
Crop Water Productivity Estimation for Resilient Agriculture Using Remote Sensing and Machine Learning Approach

by

Alfred Amboka

Project research report submitted to the department of Geomatic Engineering and Geospatial Information Systems for the award of degree of Bachelor of Science in Geomatic Engineering and Geospatial Information Systems (GEGIS), 2023.



**Department of Geomatic
Engineering and Geospatial
Information Systems (GEGIS)**

DECLARATION

I declare that this project is my work and has not been submitted by anybody else in any other university for the award of any degree to the best of my knowledge.

Sign.....

Date.....

Alfred Amboka

ENC221-0099/2018

Department of Geomatic Engineering and Geospatial Information Systems (GEGIS)

Jomo Kenyatta University of Agriculture and Technology

CERTIFICATION

This project has been submitted for examination with my approval as the candidate's supervisor.

Sign.....

Date

Mr. Moffat Magondu

Senior Lecturer, GEGIS

Department of Geomatic Engineering and Geospatial Information Systems (GEGIS)

©GEGIS 2023

Acknowledgments

The successful completion of this research project is attributable to several individuals' invaluable guidance and support. I would like to express my utmost gratitude to my supervisor, Mr. Moffat Magondu, Senior Lecturer at the Department of GEGIS, for his unwavering mentorship. His vast expertise and wealth of knowledge empowered me to explore practical solutions and real-world impacts through my research. I am deeply grateful for his encouragement, which inspired critical thinking and resilience in the face of research challenges.

In addition, I sincerely appreciate the provision of authoritative weather and irrigation data from the Kenya Meteorological Department and Kenya Irrigation Authority. Access to this crucial information formed the sound basis for validating project findings. Similarly, I acknowledge the open-source community for the multitude of free tools and satellite imagery that enabled this study's methodological approaches. This project would not have materialized without those vital contributions.

I sincerely thank the Department of GEGIS for equipping me with the foundational capacity to execute this research project. I also thank my classmates for thought-provoking discussions that stimulated new dimensions of my work.

Finally, my deepest gratitude goes to my family for their unending love, patience, and emotional support throughout the demanding research process. I especially acknowledge the financial assistance that facilitated smooth progress with the project.

The completion of this research project was a collaborative effort. I humbly acknowledge all the named and unnamed contributors for their input in this submission.

Abstract

Crop production is highly affected by water scarcity, which has a negative impact on food security. Crop Water Productivity (CWP), defined as crop yield per cubic meter of water consumption, enhances agricultural production, especially in irrigation-based farming. Therefore, CWP has been recognized as a critical performance-based evaluation indicator for resilient agriculture. Although the estimation of CWP has been achieved in previous studies, using ground-based methods, coarse spatial and temporal resolution-based imagery, statistical methods, and limited application of machine and deep learning calls for further analysis. This study aimed to: i) estimate maize yield and determine evapotranspiration (ET) values based on analyzed maize crop phenological period in Bura Irrigation Scheme; ii) estimate maize crop water productivity and analyze the spatial distribution of CWP; iii) develop machine learning models for the estimation of CWP for maize crops and use the machine learning models for CWP estimation in the study area.

The methods involve Data fusion (MODIS, Landsat, and Sentinel-2) to obtain daily high spatial resolution datasets, estimation of Evapotranspiration (ET) using the SEBAL algorithm and Penman-Monteith (P-M) equation, the estimation of yield by Biomass-Harvest index method and CWP as the ratio of yield and ET. In addition, the XGBoost model was developed to improve the estimation of CWP. The results from statistical and machine learning-based estimations were temporally and spatially consistent across the Bura Irrigation scheme (Tana River County, Kenya). Yield averaged 3.3 t/ha, below the global average of 4.9 t/h. On the other hand, the ET averaged 56mm, and CWP averaged 5.1 kg/m³ during the maize growing seasons, indicating overage productivity. Spatially, the productivity within Bura Irrigation was higher in the central and northern regions than in the southern regions throughout the study period. The XGBoost model successfully estimated CWP, Yield, and ET using imagery bands and specifically calculated indices. The model achieved an accuracy $R^2 > 75\%$ for the target variable, where ET had the highest (91%) learning and prediction rate. The machine learning (ML) estimate of CWP was 5.8, close to the statistical average. Finally, these results show the need to use CWP and ML estimation methods to enhance agricultural resilience, resulting in increased food security.

Table of contents

Contents

DECLARATION	i
CERTIFICATION	i
Acknowledgments	- 1 -
Abstract	- 2 -
Table of contents	- 3 -
List of Figures	- 6 -
List of Tables	- 7 -
Acronyms and abbreviations	- 8 -
Introduction	- 10 -
1.1 Background	- 10 -
1.2 Statement of the Problem	- 13 -
1.3 Justification	- 14 -
1.4 Research Identification and Objectives.....	- 15 -
1.4.1 Research Objectives	- 15 -
1.4.2 Research Questions.....	- 15 -
1.5 Study outline	- 16 -
2. Literature review	- 17 -
2.1 Water Use and Water Productivity.....	- 17 -
2.2 Crop Water Productivity Estimation.....	- 18 -
CWP Estimation Methods.....	- 19 -
2.2.1 Surface Energy Balance Algorithm (SEBAL).....	- 19 -
	- 3 -

2.2.2 Field-Based CWP Methods	- 20 -
2.2.3 Remote sensing and Machine learning Approaches	- 21 -
3. Materials and methods	- 23 -
3.1 Study area	- 23 -
3.2 Data	- 24 -
3.2.1 Meteorological Data	- 24 -
Optical Satellite Imagery Datasets	- 24 -
3.3 Methodology Flowchart	- 26 -
3.4 Data Fusion and Indices Calculation	- 27 -
3.4.1 Daily Imagery Dataset	- 27 -
3.4.2 Indices	- 29 -
3.5 Crop Yield	- 30 -
3.6 Evapotranspiration Estimation	- 32 -
3.7 Crop Water Productivity	- 34 -
3.8 Machine Learning	- 34 -
3.8.1 Machine learning models used to estimate Yield, ETa, and CWP	- 34 -
3.8.2 Performance Metrics and Evaluation	- 36 -
4. Results	- 37 -
4.1 Crop Phenology	- 37 -
4.2 Yield Estimation	- 39 -
4.2.1 Classification	- 39 -
4.2.2 Vegetation distribution	- 41 -
4.3 Regional Evapotranspiration Estimates	- 45 -
4.4 CWP Spatial Distribution	- 48 -
4.5 CWP, Yield and ET Trend View	- 51 -
4.6 Relative Importance of CWP Estimation Parameters	- 54 -



