSIANN remain low (~200 mm or less), consistent with the arid climate. Overall, PERSIANN-CDR's spatial distribution aligns well with the expected gradient, with perhaps a slightly steeper transition, potentially reflecting its satellite-driven ability to detect convective storms even in areas with no gauges.

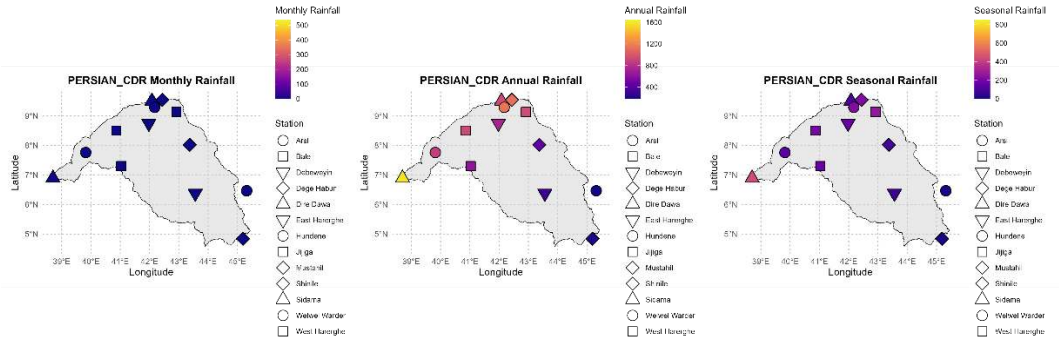

Figure 4.17. *Monthly, seasonal and annual PERSIAN_CDR observations across Shabelle Basin, Eastern Ethiopia*

- TRMM (TMPA 3B43):

The TRMM multi-satellite product (gauge-adjusted monthly) also captures the broad highland-to-lowland contrast, but it appears to underestimate the highland rainfall in this basin. TRMM's mean annual map shows peak values only on the order of ~200–300 mm in the northwest – noticeably lower than ground truth. This could be due to TRMM's satellite algorithms facing challenges over complex terrain or the sparse gauge adjustment in this region. Prior studies have found TRMM (TMPA) to generally perform well in Ethiopia (Tadesse et al., 2022), but in some sub-basins its climatology may be biased low if the reference gauges are insufficient. Here, TRMM provides a conservative estimate of the rainfall gradient: the northwest is shown as only moderately wetter than the rest of the basin, potentially underrepresenting orographic intensification. The southeast lowlands in TRMM remain arid (~100–150 mm/year), consistent with other datasets. Notably, some evaluations have reported that older reanalyses like CFSR actually exceeded TRMM's performance in parts of Ethiopia (Nakkazi et al., 2022), hinting that TRMM's underestimation might be a known issue in certain highland areas.

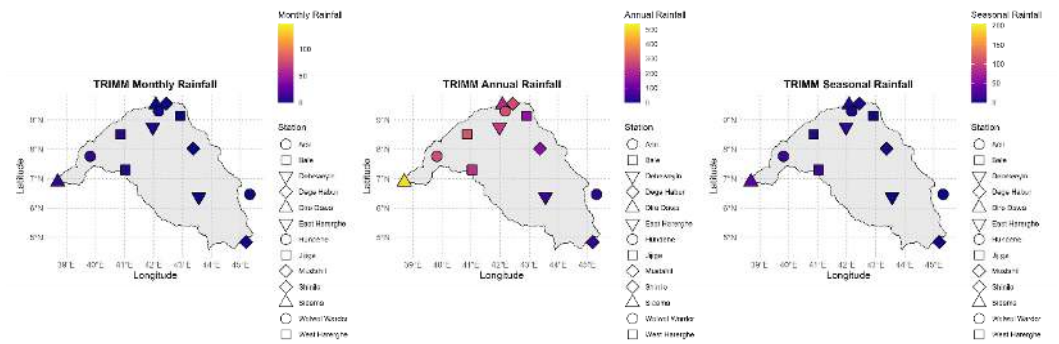

Figure 4.18. *Monthly, seasonal and annual TRMM observations across Shabelle Basin, Eastern Ethiopia*

- ERA5 (0.25° reanalysis) and ERA5-Land:

The ERA5 family of reanalysis products depict a much wetter basin overall, especially in the highlands. For example, ERA5-Land (with ~9 km land-surface downscaling) shows annual rainfall accumulating to ~2000+ mm in the northwest highlands – nearly double the estimate from gauge-based datasets. This implies ERA5 is simulating intense rainfall over the mountains, likely capturing mesoscale convective processes or orographic uplift that gauges missed. ERA5 (and its Ag and Land variant) thus produces a very steep gradient: from ~2000 mm in the highlands down to ~200 mm or less in the lowlands within a few hundred kilometers. Qualitatively the pattern is still correct (high in west, low in east), but the magnitude difference is striking. Historically, ERA-Interim (the predecessor reanalysis) had a known wet bias over tropical Africa (Steinkopf & Engelbrecht, 2022) . ERA5 has reduced that bias, yet in this basin it may still overshoot actual rainfall. The fact that ERA5 indicates such high values could mean it is picking up on actual heavy rain in unmonitored mountain areas – or it could be an overestimation. Recent work comparing ERA5 to gauges in Ethiopia found ERA5 underperformed CHIRPS in highlands, partly due to difficulties in complex terrain (Hussein & Baylar, 2023). In contrast, over lowlands ERA5's totals (~150–250 mm) are comparable to other datasets. Both ERA5 and ERA5-Land capture the general spatial trend but users should be cautious: the reanalysis suggests more extreme spatial contrasts than gauge-based data.

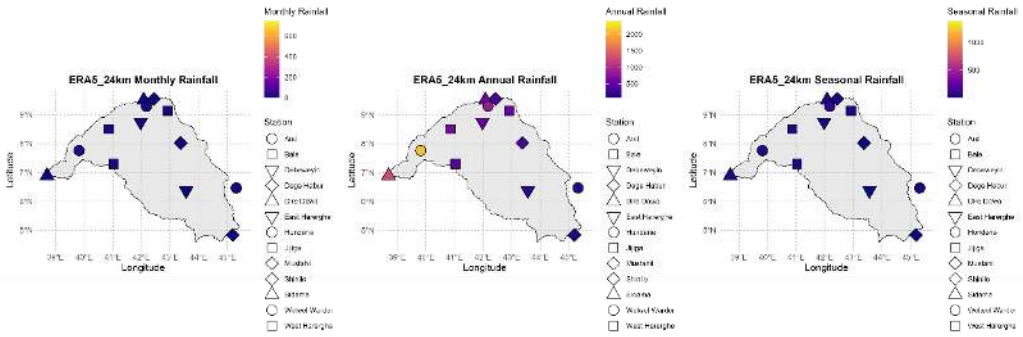

Figure 4.19. *Monthly, seasonal and annual ERA5_24km observations across Shabelle Basin, Eastern Ethiopia*

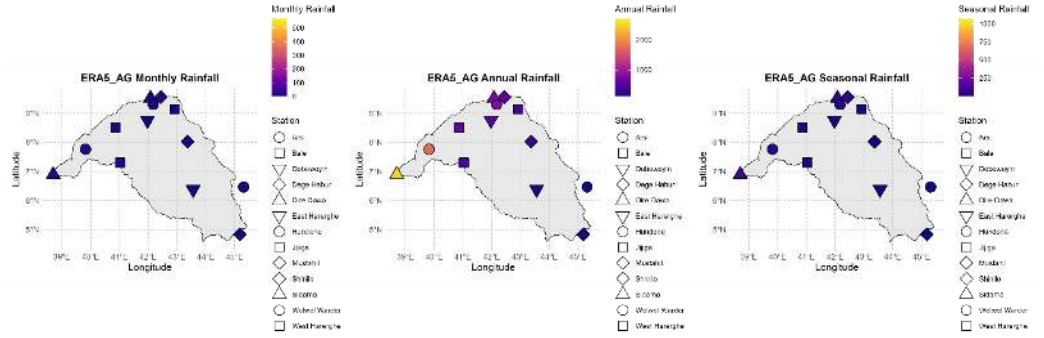

Figure 4.20. *Monthly, seasonal and annual ERA5_AG observations across Shabelle Basin, Ethiopia*

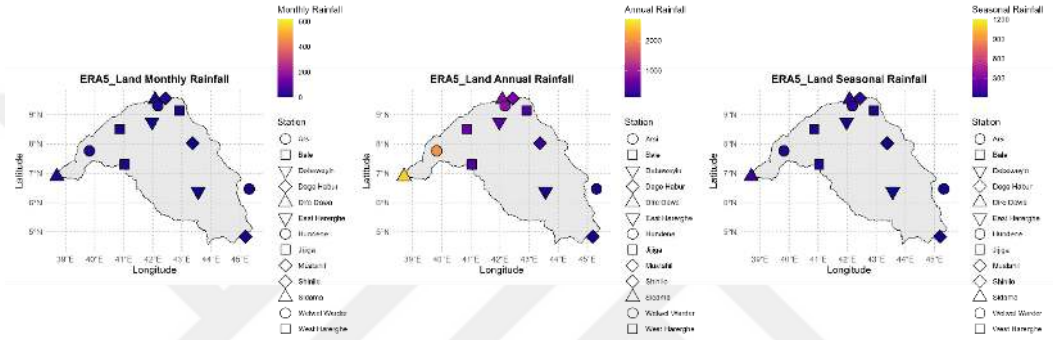

Figure 4.21. *Monthly, seasonal and annual ERA5_Land precipitation observations across Shabelle Basin*

- Other Reanalysis (CFSR and MERRA2):

The NCEP CFS (Climate Forecast System) reanalysis and NASA's MERRA-2 exhibit their own biases. CFS (and its forecast version CFSv2) is yielding high rainfall in the upper basin (~3000 mm/year regionally). This likely reflects CFS's model physics generating plentiful orographic rain. MERRA-2, which ingests bias-corrected precipitation over land, produces more moderate estimates in the Shabelle basin and aligns some ground observations. The MERRA-2 annual map maxes out around ~1500 mm in the highlands (much closer to CHIRPS and ground observation gauges) and tapers to ~150–200 mm in the southeast. This suggests MERRA-2's corrections constrain it to more realistic levels, avoiding the overestimation seen in ERA5-Land and CFS. Overall, among reanalyzing, MERRA-2 provides a spatial rainfall distribution closest to the gauge-based datasets, while CFS and ERA5 variants produce higher extremes in the highlands.

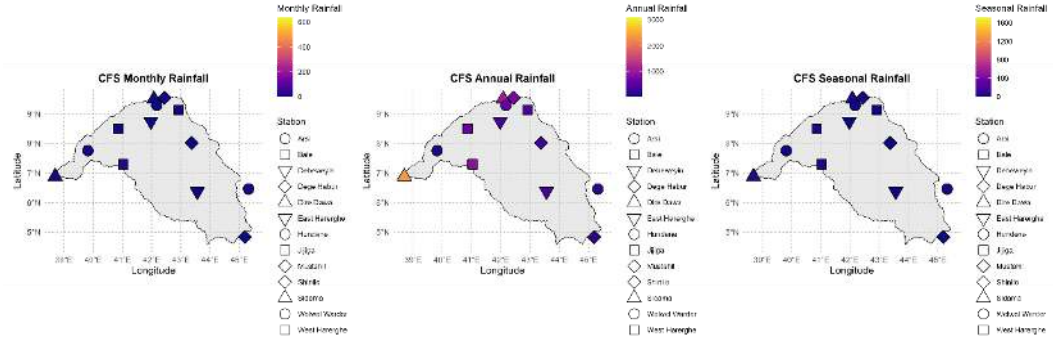

Figure 4.22. *CFSR monthly, seasonal and annual rainfall of different stations across Shabelle basin*

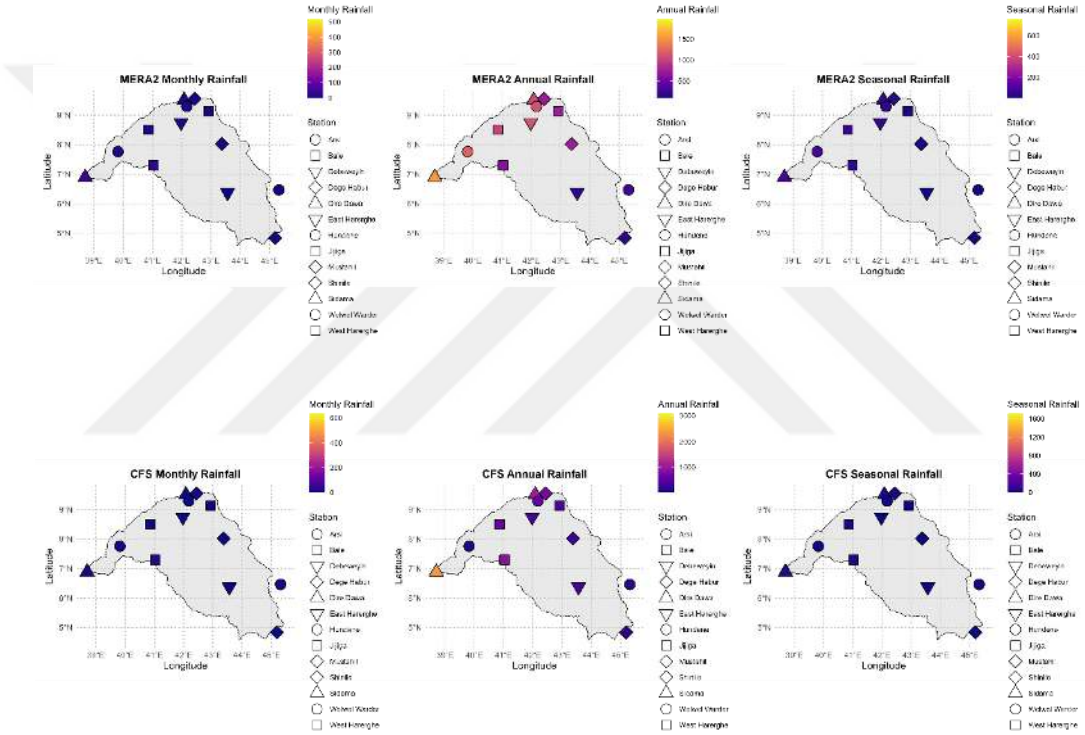

Figure 4.23. *MERRA2 Monthly, seasonal and annual rainfall of different stations across Shabelle basin*

4.3. Statistical Metric Evaluation of the Datasets

4.3.1. Regional level statistical performance

4.3.1.1. *Regional level categorical metrics (CSI, POD, FAR, HSS, SEDI, SR)*

This study used six interrelated regional categorical metrics (CSI, POD, FAR, HSS, SR, and SEDI) metric performance evaluations across all datasets in different temporal resolution. Figure (4.24) presents six heatmaps (one per categorical metric) evaluating precipitation occurrence skill for each dataset. The vertical axis lists eight datasets – for example, ERA5-Land, ERA5-AG, ERA5 (24 km), MERRA-2, PERSIANN-CDR, CPC Unified, CHIRPS, and CFS – and the horizontal axis shows four temporal scales: annual, daily, monthly, and seasonal.

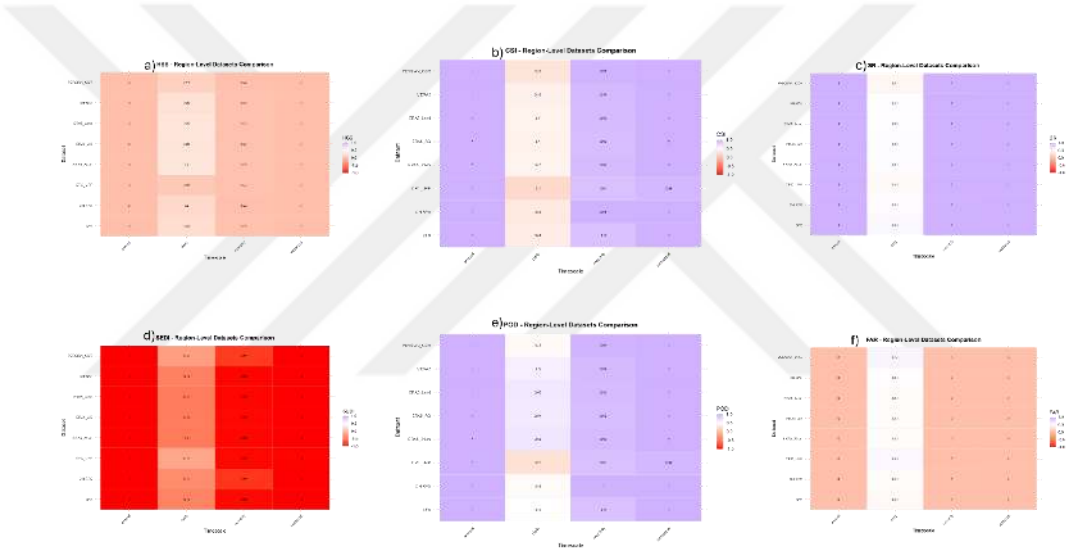

Figure 4-24 *Heatmap evaluation of categorical performance metrics for temporal region level comparison across all datasets a) HSS, b) CSI, c) SR, d) SEDI, e) POD, and f) FAR .*

As a general pattern by timescale, a clear pattern is seen across all categorical metrics: performance improves dramatically from daily to coarser time scales. At the annual and seasonal scales, most datasets attain trivial "perfect" scores for basic detection metrics. For instance, the Critical Success Index (CSI), Probability of Detection (POD), and Success Ratio (SR) are ~1.0 for every dataset on an annual basis, indicating that each product captured the occurrence of precipitation at least once per year without fail (no misses) and did not predict any year erroneously having rain when none occurred (false alarm ratio FAR = 0 annually for all) – an expected outcome since every year in the record had precipitation. Similarly, at the seasonal level, CSI and POD remain very high (≈1.0) for all datasets, with FAR ~0, reflecting that all products correctly indicated rain in each

season. However, these perfect scores at coarse scales are not indicative of forecasting skill so much as the inevitability of rain over such long intervals. In fact, the Heidke Skill Score (HSS) drops to ~0.0 at annual/seasonal scales, and the Symmetric Extremal Dependence Index (SEDI) reaches -1.0 for most datasets in those columns. In short, at coarse temporal resolutions (monthly, seasonal, annual), all datasets perform nearly identically on these categorical metrics – capturing rainfall occurrence almost perfectly – but this results from the high event prevalence rather than true model skill, as evidenced by near-zero HSS. The meaningful differences in categorical performance therefore emerge on the daily scale, where precipitation occurrence is a rarer event and more challenging to predict.

At the daily time scale, the heatmap reveals pronounced differences in how well each dataset detects rain events. No dataset achieves "good" daily detection skill in an absolute sense – for example, none have CSI or POD reaching the 0.7 threshold for *good* performance (per the given criteria). However, ERA5-based datasets consistently rank at the top for daily event detection. ERA5-Land, ERA5-AG, and ERA5 (24 km) all show POD ≈ 0.64–0.66, meaning they correctly detect roughly 64–66% of observed rain days – near the upper end of the "moderate" range (0.4–0.7) and approaching good performance. Their daily CSI (≈ 0.40–0.41) is the highest among the datasets, indicating a better balance of hits versus misses and false alarms (CSI of 0.4 is borderline between "moderate" and "poor" by the 0.4 threshold). Correspondingly, ERA5 yields the highest HSS (~0.29–0.30) at daily scale, reflecting moderate skill (HSS >0.2) in distinguishing rainy from non-rainy days – albeit still below the 0.5 threshold for *good* skill. The false alarm ratio for ERA5 datasets is ~0.48, meaning about 48% of the days they predicted rain turned out to be false alarms. This FAR is relatively low compared to other products (lower is better for FAR: 0 is ideal), underscoring ERA5's more conservative and accurate event forecasting. Indeed, the Success Ratio (SR) – the proportion of forecasted rain events that were correct (SR = 1 – FAR) – is about 0.52 for ERA5, meaning just over half of its predicted rain days were observed (SR in the 0.4–0.7 range is moderate). In summary, the ERA5 family exhibits the strongest daily categorical performance overall: moderate detection skill with the highest hit rates and among the lower false alarm rates of the group.

Other datasets achieve more modest daily results. MERRA-2 shows the next best detection capability after ERA5: daily POD ≈ 0.60 and CSI ≈ 0.38, indicating it catches

about 60% of rain days (upper-moderate range) and its overall accuracy in predicting yes/no rainfall is in the mid-0.3s. MERRA-2's HSS is 0.26 (moderate skill), slightly below ERA5. CHIRPS and PERSIANN-CDR satellite-gauge products have somewhat lower daily POD (0.46 each), meaning they detect around 46% of rainy days (mid-moderate performance). CHIRPS attains CSI ≈ 0.33 and PERSIANN ~0.29, both lower than the reanalyses – these CSI values fall in the "poor" range by strict interpretation (<0.4), suggesting these products often either miss events or issue more false alarms relative to hits than ERA5 does. Consistently, CHIRPS and PERSIANN have HSS values (~0.20 and 0.14, respectively) at or below the threshold between poor and moderate, implying minimal skill above chance for daily rain occurrence. Their false alarm ratios help explain this: FAR ≈ 0.48 for CHIRPS and 0.56 for PERSIANN. In PERSIANN's case, 56% of predicted rain days were false alarms – a rather high rate (FAR >0.5 is quite poor), which, coupled with its moderate POD, drags its CSI and HSS down. CHIRPS' FAR ~0.48 is similar to ERA5's, but because CHIRPS detects significantly fewer events (POD ~0.46 vs ERA5's 0.65), its overall CSI is lower. CFS (Climate Forecast System) reanalysis exhibits a mixed daily performance: its POD ~0.50 is moderate, and it actually has the lowest FAR (~0.45) of all datasets, meaning CFS issues comparatively fewer false alarms. This yields a daily Success Ratio ~0.55, the *highest* among the group (55% of CFS's rain predictions were correct). Thanks to that low FAR, CFS's CSI is ~0.35, on par with CHIRPS (mid-moderate), and HSS 0.23 (moderate skill). CPC Unified (gauge-only analysis) is the clear outlier on the low end. CPC attains only POD ≈ 0.27, detecting barely a quarter of the actual rainy days – a *poor* result by any standard (<0.4). Its CSI (≈ 0.20) is the lowest of all datasets, indicating very low overall accuracy in rain day prediction. CPC's HSS is ~0.08, essentially no skill (well below the 0.2 "poor" threshold), reflecting that its daily yes/no forecasts are only marginally better than random chance. Notably, CPC's FAR (0.55) is also high; in fact, CPC manages to both miss many events and also produce frequent false alarms, a combination that severely undermines its categorical scores. These daily outcomes reveal each dataset's strengths and weaknesses in event detection: ERA5 and MERRA-2 are adept at capturing rain occurrences (higher POD/HSS), CFS is conservative with fewer false alarms (high SR), CHIRPS and PERSIANN perform moderately but miss a significant fraction of events, and CPC struggles the most, with the poorest detection capability.

Table 4.1. *Daily categorical performance summary (event occurrence ranking)*

Dataset	POD	CSI	FAR	SR	HSS	Performance Summary
ERA5-L/AG	0.65	0.41	0.48	0.52	0.30	Best Overall: High hits, moderate skill
ERA5 24km	0.64	0.40	0.48	0.52	0.29	Similar to ERA5-L/AG
MERRA-2	0.60	0.38	0.50	0.50	0.26	Good: Second best hits
CFS	0.50	0.35	0.45	0.55	0.23	Strength: Lowest FAR (fewest false alarms)
CHIRPS	0.46	0.33	0.48	0.52	0.20	Moderate: Misses many events
PERSIANN	0.46	0.29	0.56	0.44	0.14	Moderate-Poor: High false alarms
CPC Unified	0.27	0.20	0.55	0.45	0.08	Poorest: Very low hits, high false alarms

According to the standard thresholds provided, we can further classify the daily scores. The ERA5-based datasets' POD (~0.65) approaches the "good" range (0.65–1.0), whereas CPC's POD (~0.27) is firmly "poor" (<0.4). Most daily CSI values (~0.3–0.4 for the better products) land in the moderate band (0.4–0.7) or just below – ERA5's CSI ~0.40 is right at the moderate/poor cusp, while CPC's 0.20 is clearly poor. All datasets have FAR in the 0.45–0.56 range; since 0 is ideal, these FAR values indicate quite imperfect forecasts. The relatively lower FAR of ERA5 (~0.48) and especially CFS (0.45) is favorable compared to PERSIANN and CPC (~0.55), but even 0.45 means 45% false alarms, highlighting that all datasets exhibit considerable over-prediction of rain at the daily scale. The Success Ratio (SR), being simply 1–FAR, mirrors this: the highest SR (CFS's ~0.55) just reaches the middle of the moderate range (0.4–0.7), while CPC's SR ~0.45 is lower-moderate. HSS values for ERA5 (~0.30) and MERRA2 (~0.26) fall in the *moderate skill* category (0.2–0.5), whereas CPC's 0.08 is *poor* (<0.2). These metrics collectively indicate that daily precipitation occurrence is difficult to model, and moderate success is the best achieved by current datasets in this regional test. All products fall short of ideal performance, but ERA5 (and to a degree MERRA2 and CFS) demonstrate comparatively better skill, whereas CPC in particular exhibits pronounced weaknesses.

In summary, ERA5 driven datasets show the highest hit rates and moderate skill scores (their categorical metrics are consistently in the moderate/good range relative to peers). They effectively capture a majority of rain events with fewer false alarms. MERRA-2 also shows solid detection ability (second-highest POD), though slightly

lower overall accuracy than ERA5. CFS has a noteworthy strength in its low false alarm frequency, indicating more confident predictions when it forecasts rain (reflected in the highest SR), but its hit rate is only moderate. CHIRPS and PERSIANN-CDR have comparable, middling categorical performance – they detect some events but miss many, and their false alarm rates are not much better than the re-analyses. These satellite-informed products thus offer only moderate skill in predicting rainy days. The CPC Unified gauge analysis is the weakest in categorical terms, with both a low detection rate and high false alarm rate. Its consistently poor POD, CSI, and HSS reveal a significant limitation in using CPC for day-to-day precipitation occurrence modeling in this region.

4.3.1.2. *Regional level continous metrics (KGE, NSE, MAE, R)*

Figure 4.25 displays heatmaps for four continuous-performance metrics – Kling-Gupta Efficiency (KGE), Nash-Sutcliffe Efficiency (NSE), Mean Absolute Error (MAE), and Pearson correlation (R) – for the same eight datasets and four time scales with y axis representing the datasets and the x axis shows time in daily, monthly, seasonal and annual.

Interms of performance trends by temporal resolution, Figure 4-30 shows a dominant trend of the sharp improvement in continuous metrics from daily to monthly to seasonal scales, mirroring what was seen with categorical metrics. Daily-scale performance is generally poor for all datasets, whereas monthly and seasonal scales show substantially higher skill. This reflects the fact that aggregating or averaging over longer periods filters out high-frequency noise and random errors (improving correlation and efficiency measures), but also indicates that the datasets struggle to capture day-to-day rainfall variability. We see that at the daily level, almost all datasets have low correlation and efficiency, and relatively high errors, whereas by the seasonal level many datasets achieve excellent correlation and good efficiency scores with much lower relative errors. The annual column (interannual variability) further highlights differences in each product's bias and long-term variability reproduction. Below, we delve into each metric and time scale in detail, comparing dataset performance.

At daily resolution, all datasets exhibit weak skill in reproducing precipitation magnitudes and variability. The Pearson correlation (R) between each dataset's daily precipitation estimates and observed daily precipitation is very low across the board: values range from roughly 0.12 (CPC, PERSIANN) to 0.36 (ERA5 variants). Every dataset's daily R falls below 0.5, which places them in the *"*weak" correlation category

(for reference, R > 0.7 would be considered good, >0.9 exScellent). The ERA5-based datasets achieve the highest daily correlations (~0.35–0.36), indicating they capture some day-to-day variability better than the others, but even these are far from strong relationships. CFS (R ~0.32) and MERRA-2 (~0.28) are the next best, while CHIRPS (~0.26) and especially PERSIANN (~0.13) and CPC (~0.12) show extremely low daily correlation to ground truth. Such weak correlations imply that on a daily basis, random errors and timing mismatches dominate, and none of the products can reliably reproduce which specific days were wet or dry beyond a low level of agreement.

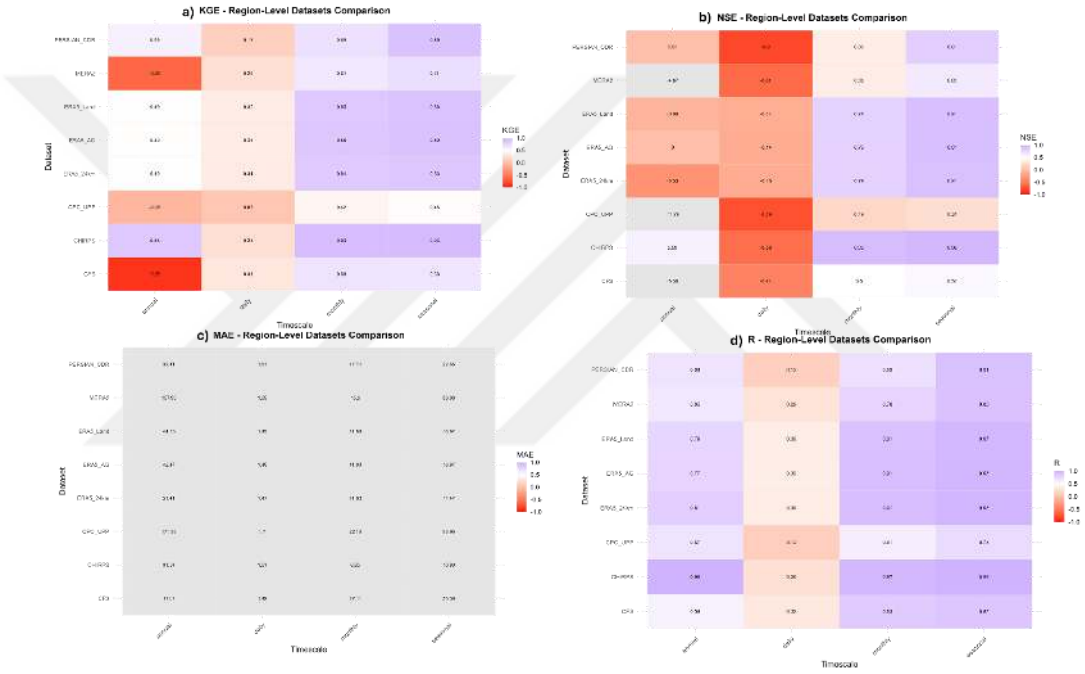

Figure 4.25. *Heatmap evaluation of continuous performance metrics for temporal region level comparison across all datasets a) KGE, b) NSE, c) MAE, d) R.*

The Nash-Sutcliffe Efficiency (NSE) values at daily scale are negative for all datasets, which is a striking indication of poor performance. Here, even the best NSE (for ERA5-Land/ERA5-AG) is about -0.14, and others range down to about -0.8, a clear sign that daily rainfall fluctuations are not being captured. Daily values of the Kling-Gupta Efficiency (KGE), which is an aggregate metric combining correlation, bias, and variability facets, are also low. ERA5-Land and ERA5-AG attain the highest daily KGE at approximately 0.34–0.35, followed by ERA5 (24 km) ~0.34 and CFS ~0.31. These are the only ones above 0.3. MERRA-2 and CHIRPS are around 0.25, PERSIANN ~0.13, and CPC is lowest at ~0.05. The low KGE and NSE alongside weak R at daily scale

highlight that the random error is very high day-to-day. Indeed, the Mean Absolute Error (MAE) at daily scale, measured in precipitation units (likely mm/day), remains significant for all. The ERA5 family again shows the smallest daily errors (1.45 mm MAE for ERA5-Land/AG, ~1.47 mm for ERA5 24 km), and CFS is similar (~1.49 mm). MERRA-2 and CHIRPS have slightly higher daily MAE (~1.61–1.65 mm), CPC ~1.70 mm, and PERSIANN is highest at ~1.81 mm/day error. These values suggest that on any given day, the typical absolute difference between the dataset's estimate and the observed rainfall is on the order of 1–2 mm. Considering that many days (especially in a semi-arid or monsoonal region) might have 0 or only a few millimeters of rain, an error of ~1.5 mm is quite large – again underscoring poor daily fidelity. In sum, no dataset exhibits strong daily performance: ERA5 and CFS have a slight edge with marginally higher correlation and lower error, while CPC and PERSIANN fare worst (lowest R and KGE, highest errors), but the overall picture is one of inadequate daily accuracy across all products.

When we aggregate to the monthly level, performance improves substantially for all datasets, revealing clearer distinctions in accuracy. Correlation (R) values climb dramatically: CHIRPS reaches about 0.97, an *excellent* correlation (well above the 0.9 threshold for excellent agreement). The ERA5-based datasets also achieve R ≈ 0.91 at monthly scale – crossing into the "excellent" range (Legates & McCabe, 1999)– indicating that ERA5 can capture month-to-month variations in precipitation very well. CFS and MERRA-2 show R ~0.83 and 0.78 respectively, which are in the *"good"* range (0.7–0.9). PERSIANN's monthly R is ~0.69, just shy of 0.7, and CPC's is ~0.61; these two lag behind, with CPC's correlation only *moderate*. Thus, at the monthly timescale, CHIRPS and ERA5 clearly stand out with the highest correlations, suggesting they track the observed seasonal cycle and monthly anomalies closely.

Table 4.2. *Continuous performance (monthly scale)*

Metric	CHIRPS	ERA5 Variants	PERSIANN-CDR	MERRA-2	CFS	CPC Unified
R	0.97 (E)	0.91 (E)	0.69 (G)	0.78 (G)	0.83 (G)	0.61 (M)
KGE	0.93 (E)	0.84-0.86 (VG)	0.69 (G)	0.64 (G)	0.68 (G)	0.42 (M)
NSE	0.93 (E)	0.78-0.79 (VG)	0.36 (P)	0.36 (P)	0.50 (G)	0.36 (P)
MAE	6.6 mm (L)	11.6-11.9 mm (L)	18-19 mm (M)	18.5 mm (M)	17.1 mm (M)	22.2 mm (H)

The KGE (Kling-Gupta Efficiency) at monthly scale likewise shows strong performance for the best datasets. CHIRPS attains KGE ≈ 0.93, indicating excellent overall reproduction of monthly precipitation (KGE > 0.5 is considered good, and 0.93 is near optimal). The three ERA5 variants have KGE in the 0.84–0.86 range, which is also firmly in the "good" category – reflecting a good balance of correlation, bias, and variability reproduction at monthly scale. MERRA-2 and CFS have moderately high KGEs (~0.64 and 0.68, respectively), comfortably above 0.5 and thus *good*, though not as high as ERA5/CHIRPS. PERSIANN's monthly KGE is around 0.69 (good), and CPC's is 0.42, which is below 0.5 and thus only *moderate*. These KGE results reinforce the ranking implied by correlation: CHIRPS is the top performer, closely followed by ERA5 products (all of which are very similar to each other), then a second tier of CFS, MERRA-2, PERSIANN with moderate-good performance, and finally CPC significantly lagging. Notably, CPC's KGE = 0.42 despite a correlation of 0.61 suggests that CPC may have substantial biases or variability errors (KGE penalizes those), which drag its score down. Conversely, PERSIANN's KGE ~0.69 is higher than its correlation might suggest, perhaps indicating it has smaller bias – but overall PERSIANN still does not reach the top tier.

The NSE (Nash-Sutcliffe Efficiency) values at monthly scale tell a very similar story. CHIRPS achieves NSE ≈ 0.93, meaning its monthly precipitation totals are extremely close to observed (93% of the variance explained, an excellent result >0.5). ERA5-Land, ERA5-AG, and ERA5 (24 km) have NSE ≈ 0.78–0.79, also indicating good performance (anything above 0.5 is considered good; ~0.8 is very good). CFS's NSE is around 0.50, right at the threshold of good – so CFS can be considered to have marginally good skill in monthly totals. MERRA-2 and PERSIANN are slightly lower with NSE ~0.36 each. The ordering across datasets is consistent: CHIRPS > ERA5 >> CFS > MERRA ≈ PERSIANN > CPC in terms of monthly efficiency.

The Mean Absolute Error (MAE) at monthly scale (likely expressed in mm of rainfall per month) quantifies the typical volume error and further highlights the differences. CHIRPS has by far the lowest monthly MAE (~6.6 mm), indicating that on average its estimated monthly rainfall deviates from observed by only about 6.6 mm. This is a very small error in many climates (for context, if a region receives e.g. ~100 mm in a typical month, a 6.6 mm error is only ~6.6%). ERA5-Land, ERA5-AG, and ERA5 (24 km) have MAE around 11.6–11.9 mm per month, roughly double CHIRPS's error –

still relatively low (perhaps ~10% error on a 100 mm month), reflecting good accuracy but not as exceptional as CHIRPS. MERRA-2 and PERSIANN's monthly MAEs are about 18–19 mm, and CFS ~17.1 mm, while CPC's is highest at ~22.2 mm. Thus, CPC's typical monthly error is more than three times CHIRPS's, underscoring CPC's deficiencies in capturing monthly rainfall totals. These MAE values align well with the efficiency metrics: CHIRPS clearly yields the smallest errors, ERA5 next best, then CFS/MERRA/PERSIANN moderate, and CPC worst. In practical terms, CHIRPS appears extremely skillful in estimating monthly precipitation amounts in this region, whereas CPC shows significant biases or miss in volume. This outcome resonates with other studies that have found CHIRPS to outperform reanalyses for monthly rainfall in certain regions (J. S. Ahmed et al., 2024a; Hussein & Baylar, 2023)(e.g. over Ethiopia, CHIRPS was found to "generally outperform" ERA5), likely due to its incorporation of local gauge data. Here, CHIRPS's advantage is evident in both correlation and error metrics at monthly scale.

At the seasonal (quarterly) aggregation, performance reaches its peak for most metrics, with many datasets showing excellent agreement with observations. The correlation (R) for top datasets is extraordinarily high: CHIRPS has R ≈ 0.99 with observed seasonal rainfall, essentially a *near-perfect correlation*. The ERA5-based datasets each achieve R ≈ 0.97, also in the *excellent* range >0.9. Even PERSIANN-CDR climbs to R ≈ 0.91 by seasonal scale, crossing the 0.9 threshold and thus exhibiting excellent correlation in seasonal totals. MERRA-2 and CFS have seasonal R values around 0.87–0.89, which are *very good* (close to the excellent range). CPC Unified's seasonal correlation improves to about 0.73, which is a notable increase from its monthly 0.61; 0.73 is solidly *good* (0.7–0.9 range), though it remains the lowest of the group.

Overall, at seasonal scale five of the eight datasets attain R > 0.85, indicating that most products reliably track seasonal precipitation patterns, with CHIRPS and ERA5 doing so exceptionally well.

The KGE scores at seasonal scale likewise are very high for the leading datasets. CHIRPS is ~0.93, ERA5-Land and ERA5-AG ~0.89, ERA5 24 km ~0.86 – all of these are well above 0.5, demonstrating good to excellent agreement in seasonal accumulations. PERSIANN's KGE rises to 0.89 (interestingly as high as ERA5's), indicating that by seasonal scale PERSIANN-CDR is capturing the total amounts nearly as well as the reanalysis, at least in terms of overall bias and variability. MERRA-2 and CFS have

seasonal KGE around 0.71 and 0.66 respectively – both above 0.5 (good), though noticeably lower than the top tier. CPC's KGE at seasonal is only 0.48, which, while an improvement from its monthly 0.42, still falls just below the "good" cutoff. Indeed, efficiency metrics reward low bias: CPC likely has large seasonal biases in this region, whereas PERSIANN (and CHIRPS/ERA5) have smaller bias and variance errors.

The NSE values confirm this pattern at seasonal scale. CHIRPS attains NSE ≈ 0.96 at seasonal scale, which is *excellent* – it indicates that CHIRPS explains 96% of the variance in seasonal rainfall totals, an extremely high skill. The ERA5 datasets have NSE ~0.90–0.91, also excellent. PERSIANN's seasonal NSE is 0.81, which qualifies as *good* performance (well above 0.5). CPC's NSE, however, is only 0.24 at seasonal scale – surprisingly low given its correlation of 0.73. An NSE of 0.24 indicates that even at the seasonal level, CPC's errors remain large – it only explains 24% of seasonal variance and is barely better than just using the mean seasonal climatology (NSE near 0). This again underscores that CPC has substantial biases or misses in certain seasons (for example, it might severely underestimate or overestimate seasonal totals for wet or dry seasons), which drag down its efficiency score. In contrast, all other datasets have NSE > 0.5 at seasonal scale, meaning they are reliably capturing seasonal rainfall quantities. The particularly high NSE for CHIRPS and ERA5 indicates these datasets excel at reproducing both the timing and magnitude of seasonal rainfall variations.

The seasonal MAE values (in mm per season) provide a tangible sense of these differences. CHIRPS's seasonal MAE is only about 10.9 mm, the lowest of all – incredibly small relative to typical seasonal rainfall (which might be hundreds of millimeters in a rainy season, depending on the region). ERA5-Land and ERA5-AG each have MAE ~16.9 mm per season, and ERA5 (24 km) ~17.5 mm. PERSIANN's seasonal MAE is ~22.7 mm, higher than ERA5's but still not very large in absolute terms. MERRA-2 and CFS are around 36–37 mm MAE per season, roughly double ERA5's error, indicating more significant deviations. CPC's seasonal MAE is by far the highest at about 53.9 mm. To put this in perspective, if a typical seasonal rainfall total is, say, 300–500 mm, CPC's error (~54 mm) could be around 10–18% of the total – whereas CHIRPS's error (~11 mm) would be only ~2–4%. These error magnitudes reinforce how much more accurate CHIRPS and ERA5 are compared to CPC (and also significantly more accurate than MERRA-2 or CFS) at capturing seasonal precipitation volumes. The

relatively low errors for CHIRPS and ERA5 correspond to their high KGE/NSE, confirming that they have both low bias and good variability matching at this scale.

Considering all continuous metrics together, CHIRPS emerges as the most consistently strong dataset for precipitation amounts in this regional evaluation. Its strengths include *exceptional correlation* at monthly to annual scales (often >0.95), *very high efficiency scores* (KGE and NSE in the 0.9 range for monthly/seasonal), and the *lowest errors* (both monthly and seasonal MAE are far smaller than those of other datasets). This indicates that CHIRPS not only tracks the timing of rainfall fluctuations well, but also gets the magnitudes right with minimal bias. ERA5-based datasets are a close second: they demonstrate high skill at monthly and seasonal scales (e.g. monthly KGE ~0.85, NSE ~0.79, R ~0.91) and relatively low errors, albeit not as low as CHIRPS. ERA5's primary weakness, as noted, might be a slight bias affecting its annual totals (reflected in moderate KGE/NSE at annual scale), but overall it performs very well from monthly scale upwards. In fact, global analyses have noted ERA5 as one of the top-performing precipitation datasets for hydrological purposes (Gebrechorkos et al., 2024), consistent with our findings of its strong metrics. CHIRPS might have a slight edge for capturing the magnitude of rainfall correctly, while ERA5 excels in timing and occurrence; together they clearly outclass the other datasets. This observation aligns with prior studies noting the strengths of reanalyses and high-quality merged products: for example, CHIRPS has been shown to outperform ERA5 in certain African regions for rainfall estimation (J. S. Ahmed et al., 2024a; Hussein & Baylar, 2023), while ERA5 is frequently among the top-ranked global datasets for precipitation and often second only to multi-source products.

In conclusion, this regional evaluation reveals that CHIRPS and ERA5-based datasets are the most reliable choices for precipitation data, with CHIRPS excelling in matching observed rainfall volumes and ERA5 excelling in capturing rainfall occurrence. MERRA-2 and CFS provide moderate skill and may be acceptable for some applications but are not outstanding. PERSIANN-CDR performs adequately at aggregate scales but less so at high resolution. CPC Unified stands out as the least dependable dataset in this analysis, with significant deficiencies across both categorical and continuous measures of performance. These findings are critical for guiding dataset selection in hydrologic and climate studies: using the best-performing datasets (CHIRPS/ERA5) can substantially.

4.3.2. Station level statistical performance

4.3.2.1. *Categorical metrics (CSI, POD, FAR, HSS, SEDI, and SR)*

The study uses both continuous variable skill metrics and categorical metrics to make assessment on station level and evaluate how the ground observed station rainfall gauges correlates to different gridded precipitation datasets. The results are structured metrics by dataset with spatio-temporal details. Across categorical event-detection metrics (CSI, POD, FAR, HSS, SEDI, Success Ratio), similar patterns emerge. Most of the categorical metrics shows better results in monthly and seasonal scales. CHIRPS yields the highest Probability of Detection (POD) and Critical Success Index (CSI) in most subregions at monthly and seasonal scales: for example, CHIRPS POD often exceeds 0.8 and CSI 0.7–0.9 in Arsi and West Hararghe, while its FAR is very low (≈0.0–0.05). This leads to Success Ratios (SR) near 1.0 at seasonal scales for CHIRPS. By contrast, daily-scale POD is very low for all datasets (<0.15), and FAR is very high (0.7–0.9), indicating poor daily event forecasting. MERRA2 and PERSIANN_CDR show good detection skill (CSI ~0.8–1.0), POD ~1 in Arsi at seasonal scale) but higher FAR than CHIRPS. For example, MERRA2 seasonal POD in Arsi is 1 (FAR of 0.01) and CSI 0.97, while all metrics shows lower results in daily and annual scale. The Heidke Skill Score (HSS) is near-zero for most cases (sometimes slightly negative), indicating that even the best products do little better than random at event timing; however, all datasets attains some positive HSS in Arsi and Mustahil (up to ~0.44 at daily scale), reflecting its strong event match there. SEDI (an extreme-event metric) is generally low for all products, with only a few positive pockets (e.g. Bale, Debeweyin, Degehbur, Diredawa, Hundene, Jigjiga and sidama) stations monthly time scale SEDI ≈0.1–0.8 for all of the datasets), indicating modest skill in capturing extremes. Overall, the categorical analysis confirms that CHIRPS most reliably detects precipitation events (high POD, low FAR) in Ethiopia, in line with other studies that rank CHIRPS and PERSIANN among top performers for rainfall event detection.FAR- all dataset faceted figure shows that all datasets shows high FAR (0.4-0.9) and low POD for daily timescale.

Heatmap of the Critical Success Index (CSI) for ground observation stations (y-axis) and four time scales (x-axis: annual, daily, monthly, seasonal) across all datasetsmultiple datasets (panels labeled CFS, CHIRPS, etc.). each panel represent datasets and in each panel, cells are colored by CSI value (white/purple high, pink low). Notably, the annual CSI is uniformly 1.00 (purple) for all stations in every dataset,

whereas daily CSI values are very low (~0.04–0.28, pink) across stations. Monthly CSI values range from about 0.36 to 1.00, and seasonal CSI are high (~0.78–1.00, purple). Some station-specific patterns emerge: for example, CHIRPS shows “Debweyin and Walwal Werder” with a low monthly CSI and relatively low seasonal (0.60), whereas stations like “Bale” exhibit very high values (monthly 0.85–0.91, seasonal 1.00). most panels across alldatasets tend to have the highest CSI at monthly/seasonal/annual scales (many values ≥0.9, deep purple). All datasets in Figure 4-22. reach near-perfect CSI at annual and seasonal scales, with the main variability observed at daily resolution (where ERA5 and MERRA slightly outperform others) and moderate differences at the monthly scale between datasets.

Figure 4.26. *Heatmap evaluation of performance metrics (CSI) for temporal station level comparison across all datasets.*

All annual FAR (Fig 4.26) values are 0.00 for every station and dataset. Seasonal FAR is also 0.00 for almost all cases , meaning no false alarms at seasonal scale. In contrast, daily FAR is uniformly high (~0.74–0.89) across all stations except Arsi in all datasets, indicating many false alarms at high temporal resolution. Monthly FAR values are moderate (mostly 0.05–0.27). An interesting anomaly is CHIRPS at Arsi, which has a monthly FAR of 1.00 – the highest in the plot – whereas CFS at Arsi has 0.00 . In

general, differences between datasets are subtle for FAR: daily FAR at Welwelawer is ~0.86–0.89 in all panels, and monthly FAR at most stations varies only in the hundredths. For example, at "Hundene" monthly FAR ranges 0.04–0.18 (pink) across datasets. Thus, while all datasets show high false alarms daily and none seasonally. No dataset uniformly outperforms others; FAR is dominated by the temporal scale (high at daily, zero at seasonal and annual).

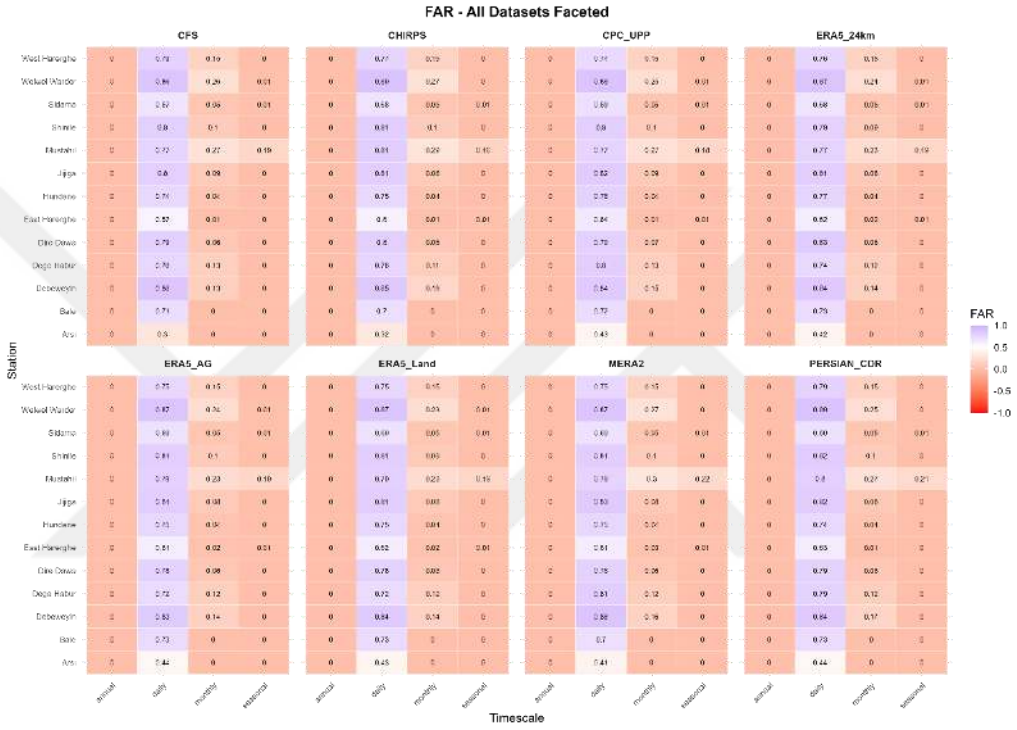

Figure 4.27. *Heatmap evaluation of performance metrics (FAR) for temporal station level comparison across all datasets.*

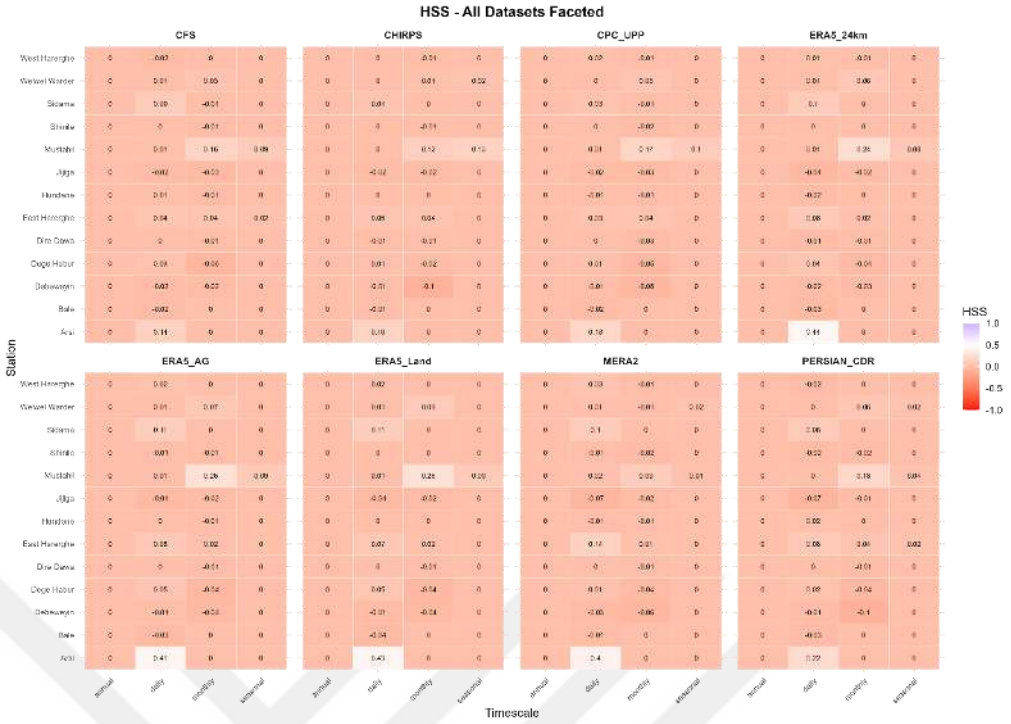

Figure 4.28. *Heatmap evaluation of performance metrics (HSS) for temporal station level comparison across all datasets.*

Figure 4.28 heatmap of the Heidke Skill Score (HSS) for each station and time scale by dataset. Nearly all values are very close to zero. Annual and seasonal HSS are exactly 0.00 for every station and dataset, indicating no skill or perfect neutrality. Daily HSS occasionally reaches small positive values (up to ~0.40) for a few station/dataset combinations: for instance, Arsi in ERA5_24km has daily HSS ≈0.40 and ERA5_AG Arsi ~0.43. Monthly HSS shows similarly tiny positives (e.g., CFS Mustahil monthly = 0.16; ERA5 Land Mustahil monthly = 0.26) and occasional slight negatives (down to about –0.06). The vast majority of HSS entries are between –0.10 and +0.10 . Differences among datasets are minimal: ERA5 panels yield the highest small values (e.g. around 0.4 at Arsi daily and Mustahil monthly) whereas CFS, CHIRPS, CPC, and PERSIAN are nearly flat at zero. In summary, HSS is essentially zero everywhere, with only a handful of off-zero cells (mostly small positive) in ERA5-derived data. This indicates uniformly low Heidke Skill across all stations and time scales, with no dataset showing markedly better HSS than the others.

As with CSI, all annual POD values (Fgiures 4.26 and 4.29) are 1.00 for every station and dataset. Daily POD is generally low-to-moderate (~0.06–0.81); for instance, Arsi has the highest daily POD (up to 0.83 in ERA5_Land) while stations like

Welwelawer remain very low (≈0.06–0.18). Monthly POD values are substantially higher (mostly 0.63–1.00, purple): ERA5 and MERRA panels often approach 0.90+ in the monthly column for many stations, whereas CHIRPS and CFS show a few lower values (e.g. CHIRPS Welwel werder ~0.47). Seasonal POD is near 1.00 across the board (purple), indicating very high detection rates at seasonal scale. In comparing datasets, ERA5_24km, ERA5_AG, ERA5_Land, and MERRA2 consistently yield high POD (≥0.8) at monthly and seasonal scales, whereas CHIRPS and CFS have slightly lower monthly POD for certain stations. Noteworthy exceptions include CHIRPS at Deboweyin (monthly 0.40, seasonal 0.60, pink shades) and CHIRPS at Welwelawer (monthly 0.47, seasonal 0.62) which are lower than the corresponding values in other datasets. Overall, detection rates increase with temporal aggregation and are highest in ERA5/MERRA products, with the largest inter-dataset differences at the daily resolution.

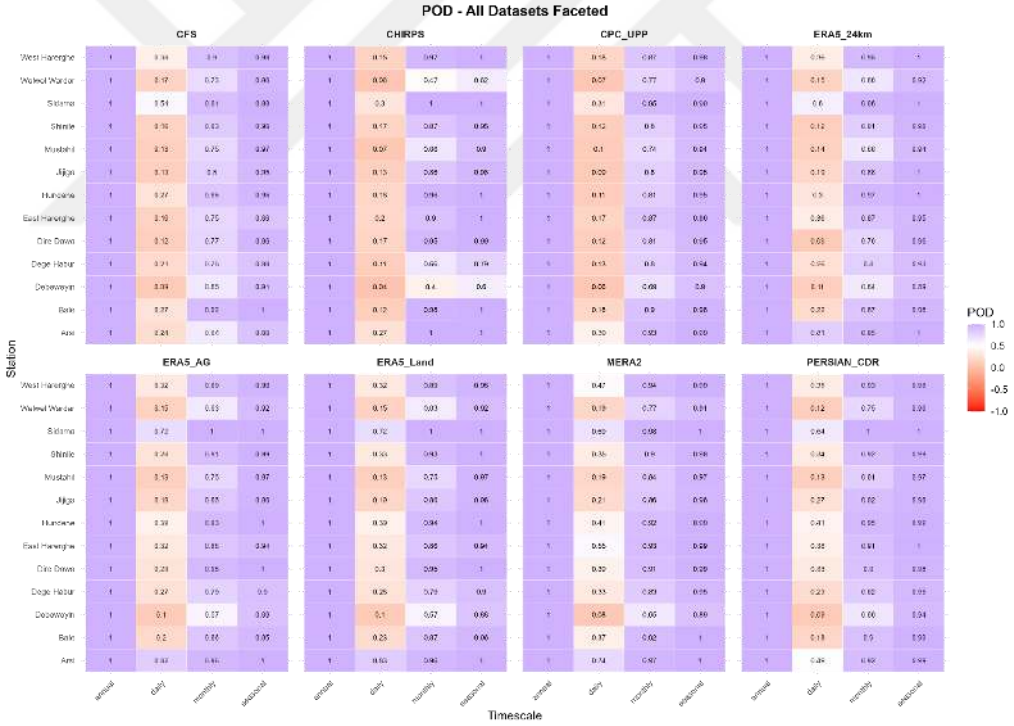

Figure 4.29. *Heatmap evaluation of performance metrics (POD) for temporal station level comparison across all datasets.*

Figure 4.30 represents heatmap of the (Symmetric Extremal Dependence Index (SEDI) where both annual and seasonal are uniformly –1.00 (deep red) in every panel ,indicating the worst possible dependence score at those scales for all stations and datasets. The daily SEDI (second column) varies around 0 (pale pink/white) for most stations, with values

typically between –0.20 and +0.10; no strong outliers. The monthly SEDI (third column) shows stark contrasts: many stations have very high SEDI (0.60–0.90, white to purple) while others are –0.99 (bright red). For example, "Sidama" has monthly SEDI ≈0.79 in CFS/ERA5, whereas CHIRPS Sikoma is ~0.00. Station "Bede" has monthly SEDI ≈0.78–0.84 across all datasets (purple), while "East Hararghe", "Deboweyin", and "Dega Habur" have monthly SEDI ≈–0.99 (red) in most panels. These differences are consistent across datasets: ERA5_Land and MERRA2 produce similarly high monthly SEDI for favorable stations (Bale, Sidama, Jijiga, Hundene, and West Harergeh), while CHIRPS and CFS also capture some of these high values. No dataset is uniformly best – instead the variability is station-specific. The key pattern is that only some station/dataset pairs achieve high extreme-dependence scores at monthly scale, whereas all annual/seasonal scores are uniformly poor (–1.0) and daily scores are near-neutral.

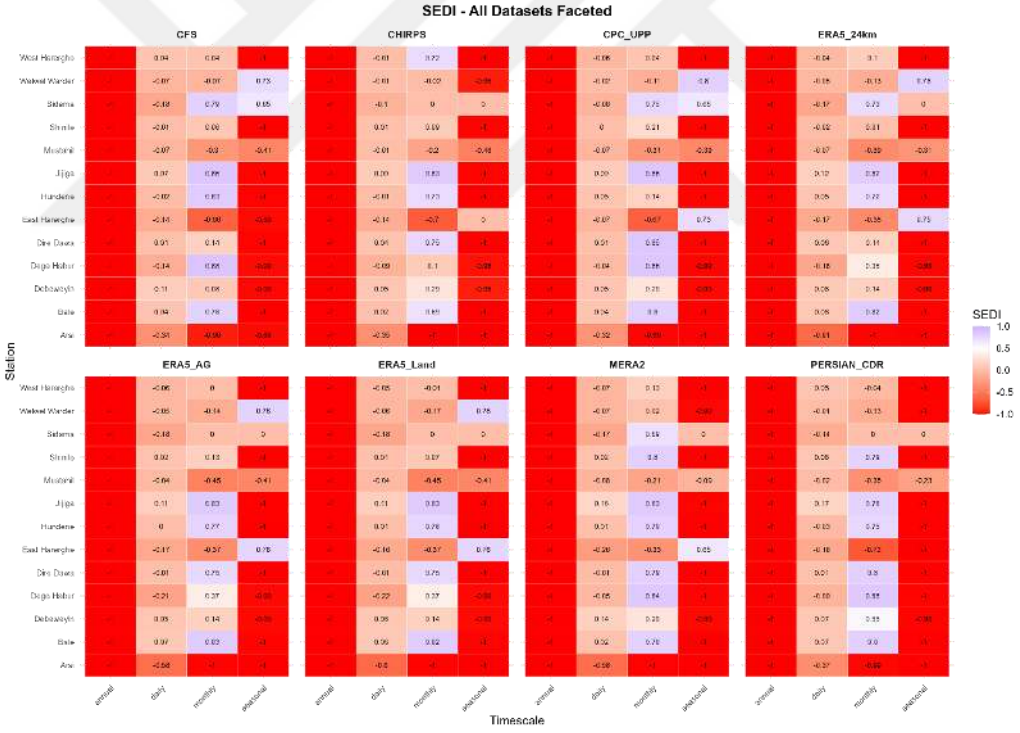

Figure 4.30. *Heatmap evaluation of performance metrics (SEDI) for temporal station level comparison across all datasets*

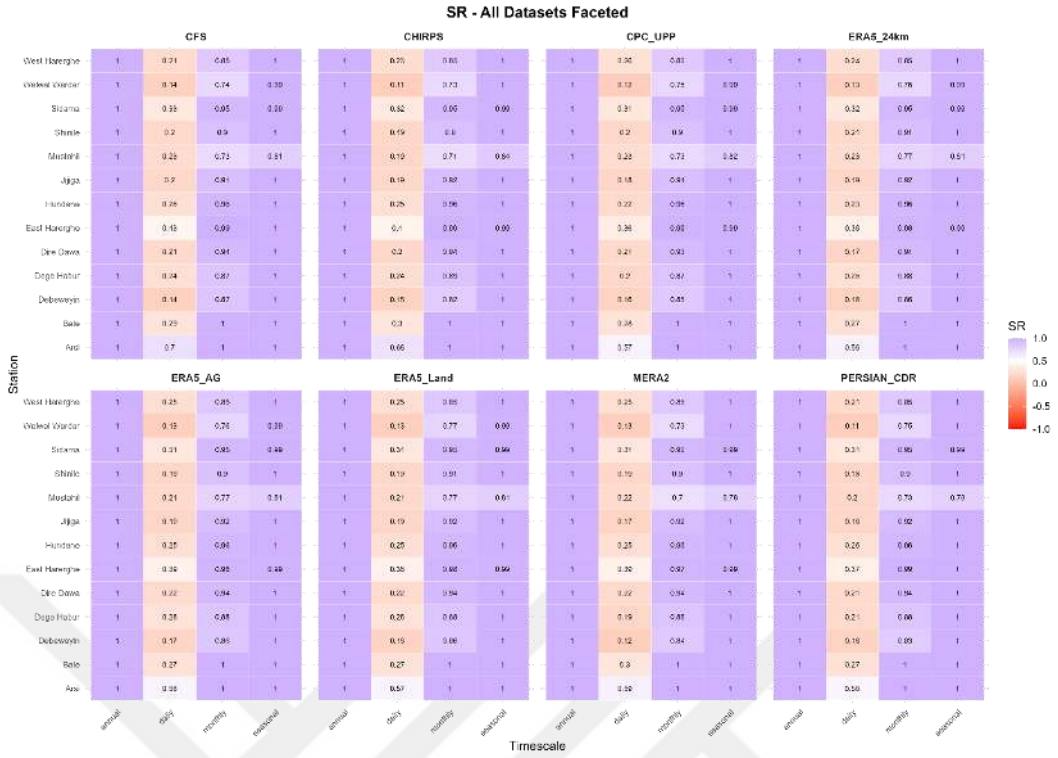

Figure 4.31. *Heatmap evaluation of performance metrics (SR) for temporal station level comparison across all datasets.*

Figure 4.31 shows heatmap of the Success Ratio (SR) for each station and time scale by dataset. Like CSI and POD, the annual SR is 1.00 (purple) for every station in all panels, and the seasonal SR is also 1.00 throughout. Daily SR values (second column) vary more: most stations are low (~0.11–0.43), except Arsi which has unusually high daily SR across datasets (e.g. 0.57–0.70, purple-ish). Indeed, Arsi's daily SR is highest in CHIRPS (0.68) and ERA5 (0.59), whereas stations like Shinle and Welwelawer have daily SR ~0.10–0.20 in every panel. Monthly SR (third column) is generally high (>0.70, purple-white) for almost all stations and datasets. The exception is Mustahil, which has the lowest monthly SR (0.71–0.77) in every panel . Overall patterns are similar across datasets: ERA5 and MERRA panels show slightly higher SR values (more purple) at some stations, but differences are small. In summary, SR is perfect at annual/seasonal scales and uniformly high at monthly scale, with the main variations at daily scale (where Arsi stands out with SR ~0.6, and Shinle stays around 0.2). All datasets produce similar SR distributions; no one dataset is consistently better for this metric.

The heatmaps reveal clear patterns in dataset performance across stations and time scales. In the categorical verification metrics (FAR, SEDI, SR, CSI, POD, HSS), high-

precipitation stations like Arsi and Bale consistently show near-perfect scores at seasonal and monthly scales, whereas dry stations (West Hararghe, Shinile, Mustahil) have much lower scores at the daily scale. Generally, CHIRPS and CPC_UPP attain the highest CSI and POD values at most stations and scales (e.g. CHIRPS CSI≥0.95 at Sidama and Debeweyin seasonally), whereas CFS has the lowest scores (often <0.5 CSI) except in the trivial annual metric. The probability of detection (POD) at seasonal/annual scales is nearly unity for all datasets and stations, but at daily resolution ERA5_Land and MERRA2 often lead with the highest POD (≈0.4–0.5 at some sites) while CHIRPS exhibits the lowest daily POD (often <0.2 at arid locations).