4.7 Xgboost Model results - 56 -

4.8 Validity of Machine Learning Model (XGBoost)..... - 57 -

5. Discussion..... - 59 -

5.1 Crop Phenology Dynamics..... - 59 -

5.2 Analysis of Yield and ET Results - 60 -

5.3 Spatio-temporal Patterns in Crop Water Productivity..... - 61 -

5.4 Machine Learning for Enhanced CWP Estimation - 62 -

6. Conclusions..... - 64 -

7. References - 66 -

Appendix..... - 78 -

1. Data - 78 -

2. Implementation Code and Models - 79 -

List of Figures

Figure 3. 1 Study area map – Bura Irrigation Scheme.....	23 -
Figure 3. 2 Flowchart Diagram.....	26 -
Figure 3. 3 Comparison of Daily MODIS and fused dataset.....	29 -
Figure 4. 1 Crop growth curve - Representation of crop phenology per growing period, using NDVI.....	38 -
Figure 4. 2 Crop growth curve - Representing crop phenology per growth period using EVI.....	38 -
Figure 4. 3 Land Use Land Cover map.....	40 -
Figure 4. 4 Visualization of vegetation cover throughout the study period (2018 - 2022).....	41 -
Figure 4. 5a Spatial Temporal Yield distribution during the short period between 2018 to 2022. . .	42 -
Figure 4. 6 Spatial Temporal Yield distribution during the long growing period between 2018 to 2022.	43 -
Figure 4. 7 Yearly maize yield regional based average.	45 -
Table 4. 2 Yield zonal statistics at season level and on yearly basis	45 -
Figure 4. 8: Short season spatial temporal distribution of Evapotranspiration.....	47 -
Figure 4. 12: Trend analysis of the changes between yield and CWP for the whole period of study.....	51 -
Figure 4. 13: Multivariable comparison and trend analysis of CWP, Yield and Evapotranspiration during the short growing season.	52 -
Figure 4. 14: Multivariable comparison and trend analysis of CWP, Yield and Evapotranspiration during the long season.	53 -
Figure 4. 15: Results of the Correlation analysis for the variables used in the machine learning modelling.	54 -
Figure 4. 16: XGBoost model results analysis for CWP, Yield and ET; Scatter plot.	56 -
Figure 4. 17: Machine learning model (XGBoost) prediction residuals.....	57 -
Figure 5. 1: Main Local weather dataset showing the seasonal variation in rainfall and temperature between 2016 to 2022 on monthly basis	59 -

List of Tables

Table 3.1 Datasets used in this study	- 25 -
Table 3.2 Spectral indices, Expressions, and references	- 29 -
Table 4.1 Classification error matrix obtained from classification of land use and land cover of Bura Irrigation Scheme using random forest, 2022.....	- 39 -
Table 4.2 Regional Zonal statistics for CWP, Eta, ETo, and Yield were obtained yearly and seasonal.....	- 53 -

Acronyms and abbreviations

ANNs	Artificial Neural Networks
APAR	Absorbed Photosynthetically Active Radiation
ASALs	Arid and Semi-Arid Lands
CWP	Crop Water Productivity
EDI	Evaporative Deficit Index
ET	Evapotranspiration
ETa	Actual Evapotranspiration
ETo	Reference Evapotranspiration
EVI	Enhanced Vegetation Index
f	Fraction of Absorbed Photosynthetically Active Radiation
FAO	Food and Agriculture Organization of the United Nations
Gas	Genetic Algorithms
GEE	Google Earth Engine
GNDVI	Green Normalized Difference Vegetation Index
HI	Harvest Index
IPC	Integrated Food Security Phase Classification
KALRO	Kenya Agriculture and Livestock Research Organization
Kc	Crop Coefficient
KMD	Kenya Meteorological Department
LST	Land Surface Temperature
LUE	Light Use Efficiency
MAE	Mean Absolute Error

ML	Machine Learning
MNDWI	Modified Normalized Difference Water Index
MODIS	Moderate Resolution Imaging Spectroradiometer
MSAVI	Modified Soil Adjusted Vegetation Index
MSE	Mean Squared Error
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIA	National Irrigation Authority
RFs	Random Forests
RMSE	Root Mean Squared Error
RS	Remote Sensing
SAI	Science for Arid Lands
SMCI	Soil Moisture Condition Index
SRM	Structural Risk Minimization
STARFM	Spatial and Temporal Adaptive Reflectance Fusion Model
SVM	Support Vector Machine
SVR	Support Vector Regression
TCI	Temperature Condition Index
TVDI	Temperature Vegetation Dryness Index
VC	Vapnik-Chervonenkis
VCI	Vegetation Condition Index
WRI	Water Ratio Index

Introduction

1.1 Background

Agriculture is the backbone of many economies, providing food and livelihoods for millions of people worldwide. However, increasing demand for food coupled with the effects of climate change has strained water resources, leading to water scarcity in many regions. Water scarcity refers to the limited availability of freshwater resources to meet the demands of various sectors, including agriculture. This significant challenge affects agricultural production and food security in many regions of the world (Booker & Trees, 2020). Agriculture depends entirely on water, which can be both rainfed or irrigated. Estimation of water used and yield produced is therefore necessary. This is enhanced through crop water productivity.

Crop water productivity (CWP) is generally defined as crop yield per cubic meter of water consumption or the ratio of yields to evapotranspiration during the growing season (H. Gao et al., 2023; Hellegers et al., 2009). Crop water productivity estimation is a crucial aspect of precision agriculture, which aims to optimize agronomic inputs like water, pesticides, and fertilizers to meet the growing demand for food while minimizing the use of natural resources like land and fresh water. The productivity of crops can vary depending on whether they are grown in rain-fed or irrigated agriculture systems.

Irrigated agriculture accounts for an estimated 70% of total freshwater withdrawals worldwide, and in many drier countries, agricultural water use accounts for more than 90% of total withdrawals (Scheierling & Tréguer, 2018). As water becomes increasingly scarce, the management of agricultural irrigation moves to the center of water management concerns. Without advances in management and more integrated policy-making in developed and developing countries, water scarcity and related water problems will significantly worsen over the next several decades. The transition from an expansionary to a maturing water economy has led to the need for more efficient water use in agriculture. The concept of water productivity, which denotes the relationship between marketable yield and the seasonal water use by the plant through evapotranspiration, is an important indicator to express the resource use efficiency and can provide an assessment of crop performance under different irrigation strategies (Hommadi & Almasraf, 2019). Rain-fed agriculture, on the other hand, is characterized by low average yields compared

to irrigated agriculture, as rainfall rarely meets the time with the required amount of water application for plant growth. The performance of rain-fed productivity remains low and stable for most crops, and crop production is undulating under rain-fed agriculture. In many countries, rain-fed areas are the critical cultivation areas with the largest concentration of rural poverty spanning several agroecological regions. The low efficiency of water uses and management in agriculture is a significant challenge, and one of the contributing factors to low crop productivity and balance is drought.

Estimations of CWP are affected by drought conditions. Therefore, understanding the climatic conditions of the region of interest is paramount for such studies. According to (Sarshad et al., 2021), drought is the most significant environmental stress in arid and semiarid regions, which has restricted agricultural development. Those agricultural activities that have to survive in such areas are exposed to extreme temperatures, variations in rainfall, long solar radiation hours, and so on. Kenya is particularly vulnerable to drought due to its geography and climate. Several droughts are common in agricultural lands in Kenya and most tropical regions. The different types of droughts, their severity, and their impacts depend on various factors, such as the duration, intensity, and spatial extent of the drought event. Meteorological drought is the most common type of drought, which occurs when there is a prolonged period of below-average precipitation. Agricultural drought refers to the impact of meteorological drought on crop production. In contrast, hydrological drought is characterized by low water availability in rivers, lakes, and groundwater as drought impact on agriculture influences crop evapotranspiration (ET) and general yield obtained, hence the need to look into how to manage available water resources for maximum production in drought conditions. This brings us to why CWP has established itself as a recognized indicator for evaluating progress toward SDG 6.4, which calls for much greater water usage efficiency (Blatchford et al., 2018; Ghorbanpour et al., 2022).

The level of CWP estimation varies with scale, including at the farm and regional levels. Several factors, including data and the extent of the study area, influence such estimation's robustness. Farm-level estimation is challenging and limited in one way, considering the high volume of estimation data required when performing time series estimation. Traditionally, CWP has relied on labor-intensive and time-consuming field-based methods, such as lysimeters and soil moisture sensors. These methods are often limited in spatial coverage and cannot provide real-

time data for large-scale agricultural areas. Moreover, they may not account for spatial variations in soil moisture and crop water requirements within a field. This has given way to remote sensing for crop water productivity estimations. Remote sensing can help estimate actual evapotranspiration (ET) and crop yield, which are important factors in CWP estimation (Ghorbanpour et al., 2022; Gao et al., 2023). This is important for improving regional agricultural water use efficiency and conservation levels. Finally, remote sensing coupled with machine learning has increased in many fields, including agriculture and water use efficiency estimation.

Machine learning is becoming an increasingly popular tool for estimating crop water productivity. Existing machine learning methods continue to prove to be more reliable through data fusion and the combination of several models (Elbeltagi et al., 2022). On remote sensing data for farming, many machine learning algorithms have been applied, including random forests (RFs), support vector machines (SVMs), artificial neural networks (ANNs), genetic algorithms (GAs), and ensemble learning (Virnodkar et al., 2020; Sadri et al., 2022). Particularly in geographic classification and remote sensing data prediction, RF applications have gained popularity for resolving data overfitting (Sadri et al., 2022; Vergopolan et al., 2021; Saini & Ghosh, 2018). Although machine learning has been used in many fields, such as yield estimation (Islam et al., 2023), weather forecasting (Patel et al., 2021), and remote sensing, it is underutilized in crop water productivity estimation. The variation is from the global level to the country and local level. Therefore, this study focuses on CWP estimation at the local level and specifically in irrigation schemes in the eastern part of Kenya - Bura Tana River Scheme.

The Bura Irrigation Scheme covers a total area of 5,360 hectares, although only 3,340 hectares are now used for irrigation due to a lack of water resources. Still, there is potential to enhance water availability, allowing irrigation of more significant areas and increasing maize output, which is now low. The plan now yields 3.5 tons per hectare of commercial maize and 4.4 tons per hectare of seed maize. (Muigai David et al., 2019). This is less than the 4.9 t ha⁻¹ average for the entire world value. Therefore, accurate water management in this area is essential, and proper estimates of how water is used and yield produced will be necessary. Using machine learning coupled with remote sensing will allow proper mitigation measures to be applied. Therefore, CWP estimation is necessary to boost production while withstanding drought conditions and help reduce the growing water scarcity.

1.2 Statement of the Problem

One of the significant challenges in agricultural systems, particularly in irrigation schemes in Kenya, is the limited availability of water resources. Water scarcity continues to increase due to the impact of climate change. According to the UN's 2022 report, over 85% of the wetlands on our planet have been lost for over 300 years. The water crisis problem has been worsened by factors such as water contamination, population increase, urbanization, and inadequate management of water resources. Therefore, food security is going to be impacted by water shortage (Mulwa et al., 2021). On the other hand, drought continues to rage in Kenya, resulting in reduced agricultural land and affecting agricultural production, especially in the supply of water for production.

The physiological and biochemical processes of plants are predicted to be affected by soil water stress, which is a significant barrier to agricultural production, particularly in arid and semi-arid lands (ASALs) (M. N. et al., 2015); Mbayaki, 2021). By managing crops and water poorly, plant quality and yields may be harmed (Fan et al., 2012; Mbayaki, 2021).

Accurate estimation of CWP requires comprehensive data on crop growth, soil moisture, and water availability. However, obtaining such data is challenging and often needs to be improved in the Bura Irrigation Scheme. Existing water productivity methods are limited, and the data used significantly contributes to this effect. Bura Irrigation scheme, being among the largest in Kenya, has faced low records of yield production, as outlined in a paper by (Muigai et al., 2019). The scheme relies on water from river Tana, which is 50 km away. Over recent years, the river Tana's water level has been reducing due to drought effects upstream. This impacts the downstream water supply for agricultural purposes, including the Bura Irrigation scheme. Optimizing water use efficiency in such scenarios is crucial for sustainable and resilient agriculture. However, conventional irrigation practices often result in inefficient water use and lower crop productivity. There is a need to address this problem by developing methods to estimate and improve crop water productivity by utilizing remote sensing (RS) and machine learning techniques. Effective water management strategies can be devised to enhance agricultural productivity while conserving water resources by accurately assessing water requirements and usage patterns.

Water scarcity continues to rage in arid and semi-arid lands due to drought. Bura Irrigation scheme, within agro-ecological zone V (semi-arid to arid), obtains its water from the Tana River, which is 50 km away, by pumping water from the river (National Irrigation Authority, 2023). In

addition, the area experiences low rainfall of about 400mm. High Temperatures are experienced all year round with slight seasonal variation. Mean maximum temperatures never fall below 31°C, and average minimum temperatures are above 20°C. All these factors make the agricultural area more vulnerable and costly, especially in the case of water pumping to meet crop water needs in the area. Water Crop water productivity is a crucial requirement for increased agricultural production because crop water is needed to substitute for water loss by transpiration and soil evaporation (Mbayaki, 2021)

Over the recent years, food insecurity has affected many parts of Kenya, with Eastern and Northern regions being highly affected. According to the Integrated Food Security Phase Classification (IPC), in 2023, an estimated 4.4 million people in ASALs will face acute food insecurity. Generally, acute food insecurity has affected around 37% of the population in Kenya between 2022 and 2023 and is expected to increase. Food insecurity continues to rise with increasing agricultural drought events, rapid population growth, water pollution reducing water use for agricultural purposes, and high demand that strains available water resources. In addition, Kenya declared In September 2021 the East African state of Kenya drought emergency. The affected drought areas (ASALs) continually received low rainfall for the season between November and December (International Committee of the Red Cross, 2022), leading to low agricultural production. With all these drought events, agricultural production can be optimized to survive water scarcity and boost productivity by managing available water resources. CWP serves as the best alternative for the estimation of crop productivity in existing irrigation schemes and may serve as a near real-time decision-making tool on water management to increase agricultural production, hence increasing food security.

1.3 Justification

Accurately estimating and optimizing crop water productivity (CWP) is crucial for agriculture resilience under water scarcity and drought conditions. This is especially important for farming regions like Kenya's arid and semi-arid lands (ASALs), which face chronic water deficit challenges (Booker & Trees, 2020; Sarshad et al., 2021). One such region is the Bura Irrigation Scheme in Tana River County. However, river flows have declined with increasing

drought severity and rainfall variability (Mulwa et al., 2021), threatening irrigation availability downstream and requiring improved water use efficiency.

This research addresses the urgent need for enhanced food security amidst increasing aridity in Bura and other ASAL croplands (Integrated Food Security Phase Classification, 2023; International Committee of the Red Cross, 2022). Insights into spatiotemporal CWP patterns can bolster resilience by targeting interventions like deficit irrigation (Shoukat et al., 2021), drainage upgrades, or alternative agronomic practices to raise productivity in struggling zones first (Jaafar & Ahmad, 2020;). If successful, the procedures could be expanded to additional Kenyan irrigation districts challenged by drought and irregular river flows (Mulwa et al., 2021). Strategically stretching limited water supplies, remote sensing, and machine intelligence (Elbeltagi et al., 2022; Khan et al., 2018) may help shield vulnerable breadbasket areas.

1.4 Research Identification and Objectives

1.4.1 Research Objectives

The main objective was to estimate crop water productivity using remote sensing and machine learning techniques from 2018 to 2022 in the Bura irrigation scheme to support irrigation management and improve agricultural resilience. Specific objectives include:

- ❖ To estimate maize yield and determine evapotranspiration (ET) values based on analyzed maize crop phenological period in the study area,
- ❖ To estimate maize crop water productivity and analyze the spatial distribution of CWP,
- ❖ To develop machine learning models estimating crop water productivity (CWP) for maize crops and use the machine learning models for CWP estimation in the study area.

1.4.2 Research Questions

The following questions are formulated for this specific study.

- How do water scarcity and drought adversely affect crop productivity within the Bura Irrigation Scheme?
- In what ways can machine learning be effectively employed to enhance Crop Water Productivity (CWP) modeling, and what are the current levels of utilization of this emerging tool in addressing water-related challenges in agricultural settings?

-
- What specific factors contribute to the low crop water productivity and frequent droughts in Kenya's Arid and Semi-Arid Lands (ASALs), particularly in agricultural regions, and how do these factors contribute to food insecurity?