As with CSI, all annual POD values (Fgiures 4.26 and 4.29) are 1.00 for every station and dataset. Daily POD is generally low-to-moderate (~0.06–0.81); for instance, Arsi has the highest daily POD (up to 0.83 in ERA5_Land) while stations like Welwelawer remain very low (≈0.06–0.18). Monthly POD values are substantially higher (mostly 0.63–1.00, purple): ERA5 and MERRA panels often approach 0.90+ in the monthly column for many stations, whereas CHIRPS and CFS show a few lower values (e.g. CHIRPS Welwel werder ~0.47). Seasonal POD is near 1.00 across the board (purple), indicating very high detection rates at seasonal scale. In comparing datasets, ERA5_24km, ERA5_AG, ERA5_Land, and MERRA2 consistently yield high POD (≥0.8) at monthly and seasonal scales, whereas CHIRPS and CFS have slightly lower monthly POD for certain stations. Noteworthy exceptions include CHIRPS at Deboweyin (monthly 0.40, seasonal 0.60, pink shades) and CHIRPS at Welwel Warder (monthly 0.47, seasonal 0.62) which are lower than the corresponding values in other datasets. Overall, detection rates increase with temporal aggregation and are highest in ERA5/MERRA products, with the largest inter-dataset differences at the daily resolution.

4.3.2.2. *Continous metrics (KGE, NSE, R and MAE)*

The study uses continuous metrics like KGE, NSE, R and MAE to evaluate the station level ground obsevation to the various datasets named above and see how each each dataset performs in a station level.

Figure 4.32 represent heatmap of the Kling–Gupta Efficiency (KGE) for each station and time scale by dataset. Most cells are red or orange (negative KGE), indicating poor performance, with occasional near-zero or positive (white/purple) values in a few panels. Notably, many CFS and ERA5_Land values are strongly negative (e.g. West

Hararge daily KGE ≈–0.92 in CFS, Bale monthly ≈–-0.95). CHIRPS shows a clear outlier at Arsi with very high KGE (daily ~0.60, monthly ~0.90, seasonal ~0.94, purple). MERRA2 and PERSIAN also have moderate positive values at some stations (e.g. Arsi seasonal 0.75–0.77 in MERA2; PERSIAN_CDR Arsi monthly ~0.71 and seasonal ~0.72). In general, ERA5-derived datasets tend to have the highest KGE (less negative) across multiple stations and scales, whereas CFS and CPC often have lowest (reddest) values. Thus, ERA5_24km/AG/Land and MERRA2 panels appear "better" (lighter colors) than CFS/CHIRPS/CPC for KGE, though no single dataset dominates in all cells. The variation is mostly spatial/temporal: stations like Arsi and Bede show much higher KGE than others in certain datasets, while other stations (e.g., Deboreweyn, Jigjiga) remain negative across the board.

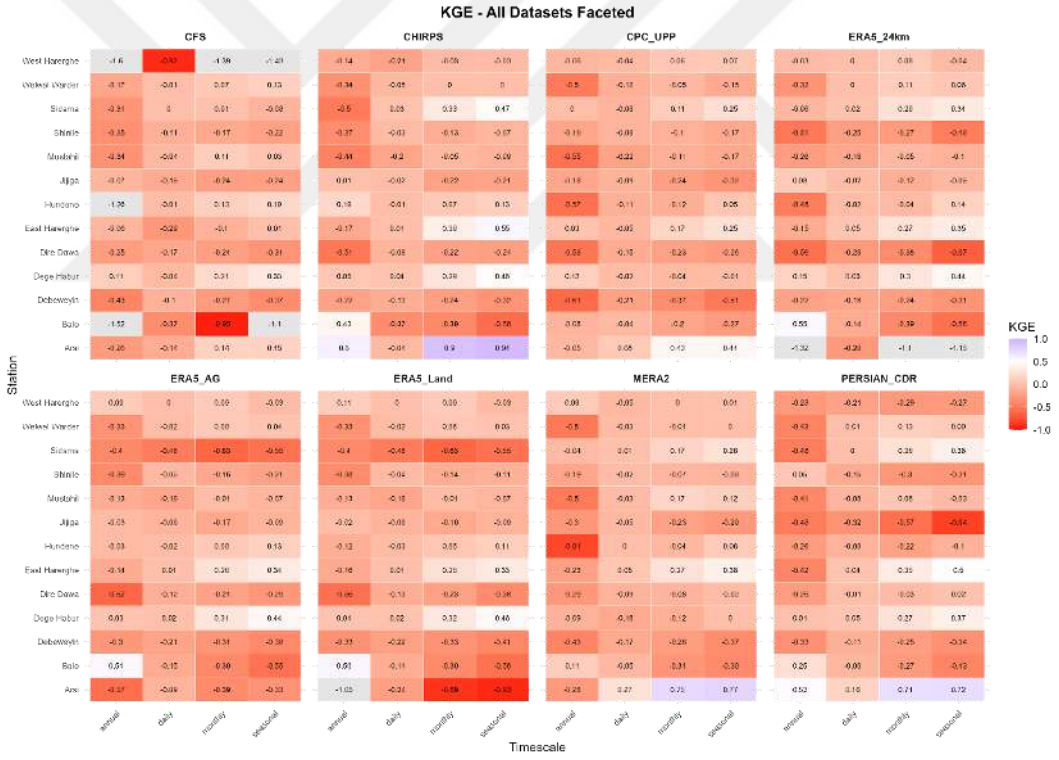

Figure 4.32. *Heatmap evaluation of continuous performance metrics (KGE) for temporal station level comparison across all datasets.*

As Fig 4.33 shows heatmap of Mean Absolute Error (MAE) for each station and time scale by dataset, it has values which are absolute errors (lower is better). The ERA5_AG and ERA5-Land panel shows extremely large annual errors (up to ~1188 at Sidama) and CFS panel shows as well extreme annual error (around 1017, at West

Harerghe, far exceeding other datasets. CPC_UPP and ERA5 panels have much smaller annual errors (hundreds): for example, West Hararge annual MAE is ~302 for CPC and ~268 for ERA5_24km , vs 1018 for CFS (red). ERA5_AG and ERA5_Land annual MAEs at Sidama are ~1188 (red) – an outlier – whereas most ERA5 errors are ~200–400. Daily MAEs are low for all: typically 1–4 in CPC/ERA5, a bit higher (up to 6.57) in CHIRPS at Sidama. Monthly MAEs are moderate: Most of the datasets generally <60). Seasonal MAEs range widely: CFS and CHIRPS often in the 100–200 range (light gray), CPC and ERA5 also ~80–180, whereas some ERA5_AG values are higher (e.g. 325 at Sikoina). Overall, CPC_UPP tend to have the lowest MAE at most scales, CFS has the highest , and CHIRPS/MERRA2 are intermediate. The station Sidama generally shows the largest errors (notably for CFS/ERA5_AG), whereas Shinile and Deboweyn tend to have lower MAE in some panels. In summary, error increases from daily to annual for all data: all datasets perform well (small errors) daily and worse annually, with CFS and ERA5-based datasets performing worst and CHIRPS often best.

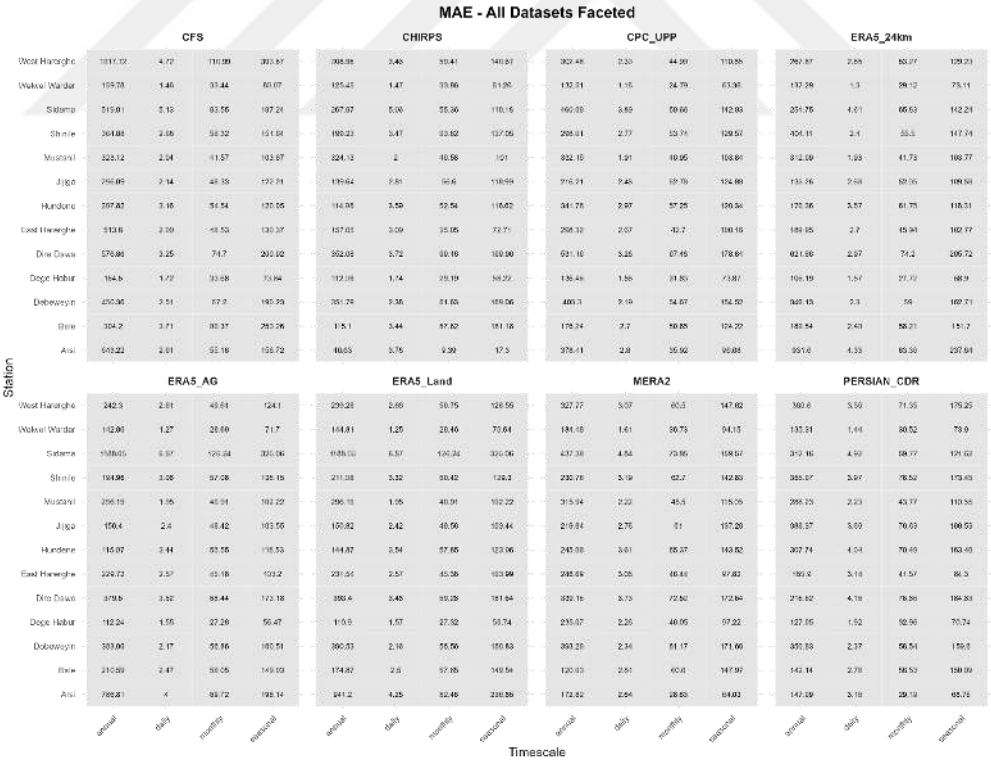

Figure 4.33. *Heatmap evaluation of continuous performance metrics (MAE) for temporal station level comparison across all datasets*

Figure 4.34 shows that most NSE values are negative, indicating poor model skill, with only a few positives. All datasets show highly negative NSE at the annual scale (column 1) across stations. Seasonal NSE is also almost entirely negative, except for a few near-zero cases (e.g. CFS East Hararge seasonal ~0). The clearest exception is CHIRPS at Arsi, which has a very high NSE (0.68 daily, 0.93 monthly, 0.97 seasonal). All CFS and CPC_UPP values are negative for every station/time. Thus, dataset performance is highly station-specific: CHIRPS (and to a lesser extent ERA5/MERRA) "succeeds" at Arsi, whereas other stations (e.g. East Hararge, Jigjiga) have NSE ≪0 across all panels. In effect, no dataset consistently achieves positive NSE except for the single standout at Arsi in CHIRPS. Overall, the heatmap reveals that NSE is generally low (worse than zero) for all datasets and stations, with a few isolated good scores in the CHIRPS panels.

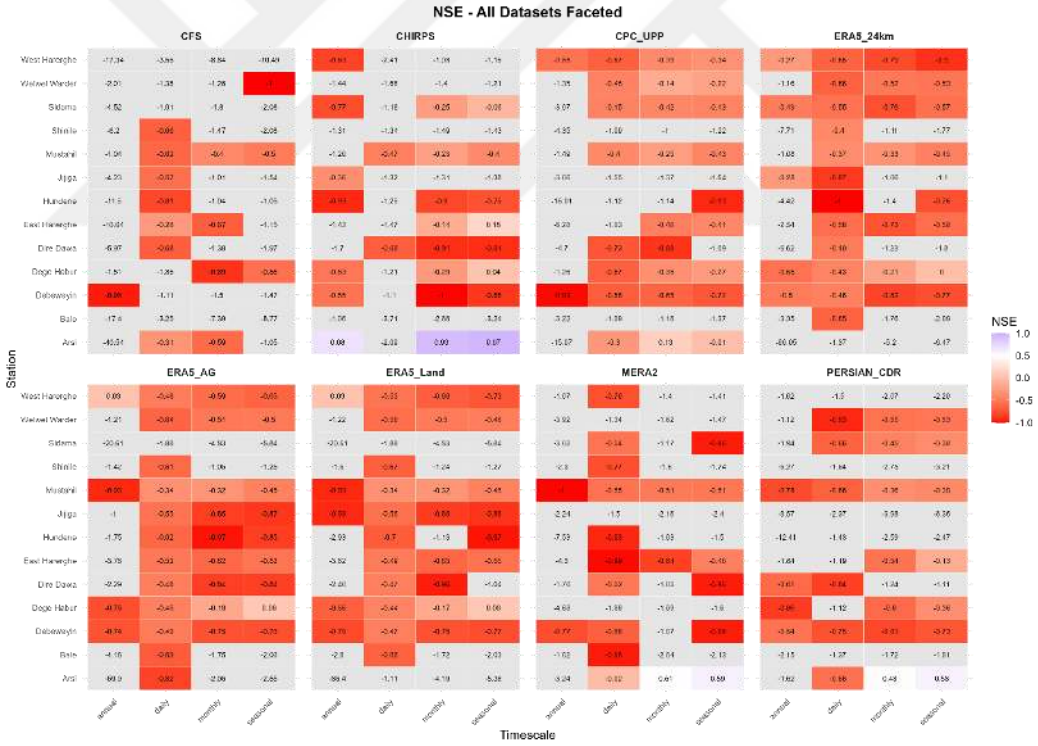

Figure 4.34. *Heatmap evaluation of continuous performance metrics (NSE) for temporal station level comparison across all datasets*

Figure 4.35 (Correlation) values vary from about –0.55 (red) to +0.99 (purple). Notably, some of the datasets achieve very high correlation at the seasonal scale (column 4): nearly every station is ≥0.70 (purple) across panels (e.g. CHIRPS Arsi = 0.99, ERA5 based Arsi = 0.87-0.90). Annual correlations (column 1) are mostly negative (red) for many

stations in CFS/CPC, but stations like Arsi and Bale shows very good across all datasets. Daily correlations (column 2) show the largest dataset differences: ERA5_based yields a moderate daily R at Arsi (0.38). The dominant pattern is that temporal aggregation boosts correlation: monthly and especially seasonal R are uniformly high for all datasets, while annual/daily R vary and highlight station-specific performance differences. Each dataset shows its own nuances (e.g. CHIRPS has the highest R for "Arsi" daily at 0.71), but seasonal R is strong across the board.

Figure 4.35. *Heatmap evaluation of continuous performance metrics (R) for temporal station level comparison across all datasets*

4.4. Spatial Performance Metrics Across Datasets

4.4.1. Region level spatial performance

4.4.1.1. *Categorical event detection metrics*

Figure (4.36 to 4.41) presents the spatio-temporal categorical performance metrics (CSI,POD,FAR, HSS, SEDI, SR) respectively. The CSI maps (Figure 4.36) indicate that most datasets achieve relatively high CSI (yellow hues) across the Shabelle Basin at multiple time scales, reflecting good overall event prediction skill. In particular, CHIRPS and ERA5-based datasets show uniformly strong CSI (bright yellow) for Annual,

Monthly, and Daily aggregations, suggesting they successfully capture a large fraction of observed rain events. CFS is also consistently high (yellow) in CSI across the basin, with minimal spatial variation – its performance is uniformly strong in both northern and southern parts of the basin. In contrast, CPC_UPP exhibits a pronounced drop in CSI at the Seasonal scale, visible as a dark purple basin-wide (very low CSI) in the Seasonal row for that dataset. This indicates that CPC_UPP struggles to capture seasonal rainfall events (likely missing many events or hits). In general, CHIRPS and ERA5-based datasets exhibit consistently high CSI (strong event capture) across all time frames, CFS also performs well, while CPC_UPP markedly under-performs in seasonal event detection, and MERRA2 is slightly lower than the top performers. This suggests that gauge-calibrated or high-resolution datasets (CHIRPS, ERA5-Land) have superior overall event prediction capability in the region, whereas CPC_UPP's ability to detect rainfall occurrences is unreliable at certain temporal scales.

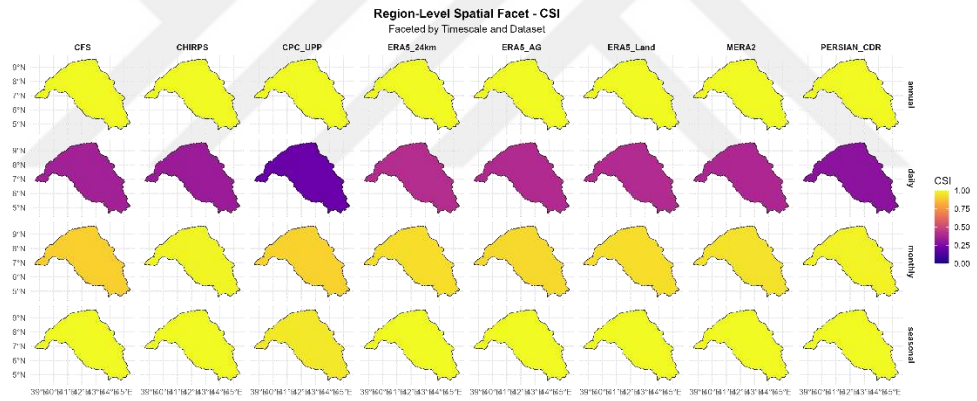

Figure 4.36. *Spatio-temporal distribution of the Critical Success Index (CSI) for all datasets*

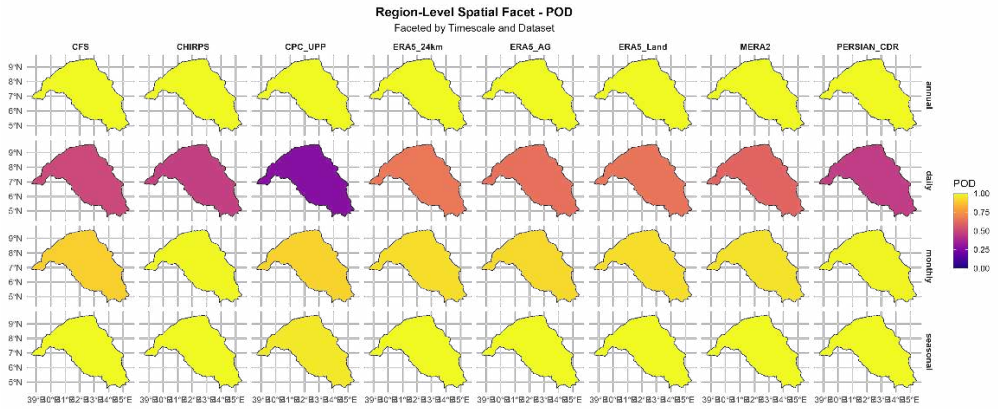

Figure 4.37. *Spatia-temporal distribution of the Probability of Detection (POD) for all datasets*

The POD maps show a pattern broadly similar to CSI, since POD is another measure of hit rate. Across Annual and Monthly scales, nearly all datasets achieve high POD (widespread yellow coloring), indicating that on aggregate they detect most precipitation events over the basin. For example, CHIRPS and ERA5-Land have uniformly yellow maps in all rows, signifying excellent detection (they capture almost all observed rain occurrences even down to the daily scale). CFS and ERA5_24km/ERA5_AG also maintain strong detection skill (mostly yellow-orange) with little spatial variation. However, CPC_UPP again stands out negatively at Seasonal scale – its Seasonal POD map is colored deep purple, meaning CPC_UPP misses a large portion of seasonal rainfall events across the entire basin. PERSIANN-CDR and MERRA2 also show somewhat reduced POD (pinker tones) at Seasonal and Daily scales compared to the leading datasets: for instance, PERSIANN's Daily POD is a bit lower (light orange) in parts of the basin, suggesting it fails to detect some daily rain events, and MERRA2's Seasonal POD is on the low side (purplish) relative to others. CHIRPS and ERA5-Land achieve the highest POD (≈yellow, near 1.0) consistently, even at daily scale, indicating very robust rain event detection. CFS and other ERA5 versions also perform well with only minor drop-offs at finer scales (remaining orange-yellow). CPC_UPP has a serious weakness in detection at Seasonal scale (very low POD), while PERSIANN-CDR and MERRA2 have moderate detection skill (notably lower than CHIRPS/ERA5 at some scales). This suggests that gauge-informed products and modern reanalyses capture rainfall occurrences more completely, whereas satellite-only (PERSIANN) and some older reanalysis/gauge products (MERRA2, CPC) may miss events, especially seasonally.

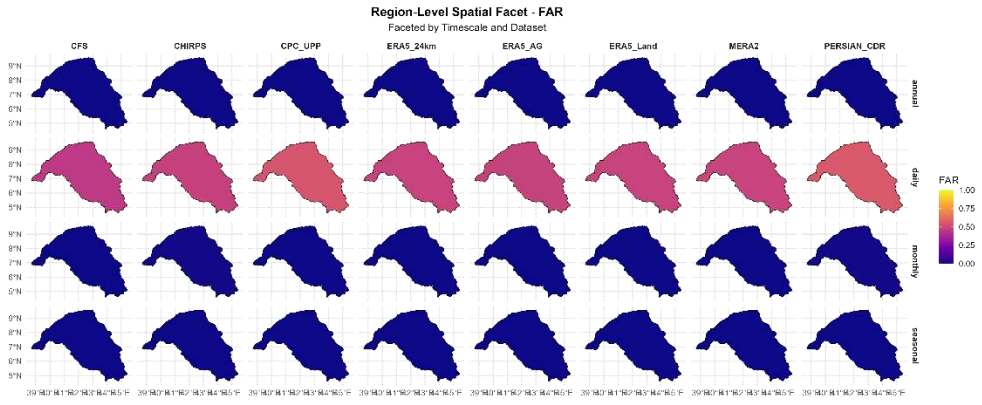

Figure 4.38. *Spatio-temporal distribution of the False Alarm Ratio (FAR) for all datasets*

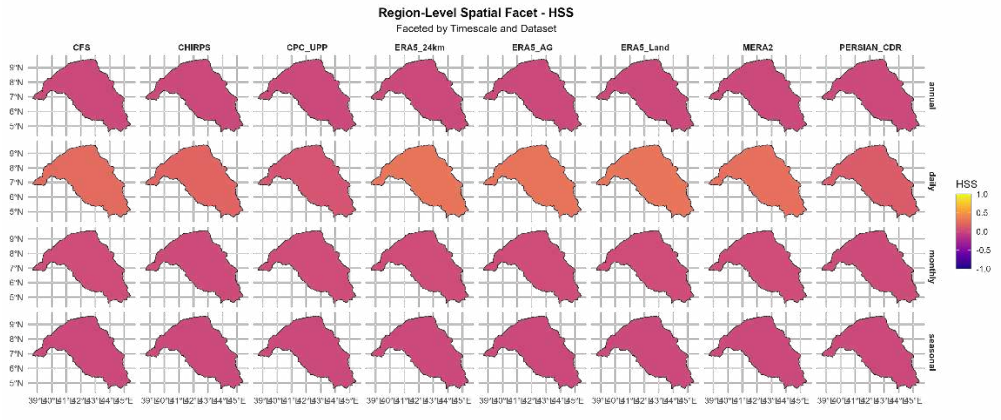

Figure 4.39. *Spatio-temporal distribution of the Heidke Skill Score (HSS) for all datasets*

The overall FAR metrics , most datasets exhibit predominantly dark blue/purple maps at Annual, Monthly, and Daily scales, implying they do not suffer from excessive false rainfall events on those scales. For example, CHIRPS and ERA5-Land are uniformly deep blue across all time frames, indicating these datasets rarely predict rain that does not occur (very few false alarms throughout the basin). CFS and ERA5_24km/AG similarly show low FAR (maps mostly blue) at most scales, with only slight increases at Seasonal or Daily scales (a hint of purple-pink in some areas). The Seasonal scale exposes clearer differences: CPC_UPP and PERSIANN-CDR have notably higher FAR during the Seasonal period, visible as broad pinkish maps, MERRA2 also shows elevated FAR at Seasonal scale (light pink), though not as extreme as CPC. CHIRPS and ERA5-based datasets demonstrate consistently low false alarm rates (FAR close to 0, dark blue) at all time scales, making their rain occurrence predictions reliable. CFS also keeps FAR low overall. On the other hand, CPC_UPP and PERSIANN-CDR show clear weaknesses with high FAR, particularly on the Seasonal scale (indicating many spurious rain events predicted), and to a lesser degree at Daily scale. MERRA2 has moderate FAR issues seasonally as well.

Spatial patterns of HSS are again quite uniform across the basin – each dataset's map is mostly a single hue – so the skill variations are largely by dataset and time scale. Seasonal HSS brings out notable contrasts: in the Seasonal row, CFS, CHIRPS, and ERA5-Land display warmer orange shades, indicating relatively higher skill (HSS perhaps 0.4–0.5) in predicting rain occurrence beyond chance. ERA5_24km and ERA5_AG also show moderate orange-pink in Seasonal HSS, suggesting decent skill. In contrast, CPC_UPP, MERRA2, and PERSIANN-CDR have cooler pink tones at Seasonal

scale (some tending toward purple), implying lower skill (HSS near 0 or only slight positive skill) for those datasets. CHIRPS and the ERA5 family (particularly ERA5-Land) attain the highest HSS values, especially at Seasonal resolution (their maps turn light orange, denoting appreciable skill). CFS also achieves a relatively good HSS at Seasonal scale. CPC_UPP, PERSIANN-CDR, and MERRA2 show weaker skill (pinkish low HSS) at nearly all scales, indicating their rain forecasts are only marginally better than random in a categorical sense. Thus, in terms of overall categorical forecast accuracy, the gauge-informed and high-res reanalysis products have a clear advantage, while CPC_UPP and PERSIANN struggle to provide skillful rain/no-rain predictions in the basin.

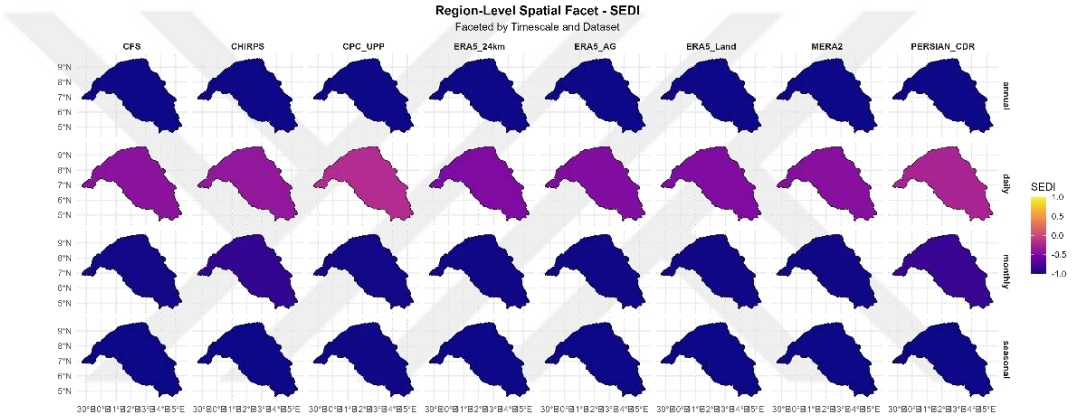

Figure 4.40. *Spatio-temporal distribution of the Symmetric Extremal Dependence Index (SEDI) for all datasets*

ERA5_24km, and ERA5_Land show uniformly deep blue SEDI maps at almost all scales, meaning these products have negligible extreme-event skill (they likely miss or randomly hit heavy rain extremes, yielding SEDI near 0). CHIRPS and ERA5-AG fare slightly better: for example, CHIRPS at Seasonal scale has a hint of purple-pink in its SEDI map, indicating marginally positive skill for extremes in that season (perhaps due to its incorporation of station data capturing some heavy events). Nearly all datasets show weak skill in predicting extreme rainfall events (SEDI close to 0, dark blue), with the exception of CPC_UPP and PERSIANN-CDR, which exhibit a modest level of skill (SEDI slightly positive, pinkish) in seasonal extreme rainfall detection. CHIRPS and ERA5-AG also have slight positive SEDI in some seasons but remain lower than CPC and PERSIANN in this metric.

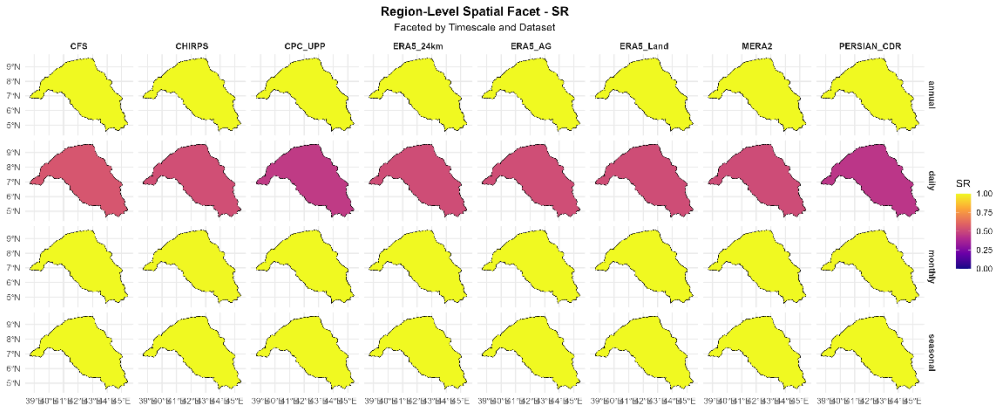

Figure 4.41. *Spatio-temporal distribution of the Succes Ratio (SR) for all datasets*

The SR maps (Figure 4.41) confirm the false alarm patterns from another angle. Since SR = 1 – FAR, a yellow SR map corresponds to high reliability (few false alarms), and purple indicates low reliability (many false alarms). As expected, datasets that had deep blue (low FAR) now show bright yellow (high SR). CHIRPS and the ERA5 family achieve the highest success ratios (near 1, yellow) throughout, indicating their rain predictions are very trustworthy with minimal false alarms. CFS also has a high SR in general. PERSIANN-CDR and CPC_UPP are the weakest in SR, especially at Seasonal scale (drops to purple, meaning nearly half or more of their predicted events did not materialize), and MERRA2 has intermediate SR (good annually/daily, but somewhat lower seasonally).

4.4.1.2. *Continous metrics*

The KGE map (Figure 4.42) shows strong differences in overall continuous performance (how well each dataset's precipitation time series matches observed values in terms of correlation, mean, and variability). CHIRPS stands out with consistently high KGE across all most all time scales and locations. ERA5-Land and ERA5-24km also perform well, especially at finer scales: their Monthly and Daily KGE are orange to yellow across the basin, indicating good correspondence with observed rainfall patterns. On the other hand, CFS and MERRA2 exhibit a striking scale-dependent pattern: at Annual scale, CFS and MERRA2 have very low KGE (dark purple maps). However, as the time scale shortens, CFS's KGE improves dramatically – by Monthly and Daily, the CFS maps turn orange-yellow, meaning that CFS captures short-term variations

reasonably well despite failing at annual totals. CHIRPS achieves the highest KGE (close to 1) consistently, followed by ERA5-Land and ERA5's other variants, which also maintain high efficiency especially at monthly/daily scales. CFS and MERRA2 are notable for poor long-term (annual) performance but reasonable short-term performance – an indication that they capture event-to-event variability but have biases in accumulated rainfall. CPC_UPP lags behind in KGE at all scales, underscoring its difficulties in matching observed magnitudes and patterns, and PERSIANN-CDR performs moderately, generally improving at finer timescales to a fair level.