1.5 Study outline

This research study is divided into 6 chapters, whereby the first chapter introduces.

The study details the background, statement of the problem, justification of the problem, and objectives. Chapter 2 contains the reviewed literature related to this topic. Further, Chapter 3 shows the data and methods used in the study, with Chapter 4 highlighting the results of the findings from the methods. Chapter 5 discusses the findings, and Chapter 6 concludes and recommends future research that might not be addressed at this level of geoscientific exploration and expertise

2. Literature review

2.1 Water Use and Water Productivity

Crop water productivity (CWP) is an important concept in agronomy that seeks to maximize viable yields per unit of water used in rain-fed and irrigated agricultural operations. CWP may be accomplished by increasing crop marketable yields per unit of transpired water and decreasing water loss from the soil water balance (Mbayaki, 2021). Water use and efficiency refer to CWP and are frequently used synonymously.

Crop water productivity (CWP) is essential in irrigated agriculture for food and the environment's security, especially when water becomes limited. (Bekchanov et al., 2012). Estimates have already been made using crop models in previous research. Crop modeling is an effective method for calculating WP and is important in water management. (Soomro et al., 2019). Commonly used crop models for maize include CERES-Maize (Crop Environment Resource Synthesis), SWAT, SWAP (Soil Water Atmosphere Plant), AQUACROP, CROPWAT, and more. The CERES-Maize model (Cuculeanu et al., 2002) is specifically designed to simulate the growth, development, and yield of maize (corn) crops under different agroclimatic and management conditions. It has been widely used to assess crop water productivity for maize and to study the impact of water management strategies on maize production (Kisekka et al., 2017; Sen et al., 2023). The model incorporates a comprehensive water balance approach, considering various inputs and outputs for the maize crop. These inputs include rainfall, irrigation, and soil water content, while outputs include evapotranspiration (ET) and drainage losses. The model estimates how efficiently the maize crop uses water to produce yield by simulating the water balance. Nonetheless, other crop models can accurately estimate crop water productivity, such as the FAO Aqua Crop model. The Aqua Crop model simulates attainable yields of major herbaceous crops as a function of water consumption under rainfed, supplemental, deficit, and full irrigation conditions. Similarly, the model has been widely used to assess water productivity (Mostafa et al., 2023); (Shan et al., 2023). Additionally, some studies have proposed a combined method for estimating the spatial and temporal variation of crop water productivity under deficit irrigation scenarios based on the Aqua Crop model (Ahmadpour et al., 2022).

2.2 Crop Water Productivity Estimation

The CWP estimation is achieved by the ratio of yield and actual evapotranspiration (ET_a) (Talpur et al., 2023; Yihun et al., 2013).

The phenological period, which refers to crop growth and development stages, is crucial in estimating ET and crop yield (H. Gao et al., 2023). Factors such as crop coefficient (K_c) and harvest index (HI), which vary during different phenological periods, are considered in estimating ET and yield. Different crops have varying crop phenology, growing periods, quantity harvested, and crop response to the environmental conditions that influence crop yield. The rate of water loss from crops through transpiration (crop transpiration) is determined by taking the reference evapotranspiration value (ET_o) for a particular region and multiplying it by a crop coefficient (K_{cTr}) that is specific to the crop type. The crop coefficient adjusts the reference evapotranspiration to account for differences between crops in transpiration rates under the same environmental conditions. According to FAO, the crop coefficient is proportional to canopy cover and varies throughout the life cycle of a crop. It is affected by water stress, which can affect canopy development and induce stomata closure, directly affecting crop transpiration. A study by (H. Gao et al., 2023) utilized the dry matter mass–harvest index, crop K_c, based on crop phenology to map crop water productivity of maize.

The evapotranspiration (ET) formula estimates the amount of water crops use. The ET formula considers temperature, humidity, wind speed, and solar radiation factors. The Penman-Monteith equation is a widely used ET formula recommended by the United Nations Food and Agriculture Organization (FAO). The Penman-Monteith equation combines energy balance and aerodynamic resistance equations and is considered the most accurate method for estimating ET. (Wang et al., 2023; (Hassan et al., 2022) and employed the ET formula in the estimation of evapotranspiration which is an important component of CWP estimation.

Several studies have been conducted using different methods and models to estimate the crop water productivity of maize. A study in Mexico used locally developed crop coefficient curves and United Nations Food and Agriculture Organization (FAO) crop coefficients to estimate maize water use and water productivity. One study proposed an ensemble approach for identifying the virtual water content (VWC) of main crops on the Korean Peninsula in past and future climates. The ensemble VWC is calculated using three types of crop yields and fifteen consumptive amounts

of water use in the past and the future (Lim et al., 2017). Future projections predicted declining yields for both crops, suggesting that these reductions could lessen future water demand.

CWP Estimation Methods

Estimation methods vary depending on the region and the crops being cultivated. They are mainly grouped into field-based, modelling and simulation, and remote sensing and machine learning methods. Every method has its advantages and disadvantages.

2.2.1 Surface Energy Balance Algorithm (SEBAL)

SEBAL has proven to be a valuable tool for water resource management, agricultural planning, and environmental monitoring, providing reliable estimates of evapotranspiration and surface energy fluxes over large areas. It has been widely adopted in various research and operational applications worldwide because it can utilize freely available satellite data and provide valuable information for water-scarce regions. The Surface Energy Balance Algorithm (SEBAL) is a method used to estimate evapotranspiration (ET) using remote sensing and the energy balance principle (Gibson et al., 2013). It was developed to analyze thermal infrared remote sensing data from satellites like the Landsat series to monitor and manage water resources and agriculture efficiently. Souza et al., 2023 carried out a study that estimated the evapotranspiration of irrigated açai plants in eastern Amazonia using SEBAL. The results showed good agreement with the Bowen ratio method, and SEBAL was useful for irrigation management and reducing water losses. Similarly, (Gao et al., 2023; Kamyab et al., 2022; Bansouleh et al., 2015; PACHAC HUERTA & CHÁVARRI VELARDE, 2019) and more recently used this model in the estimation of evapotranspiration proved to be more efficient. SEBAL is limited for accurate evapotranspiration Estimation. SEBAL relies on spatial information, such as land surface temperature and vegetation indices, which vary across different areas. This spatial dependence can introduce uncertainties in the estimation of evapotranspiration. This is due to the reliance on anchor pixels. (Prakash Mohan et al., 2020), Properly outline the limitation of anchor pixels. Another limitation of this Algorithm is wind speed observation. SEBAL requires accurate wind speed data for calculating each pixel's dry/wet endpoints. However, wind speed observations are known to have high temporal and spatial variations and may not be routinely available, especially in heterogeneous areas. SEBAL is sensitive to vegetation parameters, such as the Normalized Difference Vegetation Index (NDVI),

which estimates the vegetation cover fraction. Inaccurate or incomplete vegetation information can affect the accuracy of SEBAL estimates (Ruhoff et al., 2012).

Surface Energy Balance System (SEBS) is another remote sensing-based algorithm that estimates evapotranspiration (ET) and surface energy fluxes from satellite data. It was developed to overcome some limitations of the Surface Energy Balance Algorithm (SEBAL) and provide more accurate and physically based estimates of surface energy fluxes and ET. Similar to SEBAL, there are discrepancies in the reported accuracy of the SEBS model due to known model sensitivities. Its performance may vary depending on the input data quality and the specific characteristics of the study area. SEBS requires accurate vegetation parameters to be obtained, especially in agricultural areas where accurate vegetation parameters can be obtained, high-resolution imagery with low sensor zenith angles is available, and canopy cover is complete. This requirement may limit the applicability of SEBS in areas where such data is not readily available (Gibson et al., 2013).