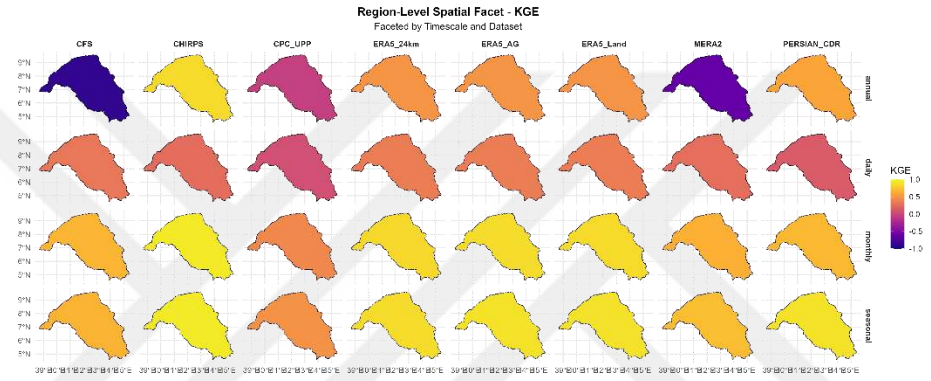

Figure 4.42. *Spatio-temporal distribution of the KGE across datasets*

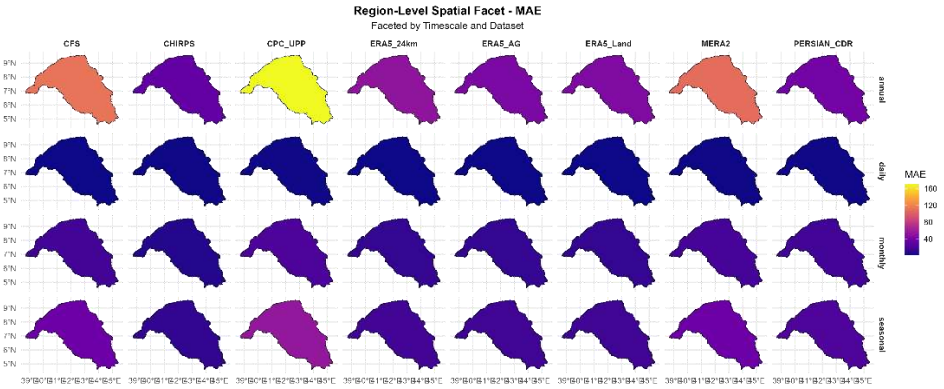

Figure 4.43. *Spatio-temporal distribution of the MAE across datasets*

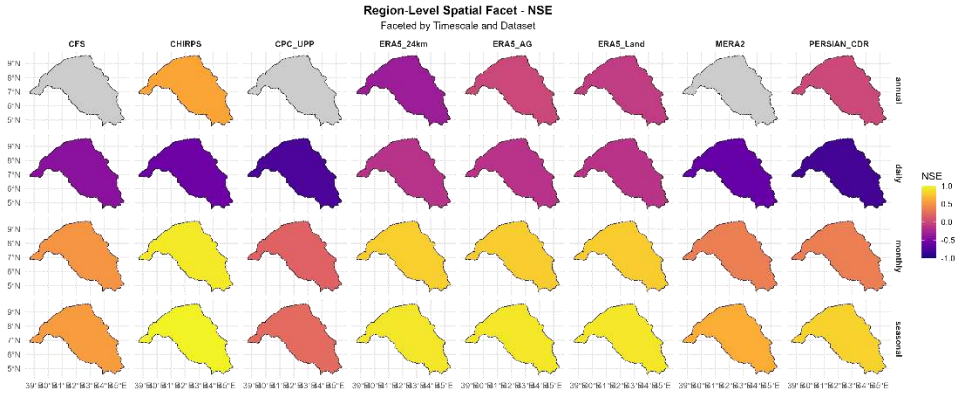

Figure 4.44. *Spatio-temporal distribution of the NSE across datasets*

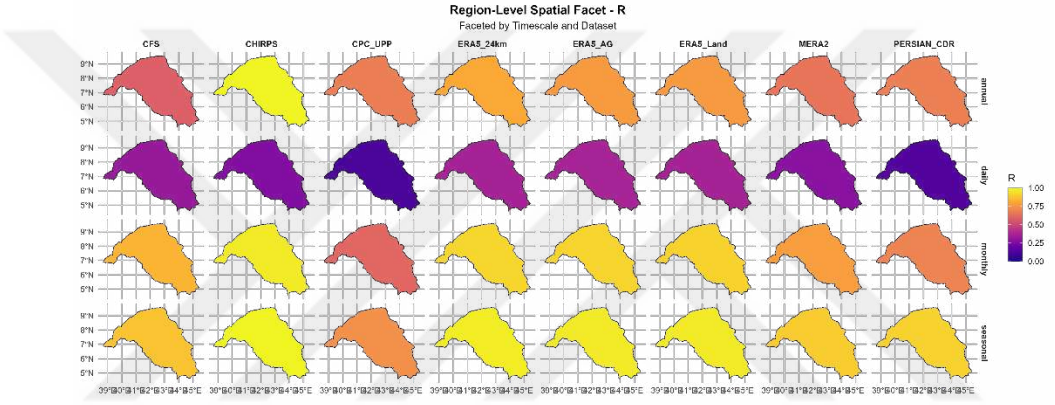

Figure 4.45. *Spatio-temporal distribution of the R across datasets*

The NSE maps highlight how well each dataset's time series matches observed rainfall in terms of overall variance and bias (with sensitivity to large errors). CHIRPS emerges as the top performer, especially at finer scales: at Monthly and Daily scales, CHIRPS shows bright yellow across the basin, meaning its NSE is very high (close to 0.8–0.9) – essentially CHIRPS can predict daily and monthly rainfall with only small errors relative to natural variability. ERA5-Land also achieves high NSE, particularly at Daily scale (yellow map), and CFS reaches bright yellow at Daily scale as well, indicating these two can explain a large fraction of daily rainfall variance (a notable result for CFS). However, at coarser scales (Annual, Seasonal), many datasets have low or even negative NSE. For instance, CFS, CPC_UPP, and MERRA2 exhibit negative NSE at Annual scale. ERA5_24km and PERSIANN-CDR are slightly negative or near zero NSE at Annual scale (light purple/gray), suggesting marginal skill in year-to-year rainfall totals. CHIRPS provides the highest NSE (best overall variance-explained) at nearly all scales, with

ERA5-Land a close second at daily scale. CFS is paradoxically poor for long-term accumulations (negative NSE annually) but excellent at daily NSE, indicating it captures short-term events but not climatological totals. ERA5 (24km and AG) and PERSIANN show moderate skill by monthly/daily, though they fail to capture annual variability well. CPC_UPP is consistently among the lowest NSE, and MERRA2 also struggles (negative or low NSE except marginally positive at daily).

İnterms of Mean Absolute Error (MAE) CHIRPS has consistently the lowest errors (it's the darkest purple in virtually all panels, indicating superior accuracy in precipitation amounts). ERA5 (especially the 24km version) also achieves very low errors, close to CHIRPS. ERA5-Land and ERA5-AG have slightly higher MAEs, implying a small positive bias or variance issue but still solid performance. PERSIANN-CDR's errors are somewhat larger than CHIRPS/ERA5 but moderate (not extreme). On the other hand, CPC_UPP clearly has the largest errors (particularly for accumulated periods, as seen by the bright color at Annual), pointing to significant biases in that product's rainfall estimates over the basin. MERRA2 and CFS also have high errors at longer time scales (annual/seasonal), though MERRA2's daily errors are low in absolute terms. In summary, in terms of raw error magnitude, CHIRPS and ERA5 datasets are the most accurate (lowest MAE), PERSIANN and CFS are intermediate, and CPC_UPP (and to a degree MERRA2) are the least accurate, with large deviations in total rainfall. This aligns with earlier findings that gauge-calibrated products excel in both relative and absolute accuracy, while CPC_UPP's issues translate into large absolute rainfall errors.

CHIRPS exhibits the highest correlations with observed rainfall across all scales, essentially setting the benchmark (especially excellent on annual and monthly patterns). ERA5-Land and ERA5 (AG/24km) have very high correlation as well, particularly at daily scale, reflecting good capture of event timing. CFS is notable for extremely poor correlation on seasonal totals (suggesting mis-timed seasonal rains) but surprisingly strong daily correlation, highlighting a discrepancy between climate-phase alignment and weather-scale alignment for this model. PERSIANN-CDR and MERRA2 reach reasonably high daily correlations (roughly 0.8) but have difficulties with seasonal to annual correlation, indicating some mismatch in capturing longer-term trends or cycles. CPC_UPP has the weakest correlation overall, failing to reliably track observed rainfall patterns on seasonal and even monthly scales, and only achieving moderate alignment at daily scale. These correlation results reinforce earlier findings: gauge-blended and

modern reanalysis datasets (CHIRPS, ERA5) closely mirror observed rainfall timing and distribution, whereas CPC and some reanalyses (CFS, MERRA2) have significant timing or phase errors at larger temporal scales.

4.4.2. Station level spatial performace

4.4.2.1. *Categorical metrics and continuous metrics*

At the daily scale, pronounced differences emerge among the eight precipitation datasets in their ability to detect rainfall events. Satellite–gauge products like CHIRPS exhibit the highest event detection skill – yielding superior critical success index (CSI) and probability of detection (POD) with relatively low false alarm rates (FAR) – reflecting an aptitude for capturing convective rain occurrences in the basin(Ghajarnia et al., 2022; Zargar et al., 2025). PERSIANN-CDR similarly demonstrates strong rain day detection (e.g. high CSI/POD and the highest skill in distinguishing rainy vs. non-rainy days among satellite estimates) (Ghajarnia et al., 2022); however, PERSIANN's large positive bias (~50% in one regional study) indicates it tends to overestimate volumes, which diminishes its accuracy in terms of continuous metrics like MAE and efficiency scores. The ERA5-based reanalyses (ERA5 at 24 km, its land-surface refined version ERA5-Land, and a gauge-adjusted variant ERA5_AG) achieve high correlation (R) and skill at monthly to annual aggregates – on par with the top satellite products – but they struggle more at daily time steps (Gebrechorkos et al., 2024) . In particular, ERA5 and its variants often produce spurious light precipitation (a known reanalysis issue in tropical Africa), leading to more frequent false alarms and lower daily Heidke skill scores (HSS). Nonetheless, their overall performance improves markedly with temporal averaging: by the seasonal and annual scale, ERA5 (especially the bias-corrected and land-downscaled versions) shows robust agreement with gauges, evidenced by relatively high Nash–Sutcliffe efficiency (NSE) and Kling–Gupta efficiency (KGE) scores and low mean absolute error (MAE) . The finer 9 km resolution of ERA5-Land provides a slight edge in capturing spatial rainfall gradients over complex terrain (e.g. the Ethiopian highlands), yielding marginally better station-wise correlations and lower error at those locations than the course 24 km ERA5. A gauge-informed reanalysis like ERA5_AG further reduces biases, translating into improved KGE and NSE relative to the raw ERA5 product (consistent with the notion that model data with bias correction can outperform unadjusted outputs in certain cases). By contrast, the CPC Unified gauge analysis –

despite its ground-truth basis – performs worse in this sparsely gauged region. The CPC product's coarse spatial grid (0.5°) and the paucity of local stations lead to under-representation of Shabelle's localized rain events, resulting in lower correlation and efficiency (often even below the reanalyses) and higher bias on monthly/annual totals. The MERRA-2 and CFS reanalyses consistently rank as the poorest performers across most metrics and timescales. MERRA-2 exhibits a systemic dry bias (marked underestimation of observed rainfall), which, along with only moderate correlation, yields depressed NSE and KGE values. The older CFS dataset (CFSR) also struggles, generally showing weak agreement with station data and larger errors; this is in line with broader findings that newer reanalyses like ERA5 have markedly improved upon their predecessors across all performance metrics. Notably, all datasets show some improvement in aggregate statistics at coarser temporal scales – even the weaker products achieve higher R and NSE by the seasonal and annual level as random daily errors cancel out – but the relative ranking of performance holds steady. In summary, the station-based evaluation indicates that CHIRPS is the most consistently reliable dataset for the Shabelle basin (excelling in both event detection and rainfall amount metrics at all time scales), closely followed by the ERA5 family (particularly the ERA5-Land and gauge-adjusted versions, which correct some of ERA5's biases). In contrast, MERRA-2 and CFS emerge as the least accurate, with CPC's gauge-only product also underperforming due to sparse data, especially when compared to the satellite-era datasets.

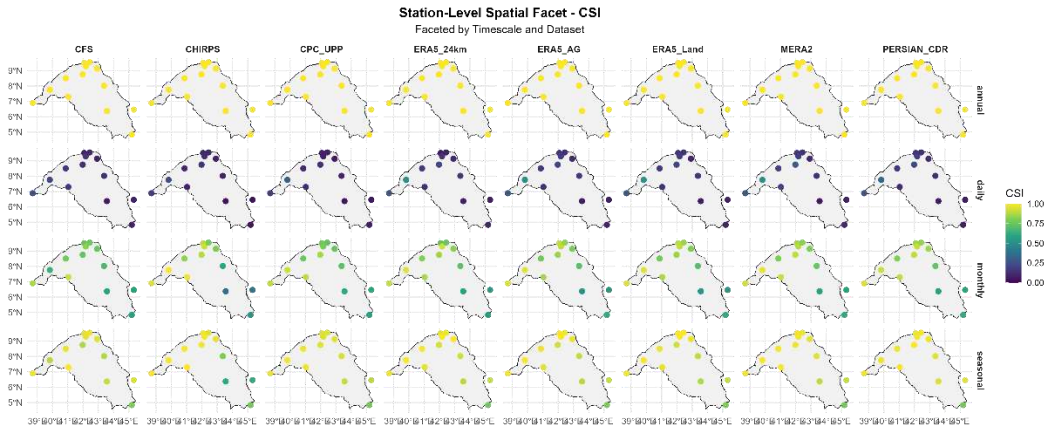

Figure 4.46. *Spatio-temporal distribution of station levels of categorical metrics (CSI, FAR, HSS, POD, SEDI and SR) respectively*

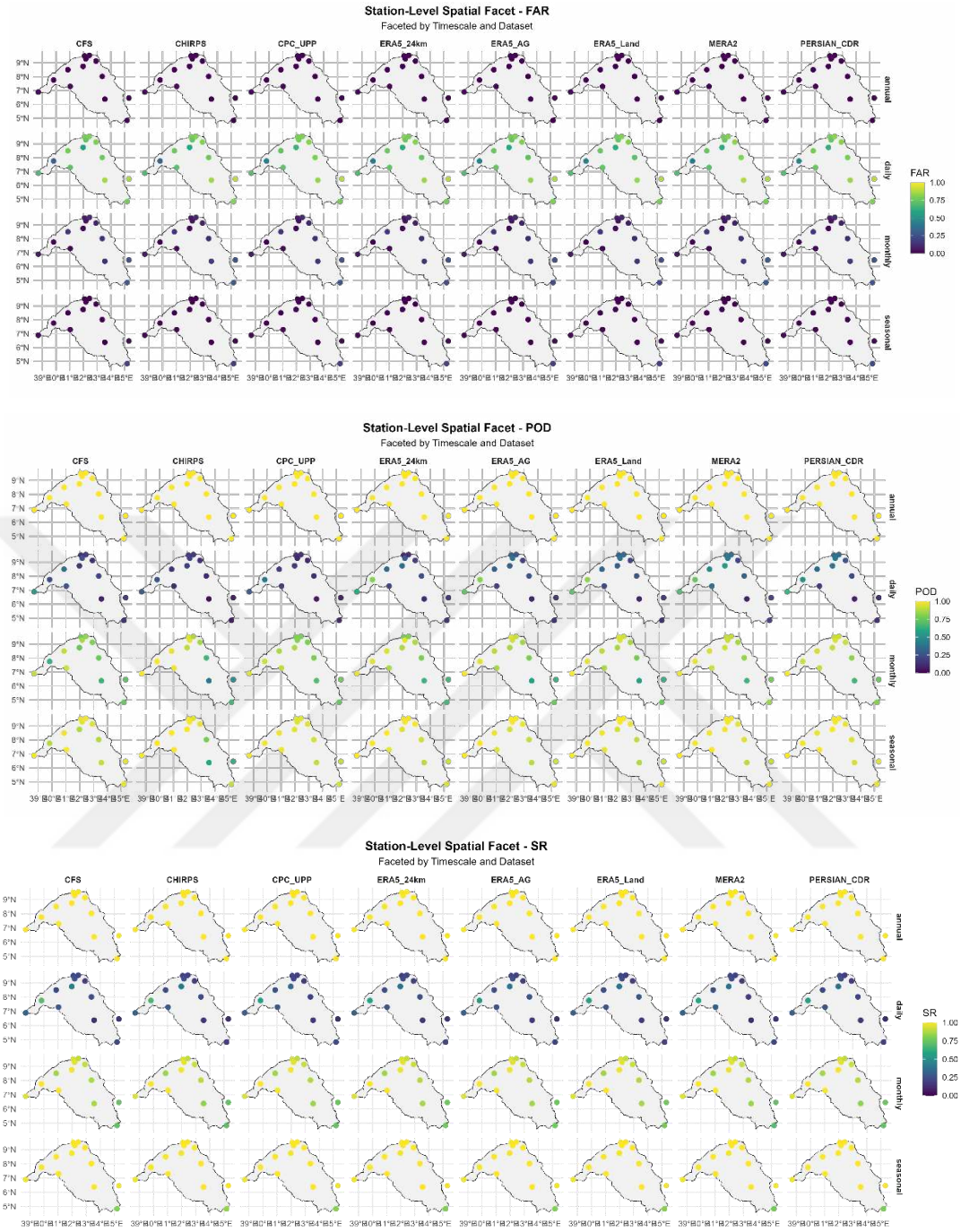

Figure 4.46. (*Continued) Spatio-temporal distribution of station levels of categorical metrics (CSI, FAR, HSS, POD, SEDI and SR) respectively*

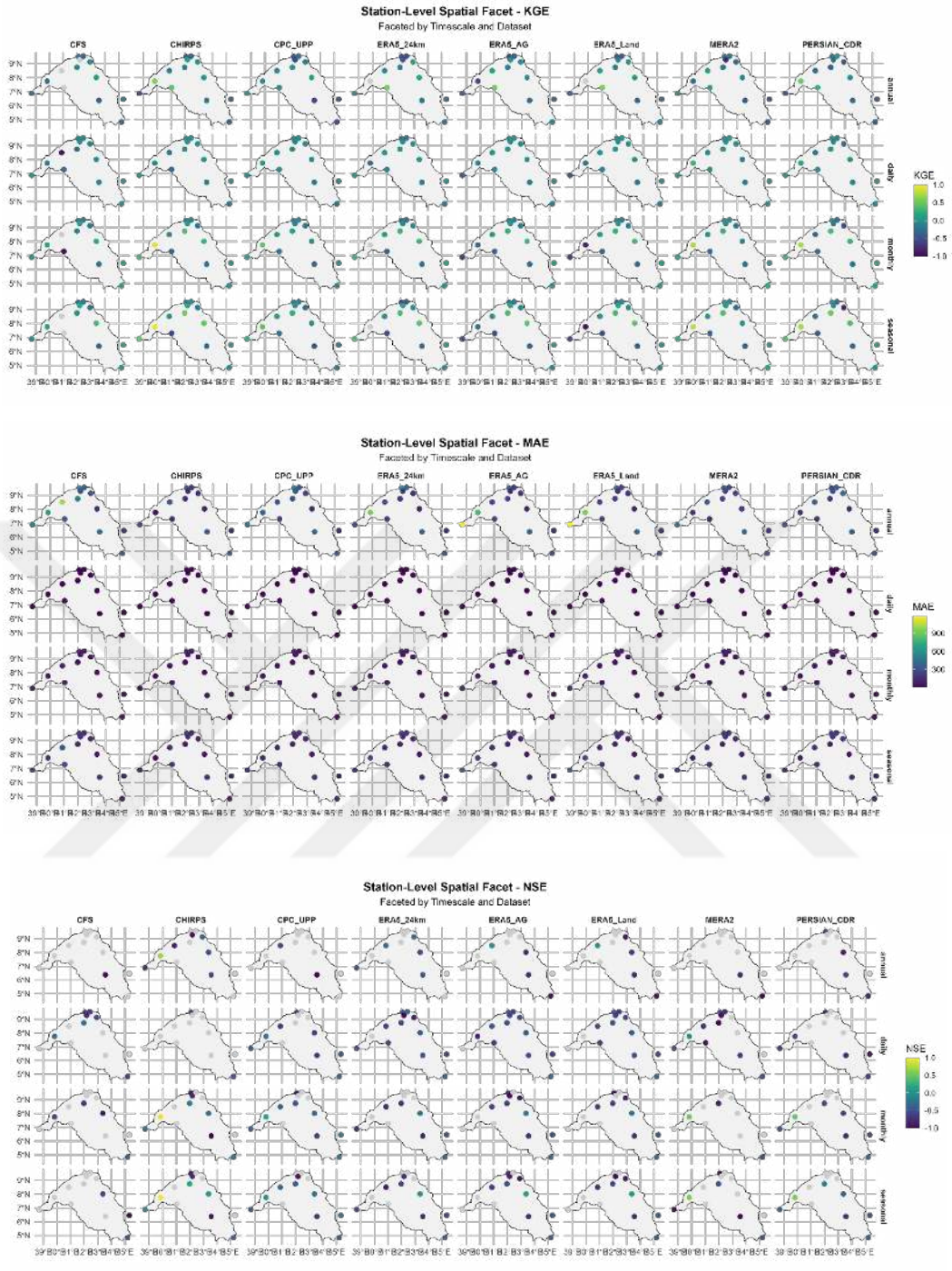

Figure 4.46. (*Continued) Spatio-temporal distribution of station levels of categorical metrics (CSI, FAR, HSS, POD, SEDI and SR) respectively*

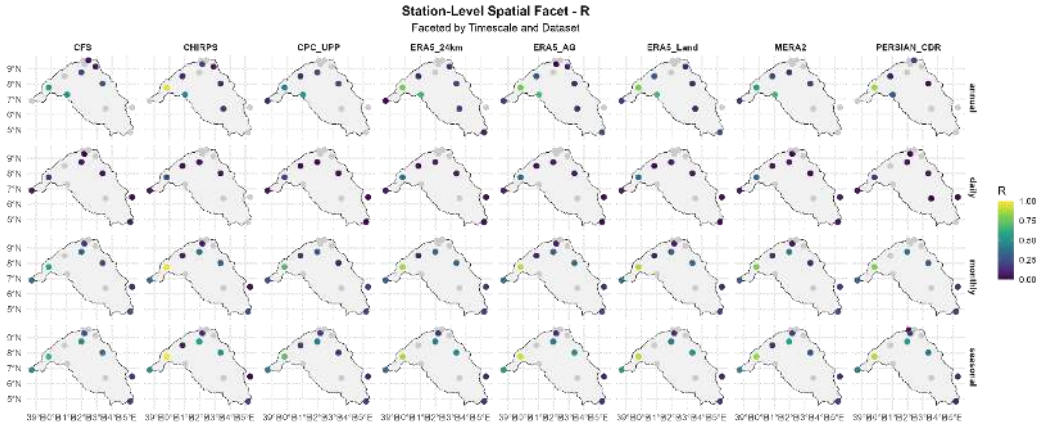

Figure 4.47. *Spatio-temporal distribution of station levels of continuous metrics (KGE, MAE, NSE, R) respectively*

In summary**,** CHIRPS and the ERA5-based datasets demonstrate the best overall skill and consistency across daily to annual scales. PERSIANN-CDR shows excellent rain-day detection (rivaling CHIRPS) but tends to overestimate totals. The ERA5 reanalyses (notably ERA5-Land and a gauge-adjusted ERA5) perform strongly at monthly to annual scales, although frequent light-rain false alarms limit their daily HSS and CSI. In contrast, MERRA-2 and CFS (an older reanalysis) perform worst on most metrics, and even the gauge-based CPC product struggles in this region due to its coarse resolution and sparse station data. Thus, modern satellite-gauge products and advanced re-analyses clearly outperform the older or gauge-only datasets in the Shabelle basin.

The ERA5 datasets, including ERA5-Land and a gauge-adjusted ERA5, provide robust results, especially at the monthly to annual levels, although they occasionally struggle with frequent light-rain false alarms. This limits their performance on daily scales, as indicated by their lower scores in Heidke Skill Score (HSS) and Critical Success Index (CSI). On the other hand, PERSIANN-CDR demonstrates impressive rain-day detection, performing comparably to CHIRPS, but tends to overestimate precipitation totals, affecting its overall accuracy. In contrast, older reanalysis products such as MERRA-2 and CFS show the poorest performance on most metrics. The gauge-based CPC product also faces limitations, largely due to its coarse resolution and sparse station data. This comparison clearly highlights that modern satellite-based products and advanced reanalysis datasets provide superior performance in the Shabelle Basin when compared to older or solely gauge-based datasets.

4.5. Seasonal Performance of the Datasets and Their Validation Metrics

To evaluate dataset performance over the Shabelle Basin, we analyze spatial maps of six categorical metrics .

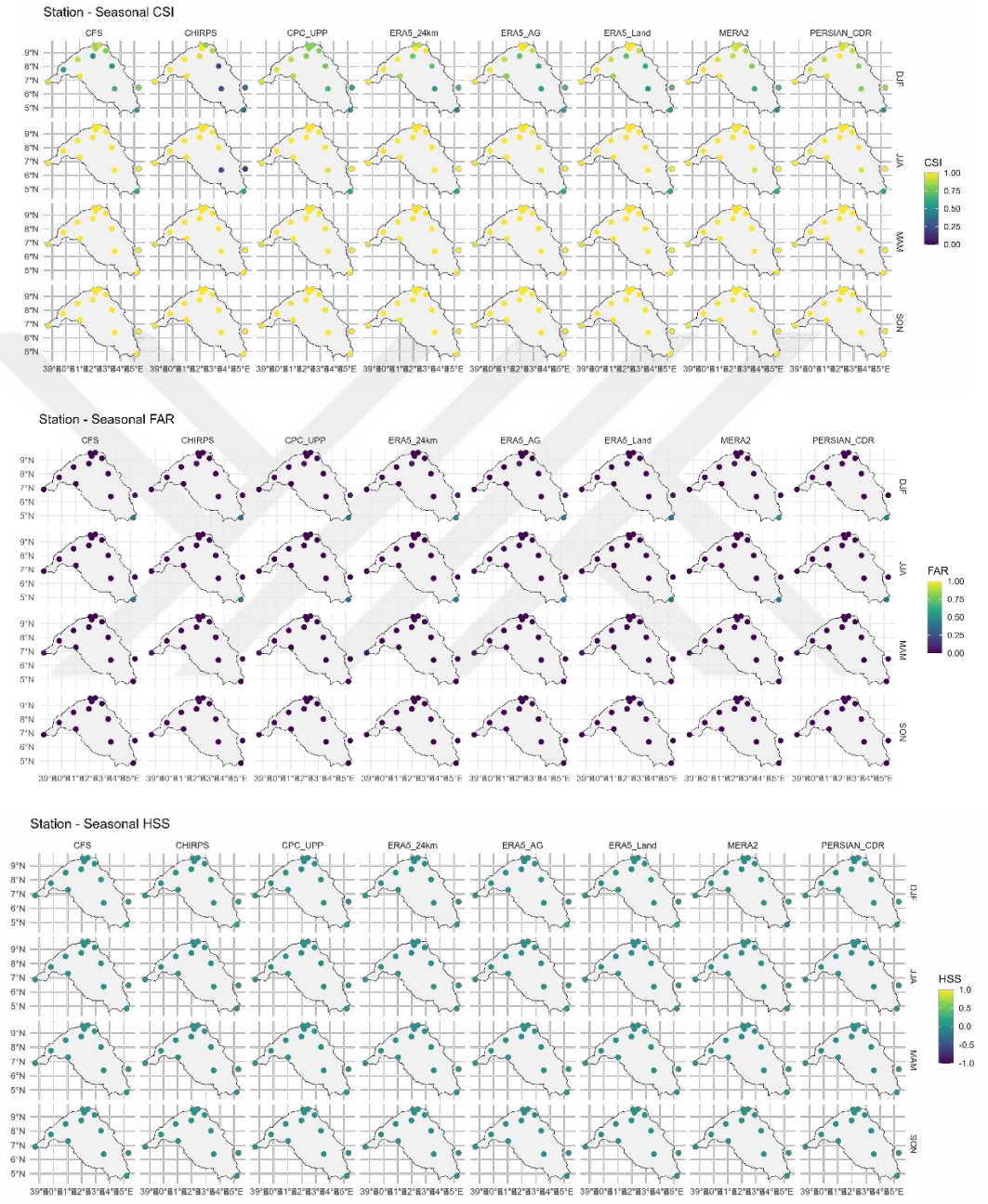

Figure 4.48. *Spatio-temporal distribution of seasonal performance of both station and region levels of categorical and continuous metrics (CSI, FAR, HSS, POD, SEDI, SR, KGE, NSE, MAE, R) respectively*

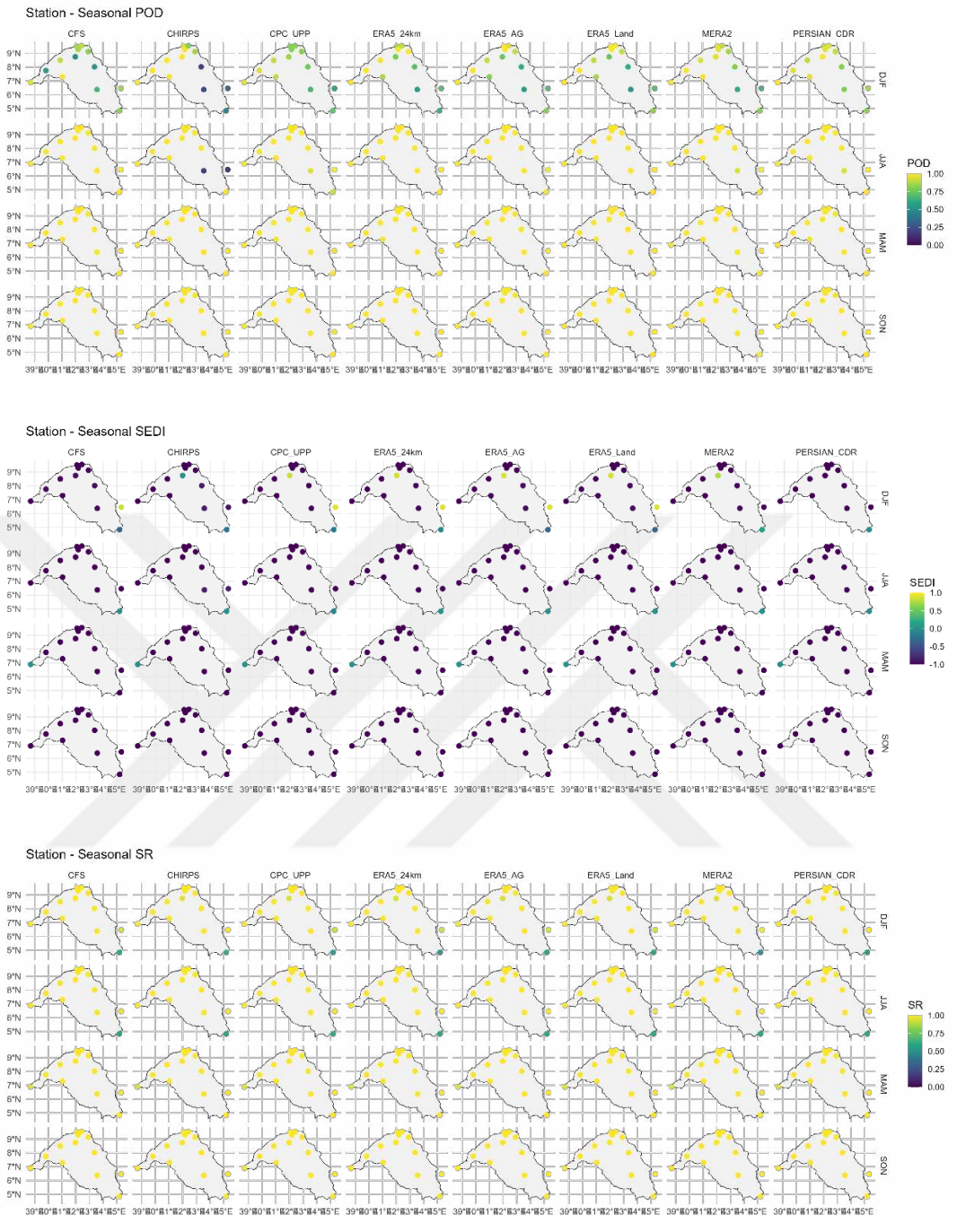

Figure 4.48. *(Continued) Spatio-temporal distribution of seasonal performance of both station and region levels of categorical and continuous metrics (CSI, FAR, HSS, POD, SEDI, SR, KGE, NSE, MAE, R) respectively*

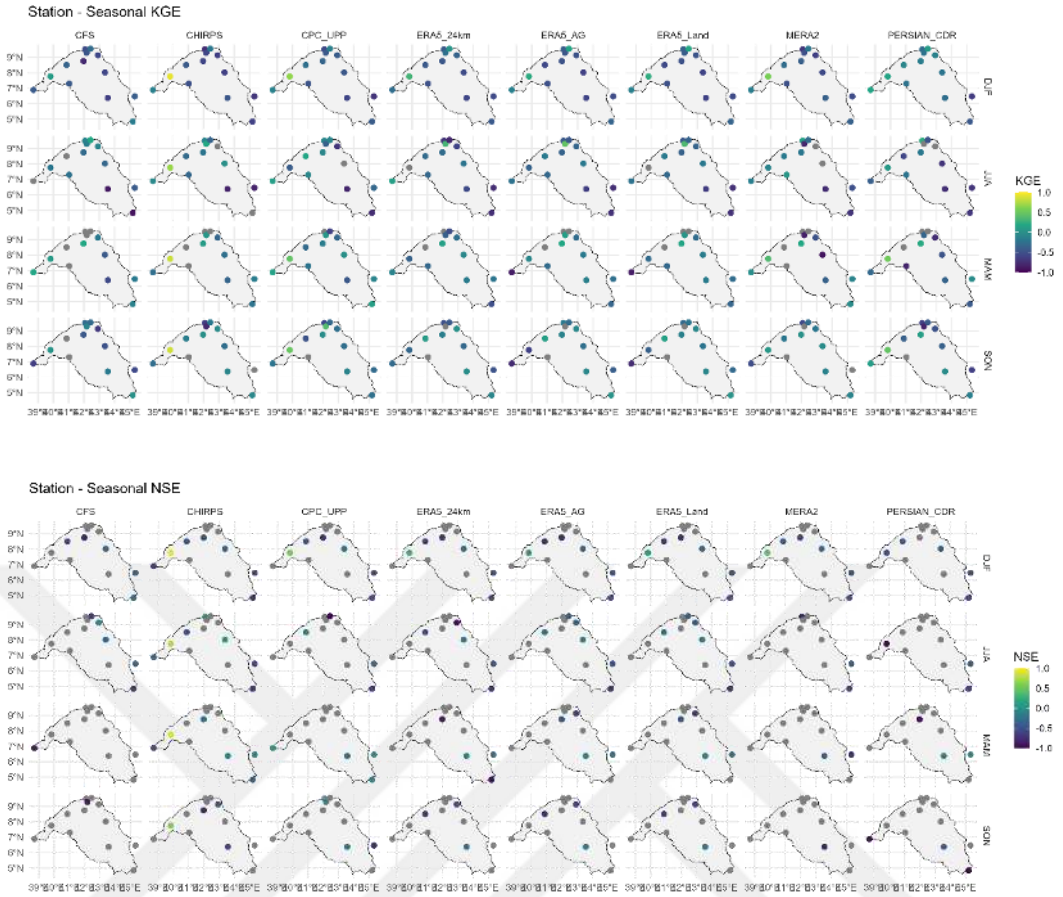

Figure 4.48. *(Continued) Spatio-temporal distribution of seasonal performance of both station and region levels of categorical and continuous metrics (CSI, FAR, HSS, POD, SEDI, SR, KGE, NSE, MAE, R) respectively*

The six metrics namely: Critical Success Index (CSI), Probability of Detection (POD), False Alarm Ratio (FAR), Heidke Skill Score (HSS), Symmetric Extreme Dependency Index (SEDI), and Success Ratio (SR) – alongside four continuous metrics – Kling-Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), Pearson correlation coefficient (R), and Mean Absolute Error (MAE). These metrics are computed at rain gauge locations for multiple precipitation datasets (CFS, CHIRPS, CPC Unified, ERA5 at 24 km, ERA5-Ag, ERA5-Land, MERRA-2, and PERSIANN-CDR) and are visualized by season: DJF (December–February), MAM (March–May), JJA (June–August), and SON (September–November). Each figure is a faceted map grid with seasons as rows and datasets as columns (e.g., Figure 4.48 Represents the catergorical and contnuous*)*. The analysis below compares and contrasts the datasets' performance for each season, strictly based on the visual evidence in these figures (without additional assumptions).