2.2.2 Field-Based CWP Methods

Several field estimation methods can improve CWP, including lysimeter measurement, nuclear techniques, modeling approaches, field observations, and genetic enhancement. Lysimeter measurement and nuclear techniques have been used to improve water management, saving water by reducing the loss of components the plants do not use and thus enhancing water productivity (WP) (Abou Zakhem et al., 2019). These techniques involve measuring soil water content, evaporation, and deep percolation to determine the amount of water plants use and the amount lost to the environment. However, this method is limited to a small scale.

Shoukat et al. (2021) conducted a field study to estimate wheat crop water productivity (CWP) under different irrigation and fertilizer regimes. Using a randomized block design with three replications, they applied three irrigation treatments - full, 20% deficit, and 40% deficit irrigation relative to the crop evapotranspiration (ETC) estimated by the CROPWAT model. Additionally, three nitrogen fertilizer treatments were applied at different wheat growth stages. By measuring the resulting wheat yields, yield components, nutrient utilization, and water use efficiencies, they could estimate CWP under different management scenarios. A commonly used field method for CWP estimation is the harvest method. The harvest method involves measuring the crop yield and the amount of water used for irrigation, then calculating crop water productivity

(CWP) as the yield-to-water use ratio. This method provides a direct measure of the productivity of a crop in terms of water use efficiency. The method faces extent limitations, especially for large spatial areas, and data availability, especially for the amount of water lost. Field-based methods may only be equally applicable to some crop types as different crops have diverse water requirements, growth patterns, and responses to water. In addition, CWP can change over time due to climate variability, seasonal fluctuations, and other factors. Field-based methods might provide only a snapshot of CWP for a specific period, limiting their ability to capture long-term trends.

2.2.3 Remote sensing and Machine learning Approaches

Remote sensing is a technique used to estimate crop water productivity by analyzing data collected from a distance, such as satellite imagery. This method has become increasingly popular in recent years due to its ability to support management, monitoring, and controlling activities at different spatial and temporal scales (Dalla Marta et al., 2018). Remote sensing has revolutionized monitoring and estimating crop water productivity (CWP) by providing valuable data on crop health, water stress, and environmental conditions over large agricultural areas. With the ability to gather information from satellites, drones, and other airborne platforms, remote sensing techniques have become essential for sustainable water management in agriculture. (Li et al., 2008; Talpur et al. 2023; Gao et al., 2023; Spiliotopoulos et al., 2023 and Darwish et al., 2023) Successfully utilized remote sensing technologies to estimate crop water use and productivity.

Remote sensing indices are utilized to estimate crop water productivity. Commonly used indices include NDVI (Mohanasundaram et al., 2023; Pandya et al., 2023; Farrell et al., 2018), EVI (Jaafar & Ahmad, 2015; (Tang et al., 2015), NDWI (Singh et al., 2021; (Z. Wang et al., 2009) and TVDI (Holzman & Rivas, 2016) and more. NDVI is one of the most widely used remote sensing indices for assessing vegetation health. It quantifies the density of green vegetation and indicates plant Vigor and growth. NDVI values near +1 indicate healthy, thriving vegetation, while values closer to -1 indicate stressed or sparse vegetation. Monitoring NDVI over time can infer changes in crop water productivity. EVI is an improvement on NDVI, designed to reduce atmospheric influences and improve sensitivity at high vegetation densities. It provides a more accurate representation of vegetation conditions, making it suitable for areas with dense vegetation cover or during periods of high atmospheric interference. TVDI combines thermal infrared data with NDVI to assess water stress in crops. It uses the difference between the daytime land surface

and potential temperatures under non-water-stressed conditions. TVDI is beneficial for monitoring crop water status, especially in regions where water scarcity is a concern.

On the other hand, machine learning (ML) techniques have been increasingly employed in the agricultural field to improve the accuracy and efficiency of crop water productivity estimation. This technique has proven to be more efficient in crop yield prediction (Ashwitha & Latha, 2022), water demand forecasting (Emami et al., 2022), remote sensing and Image analysis (Thapa et al., 2023), drought monitoring and mitigation, precision Irrigation, and agriculture. Machine learning techniques offer powerful tools to analyze complex datasets, predict crop water requirements, and enhance water productivity in agriculture.

Despite They also have limitations despite using remote sensing and machine learning in crop water productivity estimation. The limitation may be due to data availability and quality. Remote sensing relies on data from satellites and other sensors, and the availability and quality of this data can vary. Cloud cover, sensor malfunctions, and limited satellite revisit frequency can lead to gaps in data, making it challenging to get consistent and timely information for accurate estimations. Another limitation is spatial and temporal resolution. There may need to be more than the spatial resolution of remote sensing data to capture small-scale variations within fields or individual plants. Similarly, the temporal resolution may only sometimes be high enough to capture rapid changes in crop water needs, especially in rapidly evolving weather conditions. Finally, the complexity of the crop-water relationship poses a significant challenge in accurate water productivity estimation. The relationship between crop water productivity and remote sensing parameters is complex and influenced by various factors like crop type, stage of growth, and weather conditions. Capturing these complex interactions in a single model can be challenging.

3. Materials and methods

3.1 Study area

The Bura irrigation scheme is the study area located in Tana River County (1°11'39.1S, 39°50'23.0"E). The scheme is one of the largest and oldest irrigation projects in Kenya. It was established in the 1970s and covers an area of about 12,000 acres gazetted area and 10,000 acres under irrigation (National Irrigation Authority, 2023). The primary water source is from River Tana, which is 50 km away and maize is the main crop grown in the area. Other crops grown in the area include green grams, cowpeas, cotton, watermelon, sugarcane, and onions.

Low and erratic rainfall patterns characterize the scheme. The average annual rainfall in the region is relatively low, ranging from 200 to 600 millimeters (8 to 24 inches). This results in an average of about 400mm of rainfall in the region (Muigai David et al., 2019). The rainfall is highly variable, with the most precipitation occurring during the short rainy season from March to May and a shorter second rainy season from October to December (Mbayaki, 2021). Drought periods are standard, and rainfall distribution can vary significantly from year to year.

Due to the arid conditions, the Bura irrigation scheme's evapotranspiration rates are relatively high. High temperatures and low humidity contribute to increased evapotranspiration rates, which can impact water availability for crops.

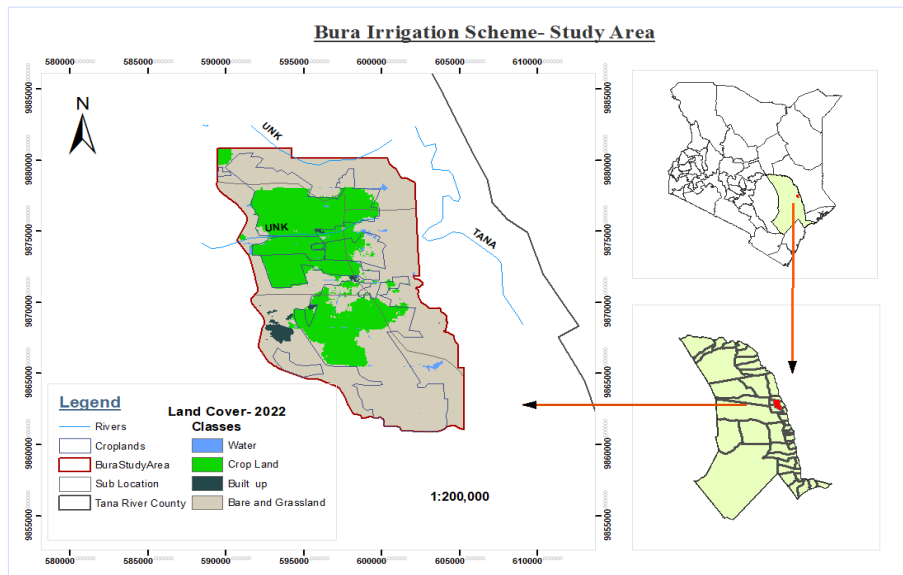


Figure 3. 1 Study area map – Bura Irrigation Scheme.

3.2 Data

This research utilizes satellite imagery and ground-based data to analyze crop productivity. Satellite data was acquired from the Google Earth Engine via its Python API for the five years between 2018 and 2022, focusing on images from the crop growing seasons.

3.2.1 Meteorological Data

The availability of consistent, long, and complete time series of meteorological data is critical for estimating crop water requirements because weather is the most important input variable for assessing crop evapotranspiration and determining soil water deficit. This study used the ERA5 daily Gridded dataset accessible from the GEE catalog and the local meteorological station dataset. ERA5 is the fifth-generation ECMWF reanalysis of the global climate and weather data of the previous eight decades. Data has been accessible since 1940 (Copernicus, 2018). The dataset contains essential atmospheric meteorological factors such as air temperature, pressure, and wind at various altitudes, as well as surface parameters such as rainfall and sea parameters such as sea-surface temperature and wave height. The daily homogenized, filtered, reviewed, and pre-processed (outliers removed) data of minimum temperature (Tmin), maximum temperature (Tmax), mean temperature (Tmean), Precipitation, incoming solar radiation, vapor pressure, and wind speed (u -component at 10m high above the ground) were acquired from Google Earth Engine (GEE) data catalog for a period of five years; between 2018 to 2022. In the case of vapor pressure, the information may not be directly available in the ERA5 dataset. However, it was estimated using atmospheric parameters such as air temperature and relative humidity, as previously implemented by (Pelosi et al., 2022) on crop water requirement studies.

Similarly, daily weather data for the Garissa weather station was acquired from the Kenya Meteorological Department on request. This was for the same period and was used as part of the ground validation data and the actual ground yield data. More accurate data could be acquired at the exact field level and in several locations, but this was limited in this study.

Optical Satellite Imagery Datasets

This research acquired several optical datasets from the GEE platform, including MODIS, Landsat 8, and Sentinel 2 datasets. MODIS provided high temporal resolution data of up to daily with an average 8-day temporal resolution. Three surface reflectance products were used for data

fusion. In this case, MOD09A1.061 Terra Surface Reflectance 8-Day Global 500m product, USGS Landsat 8 Level 2 Collection 2 Tier 1, and Sentinel-2 MSI Multispectral Instrument Level-2A dataset were accessed from the GEE data catalog using Python API. Landsat 8 dataset has a spatial resolution of 30m, and sentinel 2 has the B2 - B4 and B8 having a spatial resolution of 10m. To achieve a harmonized resolution, resampling was performed on the datasets to achieve a 10m spatial resolution. Table 3.1 below summarizes the data used in this study.

Table 3.0.1 Datasets used in this study

Data	Description/Resolution	Source
Temperature (Tmin, Tmax), Incoming Solar radiation, vapor pressure, Wind speed, and Precipitation	From ERA5_L and Daily Dataset and Actual Meteorological Local Station Data.	ECMWF/ KMD (Kenya Meteorological Department)
Evapotranspiration (ET)	8-Days at 500m resolution	MODIS
Surface reflectance (Derived Dataset: NDVI, MSAVI, EVI, etc.)	10m - Fused dataset	Sentinel 2, Landsat 8 OLI, MODIS

3.3 Methodology Flowchart

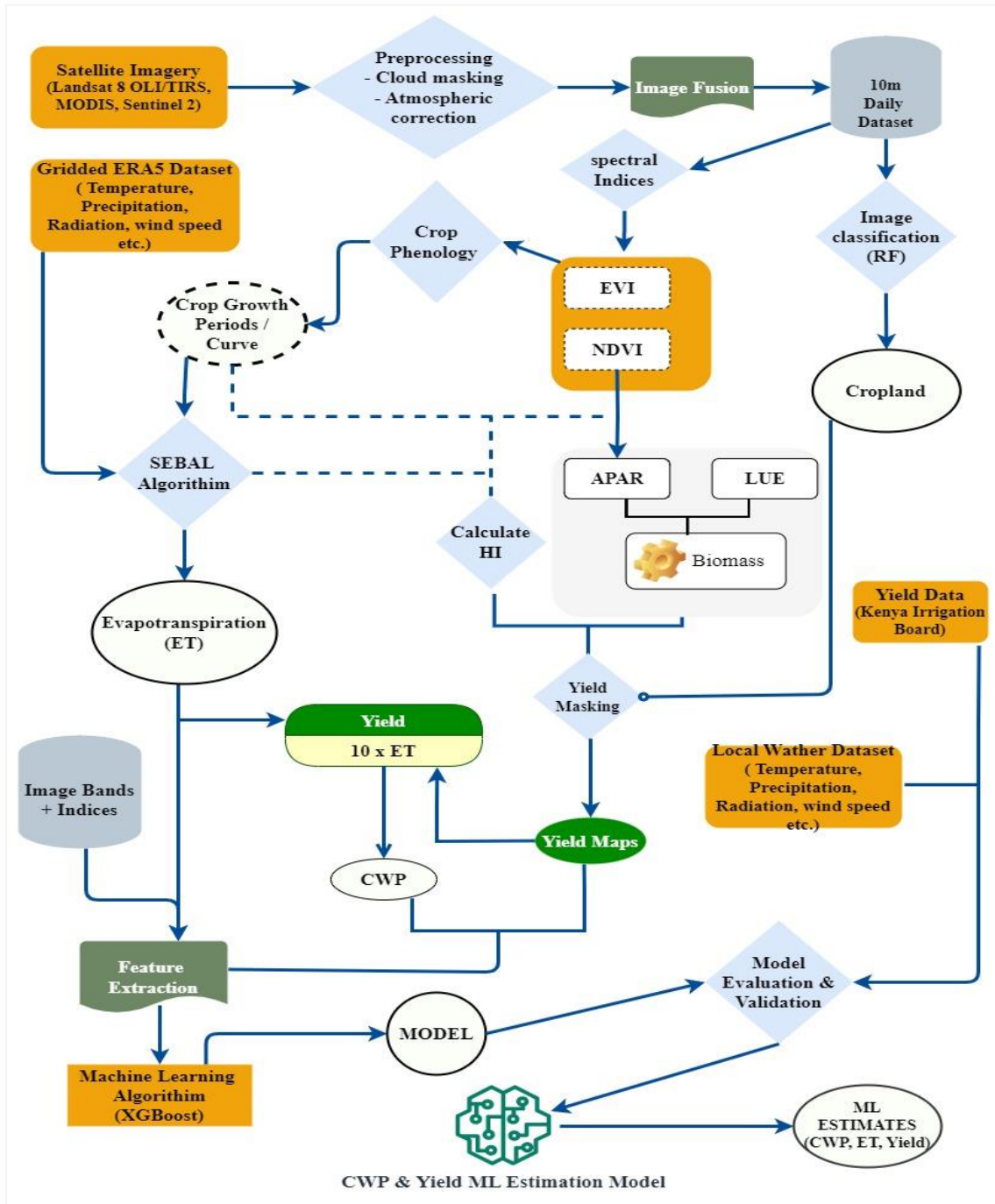


Figure 3. 2 Flowchart Diagram

3.4 Data Fusion and Indices Calculation

3.4.1 Daily Imagery Dataset

The increased availability and diversity of global satellite products and new algorithms have opened up enormous opportunities for creating new levels of data with varied geographical, temporal, and spectral resolutions (Dhillon et al., 2023). MODIS has a short return interval, and the data quality is consistent. However, because of the limited spatial resolution, it does not apply to the small-scale region. Sentinel-2 and Landsat 8 data have a better spatial resolution of about 10m and 30m, the lowest value for some bands, than MODIS data, although the data quality is unreliable