4.5.1. DJF (December–February)

During the DJF dry season, rainfall events are infrequent and generally light. Consequently, all datasets show relatively low CSI and HSS values overall (indicating limited opportunity for high skill), and differences among datasets are subtle in this season (Figure 4.48). Most products achieve moderately high POD with correspondingly low FAR, suggesting that when few rainfall events do occur in DJF, they are often detected with minimal false alarms. Indeed, the POD maps for DJF (not shown) are largely yellow, indicating detection probabilities on the order of ~0.8–1.0 for most datasets, while the FAR maps are uniformly dark (FAR near 0 for almost all cases). These uniformly low FAR values translate to high Success Ratios across datasets, meaning false alarms are rare in DJF. Such consistency implies that in the absence of significant rainfall, even simpler satellite estimates can avoid false rain detection, likely because clear-sky conditions dominate. Spatially, no pronounced gradient in skill is evident across the basin in DJF – most stations show uniformly modest performance metrics. This uniformity is expected given DJF corresponds to the dry Bega season in Ethiopia when both observed and estimated rain amounts are minimal.

Although differences are small, CHIRPS appears to exhibit the highest overall skill in DJF. For instance, CHIRPS yields slightly higher CSI values at many gauge sites compared to other datasets, and its KGE and NSE, while low in absolute terms, are less negative than those of other products. This suggests CHIRPS captures the occasional DJF rainfall events somewhat more reliably. Such superior performance of CHIRPS aligns with broader evaluations in Ethiopia that have found CHIRPS to outperform other gridded products in capturing rainfall variability (Hordofa et al., 2021; Kabite Wedajo et al., 2021). In particular, CHIRPS's blending of satellite estimates with gauge data likely helps detect even light rains when they occur (Hordofa et al., 2021). By contrast, the CFS reanalysis shows the weakest performance in DJF. Visual evidence of this is seen in its CSI and HSS panels (first column of *Figure 4-45*, DJF row), which are tinged more toward purple (lower skill) relative to others. Continuous metrics reinforce this: CFS's NSE in DJF is near or below zero at most stations (indicating no skill or worse-than-mean performance), and its KGE is among the lowest (often negative, implying poor correlation or bias issues in this dry season). This underperformance may stem from the reanalysis producing spurious drizzle or mis-timing the rare events – a known issue for some model-based datasets during dry periodsaimspress.com. Overall, for DJF the best-performing

dataset is CHIRPS, while CFS is the poorest, though it should be noted that all datasets struggle to achieve high skill due to the scarcity of rain in this season. Spatially, there are no stark regional differences in DJF performance – the entire basin sees limited rainfall, and all datasets handle the dry conditions with comparable (if modest) accuracy.

4.5.2. MAM (March–May)

The MAM season marks the spring "Belg" rains (and the Gu rains in the Somali region), bringing a notable increase in rainfall across the Shabelle Basin. During MAM, the performance metrics improve for all datasets relative to DJF, but clear differences emerge between products. CHIRPS remains the top performer in MAM, showing consistently high CSI and POD across the basin along with low FAR. Visually, CHIRPS's CSI map for MAM is dominated by green-yellow symbols (indicating CSI values generally above ~0.5 at many stations), in contrast to some other datasets that show more mixed colors. CHIRPS achieves high event detection with minimal false alarms, yielding HSS scores that are among the best of the group. This strong performance is evident in continuous metrics as well: CHIRPS's KGE is significantly higher than others at several stations (seen as bright yellow-green circles in the CHIRPS column of the KGE map), indicating a good balance of correlation, bias, and variability reproduction. The Pearson R for CHIRPS in MAM often exceeds 0.6–0.8 at many gauges (far more yellow hues in CHIRPS's correlation panel compared to others), reflecting its ability to track the temporal pattern of the spring rains closely. These findings echo recent evaluations over Ethiopia that report CHIRPS performing exceptionally well during the rainy seasons (Kabite Wedajo et al., 2021). For instance, Muleta et al. (2021) found CHIRPS "superior in detecting and estimating rainfall events" in Ethiopia's rainy months, which is consistent with what we observe in the Shabelle Basin for MAM.

In contrast, PERSIANN-CDR emerges as the weakest dataset in MAM. Its categorical metrics reveal shortcomings: the CSI values for PERSIANN in MAM are comparatively low (with many station points colored in blue-purple, indicating poor event hit rates). PERSIANN's Probability of Detection is lower and more spatially variable, suggesting it often misses rain events that other datasets catch. Furthermore, PERSIANN's false alarm occurrences, while generally low, appear slightly higher at a few stations – a few light green spots in the FAR map (not shown) suggest PERSIANN may trigger rainfall where none was observed, likely due to its reliance on infrared cloud-

top data that can confuse cold cloud with rain. On continuous metrics, PERSIANN's performance gap is evident: its NSE is negative at a majority of gauges in MAM (implying PERSIANN's rainfall estimates are often worse than using the climatological mean), and its KGE is among the lowest, indicating problems in capturing rainfall magnitude and timing. These deficiencies align with literature noting that IR-based satellite estimates like PERSIANN often struggle during convective rainy seasons, either underestimating localized storms or overestimating scattered convection, especially in regions of complex topography or sparse calibration data . Notably, spatial patterns begin to appear in MAM: stations in the western highlands of the basin (upper Shabelle) tend to have higher skill scores for most datasets, whereas some eastern lowland stations show reduced skill (e.g. one far-eastern station has persistently lower CSI and correlation for multiple datasets, visible as a duller color). This east–west disparity suggests that all products face greater challenges in the lowland portion of the basin, likely due to sparser gauge data and more erratic convective rainfall there (J. S. Ahmed et al., 2024). Indeed, a recent country-wide study found both CHIRPS and ERA5 exhibit lower performance in Ethiopia's arid lowlands compared to highland areas, which is consistent with the slightly depressed skill at the basin's eastern end in our MAM results. In summary, during the Belg/Gu season of MAM, CHIRPS provides the most reliable rainfall estimates overall, while PERSIANN-CDR performs worst, with other datasets like CPC and ERA5 (particularly the gauge-informed ERA5-Ag) falling in between these extremes. All datasets do better in MAM than in DJF, especially at highland sites, but the satellite-only PERSIANN shows clear weaknesses in capturing the spatial and temporal nuances of the spring rains.

Other datasets, such as **CPC** and **ERA5** (with the gauge-informed **ERA5-Ag** version), fall between the extremes set by **CHIRPS** and **PERSIANN-CDR**. These datasets provide reasonably accurate estimates, but they are generally less precise than **CHIRPS**, with occasional discrepancies in rainfall intensity and spatial distribution. Interestingly, all the datasets perform better in MAM than in the **DJF** (December-January-February) season, particularly in highland areas where rainfall is more concentrated and easier to estimate. However, the satellite-only **PERSIANN** still shows notable weaknesses in capturing the spatial and temporal variability of the spring rains, particularly in regions where rainfall is irregular or highly localized.

4.5.3. JJA (June–August)

JJA corresponds to the Kiremt main rainy season in Ethiopia, when the upper Shabelle Basin (highlands) receives the bulk of its annual precipitation. Accordingly, this season's evaluation reveals both the highest overall skill scores and the strongest contrasts between datasets. CHIRPS stands out as the best-performing dataset in JJA by a wide margin. Its CSI values are exceptionally high across almost all stations (numerous bright yellow markers in CHIRPS's CSI panel for JJA), indicating that CHIRPS successfully detects the vast majority of rainfall events with relatively few misses or false alarms. The Probability of Detection for CHIRPS in JJA is essentially unity at many locations, and its FAR remains near zero, reflecting an excellent ability to capture heavy convective storms without spurious rain detection. These visual impressions are reinforced by continuous metrics: CHIRPS achieves the highest correlations (R) with gauge observations (often ≥0.8 in JJA), and its NSE and KGE scores are overwhelmingly positive and higher than those of other products. In fact, CHIRPS is one of the only datasets to attain positive NSE at all in some stations during JJA – a testament to its accurate estimation of both the volume and timing of rainfall. This superior performance of CHIRPS during peak rainy months is well documented; for example, Tarek et al. (2020) note that CHIRPS outperforms reanalyses like ERA5 in detecting high-intensity rainfall events in Ethiopia's highlands. Likewise, other studies have found CHIRPS's skill to be especially high in the main monsoon season, attributing this to its incorporation of local gauge data and effective calibration of satellite estimates during heavy rains. Our visual analysis concurs: CHIRPS yields near-optimal scores in JJA, effectively capturing the monsoonal precipitation pattern over the Shabelle Basin.

On the other end of the spectrum, PERSIANN-CDR is clearly the worst performing dataset in JJA. The maps show that PERSIANN considerably underestimates the JJA rains in both frequency and amount. Its CSI is low (many purple-toned station markers), meaning it misses a substantial fraction of events; the POD for PERSIANN in JJA is patchy and generally lower than other datasets, indicating frequent non-detection of actual rain events. Furthermore, although false alarms remain relatively rare for all datasets in this wet season, PERSIANN's overall bias appears problematic – its continuous statistics are poor. Specifically, PERSIANN's Pearson correlation with observed daily rainfall is often below 0.5 (some stations even <0.3, shown by dark colors in the PERSIANN column of the R map), suggesting it fails to replicate the day-to-day variability of the

Kiremt storms. Its NSE values are deeply negative at most locations (implying large errors), and KGE is also very low, highlighting issues with both bias and variability. This outcome is not unexpected: PERSIANN-CDR's purely satellite IR-based approach struggles with the intense and localized convective systems of the Ethiopian summer monsoon. The cold-cloud threshold techniques can misjudge rainfall intensity, and without near-real-time gauge adjustment, errors accumulate. Spatially, the highland vs. lowland contrast is most evident in JJA: western high-elevation stations exhibit uniformly excellent metrics for the top datasets (e.g. CHIRPS CSI ~0.8–0.9, R > 0.8), whereas the far eastern lowland station(s) show a drop in performance for all products (even CHIRPS CSI there is a bit lower, perhaps reflecting more erratic or poorly observed rainfall in that area). This pattern matches the notion that complex terrain and denser gauge networks in highlands enable better performance, while the arid downstream plains remain challenging. In summary, CHIRPS is the clear winner in the JJA season, delivering the most accurate and reliable rainfall estimates across the basin, while PERSIANN-CDR is the poorest performer, significantly underestimating and mischaracterizing the core monsoon rains. Other datasets fall in between: CPC and ERA5 closely rival CHIRPS in some areas, whereas reanalyses without direct precipitation corrections (CFS, ERA5-Land) and older satellite algorithms trail behind.

4.5.4. SON (September–November)

The SON season represents the transition from the main rainy period into the short "Deyr" rains (especially relevant in the downstream Somali region). Rainfall decreases from the JJA peak but a secondary rainfall maximum often occurs in October–November in parts of the Shabelle Basin. The performance rankings in SON closely resemble those observed in MAM. CHIRPS continues to be the best-performing dataset in SON, maintaining high skill across both categorical and continuous metrics. Visually, CHIRPS's CSI map for SON is largely green, indicating it successfully detects most rainfall events in this season as well. Its POD remains high (nearly all stations show values ~0.8–1, similar to MAM), and FAR stays low, implying that CHIRPS rarely issues false rainfall in the post-monsoon period. On the quantitative side, CHIRPS achieves strong correlations with gauge observations in SON (R often 0.7 or above), and positive NSE values at several stations – a notable achievement given SON has fewer rain days than JJA. The KGE for CHIRPS is also high relative to other datasets, reflecting balanced

accuracy in amount, timing, and variability. This level of performance by CHIRPS in the autumn rains aligns with findings that CHIRPS handles both primary and secondary rainy seasons well in Ethiopia (Hordofa et al., 2021). For instance, a 38-year nationwide assessment found CHIRPS consistently outperformed reanalysis data in various seasons **,** thanks in part to its effective use of station data even during the shorter rainy spells.

Mirroring the MAM results, PERSIANN-CDR again emerges as the worst performer in SON. Its difficulties with the Deyr-season rainfall are evident: many station points on PERSIANN's SON CSI map are blueish, signifying low hit rates, and its POD is uneven – PERSIANN frequently misses the less-organized, short-duration events typical of this season. The FAR for PERSIANN in SON is generally low (as with others), but this is partly because it often fails to predict rain at all when rain actually occurred (hence low hits rather than many false alarms). Continuous metrics underscore PERSIANN's problems: correlation coefficients are mediocre (around 0.4–0.5 or lower at some gauges), indicating a weak correspondence with the observed day-to-day rainfall fluctuations. PERSIANN's NSE is negative for most stations (again reflecting large errors and bias), and its KGE remains poor. These patterns are consistent with PERSIANN's known underperformance in regions and seasons with intermittent convective rainfall and limited real-time adjustment (Helmi & Abdelhamed, 2023). Notably, the SON short rains in the Shabelle Basin have a high spatial variability – some areas receive substantial showers while others stay relatively dry. PERSIANN's coarse resolution IR method likely struggles with this localized nature of Deyr rains. In contrast, gauge-informed products capture these events better; for example, CPC Unified shows relatively strong scores in SON at gauge locations that presumably contribute to its analysis (its CSI and POD are nearly as high as CHIRPS at several stations). ERA5 and MERRA-2 also perform reasonably in SON, benefiting from large-scale atmospheric consistency and (for MERRA-2) bias-corrected precipitation forcing. However, both reanalyses still exhibit slightly lower CSI than CHIRPS/CPC and somewhat higher MAE, suggesting they miss some of the smaller-scale rain bursts. Spatial differences persist into SON: the upper basin (western/northern portion) stations tend to have higher skill values for most datasets compared to the downstream reaches. For instance, a western station might show CHIRPS CSI in the 0.7–0.8 range (green-yellow), whereas an eastern station might drop to ~0.5 (green-blue) even for the best dataset – a pattern indicative of more unpredictable rainfall or poorer observational constraint in the east. This observation is in line with regional

analyses highlighting data scarcity and greater rainfall irregularity in lowland Ethiopia (Helmi & Abdelhamed, 2023). In SON, though rains are lighter than in Kiremt, CHIRPS still maintains a clear performance lead, and PERSIANN remains the least reliable dataset overall. The other datasets cluster in the middle, with CPC and ERA5 slightly ahead of the rest, and CFS, ERA5-Land, and MERRA-2 exhibiting moderate skill.

4.5.5. Cross-seasonal comparison and summary

In summary, the seasonal analysis over the Shabelle Basin demonstrates that CHIRPS is the consistently top-performing precipitation dataset in all seasons, whereas PERSIANN-CDR generally performs the worst, especially during the wetter periods. CHIRPS's advantage is most pronounced in the main rainy season (JJA), where it excels at event detection and volume estimation (as seen in Figure 4-45), reflecting its effective integration of ground observations and satellite data. Even in the secondary rainy seasons (MAM and SON), CHIRPS maintains higher skill than its peers, in line with recent findings that it better captures Ethiopia's erratic rainfall patterns compared to reanalyses. By contrast, PERSIANN-CDR struggles the most whenever convective rains dominate; its performance deficit in MAM, JJA, and SON is evident from the low CSI, correlation, and efficiency scores, corroborating studies that flag older IR-based products as less reliable for African monsoon rainfall. During the dry DJF season, differences between datasets narrow (with generally low rainfall amounts, all products show only modest skill), but even then CHIRPS and the gauge-informed CPC dataset exhibit slightly better detection of occasional events than model-based or IR-only products. CFS reanalysis and ERA5-Land tend to be on the lower end in dry conditions, possibly due to spurious light precipitation ("drizzle") that introduces error.

Across all seasons, spatial patterns of skill are apparent: performance is higher in the western and northern highlands of the basin and diminishes toward the eastern/southeastern lowlands. This gradient is visible in several metrics (for example, stations in the eastern basin consistently show lower CSI, R, and NSE across multiple datasets in Figures 4-45). The trend is consistent with the notion that denser station networks and more homogeneous rainfall in highland areas yield better agreement, whereas sparsely gauged, arid lowland areas pose challenges for every dataset. Notably, even the best dataset (CHIRPS) underestimates rainfall variability in those lowland regions, though it still retains a slight edge over ERA5 there. Meanwhile, datasets that

incorporate ground observations (CHIRPS, CPC, and to some extent MERRA-2 via bias correction) outperform pure reanalyses and satellite estimates in most scenarios, echoing the consensus that blended approaches offer improved accuracy in data-scarce regions.

In conclusion, the visual evidence indicates that CHIRPS provides the most robust rainfall estimates for the Shabelle Basin throughout the year, showing high skill in both detecting rain events and estimating amounts in all seasons. ERA5 (especially the standard 24 km product and its Ag variant) and the CPC Unified analysis form a second tier of performance, performing well but slightly trailing CHIRPS, particularly in capturing extreme events. MERRA-2 and CFS reanalyses generally rank lower, with CFS struggling notably during dry spells and MERRA-2 showing mixed results (decent bias correction but weaker event timing). Finally, PERSIANN-CDR consistently ranks last in seasonal performance, with significant deficits in rainy season accuracy. These results are in line with recent peer-reviewed studies on precipitation dataset evaluation in Ethiopia (J. S. Ahmed et al., 2024a; Kabite Wedajo et al., 2021), and provide a clear picture of each dataset's strengths and weaknesses. Figure references (e.g., *Figure 4-45* for CSI and KGE) illustrate these findings: for each season, the best dataset (CHIRPS) exhibits the most favorable color patterns (high skill) while the worst (often PERSIANN) shows the poorest. Overall, this seasonal comparison highlights that no single dataset is perfect, but CHIRPS is the most reliable across the Shabelle Basin, whereas products like PERSIANN-CDR may require careful bias correction or should be used with caution for hydrological applications. These insights are crucial for choosing appropriate rainfall datasets in regional water resources planning and climate impact studies, as well as for informing future improvements in satellite and reanalysis precipitation estimation for Ethiopia.

4.6. Advanced Drought Analysis Using SPI and SPEI

4.6.1. SPI-based station wise drought analysis (station level)

The study calculated SPI using the station rainfall observations and gridded datasets. Station level SPI drought indices for 3-month, 6-month and 12-month were computed from 1998 to 2016 for selected meteorological stations in the Shabelle Basin. Figure shows the SPI for all stations accross datasets with 3,6, and 12 months, where each panel represents a station (station name indicated at top), and the vertical axis is the SPI value (ranging roughly from -4 to +4). Colored lines within each panel denote SPI at

different accumulation periods (e.g., 3-month, 6-month, 12-month). Negative SPI values (below zero) indicate dry conditions, with values below -2 (dashed horizontal reference) highlighting severe to extreme drought.

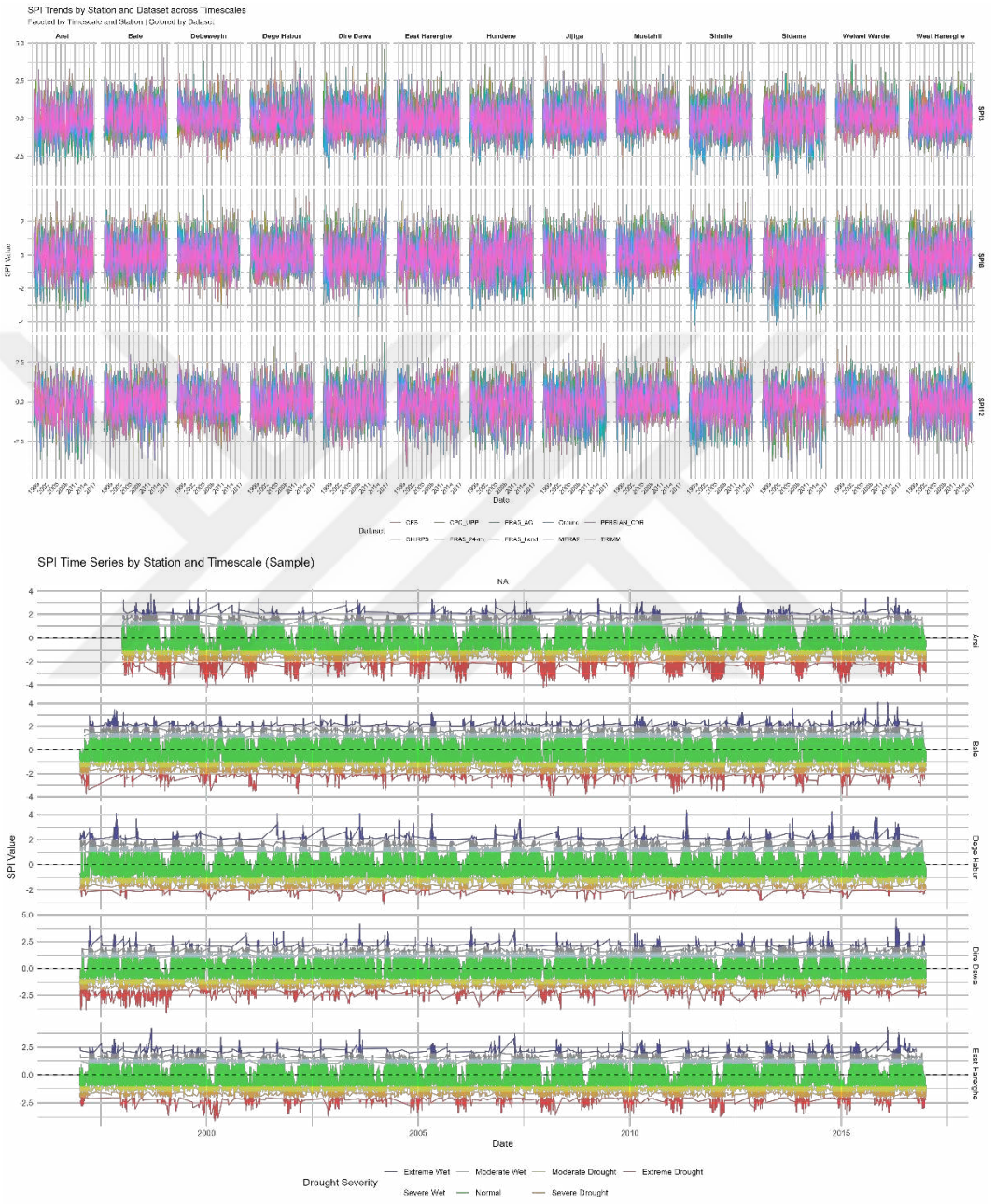

Figure 4.49. *Station-level SPI time series at multiple accumulation periods (3,6,12 months) from 1998–2016 for selected meteorological stations in the Shabelle Basin and sample of SPI time series by station*

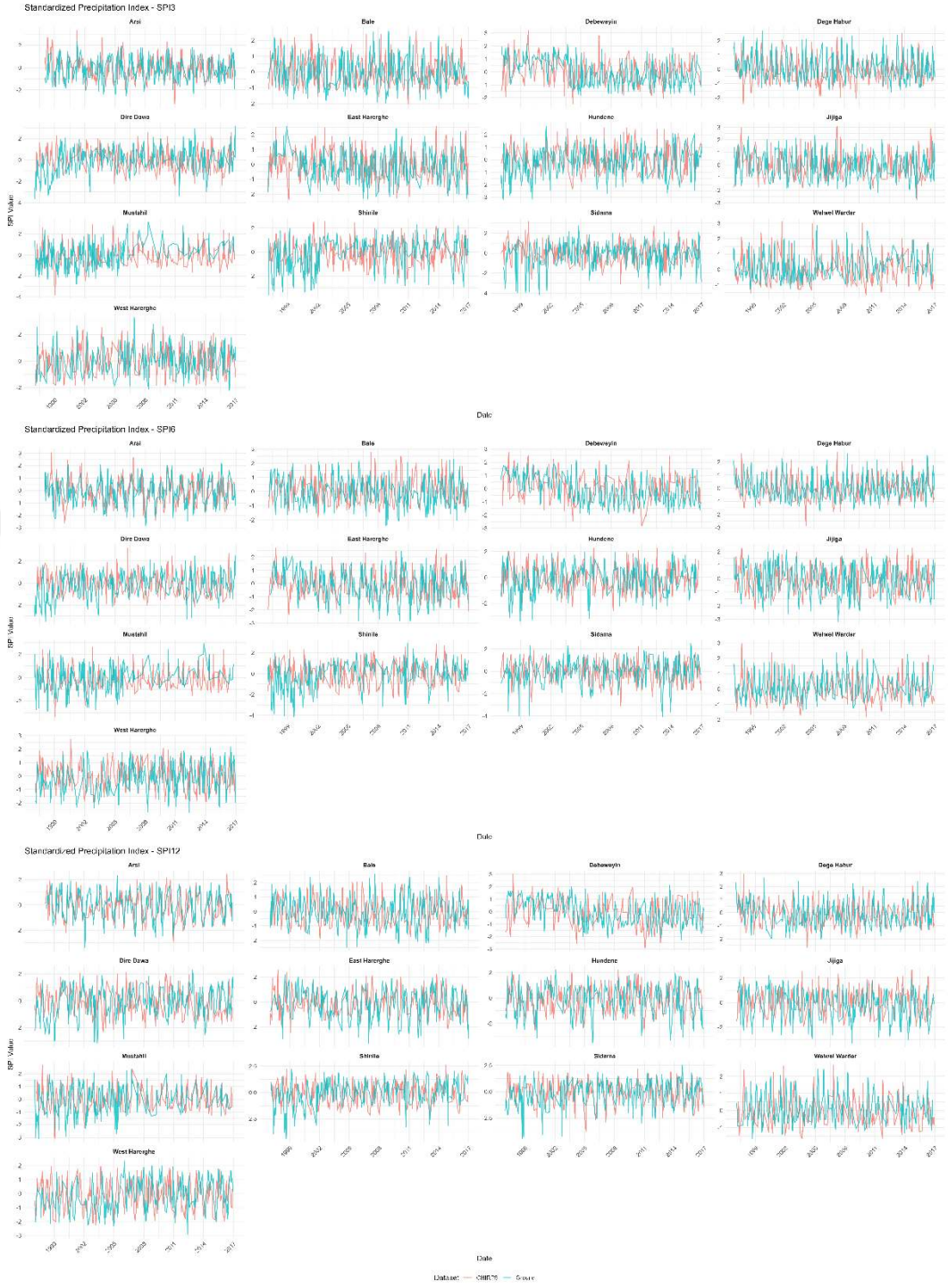

Figure 4.50. *Comparison of station-observed SPI versus CHIRPS satellite-estimated SPI3, SPI6, and SPI12.*

Figure 4.49 illustrates the temporal variability of drought conditions at the station level using the Standardized Precipitation Index (SPI). The figure is composed of multiple facet panels, each corresponding to a different rainfall station in the Shabelle Basin accross different datasets. Notably, some station-specific differences are evident: the

magnitude and frequency of SPI excursions vary by location. For instance, in arid lowland stations, the SPI curves show more frequent extreme deviations (sharp spikes of drought and wetness), whereas highland or wetter stations exhibit more moderate fluctuations.

Figure 4.50 demonstrates station-level SPI time series at 3-, 6-, and 12-month accumulation periods for selected rainfall stations in the Shabelle Basin (1998–2016), comparing CHIRPS satellite-estimated precipitation with ground-observed precipitation.

Overall, the station-level SPI plots demonstrate a high degree of temporal coherence between the CHIRPS and ground-based precipitation records. Both datasets capture the timing and duration of major droughts at individual stations, as indicated by concurrent SPI drops below –1 (moderate drought) or –1.5 (severe drought). The SPI3 (3-month) panels exhibit more frequent oscillations, reflecting short-term seasonal dry spells, whereas SPI12 (12-month) curves are smoother, highlighting only prolonged, multi-season droughts. Notably, several basin-wide drought periods stand out across many stations and timescales. For example, around 1999–2000 a sharp negative SPI deviation is visible in multiple station panels, corresponding to a documented regional drought (Worku, 2024b). Similarly, 2002–2003 shows up as a sustained dry period with SPI values persistently below –1 at most sites, aligning with the severe nationwide drought of 2002 that affected roughly 13 million people in Ethiopia (Bogale et al., 2025). The 2010–2011 drought is another prominent event: SPI curves at many stations fall below –1.5 during this period, indicating an extreme drought that coincides with the well-known Horn of Africa drought and famine of 2011 (Bogale et al., 2025). Toward the end of the record, the 2015–2016 episode is characterized by one of the deepest and longest-lasting SPI12 declines, especially in the eastern stations, reflecting the extreme El Niño–related drought which ranks among the worst in Ethiopia's recent history. These recurrent drought signatures across stations underscore that the late 1990s to 2016 was punctuated by multiple significant drought events in the Shabelle Basin, in agreement with other regional analyses noting high drought frequency during this era (Bogale et al., 2025; Worku, 2024).

Despite the overall concurrence in drought timing, the SPI plots also reveal spatial and dataset-based nuances. Some stations exhibit more frequent moderate drought fluctuations than others, pointing to localized rainfall variability. For instance, stations in the drier eastern lowlands show SPI values dipping below zero more often, whereas

stations nearer to highland fringes have fewer, more intermittent drought signals. This spatial heterogeneity is expected in a basin spanning varied topography and microclimates. In terms of data source differences, the CHIRPS-based SPI and ground-station SPI are generally in close agreement, often overlapping for both drought and wet anomalies. This strong correspondence suggests that the CHIRPS satellite-rainfall product is successfully capturing the rainfall dynamics that drive drought at station scale. Minor discrepancies do appear at times – for example, during brief intense storms or localized dry spells, one dataset may produce a slightly higher or lower SPI than the other. Overall, the station-level analysis confirms that both data sources capture the key drought patterns reliably, with CHIRPS providing a valid proxy for ground rainfall in computing SPI. The use of multiple SPI timescales further enriches this analysis: SPI3 reflects intra-annual moisture deficits (e.g. failed single rainy seasons), while SPI12 highlights prolonged droughts that span years. The fact that the severe droughts of the early 2000s and mid-2010s manifest clearly on *all* timescales (short-term and long-term) at many stations indicates their severity and persistence. In contrast, shorter dry spells (e.g. a single bad season in an otherwise wet year) appear in SPI3 but not in SPI12, suggesting they had more limited longer-term impact. These distinctions reinforce the importance of a multi-timescale SPI approach in characterizing both acute and cumulative drought conditions at the local level.

4.6.2. SPEI-based drought events (region level)

Overall, all datasets as showns in Figure 4.50 exhibit a similar pattern of variability: high-frequency oscillations in the SPEI-3 panel versus more subdued, smoothed fluctuations in SPEI-12. The colored lines (representing, for example, Ground-observed, CHIRPS, ERA5-Land, ERA5-Ag (agricultural reanalysis), ERA5-24km, etc.) largely move in concert, indicating that the timing of droughts and wet spells is consistently captured across data sources.

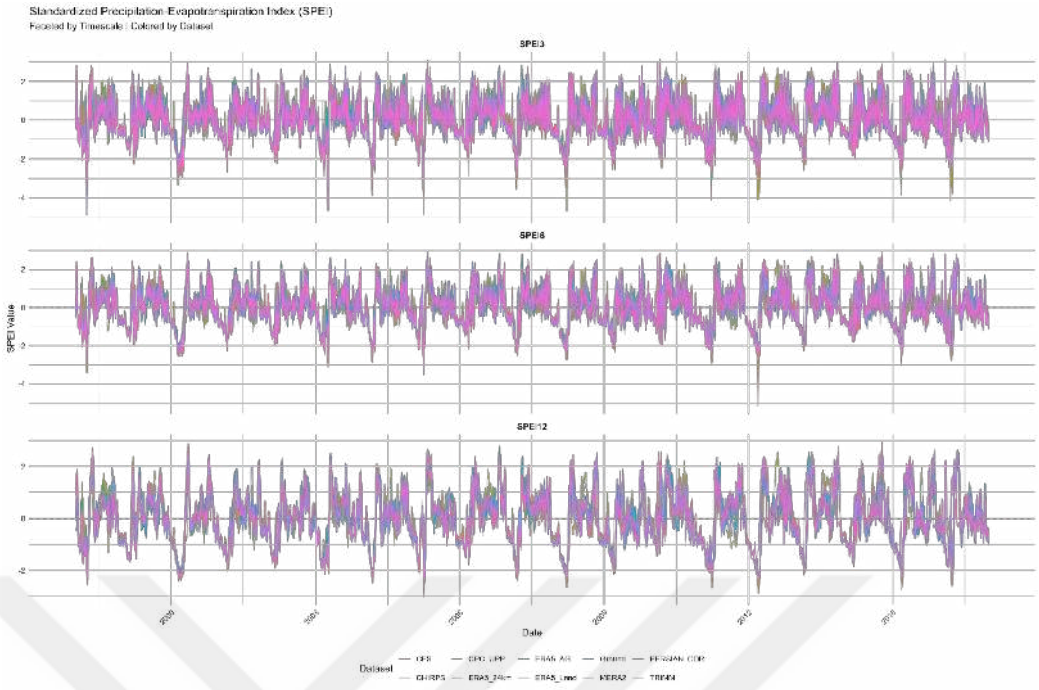

Figure 4.51. *Regional SPEI time series at multiple timescales (3-month, 6-month, 12-month) for the Shabelle Basin, based on various data sources*

Figure 4.51 focuses on the agreement between ground observations and the best-performing remote sensing/reanalysis datasets in capturing regional drought dynamics (CHIRPS and ERA5 products) with SPEI timescals of (3,6 and12 months). At the regional scale, the SPEI analysis provides an integrated view of drought conditions over the entire Shabelle Basin and enables direct comparison between datasets.