Implementing the fusion algorithm for the three datasets was based on a method implemented by (H. Gao et al., 2023) by extending the formula used with one for Landsat. However, several fusion algorithms do exist. An example is the improved spatial and temporal data fusion approach (ISTDFA) used by (Wu et al., 2018) to fuse MODIS and Landsat. Another algorithm used was STAIR 2.0 to fuse MODIS, Sentinel 2, and Landsat. Other commonly used algorithms similar to the ones employed here are the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) and two extended data fusion techniques, STAARCH and ESTARFM, which have been utilized in prior studies to integrate MODIS and Landsat data (F. Gao et al., 2015). In this study, the fusion algorithm used the formula below to obtain a daily 10 m resolution dataset.

$$\mathbf{V}_H(t_i) = \mathbf{V}_M(t_i) + \mathbf{K} \cdot \frac{\sum_{j=1}^n \left[\omega(t_i, t_j) (\mathbf{V}_T(t_j) - \mathbf{V}_M(t_j)) \right]}{\sum_{j=1}^n \omega(t_i, t_j)} + \mathbf{K} \cdot \frac{\sum_{k=1}^m \left[\omega(t_i, t_k) (\mathbf{V}_L(t_k) - \mathbf{V}_M(t_k)) \right]}{\sum_{k=1}^m \omega(t_i, t_k)} \quad (1)$$

$$\omega(t_i, t_j) = \frac{1}{|t_i - t_j| + 1} \quad (2)$$

$$\omega(t_i, t_k) = \frac{1}{|t_i - t_k| + 1} \quad (3)$$

Where:

- $V_H(t_i)$ - Value of the fused image pixel at time t_i
- $V_M(t_i)$ - Value of the MODIS image pixel at a time t_i
- $V_T(t_j)$ - Value of the Sentinel-2 image pixel at time (t_j)
- $V_L(t_k)$ - Value of the Landsat 8 image pixel at time t_k
- N - Number of Sentinel-2 images
- M - Number of Landsat 8 images
- t_i - Day of Year (DOY) corresponding to the MODIS data.
- t_j - DOY corresponding to the Sentinel-2 data
- t_k - DOY corresponding to the Landsat 8 data.
- $\omega(t_i, t_j)$ - Weight of the Sentinel-2 image at a time t_i, t_j
- $\omega(t_i, t_k)$ - Weight of the Landsat 8 image at a time t_i, t_k
- K - The empirical coefficient accounts for the influence of missing Sentinel-2 and Landsat 8 images on data fusion.

The fusion process involved data preprocessing, such as performing atmospheric corrections on the Landsat 8 and Sentinel 2 images. Cloud masking and filtering were performed to ensure that cloud-free photos were obtained for fusion. This methodology adopted a procedure utilized by (Luo et al., 2020) in the actualization of the STAIR 2.0 algorithm. This involved cloud removal, resembling the MODIS dataset from 500m to 30m spatial resolution, and fusing MODIS with Landsat, as shown in equation (1) above. Similarly, atmospheric corrections, co-registration, and cloud removal and filtering were performed on the Sentinel-2 dataset, fused with a resampled 10 m resolution MODIS-Landsat 8 dataset to 10 m resolution daily dataset collection between 2018 and 2022.

Figure 3.3 below shows a sample NDVI map comparing the output from the fused dataset and coarse 500m resolution MODIS NDVI between January and December 2022. This helps achieve high spatial and temporal resolution for the fused dataset, ensuring improved crop monitoring and metric generation within the study area.

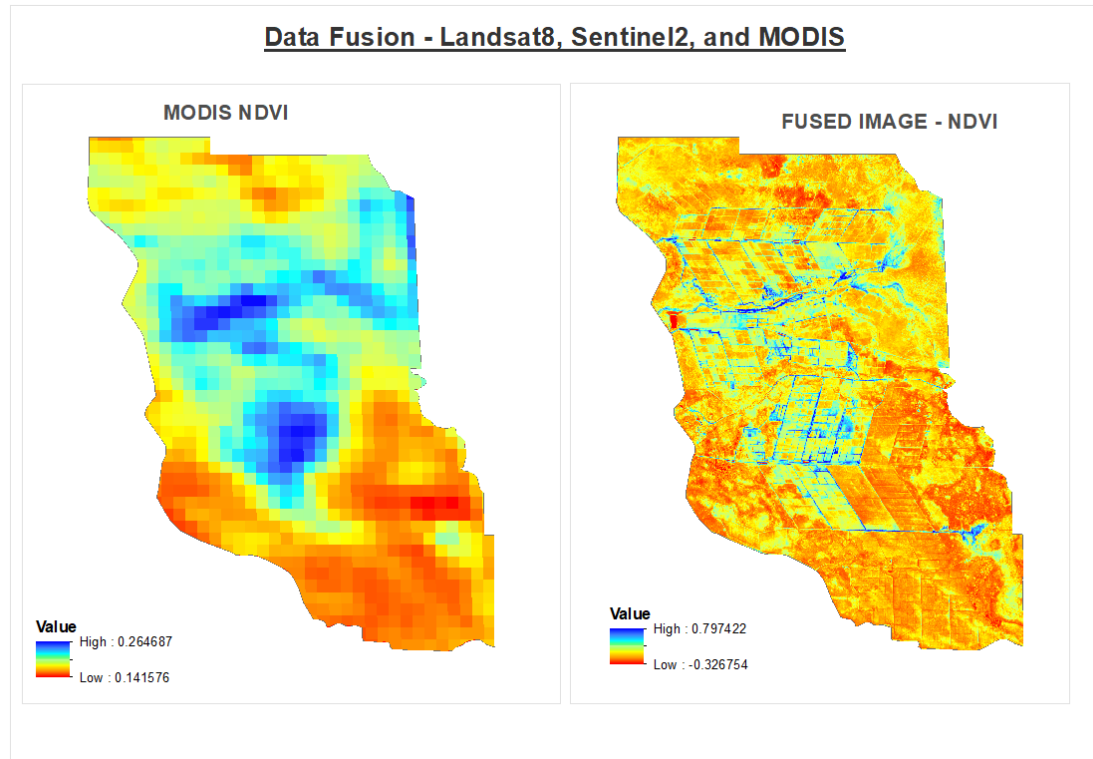


Figure 3. 3 Comparison of Daily MODIS and fused dataset

3.4.2 Indices

The study calculated several indices to support the analysis of CWP and machine learning modeling. The indices are drought-based, vegetation-based, soil-based, and water-based. They include NDVI, VCI, and more.

Table 3.2 below summarizes all the indices calculated and used in the study.