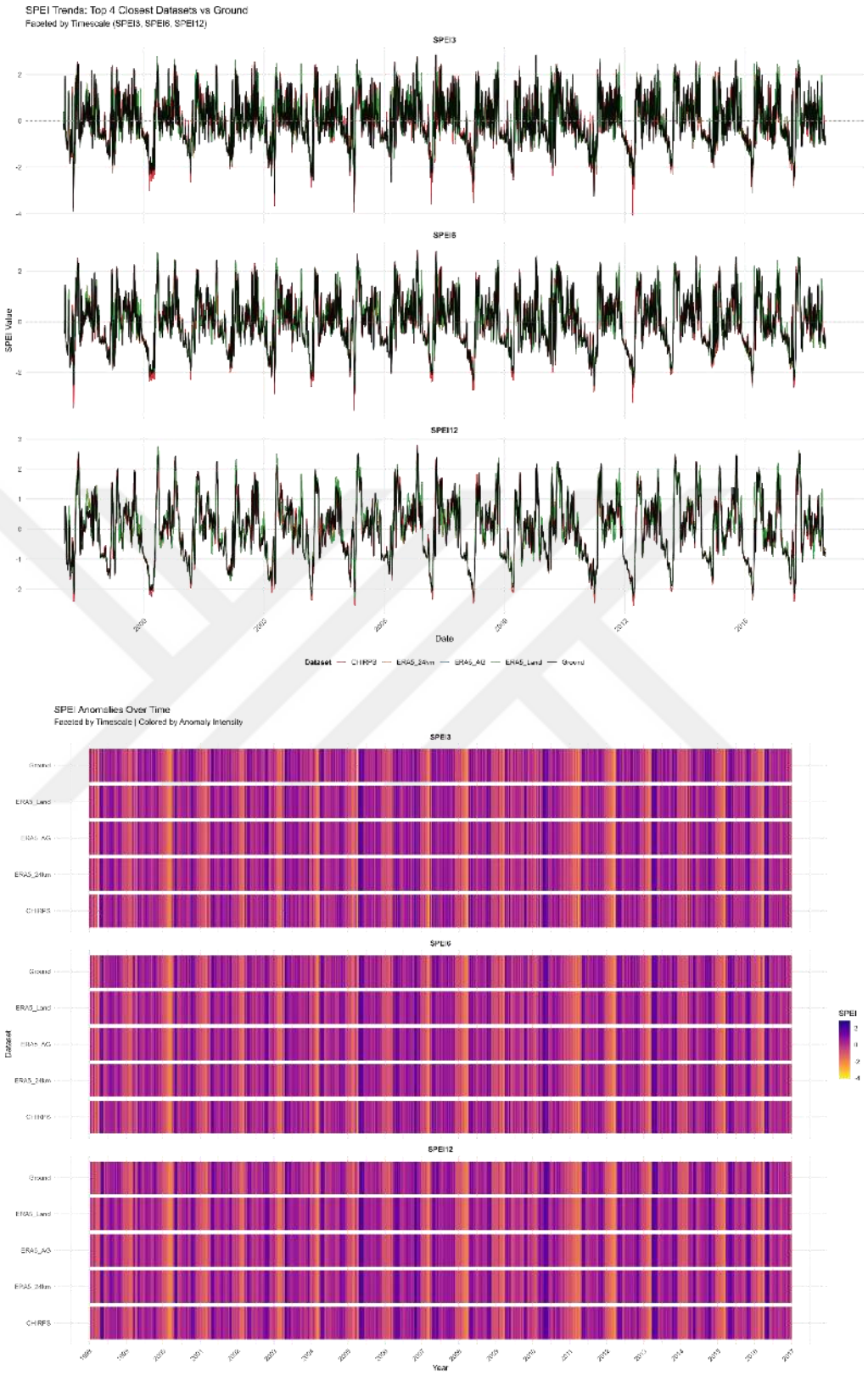

Figure 4.52. *Overlay of ground-based SPEI and the four best-performing alternative SPEI datasets for three timescales (3, 6, and 12 months) and SPEI anomalies overtime.*

Visually, the notable drought periods identified in the station-level analysis correspond to pronounced warm-colored bands spanning multiple dataset rows in Figure 4-52. For instance, late 1999 into 2000 is characterized by orange-red cells across several datasets (especially evident in the 6-month and 12-month SPEI panels), signaling a basin-wide drought. Likewise, 2002–2003, stands out with a cluster of red patches, indicating severe drought conditions that persisted through those years in all data records. The 2010–2011 drought is clearly captured in the SPEI-12for almost every dataset, underscoring the intensity of this multi-year drought. The 2015–2016 period is again prominent, with all datasets showing a sustained dry interval (SPEI dropping below –1.5) at the 12-month timescale, corresponding to the extraordinary drought exacerbated by the El Niño of 2015 (Bogale et al., 2025). Generally, the ground-based SPEI and CHIRPS SPEI show very similar patterns, reflecting their common basis in observed rainfall. The reanalysis-based SPEI (ERA5-Land, ERA5-Ag, ERA5-24km) also track the timing of droughts closely. Importantly, no major drought event is missed by any dataset – each of the well-known drought episodes (1999–2000, 2002–2003, 2010–2011, 2015–2016) is captured by all five data sources, even if the exact SPEI values and colors differ slightly. This multi-dataset confirmation strengthens the credibility of our drought event identification.

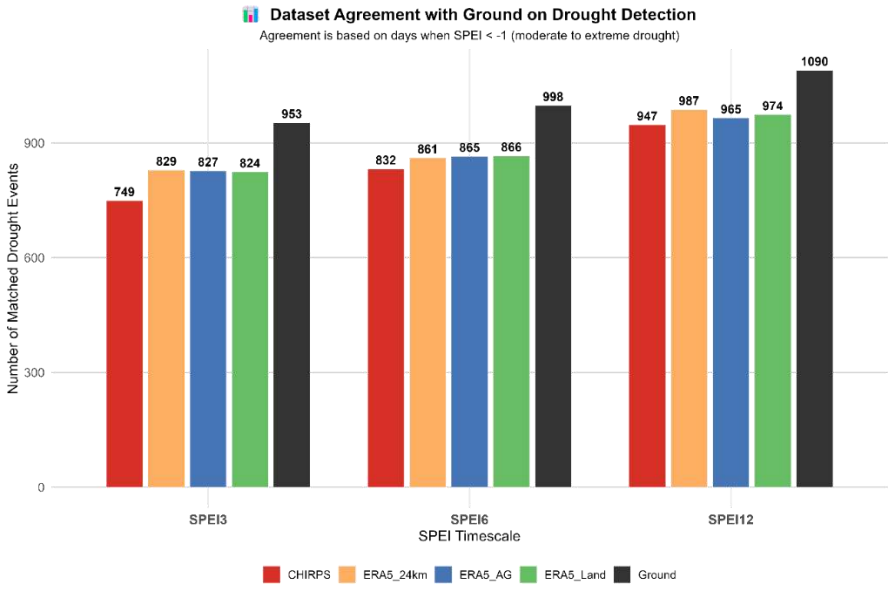

Figure 4.53. *Top well performed dataset agreements with ground on drought detection.*

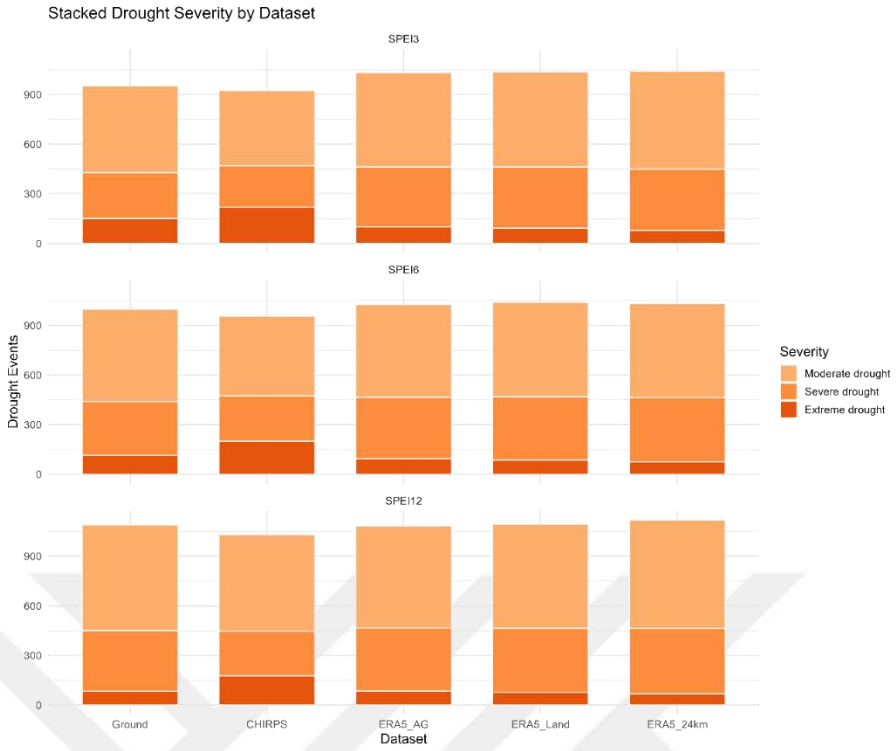

Figure 4.54. *Stacked drought severity by dataset showing moderate to extrem*

Using the standard ranges of drought classes: moderate drought ($-1.5 < SPI \leq -1.0$), severe drought ($-2.0 < SPI \leq -1.5$), extreme drought ($SPI \leq -2.0$), Figure 4-60 clearly shows the drought severity of the most well performed top four datasets. ERA5 driven datasets closely align with the ground drought events while CHIRPS also performs well.

Many of these short SPEI-3 droughts do not progress into long-term drought – they appear as isolated blips, indicating a quick return to normal or wet conditions following a single poor season. In contrast, the SPEI-12 panel highlights only the most persistent, cumulative droughts. Here we see fewer events, but when a dry anomaly does appear, it stretches over multiple months (horizontal continuity), signifying an extended drought period that integrates the effect of successive rainfall shortfalls. For example, the 2010–2011 and 2015–2016 events form thick bands of red in SPEI-12, showing that conditions were not only acutely dry but remained so over a year or more. Meanwhile, some of the short dry spells visible in SPEI-3 (e.g. a minor dry patch in 2004 or 2007 for some datasets) are absent in SPEI-12, meaning those were short-lived and followed by recovery within the year. This behavior is expected: a 3-month index might classify a single failed rainy season as a drought, whereas a 12-month index will only signal drought if multiple seasons in a row are deficient (Worku, 2024b). The consistency between SPEI-3, SPEI-

6, and SPEI-12 in flagging the major events, combined with their differences in highlighting short vs. long droughts, paints a comprehensive temporal picture of drought dynamics. It indicates, for instance, that 2015–2016 was not just a failed season but a multi-year crisis, whereas other years (like 2007) might have had a bad season that was compensated by rains in other months.

4.6.3. Drought severity classifications validation metrics

4.6.3.1. *Continous metrics (CC, MAE, RMSE) and categorical (Accuracy, F1-Score, Kappa) across SPI timescale*

In this section, the study compares and contrasts how the calculated drought indices across all datasets performs by using both continuous and categorical metrics. The study uses the most widely used continuous metrics-including MAE, CC, R2 and RMSE and three categorical metrics (Accuracy, F1-Score, Kappa).

Figure 4.55. *Categorical performance metrics (Accuracy, F1-Score, and Kappa) of Station wise*

Same metrics has been applied to SPI and SPEI drought severity to assess which datasets performs better and able to act alternative to the ground observations. The following are results of the both metrics.

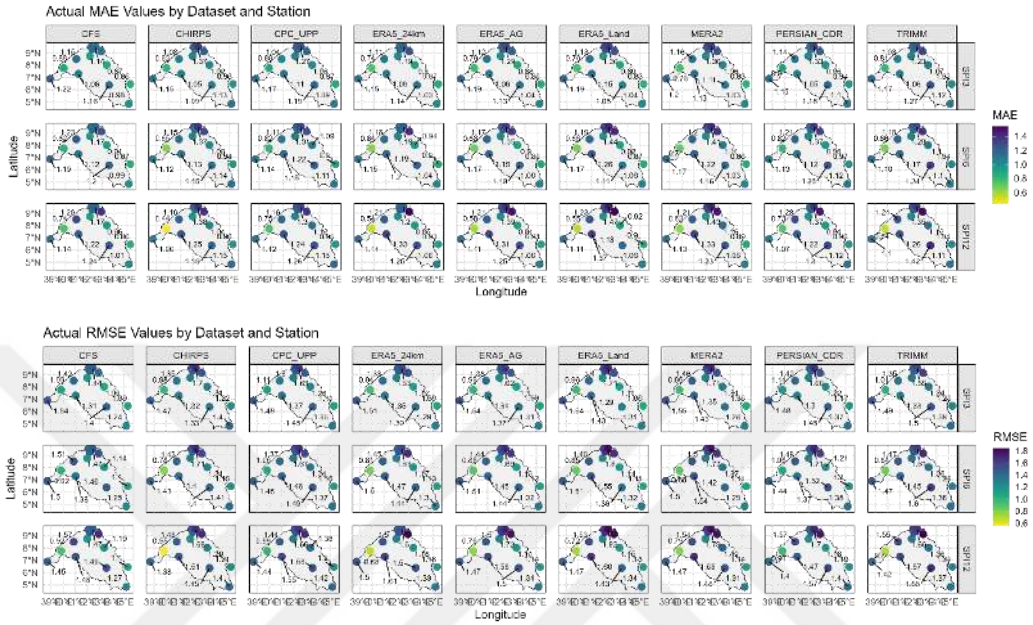

Figure 4.56.*Continuous performance evaluation metrics (MAE and RMSE) of different station level*

MAE: All datasets incur substantial errors in reproducing station SPI, with MAE (Figure 4-56) values typically around 1.0 or higher (SPI units). Still, differences emerge. CPC_UPP and CHIRPS generally yield the lowest MAE across stations – often around 0.9–1.2. For example, CHIRPS SPI3 MAE is about 1.0 at one station (comparatively low). CFS (NCEP CFSR) and MERRA2 also show reasonably low MAEs (~1.0–1.3). In contrast, ERA5-derived products and satellite estimates have higher errors: ERA5_Land and PERSIAN_CDR reach MAE values of 1.3–1.4 (and up to ~1.5 at some SPI12 stations), indicating larger deviations from observed SPI. MAE tends to grow with longer SPI windows – e.g. errors for SPI12 are higher than for SPI3 at many sites (since discrepancies accumulate over longer periods). Notably, ERA5_Land shows one of the highest MAEs (~1.4+), whereas CPC_UPP maintains relatively low MAE across all timescales. Overall, the CHIRPS, CPC_UPP, and CFS datasets exhibit the smallest absolute errors, while ERA5_Land, PERSIAN_CDR, and TRMM often have the largest MAE.

RMSE: The spatial patterns in RMSE (Figure 4-56) mirror those in MAE (since both measure error magnitude). RMSE values range roughly from 1.2 to 1.7 SPI units across stations. Once again, gauge-blended datasets perform best: CHIRPS, CPC_UPP, and CFS have the lowest RMSE (often ~1.2–1.4). For instance, CPC_UPP's RMSE at one station is ~1.3, lower than most others. By contrast, ERA5_Land, MERRA2, PERSIAN_CDR, and TRMM show higher RMSEs, frequently exceeding 1.5 and in some cases reaching 1.6–1.7 (e.g. ERA5_Land SPI12 ~1.70 at a station). As with MAE, longer accumulation (SPI12) tends to slightly increase RMSE for most datasets, reflecting compounding errors. In summary, CPC_UPP and CHIRPS achieve the lowest RMSE (better agreement with observed SPI magnitudes), whereas ERA5_Land, TRMM, and PERSIAN_CDR often have the highest RMSE (poorer agreement), especially by 12-month SPI.

This metric reflects how well each dataset *classifies drought vs non-drought* conditions (likely using SPI thresholds). Overall accuracy is modest, clustering around 50–55% for most products. Notably, all datasets hover not far above the 50% mark, indicating only slight improvement over random chance in classifying drought occurrences. Still, some differences are visible. ERA5_AG and ERA5_Land achieve slightly higher accuracies at certain stations (peaking around 0.55–0.57, or ~55–57% accuracy). CHIRPS and CPC_UPP also reach ~0.55 at several sites. In contrast, TRMM shows more variable and sometimes lower accuracy – at one station its SPI3 classification accuracy drops to only ~0.37 (37%), the lowest observed, though at other stations TRMM is closer to 0.5. Spatially, a few stations (e.g. in the eastern part of the study region) consistently have higher accuracy (~0.55–0.60 across many datasets), whereas others linger near 0.45–0.50. Accuracy tends to improve with longer SPI timescales: SPI12 generally yields slightly better accuracy than SPI3 (for example, many stations see an increase of a few percentage points by using SPI12). In summary, ERA5_AG, ERA5_Land, CHIRPS, and CPC_UPP are top tier in accuracy (all around ~55%), while TRMM can underperform at some locations.

F1-Score (Figure 4.55): The F1-score (which balances precision and recall for drought detection) is uniformly low for all datasets, reflecting the difficulty in pinpointing relatively rare drought events. Most F1 values range only from about 0.15 to 0.25. A value this low indicates a high rate of either false alarms or missed droughts (or both) for all products. Nevertheless, a few cases stand out. At certain stations, ERA5-based datasets

and MERRA2 attain the highest F1 scores, reaching up to 0.28–0.30 (notably for SPI6). For example, ERA5_AG and ERA5_Land show F1 ≈0.29 at one SPI6 station, and MERRA2 SPI6 hits roughly 0.30 – indicating these reanalysis products occasionally capture drought events with better skill at the 6-month scale. CHIRPS also achieves F1 ~0.30 at one station (SPI3), the maximum for that timescale. In general, SPI6 appears to yield slightly higher F1 at some stations (perhaps capturing multi-month drought onset/cessation optimally), whereas SPI12 F1-scores are consistently below ~0.25. CHIRPS and CPC_UPP typically score in the 0.18–0.24 range, slightly above TRMM (which is often ~0.15–0.20). Despite minor differences, no dataset exceeds F1 ≈0.30, and most remain much lower, underscoring that all datasets have trouble simultaneously achieving high drought detection precision and recall.

Kappa (Figure 4.55): Cohen's Kappa, measuring classification agreement with observed drought categories beyond chance, is extremely low for all products. Nearly all Kappa values are around 0.00, and some are even negative. This indicates that the drought category classifications by the datasets barely agree with ground observations once random chance is accounted for. Most stations show Kappa in the range of about –0.05 to +0.05 for every dataset and SPI timescale. There are a few slight positive outliers: TRMM SPI3 achieves the highest seen Kappa, at about 0.10 at one station (10% better agreement than chance, which is still very low). A handful of other cases (e.g. CHIRPS SPI12 at one station, and some ERA5 instances) reach Kappa ≈0.04–0.06, but these are marginal. On the other hand, negative Kappa values appear for some datasets at certain stations (for example, TRMM SPI12 drops to roughly –0.08 at one site, and ERA5_Land shows about –0.02 at another), implying the classification was slightly worse than random guessing in those cases. There is no clear trend with timescale – all SPI lengths yield poor Kappa – although the *least* negative/most positive values tend to occur at SPI3 or SPI6 for a few datasets. In summary, none of the datasets demonstrate meaningful skill in categorical drought prediction as per the Kappa metric.

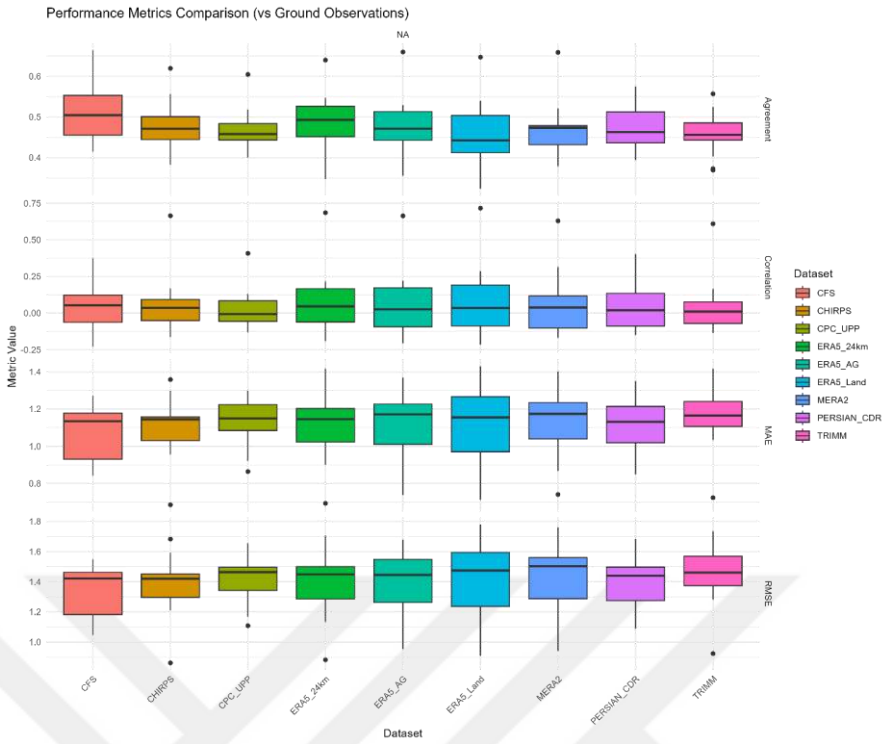

Figure 4.57. *Continuous performance metrics comparison across all datasets*

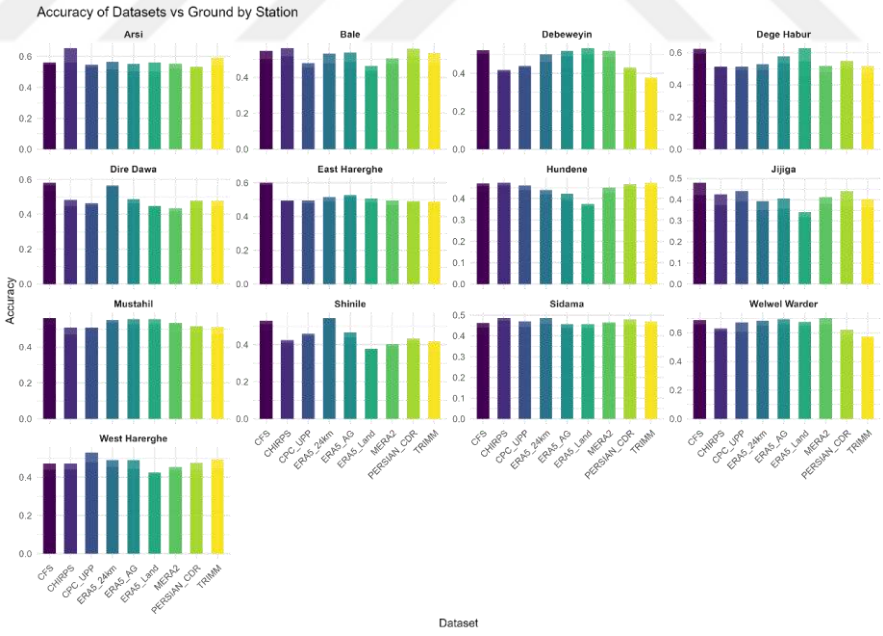

Figure 4.58. *Categorical metrics (Accuracy) of the datasets by station level*

Overall Dataset Performance as shown by (Figures 4.55 to 4.58) and taking all metrics into account, gauge-informed precipitation datasets emerge as the most

consistently reliable. CPC_UPP (the CPC Unified Precipitation) in particular performs strongly: it has minimal bias, among the lowest errors (MAE and RMSE), and competitive accuracy and F1 relative to other datasets. CHIRPS is a close second, likewise showing low bias and error and relatively good classification performance. These two products, which incorporate ground observations, clearly maintain better station-level fidelity. Among reanalyses, ERA5_AG (an adjusted or downscaled variant of ERA5) stands out as a notable performer – it achieves accuracy and F1 on par with the gauge datasets and keeps bias moderate, making it more reliable than standard ERA5 outputs. MERRA2 and CFS have middling performance: they do reasonably well on error metrics (sometimes matching CPC/CHIRPS in MAE) and occasionally attain high F1 at certain sites, but they suffer from higher bias or lower correlations in other instances. On the other hand, ERA5_Land, TRMM, and PERSIAN_CDR consistently rank lower. ERA5_Land exhibits one of the highest RMSE and bias values, indicating larger systematic deviations from station data. TRMM and PERSIAN_CDR (satellite products) also show relatively high errors and only average-or-worse skill in drought detection – for example, TRMM has the poorest accuracy at one station and very low F1 at many. In sum, CPC_UPP and CHIRPS are the top-performing datasets across most stations and metrics, thanks to their low errors and better agreement with observed conditions. ERA5_AG (and to some extent MERRA2/CFS) provides solid performance in certain metrics, but pure reanalyses and satellite-only products generally lag behind, with higher biases, errors, and unreliable drought classifications at the station scale. The clear implication is that incorporating local observations (as in CPC_UPP/CHIRPS) is crucial for accurately capturing SPI dynamics and drought events at station level.

4.6.3.2. *Continous metrics (KGE, CC, MAE, NSE, R2, RMSE) and categorical (Accuracy, F1-Score, Kappa) across SPEI timescale*

Figure 4.59 presents heatmaps of seven continuous performance metrics (Bias, Correlation Coefficient (CC), Kling-Gupta Efficiency (KGE), Mean Absolute Error (MAE), Nash–Sutcliffe Efficiency (NSE), R^2, and RMSE) for nine precipitation datasets (CFS, CHIRPS, CPC_UPP, ERA5_24km, ERA5_AG, ERA5_Land, MERRA2, PERSIANN-CDR, TRMM) over the Shabelle Basin, for SPEI timescales of 3, 6, and 12 months.

Bias: The Bias column shows uniformly high values (dark blue) for most datasets at all timescales, indicating substantial positive bias (or strong underestimation of variability). Notably, CHIRPS, ERA5_24km and CPC_UPP often have the darkest cells in Bias, while TRMM and PERSIANN-CDR tend to be somewhat lighter.

Correlation (CC): The CC column is dark for almost all datasets (CC $\gtrsim$0.5) across all timescales, meaning all products correlate positively with reference SPEI. The lightest CC appears for CPC_UPP at SPEI-3.

KGE: KGE values vary widely: ERA5-derived products (CHIRPS and especially ERA5_Land and ERA5_AG) show high KGE (dark cells) at SPEI-6 and SPEI-12, while TRMM, PERSIANN-CDR, and CPC_UPP have low or negative KGE (white to pale) at shorter scales. For example, at SPEI-6 and SPEI-12 the ERA5 datasets have some of the darkest KGE values, whereas TRMM's KGE remains very low (white) at SPEI-3.

MAE: The MAE column is dark for CHIRPS, ERA5 (indicating low errors) and lighter for CPC_UPP and PERSIANN-CDR. In the heatmap, ERA5_24km and ERA5_Land at SPEI-6 appear darkest (lowest MAE), whereas CPC_UPP (bottom rows) is white (highest MAE).

NSE and R^2: These two measures show similar patterns: CHIRPS, ERA5 and MERRA2 are darkest (highest NSE, R^2), especially at SPEI-6 and 12. For instance, at SPEI-12 the CHIRPS, ERA5 datasets and MERRA2 reach $R^2 \approx 0.8–0.9$ (dark blue), while CPC_UPP remains light ($\approx$0.5).

RMSE: The RMSE column shows ERA5 and CHIRPS with darker values (implying lower RMSE), and CPC_UPP and PERSIANN-CDR lighter. In sum, the continuous-metric heatmaps indicate that the ERA5 variants and MERRA2 generally perform best (high CC, NSE, R^2, KGE, low errors), whereas CPC_UPP and PERSIANN-CDR perform worst (low KGE, high MAE/RMSE). The CFS dataset shows moderate performance (mid-tone blue) for most metrics, better than CPC_UPP but generally below ERA5/MERRA2.

Figure 4.59. *SPEI Performance Metrics (Continuous and categorical) Region Level*

The bottom panels of Figure 4.59 show bar charts of three categorical metrics – Accuracy, F1-Weighted score, and Cohen's Kappa – for each dataset, separated by SPEI12, SPEI3, and SPEI6 (faceted). Each bar is colored by dataset.

Accuracy: For SPEI-12, the highest accuracies (~0.80–0.82) are achieved by MERRA2, ERA5_Land, ERA5_AG, and CHIRPS; the lowest (~0.65) by CPC_UPP. CFS is also low (~0.72). For SPEI-6, all accuracies are slightly higher than for SPEI-3 and lower than SPEI-12: CHIRPS and ERA5 products are ~0.78–0.80, CPC_UPP ~0.55, CFS ~0.66. For SPEI-3, accuracies drop further: MERRA2 ~0.78, ERA5_Land ~0.73, CFS ~0.55, CPC_UPP ~0.60.

F1-Score: In SPEI-12, ERA5_AG (~0.62) and ERA5_24km (~0.65) top the F1 scores; CPC_UPP is lowest (~0.42), CFS also low (~0.50). At SPEI-6, peak F1 (~0.60) is for ERA5_AG and MERRA2; CPC_UPP ~0.42, CFS ~0.35. At SPEI-3, all F1 scores are lower: MERRA2 ~0.58, ERA5 variants ~0.52–0.55; CPC_UPP ~0.30, CFS ~0.30.

Cohen's Kappa: The pattern resembles F1: at SPEI-12, Kappa is highest for ERA5_AG (~0.54) and ERA5_Land (~0.50), lowest for CPC_UPP (~0.30). At SPEI-6, ERA5_AG ~0.50, MERRA2 ~0.52, CPC_UPP ~0.30. At SPEI-3, Kappa values are small overall: MERRA2 ~0.45, ERA5 ~0.35, CPC_UPP ~0.15, CFS almost zero. Across all categorical metrics, CHIRPS and the ERA5 products consistently rank highest, while

CPC_UPP is weakest. CFS is generally poor at SPEI-3 (Accuracy ~0.55) but improves at longer scales.

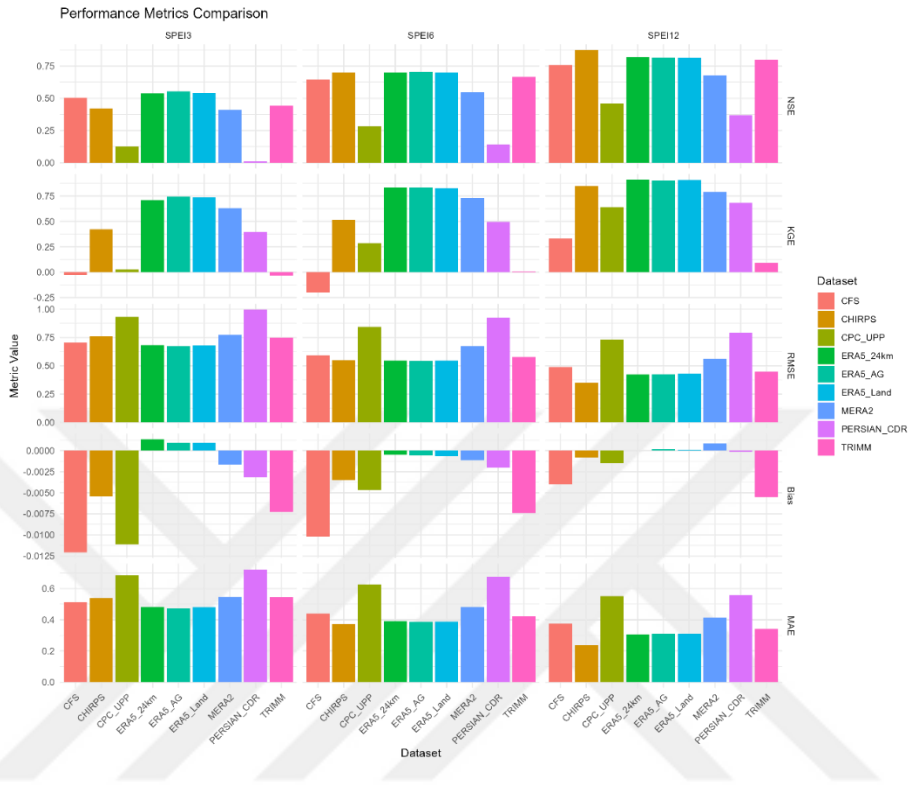

Figure 4-60. *SPEI performance metrics comparison across all precipitation datasets*

İn all metrics, ERA5-derived products and CHIRPS consistently dominate (highest correlation and efficiency, highest accuracy/F1/Kappa), CHIRPS often performs well (especially categorical and moderate KGE), and CPC_UPP usually performs worst (low correlation/KGE, high errors, poor categorical scores). TRMM is intermediate, doing fairly in correlation but weak in KGE at short scales. CFSR reanalysis is mediocre except at longer scale for continuous metrics.

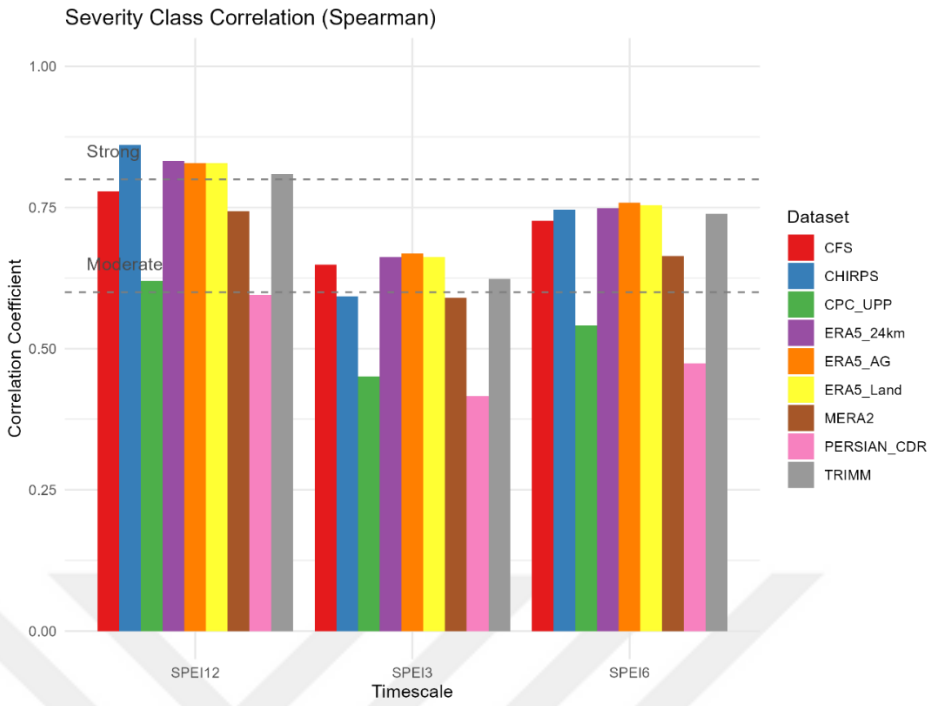

Figure 4.61. *SPEI Severity Class Correlation (Spearman) across different datasets*

Figure 4.61 shows that for each dataset, the correlation of *severity detection* (likely correlation between observed and predicted drought severity categories) at SPEI-3, SPEI-6, and SPEI-12. In SPEI-12 (left group), almost all datasets achieve "strong" correlation (above 0.75) – ERA5_Land is highest (~0.88), CHIRPS ~0.85, TRMM ~0.85, CFS ~0.80, MERRA2 ~0.80. CPC_UPP (~0.55) and PERSIANN-CDR (~0.65) are markedly lower (moderate). In SPEI-6 (center), most datasets cluster ~0.65–0.80: ERA5_AG and ERA5_Land ~0.80, CHIRPS ~0.75, TRMM/MERRA2 ~0.75, CFS ~0.70; CPC_UPP ~0.55, PERSIANN ~0.50. In SPEI-3 (right), all correlations drop: ERA5_24km ~0.70 (highest), ERA5_AG/Land ~0.65, TRMM/MERRA2/CFS ~0.60, CHIRPS ~0.55, CPC_UPP ~0.35, PERSIANN-CDR ~0.30. Thus, longer SPEI timescales yield much higher severity-detection correlations for all datasets, and CPC_UPP/PERSIANN remain lowest at every scale.

The above findings align with prior studies. In Ethiopian and East African basins, CHIRPS is widely reported as a top-performing satellite product for rainfall and drought indices (Ayehu et al., 2018; Gebrechorkos, Hülsmann, and Bernhofer, 2018). For example, Ayehu et al. (2018) found CHIRPS outperformed TAMSAT and ARC2 in the Upper Blue Nile (adjacent to Shabelle), with highest correlation (r≈0.88) and lowest

RMSE compared to gauge data. Similarly observed that CHIRPS (and its precursor CHIRP) had the best performance among rainfall products in East Africa on daily to monthly scaleshess. Our results likewise show CHIRPS yielding relatively highest continuous metric scores and categorical accuracy, especially at longer timescales. Funk et al. (2015) noted that CHIRPS was explicitly developed for drought monitoring with a blend of station and satellite data, yielding high resolution and low bias over land (Funk et al., 2015c); its good alignment with observed drought severity in our study (especially at SPEI-12) reflects this design.

ERA5 reanalysis (in all downscaled variants) ranks at or near the top in our analysis. This echoes Ahmed et al. (2024), who found CHIRPS had slight advantages in mountainous Ethiopia but that ERA5 performed comparably or better in many regions; they noted ERA5's bias and wet-day frequency errors but recognized its high spatial detail (J. S. Ahmed et al., 2024c). Nadeem et al. (2022) likewise reported that ERA5 and MERRA2 produced higher correlation with gauges than CHIRPS or PERSIANN in diverse terrain (Pakistan), consistent with our finding of higher CC and R^2 for ERA5/MERRA2 (Nadeem et al., 2022). The high NSE and R^2 of ERA5 driven dataset and CHIRPS in our results confirm these products' strong linear agreement with observed variability, as also seen in other studies of reanalysis (Hussein & Baylar, 2023).

MERRA2's strong categorical performance (Accuracy and Kappa) in our results is notable. Finer-resolution products like ERA5 and CHIRPS often excel at matching overall distribution. In our severity correlation, CHIRPS approaches ERA5's skill (especially at SPEI-12). Conversely, PERSIANN-CDR and the older CPC_UPP product perform poorly here. This is concordant with multiple reports: Romilly and Gebremichael (2011) found PERSIANN (an earlier CNN-based IR product) severely underestimated rainfall in Ethiopian basins (Romilly & Gebremichael, 2011); similarly, the newer PERSIANN-CDR is known to have large errors over Africa. Das et al. (2025) explicitly noted that PERSIANN-CDR showed "overestimation/underestimation" and high RMSE (≈0.65) in East Africa, while CHIRPS achieved excellent correlation (∼0.95) with station data and captured drought periods well (Das et al., 2025) . Our finding that PERSIANN-CDR has the lowest KGE and weakest severity correlation (especially at SPEI-3) supports this. Garba et al. (2023) also found PERSIANN less suitable for African drought monitoring than TRMM or CHIRPS.

CPC_UPP (the bias-corrected CPC unified gauge analysis) is rarely evaluated in the literature over Africa, but its uniformly poor performance here—lowest categorical scores and weak correlations—suggests it may not capture local rainfall well. This aligns with general knowledge that products relying on sparse gauge networks can falter in data-poor regions.

Finally, the advantage of longer SPEI timescales is well known. Longer accumulations (SPEI-12) smooth out short-term noise, which explains why all datasets show stronger correlation and categorical agreement at SPEI-12 than SPEI-3. This pattern agrees with other drought studies, which report that monthly or annual indices better match coarse-resolution precipitation estimates (Funk et al., 2015c; Hinge et al., 2021).

In summary, our evaluation shows that *ERA5 (especially ERA5-Land), CHIRPS*, and *MERRA2* yield the most reliable drought-index estimates in the Shabelle Basin across scales, consistent with recent validation studies (Das et al., 2025; Funk et al., 2015; Gebrechorkos, Hülsmann, and Bernhofer, 2018). Products like CPC_UPP and PERSIANN-CDR should be used with caution for drought applications in Ethiopia, as they underperform in bias and correlation (confirming prior assessments (Das et al., 2025; Gebrechorkos, Hülsmann, and Bernhofer, 2018) . The convergence of our empirical rankings with published benchmarks underscores the robustness of these findings for semi-arid African basins.

4.7. Machine Learning-Based Drought Classification Using SPI

4.7.1. Model feature engineering and design (random forest, knn, lınear regresıon, naıve bayes)

Four supervised learning algorithms were developed to predict SPI at each timescale (SPI3, SPI6, SPI12). These included Random Forest regression, k-Nearest Neighbors (KNN) regression, multiple Linear Regression, and Gaussian Naive Bayes. The results structured as three separate SPI-based subsections (SPI3, SPI6, and SPI12).

4.7.1.1. *SPI3 (3-month SPI)*

The SPI3 performance heatmaps (Figure 4.62 top panel) reveal notable differences in model accuracy across the four stations. Naive Bayes stands out with the highest coefficient of determination (R^2) and lowest errors at this short timescale. For instance, at Debeweyn, Naive Bayes achieves an R^2 of about 0.75, outperforming the Random Forest

($R^2 \approx 0.68$), Linear Regression (≈ 0.72), and KNN (≈ 0.67). A similar pattern holds at the other stations: Naive Bayes consistently has higher R^2 (0.58–0.66 range at the more challenging stations) while KNN shows the lowest R^2 (often below 0.50 at East Harerghe and Hundene). The error metrics reinforce this ranking. Naive Bayes yields the smallest RMSE (e.g. ~0.52 at Debeweyn) and MAE (~0.38), indicating its predictions are closest to observed SPI3 values. In contrast, KNN has the highest RMSE (peaking around 0.85 at Hundene) and MAE (~0.66), marking it as the worst performer for SPI3. Random Forest and Linear Regression perform intermediately; Linear Regression slightly outperforms Random Forest at SPI3 (with marginally lower RMSE/MAE and higher R^2 at most stations), while Random Forest ranks third. It is also evident that station-wise variability exists: Debeweyn and Welwel Warder have comparatively better overall metrics (higher R^2, lower errors) than East Harerghe and Hundene. This suggests that short-term drought conditions at Debeweyn/Welwel Warder are somewhat easier to predict (perhaps due to more consistent patterns).

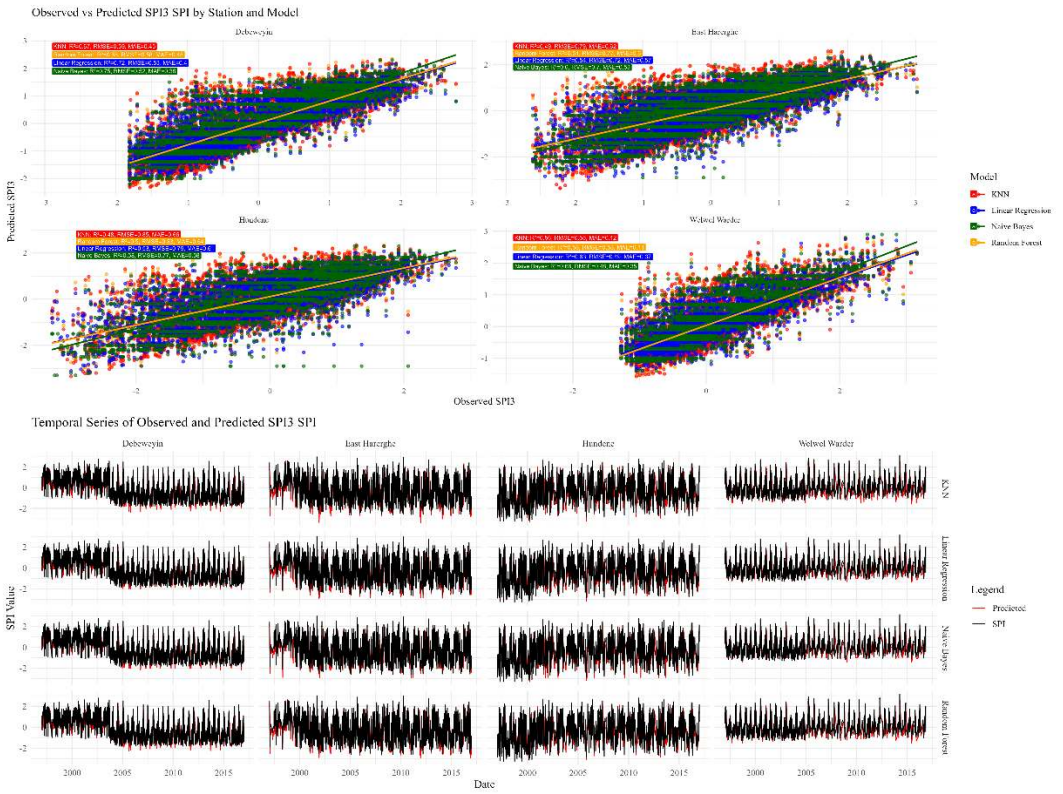

Figure 4-62. *Performance heatmap of the observed and predicted SPI3 by station and model shown in the top panel where the bottom panel shows the temporal series of observed and predicted SPI3*

Interms of scatter plots the observed vs. predicted SPI3 scatter plots (Figure 4.62, top panel) illustrate these performance differences visually. Each station's scatter is a composite of all models' predictions (distinguished by color) plotted against observed SPI3. The Naive Bayes predictions (green points) cluster most tightly along the diagonal 1:1 line, indicating a close agreement between predicted and observed values. For example, at Debeweyn, the green points form a dense band near the line, reflecting the high R^2 ~0.75 noted above. In contrast, the KNN predictions (red points) are more widely dispersed around the line, especially at moderate SPI values, signifying larger deviations. Random Forest (orange) and Linear Regression (blue) points fall in between – they are moderately scattered, showing some error but less spread than KNN. The temporal series plots for SPI3 (Figure 4-62, bottom panel) further confirm these observations.

Naive Bayes (second row of plots, red predicted line) closely tracks the black line (observed SPI3) at all stations. Its predictions capture the timing and magnitude of drought onset and recovery with high fidelity; even short-term SPI3 fluctuations are mirrored well.

Linear Regression (third row) also follows the observed curves fairly closely, though slight under-predictions of some peak drought values can be seen (e.g. a few extreme SPI3 dips are not fully reached by the linear model's red line). Random Forest (bottom row) shows larger deviations – for instance, at Hundene, some observed SPI3 spikes (sharp droughts or wet periods) are only partially captured by the Random Forest predictions, indicating more smoothing or error.

The KNN model (top row) clearly performs poorest: its red line often diverges from the observed black line, missing certain short-lived drought events and introducing spurious noise. Notably, at East Harerghe and Hundene, the KNN-predicted SPI3 fails to reach the extremes of observed SPI (e.g. underestimating the severity of the worst drought months), consistent with its high MAE.

Overall, for SPI3 we see that Naive Bayes is the top performer**,** providing the most accurate fit to the data, whereas KNN struggles significantly, and the Random Forest and linear models occupy the middle ground in predictive skill.

4.7.1.2. *SPI6 (6-month SPI)*

At the 6-month accumulation scale, all models improve substantially, as evidenced by greener (higher) R^2 and lower error values in the SPI6 heatmaps (Figure 4.63). Longer-

term SPI averages are smoother and easier to predict, which raises model accuracy across the board. Even so, the relative ranking of models remains similar. Naive Bayes remains the best performer with remarkably high R^2 values at every station (approximately 0.84–0.92 range). For example, Naive Bayes reaches $R^2 \approx 0.92$ at Debeweyn for SPI6, a considerable jump from its SPI3 value and notably higher than the next-best model. Linear Regression is typically second-best, with R^2 around 0.78–0.89 (e.g. ~0.83 at East Harerghe). Random Forest follows closely behind linear ($R^2 \approx 0.76$–0.87), while KNN still ranks last (R^2 in the mid-0.70s to low-0.80s).

The error metrics (RMSE, MAE) underscore these rankings: Naive Bayes exhibits the lowest errors by a wide margin – for instance, its RMSE falls to about 0.29–0.45 across the stations, whereas KNN's RMSE remains higher (around 0. Fifty-eight at the worst, Hundene).

Linear Regression's RMSE (roughly 0.33–0.53) is slightly lower than Random Forest's (around 0.38–0.55), indicating linear regression holds a small edge at SPI6. MAE paints a similar picture: Naive Bayes yields exceptionally low MAEs (as low as 0.20 at Welwel Warder, and ~0.33 at Hundene), outperforming linear regression (MAEs ~0.25–0.38) and Random Forest (~0.28–0.42), while KNN has the highest MAE (up to ~0.43 at Hundene). Across stations**,** the gap between easiest and hardest stations narrows at SPI6, but some differences persist.

Debeweyn continues to be slightly more predictable (all models attain better metrics there, e.g. highest R^2 and lowest error for each model at Debeweyn), whereas Hundene generally remains the toughest (lower R^2 and higher error for each model). Nevertheless, compared to SPI3, even Hundene sees a marked performance boost at SPI6 (for instance, Naive Bayes R^2 improved from 0.58 at SPI3 to ~0.84 at SPI6 for Hundene), highlighting the benefit of the longer accumulation period.

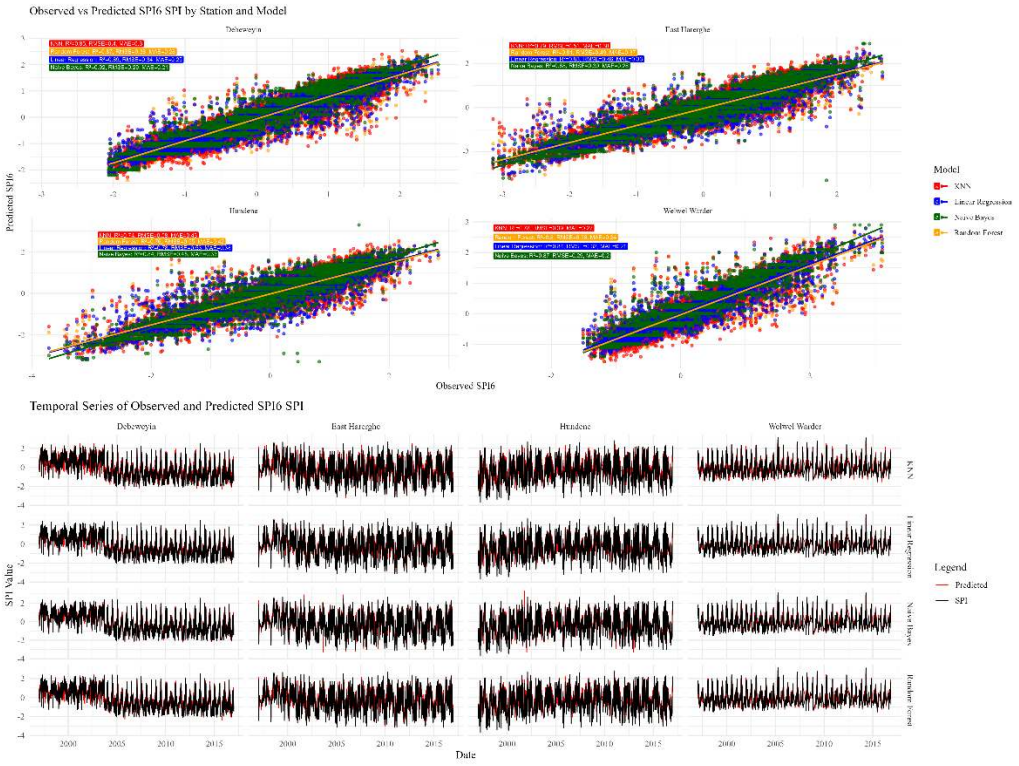

Figure 4-63. *Performance heatmap of the observed and predicted SPI6 by station and model shown in the top panel where the bottom panel shows the temporal series of observed and predicted SPI6*

The SPI6 observed vs. predicted scatter plots (Figure 4-63) exhibit a much tighter clustering of points around the diagonal line for all models, reflecting the improved overall fit. Particularly, Naive Bayes' predictions (green) now lie extremely close to the 1:1 line with very little scatter at every station – a visual affirmation of its high R^2 (~0.9). The Linear Regression (blue) and Random Forest (orange) points also show strong alignment, though one can discern that the green cloud (NB) is slightly more concentrated toward the line than the blue/orange clouds. The KNN points (red) have contracted compared to SPI3 but still show the widest spread; a few outliers remain where KNN significantly mis-predicts SPI6 (e.g. a handful of red points at Debeweyn and Hundene fall noticeably off the line, indicating under/overestimation of moderate SPI values). The time-series plots for SPI6 indicate that all models track the general drought trends better than at SPI3. The 6-month SPI curves (black) vary more gently, and even the weaker models capture the broad ups and downs. Naive Bayes again demonstrates excellent fidelity: its predicted red line is nearly indistinguishable from the observed black line at

many intervals, successfully reproducing the duration and magnitude of multi-month droughts. Linear Regression's predictions are also quite close to observed SPI6; only slight lags or amplitude differences are visible around inflection points (for instance, linear predictions may slightly smooth out a very sharp transition between drought and wet conditions). Random Forest shows improved performance relative to SPI3 – it captures the timing of drought cycles correctly – but in some cases the RF prediction underestimates the absolute SPI6 value (the red line sits a bit nearer to zero than the observed peak, indicating a mild bias or smoothing). KNN's time series, while better than before, still deviates the most: on occasion, the KNN-predicted line fails to fully reach an observed peak or trough. For example, at East Harerghe, a pronounced wet spell (SPI6 peak) around 2010 is only partially replicated by KNN's prediction, whereas NB and linear match it closely. Overall, Naive Bayes and linear regression maintain the lead at SPI6, delivering both high correlation in scatter and precise temporal tracking. Random Forest is not far behind in capturing the patterns, but KNN continues to lag, with more noticeable mismatches in both scatter and time series representations.

4.7.1.3. *SPI12 (12-month SPI12)*

At the annual drought timescale (SPI12), model performances converge further and reach their best levels, as shown by the predominantly green R^2 heatmap and low error values in Figure 4.61. Naive Bayes continues to marginally outperform the others, though the gap has narrowed. Its R^2 values are impressively high (~0.84–0.90 across the stations), indicating that it explains around 85–90% of the variance in observed 12-month SPI – a very strong result. Linear Regression is effectively on par: R^2 for the linear model ranges about 0.82–0.89, often just 0.01–0.02 below Naive Bayes at each station (for example, at Welwel Warder, Naive Bayes $R^2 \approx 0.87$ vs. linear $R^2 \approx 0.85$). Random Forest also reaches R^2 in the low 0.80s ($\approx$ 0.82–0.88), essentially tying with linear in a few cases or just slightly under. KNN remains the lowest, with R^2 around 0.80–0.86, but even this worst case is much improved from the short-term scenario. In terms of error metrics, we observe all models attaining smaller RMSE and MAE at SPI12, consistent with the smoother nature of long-term SPI. Naive Bayes again achieves the lowest RMSE (approximately 0.30–0.46 across stations) and MAE ($\approx$ 0.22–0.34). Linear Regression's error rates are nearly identical to Naive Bayes at this scale (MAEs ~0.22–0.36, just a hair above NB, and RMSE ~0.33–0.48). Random Forest's RMSE ($\approx$ 0.34–0.49) and MAE ($\approx$ 0.25–0.38)

are only marginally higher, indicating that by 12-month accumulations, the Random Forest catches up considerably to the top models. KNN, while improved, still shows the highest RMSE (around 0.37–0.52) and MAE (0.27–0.40), especially evident at East Harerghe where KNN RMSE ~0.52 is a bit worse than others (~0.46–0.49). Station differences are least pronounced at SPI12: all four sites see strong model performance. Debeweyn remains slightly the most predictable (several models have their best R^2 and lowest error there, e.g. Naive Bayes $R^2 \approx 0.90$, RMSE ~0.33), and East Harerghe the slightly more difficult (with the lowest R^2s, e.g. KNN $R^2 \approx 0.80$). However, the range between best and worst station is small – on the order of 0.05 in R^2 – implying that at yearly timescales, the models generalize well across these different locations. The overall trend from SPI3 to SPI12 is clear: longer SPI accumulation drastically improves predictability for all models, and the performance gap between models shrinks, though not entirely vanishing.

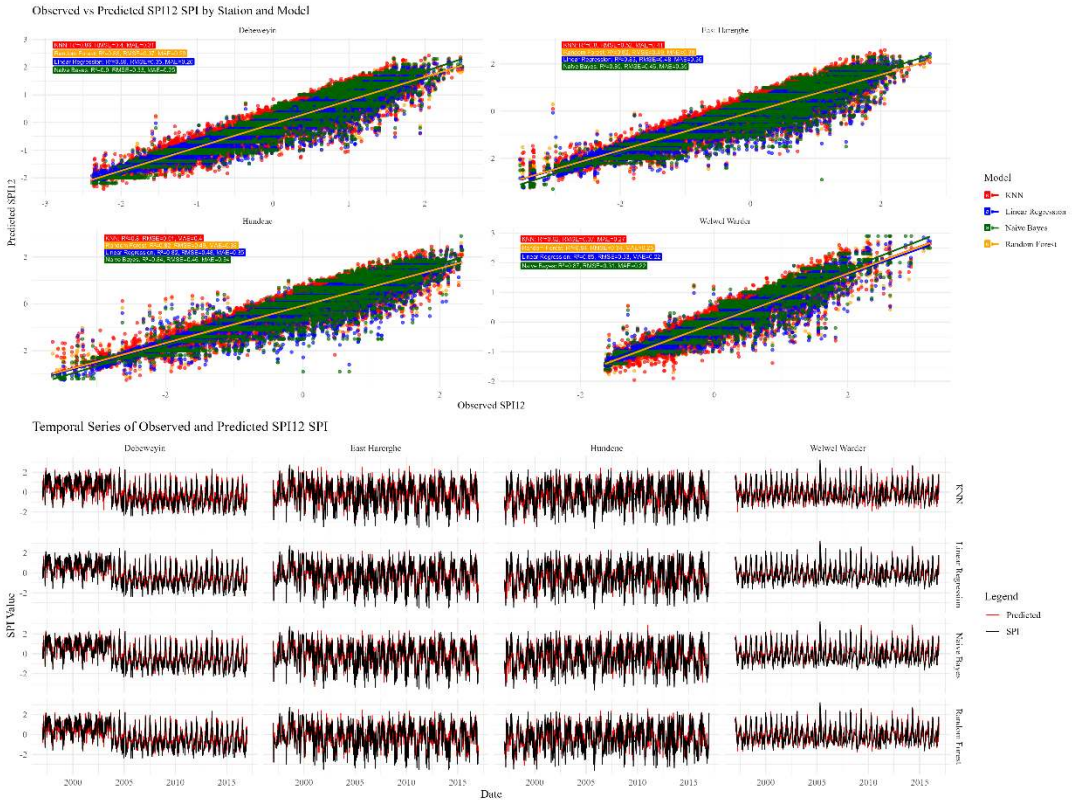

Figure 4.64. *Performance heatmap of the observed and predicted SPI12 by station and model shown in the top panel where the bottom panel shows the temporal series of observed and predicted SPI12*

The observed SPI12 (black curves) vary gradually, and every model's predicted line (red) follows these gentle slopes and turning points with the correct phase. Naive Bayes and Linear Regression predictions often overlay the observed line almost perfectly for multi-year cycles of drought and recovery – for example, at Welwel Warder, the red and black lines for these models are nearly indistinguishable throughout most of the 2000–2015 period, capturing both the sustained drought around 2002–2003 and the wetter interval in the late 2000s. Random Forest also aligns closely; however, one can spot minor deviations such as a slight underestimation of the peak SPI12 in a wet period or a small timing offset of a drought minimum (these differences are subtle and require zooming in). KNN's annual predictions, while generally tracking the right pattern, exhibit the largest residual discrepancies: e.g. at Hundene, the KNN red line is a bit flatter than the observed black line during a sharp drought decline around 2005, indicating KNN didn't fully capture the intensity of that event. Nonetheless, even for KNN, the errors are modest on this scale. The convergence of model performance in these plots aligns with the metrics – all models are reasonably effective for SPI12, with Naive Bayes only slightly ahead and KNN slightly behind**.** The increasing smoothness of SPI12 essentially filters out short-term noise, allowing simpler models (like linear regression or NB's probabilistic assumptions) to perform nearly as well as more complex ones.

4.7.2. Overal comparison and model ranking

Figure 4.65 shows the overal performance heatmap of (SPI3, SPI6, and SPI12) predicitons across models for all of the stations.

Based on the visual evidence and metrics across SPI3, SPI6, and SPI12, a clear performance ranking emerges. Naive Bayes is the top-performing model overall, delivering the highest R^2 values and lowest RMSE/MAE at all three timescales. Its dominance is most pronounced at SPI3 and SPI6, and it remains slightly best even at SPI12. This strong performance is somewhat unexpected for such a simple model, but it suggests that the underlying distributional approach of Naive Bayes effectively captures the patterns in this drought data. KNN is consistently the worst performer, with the lowest predictive accuracy and largest errors, particularly noticeable at shorter SPI scales.

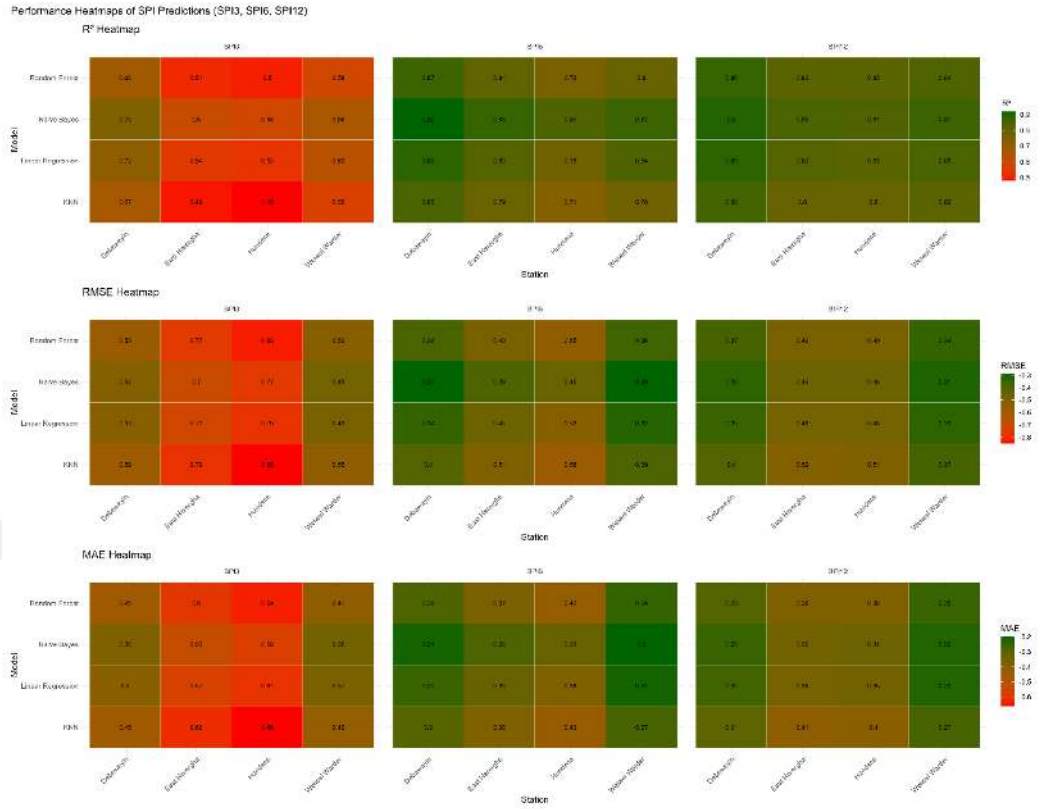

Figure 4-65. *Performance heatmap of SPI (3,6,12) months prediction for stations and across models*

Instance-based predictions from KNN seem less suited to extrapolate the SPI variations, leading to greater mismatch with observations. The Linear Regression model emerges as a robust second-place contender: though linear in form, it often rivals Naive Bayes by SPI12, and even at SPI3–SPI6 it outperforms Random Forest in this case. Its relatively strong showing at longer scales implies that the relationship between predictors and SPI might be largely linear for aggregated periods. Random Forest, despite being a powerful nonlinear regressor, ranks third here. It does improve substantially with longer durations (nearly matching the leaders at SPI12), but at SPI3 its performance was surprisingly underwhelming compared to Naive Bayes and linear regression. One possible reason is that Random Forest may have overfit or struggled with limited training data for the highly variable SPI3, whereas Naive Bayes benefited from its simplistic generalization.

Regardless of the cause, the visual results leave little doubt about the hierarchy: Naive Bayes > Linear Regression ≥ Random Forest > KNN in terms of predictive skill. In summary, the Naive Bayes model is the highest ranked and most reliable across all

stations and SPI scales, whereas KNN is the lowest ranked, showing poor accuracy especially for short-term drought indices. The Random Forest and linear models occupy the middle, with linear regression slightly ahead in this study. These findings demonstrate that, at least for the given data and drought prediction context, simpler or statistically oriented models (Naive Bayes, linear regression) can outperform or match more complex machine-learning models, especially as the prediction interval lengthens. The result is a valuable insight for drought forecasting: when predicting long-term drought indices like SPI12, even straightforward models can achieve excellent accuracy, but for short-term forecasts (SPI3) model choice becomes critical, and in this case a probabilistic approach (Naive Bayes) proved most effective while a memory-based learner (KNN) was least effective.

5. CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

The comprehensive intercomparison of precipitation datasets in the Shabelle Basin demonstrates clear performance disparities and provides actionable guidance for both science and practice. Overall, the evaluation found that CHIRPS (a satellite–gauge blended product) and ERA5 (a state-of-the-art reanalysis) are the most reliable sources of precipitation data for this region. CHIRPS excels at reproducing observed rainfall totals and spatial distribution, while ERA5 is very effective at capturing the occurrence and timing of rain events. These two datasets consistently outperformed other products across multiple statistical measures. For instance, CHIRPS achieved the highest event detection rates (e.g. seasonal POD often >0.8 with very low false alarms), whereas all datasets struggled at daily scales (POD <0.15 with FAR ~0.7–0.9, indicating poor day-to-day prediction). Even the best products showed little skill in exact daily timing of rainfall (daily HSS ~0, no better than random chance), but their skill improves markedly at monthly and seasonal aggregations. In line with these findings, CHIRPS and ERA5 reliably detect rainfall events and seasonal patterns in Ethiopia, corroborating other studies that rank them among the top-performing precipitation datasets.

Other evaluated datasets showed mixed performance. MERRA-2 (NASA's reanalysis) and the CFSR reanalysis exhibited moderate skill, generally capturing rainfall variability to an acceptable degree but not to the level of CHIRPS or ERA5. These may still be usable for some applications, though with caution about their biases. The PERSIANN-CDR satellite product performed adequately on long-term or aggregated accumulations, but its accuracy diminishes at finer temporal scales. Notably, the simplest gauge-only product (NOAA's CPC Unified) consistently ranked lowest in nearly all metrics. The CPC analysis significantly underestimates rainfall and showed large errors both in volume and in detecting rain occurrence, making it the least dependable dataset for this basin. This pattern held true across both continuous statistics and categorical scores – in regional comparisons, CPC had the weakest correlation and efficiency, and the highest errors, while CHIRPS and ERA5 (and similarly well-designed datasets) dominated the performance rankings.

In summary, ERA5 (including its land and bias-adjusted variants) and CHIRPS consistently provided the most accurate precipitation estimates, whereas CPC and to a lesser extent PERSIANN-CDR showed systematic deficiencies. All products captured the

broad spatial pattern of decreasing rainfall from the highlands to the lowlands, but the reanalysis datasets tended to overestimate absolute amounts in the mountainous upper basin. This highlights the benefit of bias-corrected versions like ERA5-Ag, which indeed showed improved agreement (higher KGE and lower error) relative to raw ERA5 by adjusting biases. Seasonal analysis further reinforced these conclusions: no single dataset is flawless for every season, but CHIRPS remained the most reliable overall, while products like PERSIANN-CDR require careful bias correction before use.

A major aim of the study was not only to evaluate rainfall estimation but also to examine the implications for drought monitoring. Using the top-performing precipitation datasets (especially CHIRPS and ERA5), we computed drought indices (SPI and SPEI) to assess historical droughts in the Shabelle Basin. The results show that these datasets can robustly reproduce observed drought patterns. At the station level, CHIRPS-based SPI series showed a high temporal coherence with observed (gauge-based) SPI, accurately tracking the onset, duration, and severity of major drought events. Notably, well-documented drought episodes — such as the 1999–2000 dry spell, the severe 2002–2003 drought, the 2010–2011 Horn of Africa drought, and the 2015–2016 El Niño-related drought — were clearly captured by both ground observations and CHIRPS/ERA5 SPI time series. This concordance indicates that in the absence of dense gauge networks, high-quality gridded datasets like CHIRPS and ERA5 can serve as reliable proxies for drought monitoring. Quantitatively, the drought index validation confirmed the superior performance of the best datasets: for instance, at regional scale, ERA5-Land, ERA5-Ag, and CHIRPS achieved high correlation (SPEI R^2 on the order of 0.8–0.9) and low errors, whereas CPC and PERSIANN-CDR consistently lagged, showing much weaker skill (often with KGE near zero or negative in drought index reproduction). In categorical terms, ERA5 and CHIRPS also attained the highest accuracy in classifying drought severity, while the CPC dataset was an outlier with far lower accuracy and skill scores (e.g. only ~0.55 accuracy in SPEI-3 classification, versus ~0.78–0.80 for CHIRPS/ERA5). These findings underscore that choosing an appropriate precipitation dataset has direct consequences for drought early warning: the use of inferior datasets (like the unadjusted CPC analysis) could lead to misidentification of drought events or false alarms, whereas using the best-performing data provides a much more trustworthy depiction of drought conditions.

In addition to evaluating existing datasets, the study explored a machine-learning approach to drought prediction using SPI as a predictand. We tested four modeling techniques (Random Forest, k-Nearest Neighbors, multiple linear regression, and Gaussian Naïve Bayes) to predict short-term SPI values using antecedent climate information. The comparison revealed that simpler models can often be as effective as more complex ones in this context. The Naïve Bayes model emerged as the top performer, yielding the highest overall predictive skill for SPI at 3-, 6-, and 12-month timescales. It consistently achieved the lowest error (e.g. lowest RMSE and MAE) and highest explained variance (R^2) among the models, particularly excelling at the shortest accumulation (SPI-3) where drought onset and cessation are most rapid. In contrast, the more complex Random Forest and the instance-based KNN algorithm did not outperform the simpler approaches; in fact, KNN performed the worst overall, struggling to generalize and showing the largest prediction errors. Linear regression offered intermediate skill, slightly behind Naïve Bayes but ahead of Random Forest in some cases. These outcomes suggest that, given the relatively short data record and the stochastic nature of drought index time series, a straightforward probabilistic model like Naïve Bayes can capture the essential patterns without overfitting, whereas more complex nonlinear models may not yield added benefit. This insight is theoretically important for drought forecasting: when data are limited, simpler statistical models (which are also more interpretable) may suffice or even excel, an observation aligned with the principle of parsimony in model selection.

5.2. Recommendations

Building on these conclusions, several practical and theoretical recommendations can be made for hydrological analysis and drought management in data-sparse regions like the Shabelle Basin:

- Adopt high-performing datasets for operational use: Water and climate stakeholders should utilize the best-performing gridded datasets identified by this study – notably CHIRPS for its bias-minimized rainfall estimates and ERA5 (or its land-adjusted versions) for their event timing and consistency – as primary data sources for precipitation monitoring and drought. Relying on these products will improve the fidelity of hydrological models and drought early warning systems in the region. By contrast, caution is advised against using low-skill products like

the CPC Unified precipitation dataset for decision-making. If gauge-only analyses (e.g. CPC) must be used, they require rigorous bias correction and validation, as their raw form showed significant underestimation and could introduce large errors into drought.

- Leverage complementary strengths via data blending: The two top datasets offer complementary advantages – CHIRPS excels in rainfall magnitude and spatial detail, while ERA5 provides strength in capturing temporal Future operational frameworks should consider blending or combining these sources (and similar high-quality products) to harness their strengths. For example, merging gauge-calibrated satellite estimates with reanalysis fields could further improve rainfall representation in heterogeneous terrain. Such multi-source integration, as hinted by our findings, may enhance robustness for drought monitoring beyond what any single product can achieve. Even without formal merging, cross-validation between independent datasets (satellite vs. reanalysis) is recommended to increase confidence in detected drought signals.
- Apply the evaluation methodology to other regions: The approach developed in this research – combining a thorough multi-scale statistical evaluation with drought index verification – is broadly applicable. It is recommended that similar data-sparse basins in Africa (and elsewhere) undertake analogous intercomparisons of available precipitation products before data selection. As demonstrated, performance can vary greatly by region and; a dataset performing well in one area may falter in another. By benchmarking products against local observations and through drought index correlations, researchers and practitioners can identify which dataset (or combination) is most suitable for a given basin's climate regime. This practice will improve the reliability of hydrological simulations and hazard analyses across diverse environments.
- Improve drought monitoring and early warning practices: Given that CHIRPS and ERA5 reliably capture historical drought events in the Shabelle Basin, these datasets (and similarly validated products) should be integrated into regional drought monitoring and early warning systems. Humanitarian agencies and water managers operating in Ethiopia's Somali Region can use CHIRPS/ERA5-based SPI and SPEI analyses to fill information gaps where ground stations are sparse. The strong agreement between CHIRPS/ERA5 and observed drought indices in

this study confirms that these products can trigger timely alerts for emerging droughts with reasonable confidence. However, agencies should also be aware of each dataset's limitations: for instance, all datasets had difficulty with daily rainfall prediction, so short-term forecasts should be interpreted with caution. Emphasizing monthly to seasonal drought indicators will yield more reliable results. In addition, lower-performing datasets like CPC or certain satellite-only estimates should not be used in isolation for critical drought decisions; if they are the only available source, their indications should be cross-checked against better-performing proxies or ground data to avoid false alarms or missed events.

- Optimize the use of machine learning for drought prediction: The modeling results suggest that simple, robust algorithms (such as Naïve Bayes) can outperform more complex machine-learning models in predicting drought indices, especially with limited training data. Therefore, it is recommended that drought forecasting efforts begin with and give preference to such parsimonious models. They offer easier interpretation and lower risk of overfitting while still capturing the essential drought signal. Complex models (e.g. sophisticated neural networks or ensemble methods) should only be introduced if there is evidence of substantial improvement and if longer records or additional predictors (such as large-scale climate indices) are available to justify their complexity. In operational terms, maintaining a simpler modeling approach could be advantageous for national meteorological agencies in data-scarce regions, as it requires fewer resources and can be more easily updated or understood by decision-makers. The success of the Naïve Bayes model in this study illustrates that effective drought prediction can be achieved with relatively basic statistical tools when they are applied thoughtfully.

In conclusion, this work provides a rigorous assessment of precipitation data options and a pathway to improved drought resilience in the Shabelle Basin. By identifying which gridded products most closely approximate reality, it fills a critical knowledge gap for one of East Africa's important but understudied river basins. The insights gained – particularly the recommendation to use blended satellite–gauge data and modern reanalyses for hydro-climatic applications – can directly inform operational water management and disaster risk reduction. Moreover, the innovative combination of validation techniques demonstrated here (spatio-temporal accuracy assessment coupled

with drought index evaluation and machine-learning experimentation) serves as a template for future studies. Employing such comprehensive evaluation frameworks will enhance confidence in using gridded climate datasets in other data-sparse regions, ultimately supporting more effective planning for floods, droughts, and water resources under changing climatic conditions.

REFERENCES

Aghakouchak, A., Feldman, D., and Hsu, K. (2011). Evaluation of satellite-retrieved extreme precipitation rates across the central United States. J Geophys Res, 116(2), D02115.

Abbasi, A., Khalili, K., Behmanesh, J., and Shirzad, A. (2019). Drought monitoring and prediction using SPEI index and gene expression programming model in the west of Urmia Lake. *Theoretical and Applied Climatology*, *138*(1–2), 553–567.

Abdelwares, M., Lelieveld, J., Zittis, G., Haggag, M., and Wagdy, A. (2020). A comparison of gridded datasets of precipitation and temperature over the Eastern Nile Basin region. *Euro-Mediterranean Journal for Environmental Integration*, *5*(1), 1–16.

Ahmed, J. S., Buizza, R., Dell'Acqua, M., Demissie, T., and Pè, M. E. (2024a). Evaluation of ERA5 and CHIRPS rainfall estimates against observations across Ethiopia. *Meteorology and Atmospheric Physics*, *136*(3).

Ahmed, J. S., Buizza, R., Dell'Acqua, M., Demissie, T., and Pè, M. E. (2024b). Evaluation of ERA5 and CHIRPS rainfall estimates against observations across Ethiopia. *Meteorology and Atmospheric Physics*, *136*(3).

Ahmed, J. S., Buizza, R., Dell'Acqua, M., Demissie, T., and Pè, M. E. (2024c). Evaluation of ERA5 and CHIRPS rainfall estimates against observations across Ethiopia. *Meteorology and Atmospheric Physics*, *136*(3).

Ahmed, K., Shahid, S., Wang, X., Nawaz, N., and Najeebullah, K. (2019). Evaluation of Gridded Precipitation Datasets over Arid Regions of Pakistan. *Water 2019, Vol. 11, Page 210*, *11*(2), 210.

Asfaw, A., Simane, B., Hassen, A., and Bantider, A. (2018). Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. *Weather and Climate Extremes*, *19*, 29–41.

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., and Prat, O. P. (2015). PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, *96*(1), 69–83.

Ayehu, G. T., Tadesse, T., Gessesse, B., and Dinku, T. (2018). Validation of new satellite rainfall products over the Upper Blue Nile Basin, Ethiopia. *Atmospheric Measurement Techniques*, *11*(4), 1921–1936.

Basheer, M., and Elagib, N. A. (2019). Performance of satellite-based and GPCC 7.0 rainfall products in an extremely data-scarce country in the Nile Basin. *Atmospheric Research*, *215*, 128–140.

Bayissa, Y., Tadesse, T., Demisse, G., and Shiferaw, A. (2017). Evaluation of satellite-based rainfall estimates and application to monitor meteorological drought for the Upper Blue Nile Basin, Ethiopia. *Remote Sensing*, *9*(7).

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F. (2019). Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrology and Earth System Sciences*, *23*(1), 207–224.

Beck, H. E., Van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and De Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217.

Beguería, S., Vicente-Serrano, S. M., Reig, F., and Latorre, B. (2014). Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, *34*(10), 3001–3023.

Belay, A. S., Fenta, A. A., Yenehun, A., Nigate, F., Tilahun, S. A., Moges, M. M., Dessie, M., Adgo, E., Nyssen, J., Chen, M., Van Griensven, A., and Walraevens, K. (2019). Evaluation and Application of Multi-Source Satellite Rainfall Product CHIRPS to Assess Spatio-Temporal Rainfall Variability on Data-Sparse Western Margins of Ethiopian Highlands. *Remote Sensing 2019, Vol. 11, Page 2688*, *11*(22), 2688.

Belayneh, A., Sintayehu, G., Gedam, K., and Muluken, · Tirunesh. (2020). *Evaluation of satellite precipitation products using HEC-HMS model*. *6*, 2015–2032.

Bitew, M. M., Gebremichael, M., Ghebremichael, L. T., and Bayissa, Y. A. (2012). Evaluation of high-resolution satellite rainfall products through streamflow simulation in a hydrological modeling of a small mountainous watershed in Ethiopia. *Journal of Hydrometeorology*, *13*(1), 338–350.

Bogale, T., Degefa, S., Dalle, G., and Abebe, G. (2025). Spatio-temporal variations of drought in the Welmel watershed, southeast of Ethiopia using the vegetation condition index and standardized precipitation index. *Ecological Processes*, *14*(1), 1–21.

Cao, B., Guan, W., Li, K., Wen, Z., Han, H., and Pan, B. (2020). Area and Mass Changes of Glaciers in the West Kunlun Mountains Based on the Analysis of Multi-Temporal Remote Sensing Images and DEMs from 1970 to 2018. *Remote Sensing 2020, Vol. 12, Page 2632*, *12*(16), 2632.

Ceferino-Hernández, L., Magaña-Hernández, F., Campos-Campos, E., Morosanu, G. A., Torres-Aguilar, C. E., Mora-Ortiz, R. S., and Díaz, S. A. (2024). Assessment of PERSIANN Satellite Products over the Tulijá River Basin, Mexico. *Remote Sensing 2024, Vol. 16, Page 2596*, *16*(14), 2596.

Chen, C., Li, Z., Song, Y., Duan, Z., Mo, K., Wang, Z., and Chen, Q. (2020). Performance of multiple satellite precipitation estimates over a typical arid mountainous area of China: Spatiotemporal patterns and extremes. *Journal of Hydrometeorology*, *21*(3), 533–550.

Chen, S., Li, Q., Zhong, W., Wang, R., Chen, D., and Pan, S. (2022). Improved Monitoring and Assessment of Meteorological Drought Based on Multi-Source Fused Precipitation Data. *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 1542*, *19*(3), 1542.

Citakoglu, H., and Coşkun, Ö. (2022). Comparison of hybrid machine learning methods for the prediction of short-term meteorological droughts of Sakarya Meteorological Station in Turkey. *Environmental Science and Pollution Research*, *29*(50), 75487–75511. https://doi.org/10.1007/S11356-022-21083-3,

Dahri, Z. H., Ludwig, F., Moors, E., Ahmad, S., Ahmad, B., Shoaib, M., Ali, I., Iqbal, M. S., Pomee, M. S., Mangrio, A. G., Ahmad, M. M., and Kabat, P. (2021). Spatio-temporal evaluation of gridded precipitation products for the high-altitude Indus basin. *International Journal of Climatology*, *41*(8), 4283–4306.

Das, P., Ghosh, S., Zhang, Z., and Feng, S. (2025). Assessment of CHIRPS and PERSIANN-CDR products for meteorological drought monitoring and future trends in East Africa. *Environmental Research: Climate*.

Dembélé, M., and Zwart, S. J. (2016). Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa. *International Journal of Remote Sensing*, *37*(17), 3995–4014. https://doi.org/10.1080/01431161.2016.1207258

Dinku, T. (2019). Challenges with availability and quality of climate data in Africa. *Extreme Hydrology and Climate Variability: Monitoring, Modelling, Adaptation and Mitigation*, 71–80.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., and Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Quarterly Journal of the Royal Meteorological Society*, *144*, 292–312.

El Ibrahimi, A., and Baali, A. (2018). Application of several artificial intelligence models for forecasting meteorological drought using the standardized precipitation index in the Saïss Plain (Northern Morocco). *International Journal of Intelligent Engineering and Systems*, *11*(1), 267–275.

Elbeltagi, A., Pande, C. B., Kumar, M., Tolche, A. D., Singh, S. K., Kumar, A., and Vishwakarma, D. K. (2023). Prediction of meteorological drought and

standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. *Environmental Science and Pollution Research*, *30*(15), 43183–43202.

Elmi Mohamed, A. (2013). Managing Shared Basins in the Horn of Africa – Ethiopian Projects on the Juba and Shabelle Rivers and Downstream Effects in Somalia. *Natural Resources and Conservation*, *1*(2), 35–49.

Fenta, A. A., Yasuda, H., Shimizu, K., Ibaraki, Y., Haregeweyn, N., Kawai, T., Belay, A. S., Sultan, D., and Ebabu, K. (2018). Evaluation of satellite rainfall estimates over the Lake Tana basin at the source region of the Blue Nile River. *Atmospheric Research*, *212*, 43–53.

Ferro, C. A. T., and Stephenson, D. B. (2011). Extremal dependence indices: Improved Verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, *26*(5), 699–713.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015a). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data 2015 2:1*, *2*(1), 1–21.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015b). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data 2015 2:1*, *2*(1), 1–21.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J. (2015c). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data 2015 2:1*, *2*(1), 1–21.

Gao, Z., Huang, B., Ma, Z., Chen, X., Liu, D., and Qiu, J. (2020). Comprehensive comparisons of state-of-the-art gridded precipitation estimates for hydrological applications over southern China. *Remote Sensing*, *12*(23), 1–20.

Gebrechorkos, S. H., Hülsmann, S., and Bernhofer, C. (2018). Evaluation of multiple climate data sources for managing environmental resources in East Africa. *Hydrology and Earth System Sciences*, *22*(8), 4547–4564.

Gebrechorkos, S. H., Leyland, J., Dadson, S. J., Cohen, S., Slater, L., Wortmann, M., Ashworth, P. J., Bennett, G. L., Boothroyd, R., Cloke, H., Delorme, P., Griffith, H., Hardy, R., Hawker, L., McLelland, S., Neal, J., Nicholas, A., Tatem, A. J., Vahidi, E., … Darby, S. E. (2024). Global-scale evaluation of precipitation datasets for hydrological modelling. *Hydrology and Earth System Sciences*, *28*(14), 3099–3118.

Gebremicael, T. G., Mohamed, Y. A., Zaag, P. van der, Gebremedhin, A., Gebremeskel, G., Yazew, E., and Kifle, M. (2019). Evaluation of multiple satellite rainfall products over the rugged topography of the Tekeze-Atbara basin in Ethiopia. *International Journal of Remote Sensing*, *40*(11), 4326–4345.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., … Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454.

Ghajarnia, N., Akbari, M., Saemian, P., Ehsani, M. R., Hosseini-Moghari, S. M., Azizian, A., Kalantari, Z., Behrangi, A., Tourian, M. J., Klöve, B., and Haghighi, A. T. (2022). Evaluating the Evolution of ECMWF Precipitation Products Using Observational Data for Iran: From ERA40 to ERA5. *Earth and Space Science*, *9*(10).

Gumus, V. (2023). Evaluating the effect of the SPI and SPEI methods on drought monitoring over Turkey. *Journal of Hydrology*, *626*, 130386.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology, 377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Haile, G. G., Tang, Q., Hosseini-Moghari, S. M., Liu, X., Gebremicael, T. G., Leng, G., Kebede, A., Xu, X., and Yun, X. (2020). Projected Impacts of Climate Change on Drought Patterns Over East Africa. *Earth's Future*, *8*(7), e2020EF001502.

Helmi, A. M., and Abdelhamed, M. S. (2023). Evaluation of CMORPH, PERSIANN-CDR, CHIRPS V2.0, TMPA 3B42 V7, and GPM IMERG V6 Satellite Precipitation Datasets in Arabian Arid Regions. *Water (Switzerland)*, *15*(1), 92.

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. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Hinge, G., Mohamed, M. M., Long, D., and Hamouda, M. A. (2021). Meta-Analysis in Using Satellite Precipitation Products for Drought Monitoring: Lessons Learnt and Way Forward. *Remote Sensing 2021, Vol. 13, Page 4353*, *13*(21), 4353.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487.

Hordofa, A. T., Leta, O. T., Alamirew, T., Kawo, N. S., and Chukalla, A. D. (2021). Performance Evaluation and Comparison of Satellite-Derived Rainfall Datasets over the Ziway Lake Basin, Ethiopia. *Climate 2021, Vol. 9, Page 113*, *9*(7), 113.

Houngnibo, M. C. M., Minoungou, B., Traore, S. B., Maidment, R. I., Alhassane, A., and Ali, A. (2023). Validation of high-resolution satellite precipitation products over West Africa for rainfall monitoring and early warning. *Frontiers in Climate*, *5*, 1185754.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F., and Wolff, D. B. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55.

Hussein, A. A., and Baylar, A. (2023). Hydrological Model Evaluation of Ground, GPM IMERG, and CHIRPS precipitation data for Shabelle Basin in Ethiopia. *Journal of Electronics,Computer Networking and Applied Mathematics(JECNAM) ISSN : 2799-1156*, *3*(01), 41–60.

Kabite Wedajo, G., Kebede Muleta, M., and Gessesse Awoke, B. (2021). Performance evaluation of multiple satellite rainfall products for Dhidhessa River Basin (DRB), Ethiopia. *Atmospheric Measurement Techniques*, *14*(3), 2299–2316.

Kalisa, W., Zhang, J., Igbawua, T., Ujoh, F., Ebohon, O. J., Namugize, J. N., and Yao, F. (2020). Spatio-temporal analysis of drought and return periods over the East African region using Standardized Precipitation Index from 1920 to 2016. *Agricultural Water Management*, *237*, 106195.

Katsanos, D., Retalis, A., Tymvios, F., and Michaelides, S. (2016). Analysis of precipitation extremes based on satellite (CHIRPS) and in situ dataset over Cyprus. *Natural Hazards*, *83*, 53–63.

Dan. R., and McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241.

Li, L., She, D., Zheng, H., Lin, P., and Yang, Z.-L. (2020). Elucidating Diverse Drought Characteristics from Two Meteorological Drought Indices (SPI and SPEI) in China. *Journal of Hydrometeorology*, *21*(7), 1513–1530.

Li, Z., Zhou, Z., Chen, S., Li, Y., and Wei, C. (2024). Hydrological Evaluation of CRA40 and ERA5 Reanalysis Precipitation Products over Ganjiang River Basin in Humid Southeastern China. *Water 2024, Vol. 16, Page 2774*, *16*(19), 2774.

Maghsood, F. F., Hashemi, H., Hosseini, S. H., and Berndtsson, R. (2020). Ground validation of GPM IMERG precipitation products over Iran. *Remote Sensing*, *12*(1).

Mulualem, G. M., Raju, U. J. P., Stojanovic, M., and Sorí, R. (2024). The phenomenon of drought in Ethiopia: Historical evolution and climatic forcing. *Hydrology Research*, *55*(6), 595–612.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383.

Nadeem, M. U., Ghanim, A. A. J., Anjum, M. N., Shangguan, D., Rasool, G., Irfan, M., Niazi, U. M., and Hassan, S. (2022). Multiscale Ground Validation of Satellite and Reanalysis Precipitation Products over Diverse Climatic and Topographic Conditions. *Remote Sensing 2022, Vol. 14, Page 4680*, *14*(18), 4680.

Nakkazi, M. T., Sempewo, J. I., Tumutungire, M. D., and Byakatonda, J. (2022). Performance evaluation of CFSR, MERRA-2 and TRMM3B42 data sets in simulating river discharge of data-scarce tropical catchments: a case study of Manafwa, Uganda. *JWCC*, *13*(2), 522–541.

Navidi Nassaj, B., Zohrabi, N., Nikbakht Shahbazi, A., and Fathian, H. (2021). Evaluation of three gauge-based global gridded precipitation datasets for drought monitoring over Iran. *Hydrological Sciences Journal*, *66*(14), 2033–2045.

Navidi Nassaj, B., Zohrabi, N., Nikbakht Shahbazi, A., and Fathian, H. (2022). Evaluating the performance of eight global gridded precipitation datasets across Iran. *Dynamics of Atmospheres and Oceans*, *98*.

Nwayor, I. J., and Robeson, S. M. (2024). Exploring the relationship between SPI and SPEI in a warming world. *Theoretical and Applied Climatology*, *155*(4), 2559–2569.

Omondi, O. A., and Lin, Z. (2023). Trend and spatial-temporal variation of drought characteristics over equatorial East Africa during the last 120 years. *Frontiers in Earth Science*, *10*, 1064940.

Pei, Z., Fang, S., Wang, L., and Yang, W. (2020). Comparative Analysis of Drought Indicated by the SPI and SPEI at Various Timescales in Inner Mongolia, China. *Water 2020, Vol. 12, Page 1925*, *12*(7), 1925.

Philip, S., Kew, S. F., van Oldenborgh, G. J., Otto, F., O'Keefe, S., Haustein, K., King, A., Zegeye, A., Eshetu, Z., Hailemariam, K., Singh, R., Jjemba, E., Funk, C., and Cullen, H. (2018). Attribution analysis of the Ethiopian drought of 2015. *Journal of Climate*, *31*(6), 2465–2486.

Piraei, R., Niazkar, M., Gangi, F., Eryılmaz Türkkan, G., and Afzali, S. H. (2024). Short-Term Drought Forecast across Two Different Climates Using Machine Learning Models. *Hydrology*, *11*(10), 163.

Polong, F., Chen, H., Sun, S., and Ongoma, V. (2019). Temporal and spatial evolution of the standard precipitation evapotranspiration index (SPEI) in the

Tana River Basin, Kenya. *Theoretical and Applied Climatology*, *138*(1–2), 777–792.

Popovych, V., and Dunaieva, I. (2021). Assessment of the GPM IMERG and CHIRPS precipitation estimations for the steppe part of the Crimea. *Meteorology Hydrology and Water Management*.

Poudel, B., Dahal, D., Banjara, M., and Kalra, A. (2024). Assessing Meteorological Drought Patterns and Forecasting Accuracy with SPI and SPEI Using Machine Learning Models. *Forecasting 2024, Vol. 6, Pages 1026-1044*, *6*(4), 1026–1044.

Prakash, S., Mitra, A. K., Pai, D. S., and AghaKouchak, A. (2016). From TRMM to GPM: How well can heavy rainfall be detected from space? *Advances in Water Resources*, *88*, 1–7.

Ramadhan, R., Yusnaini, H., Marzuki, M., Muharsyah, R., Suryanto, W., Sholihun, S., Vonnisa, M., Harmadi, H., Ningsih, A. P., Battaglia, A., Hashiguchi, H., and Tokay, A. (2022). Evaluation of GPM IMERG Performance Using Gauge Data over Indonesian Maritime Continent at Different Time Scales. *Remote Sensing*, *14*(5).

Romilly, T. G., and Gebremichael, M. (2011). Evaluation of satellite rainfall estimates over Ethiopian river basins. *Hydrology and Earth System Sciences*, *15*(5), 1505–1514.

Sadeghi, M., Nguyen, P., Naeini, M. R., Hsu, K., Braithwaite, D., and Sorooshian, S. (2021). PERSIANN-CCS-CDR, a 3-hourly 0.04° global precipitation climate data record for heavy precipitation studies. *Scientific Data*, *8*(1), 157.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y. T., Chuang, H. Y., Juang, H. M. H., Sela, J., … Goldberg, M. (2010). The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y. T., Chuang, H. Y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., Van Den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2014). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208.

Samim, A. T., Nayyer, F., Hussainzada, W., and Lee, H. S. (2024). Intercomparison of gridded global precipitation data for arid and mountainous regions: A case study of Afghanistan. *Journal of Hydrology: Regional Studies*, *53*.

Satgé, F., Defrance, D., Sultan, B., Bonnet, M. P., Seyler, F., Rouché, N., Pierron, F., and Paturel, J. E. (2020). Evaluation of 23 gridded precipitation datasets across West Africa. *Journal of Hydrology*, *581*.

Schneider, D. P., Deser, C., Fasullo, J., and Trenberth, K. E. (2013). Climate data guide spurs discovery and understanding. *Eos*, *94*(13), 121–122.

Steinkopf, J., and Engelbrecht, F. (2022). Verification of ERA5 and ERA-Interim precipitation over Africa at intra-annual and interannual timescales. *Atmospheric Research*, *280*.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K. L. (2018). A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics*, *56*(1), 79–107.

Tadesse, K. E., Melesse, A. M., Abebe, A., Lakew, H. B., and Paron, P. (2022). Evaluation of Global Precipitation Products over Wabi Shebelle River Basin, Ethiopia. *Hydrology*, *9*(5).

Taye, M., Sahlu, D., Zaitchik, B. F., and Neka, M. (2020). Evaluation of Satellite Rainfall Estimates for Meteorological Drought Analysis over the Upper Blue Nile Basin, Ethiopia. *Geosciences 2020, Vol. 10, Page 352*, *10*(9), 352.

Thiemig, V., Pappenberger, F., Thielen, J., Gadain, H., de Roo, A., Bodis, K., Del Medico, M., and Muthusi, F. (2010). Ensemble flood forecasting in Africa: A feasibility study in the Juba-Shabelle river basin. *Atmospheric Science Letters*, *11*(2), 123–131. https://doi.org/10.1002/ASL.266

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., and De Roo, A. (2012). Validation of satellite-based precipitation products over sparsely Gauged African River basins. *Journal of Hydrometeorology*, *13*(6), 1760–1783.

Uysal, G., and Şorman, A. Ü. (2021). Evaluation of PERSIANN family remote sensing precipitation products for snowmelt runoff estimation in a mountainous basin. *Hydrological Sciences Journal*, *66*(12), 1790–1807.

Wagesho, N., and Claire, M. (2016). Analysis of Rainfall Intensity-Duration-Frequency Relationship for Rwanda. *Journal of Water Resource and Protection*, *08*(07), 706–723.

Wang, X., Ding, Y., Zhao, C., and Wang, J. (2019). Similarities and improvements of GPM IMERG upon TRMM 3B42 precipitation product under complex topographic and climatic conditions over Hexi region, Northeastern Tibetan Plateau. *Atmospheric Research*, *218*, 347–363.

Wang, Z., Zhong, R., Lai, C., and Chen, J. (2017). Evaluation of the GPM IMERG satellite-based precipitation products and the hydrological utility. *Atmospheric Research*, *196*, 151–163.

Ware, M. B., Matewos, T., Guye, M., Legesse, A., and Mohammed, Y. (2023). Spatiotemporal variability and trend of rainfall and temperature in Sidama Regional State, Ethiopia. *Theoretical and Applied Climatology*, *153*(1–2), 213–226.

Wei, L., Jiang, S., Ren, L., Wang, M., Zhang, L., Liu, Y., Yuan, F., and Yang, X. (2021). Evaluation of seventeen satellite-, reanalysis-, and gauge-based precipitation products for drought monitoring across mainland China. *Atmospheric Research*, *263*.

Zargar, M., Bronstert, A., Francke, T., Zimale, F. A., Worku, K. B., Wiegels, R., Lorenz, C., Hageltom, Y., Sawadogo, W., and Kunstmann, H. (2025). Comparison and hydrological evaluation of different precipitation data for a large tropical region: the Blue Nile Basin in Ethiopia. *Frontiers in Water*, *7*.

CURRICULUM VITAE

ORCID ID: 0000-0002-1622-7854

Name Surname : Abdinour Hussein

Foreign Language : Somali, English, Turkish, Arabic, and Amharic.

Education:

- 2025, Eskisehir Technical University, Institute of Graduate Programs, Deprtment of Civil Engineering, PhD
- 2020, Anadolu Universtiy, Institute of Graduate Programs, Department of Civil Engineering, Msc.
- 2016, Arbaminch University, Department of Hydraulic and Water Resource Engineering, Bsc

Professional Background:

- 2025, Civil Engineer, Minnesota Department of Transportation
- 2023, National Engineer, United Nations Office of Project Services
- 2023, Lecturer, Jigjiga University,
- 2017, Infrastructure Engineer, World Vision International

Publications and/or Scientific/Artistic Activities:

1. Hussein, A.A., & Baylar, A. (2023). Hydrological Model Evaluation of Ground, GPM IMERG, and CHIRPS precipitation data for Shabelle Basin in Ethiopia. Journal of Electronics, Computer Networking and Applied Mathematics. DOI:10.55529/jecnam.31.41.60

Awards:

1. 2020 Ph.D. Fellowship award, Presidency for Turks Abroad and Related Communities, Turkey
2. 2017 MSc Fellowship Award, Presidency for Turks Abroad and Related Communities, Turkey
3. 2011, Ethiopian Higher Education Scholarship, Ethiopia.

Professional Union/Association/Organization Memberships:

- 2025, Minnesota Engineering Union, Minnesota, USA
- 2025, American Society of Engineer, Minnesota, USA