Table 3.2 Spectral indices, Expressions, and references

Indices	Expression	Use Case/ Reference
NDVI	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	(Dhau et al., 2021; Bolfe et al., 2023)
EVI	$2.5 * ((\text{NIR} - \text{Red}) / (\text{NIR} + 6 * \text{Red} - 7.5 * \text{Blue} + 1))$	(Dhau et al., 2021)

GNDVI	$(\text{NIR} - \text{Green}) / (\text{NIR} + \text{Green})$	(Dhau et al., 2021)
MSAVI	$(2 * \text{NIR} + 1 - \text{sqrt}((2 * \text{NIR} + 1)^2 - 8 * (\text{NIR} - \text{Red}))) / 2$	(Dhau et al., 2021)
MNDWI	$(\text{Green} - \text{SWIR}) / (\text{Green} + \text{SWIR})$	(Agilandeewari et al., 2022; Laonamsai et al., 2023)
EDI	$(\text{PET} - \text{ET}) / \text{PET}$	(Khan et al., 2018)
VCI	$(\text{NDVI}_i - \text{NDVI}_{\text{min}}) / (\text{NDVI}_{\text{max}} - \text{NDVI}_{\text{min}})$	(Sánchez et al., 2016; Zhao et al., 2022)
TCI	$(\text{LST}_{\text{max}} - \text{LST}_i) / (\text{LST}_{\text{max}} - \text{LST}_{\text{min}})$	(Sánchez et al., 2016; Zhao et al., 2022)
SMCI	$(\text{SM}_i - \text{SM}_{\text{min}}) / (\text{SM}_{\text{max}} - \text{SM}_{\text{min}})$	(Sánchez et al., 2016; Zhao et al., 2022)

3.5 Crop Yield

Yield calculation in this study is achieved by determining total dry matter biomass and harvest index. Harvest index is a critical agricultural term relating to crop plants' capacity to convert photosynthetically fixed carbon into edible yield, frequently the crop's harvested component such as grains, fruits, or vegetables. Combined product of harvest index and summation of biomass results to yield estimate. In this case, the Biomass-Harvest index method is used. The harvest index method is inherited from a study by (Moriondo et al., 2007).

$$HI = HI_{max} - HI_{range} \left(1 - \frac{NDVI_{POST}}{NDVI_{PRE}} \right) \quad (4)$$

Where:

HI - Actual Harvest index for C4 crop (maize/corn)

HI_{max} - Maximum harvest index for maize. In C4 crops have been chosen to be 60% (0.6)

HI_{range} - The range selected is 0.2 for C4 crops.

$NDVI_{POST}$ - The mean value of NDVI between flowering and maturity

$NDVI_{PRE}$ - The mean value of NDVI between planting and flowering for maize crop.

Dry matter Biomass estimation utilizes NDVI values and solar radiation as major input parameters. According to Bastiaanssen & Ali, 2003, Biomass can be estimated using the following formula.

$$\text{Biomass} = \Sigma (0.864 \times \varepsilon \times \text{APAR}) \quad (5)$$

Where:

ε is the light use efficiency (LUE), and APAR is the 24-hour absorbed photosynthetically active radiation (W/m^2). APAR is given as shown;

$$\text{APAR} = 0.48 \times f \times S\downarrow \quad (6)$$

Where f refers to the APAR fraction which changes with respect to the leaf area index. $S\downarrow$ Represents the incoming solar radiation. This component is obtained from weather data. The value of 0.48 is an average or typical value that has been found to represent the efficiency of plants in converting PAR into chemical energy through photosynthesis. It is also important to note that the PAR value describes the total radiation available for photosynthesis assuming leaves intercept all sunlight. This is a very speculative estimate since leaves both transmit and reflect solar radiation.

Determination of the fraction of APAR (f) depends on vegetation indices of the crop at a specific point in time. In this case, NDVI is specifically used as the main index in the calculation of the fraction of APAR. Given by the expression;

$$f = -0.161 + 1.257 NDVI_t \quad (7)$$

Another important component used in the determination of Biomass is the LUE. LUE quantifies the plants' *efficiency* in utilizing solar radiation for photosynthetic carbon fixation. Given by the expression;

$$\varepsilon = \varepsilon_{max} \times g(T) \times g(D) \times \lambda \quad (8)$$

ε_{max} represent the max value for C4 crops like corn/maize. A study determined ε_{max} to be approximately 2.5 gMJ^{-1} (Huang et al., 2022). Similarly, $g(D)$ accounts for vapour pressure, and $g(T)$ accounts for crop heat stress. Both $g(T)$ and $g(D)$ and scalar quantities. Finally, λ represents the water stress mainly obtained as an evaporative fraction. Utilizing results from biomass and harvest index (HI), yield is then obtained by the following expression;

$$\text{Yield (Y)} = \text{Biomass} \times \text{HI} \quad (9)$$

3.6 Evapotranspiration Estimation

The estimation of actual ET utilizing the SEBAL algorithm parameters was achieved by the following formular, which is based on thermal and multispectral remote sensing datasets that compute latent heat flux (LE) as a residual by subtracting the soil heat flux and sensible heat flux from the instantaneous surface energy balance equation's net radiation (Rn) (Gonçalves et al., 2022). In the computation of ET, both weather data and satellite images are used. LE is expressed as shown in equation (10).

$$\text{LE} = \text{Rn} - \text{G} - \text{H} \quad (10)$$

H is the instantaneous sensible heat flux (W/m^2), and G is the soil heat flux (W/m^2). Furthermore, Rn and G are determined as follows;

$$\text{Rn} = (1 - \alpha) \text{Rs}\downarrow + \text{Rl}\downarrow - \text{Rl}\uparrow - (1 - \varepsilon_0) \text{Rl}\downarrow \quad (11)$$

$$\frac{\text{G}}{\text{Rn}} = \alpha(T_s - 273.15) (0.0038\alpha + 0.0074\alpha^2) (1 - 0.98\text{NDVI}^4) \quad (12)$$

Where α is the surface albedo calculated from satellite image bands according to (Tasumi et al., 2008) that demonstrated this using Landsat image bands? $\text{Rs}\downarrow$ represents the incoming short-wave radiation, $\text{Rl}\downarrow$ is the incoming long wave radiation, $\text{Rl}\uparrow$ the outgoing longwave radiation, and ε_0 is the surface thermal emissivity. According to (Jaafar & Ahmad, 2020), T_s, corr is the corrected land surface temperature (T_s) in K based on the DEM map and the difference between extraterrestrial solar radiation on sloping and flat terrains to account for temperature variations owing to common elevation data and slope.

The automated statistical technique used to choose the hot and cold endmembers is a reduced version of the CIMEC algorithm used in METRIC. Endmember candidates are chosen based on percentiles of LST and normalized difference vegetation index (NDVI) readings. To calibrate the dT function, the CIMEC process is employed. Surface temperature, wind speed,

surface roughness, and surface-to-air temperature gradients are all used to determine sensible heat flow (H).

$$H = \frac{\rho_a C_p dT}{r_{ah}} \quad (13)$$

Where C_p is the specific heat capacity and is the aerodynamic resistance of turbulent heat transport from the evaporating surface at height z_1 to the air above the evaporating surface z_2 , to solve the iteration process, selecting hot and cold endmembers is necessary. In this case linear relationship between T_s and (dT) is assumed; a and b coefficients are empirically determined for each image.

$$dT = a + bT_s \quad (14)$$

More details from a research paper done by Laipelt et al., 2021.

The evaporative fraction (Λ) is expressed as

$$\Lambda = \frac{LE}{Rn - G} \quad ; \quad (16)$$

To obtain daily ET, the following expression is used within the SEBAL algorithm.

$$ET_{a24h} = \frac{\Lambda Rn_{24}}{\lambda} \quad (16)$$

Where is the latent heat of vaporization (MJ·kg⁻¹).

In addition, the maize crop coefficient was calculated utilizing the Penman-Monteith equation (Wang et al., 2023) which FAO recommends. ET_a is given by the expression;

$$ET_a = ET_0 \times K_c \quad (17)$$

ET_0 is the daily potential evapotranspiration and K_c is the crop coefficient for maize at the corresponding growth stage.

$$ET_0 = \frac{0.408\Delta (Rn - G) + \gamma \frac{900}{T_{mean} + 273} u_{10} (e_s - e_a)}{\Delta + \gamma (1 + 0.34 u_{10})} \quad (18)$$

Where Δ represent the slope of the saturation of water pressure curve, T_{mean} is the mean of daily air temperature at 10m height in degrees Celsius, u_{10} is the speed of wind at 10m height in meters per second, e_a represent the actual vapour pressure, and is the saturation vapour pressure. The difference of and gives the saturation vapour pressure deficit. γ is the psychrometric

constant. G is as previously defined—more on the calculation of each component follows (Gebremedhin et al., 2022).

Crop coefficient (K_c) is a spatially and temporary variable depending on the crop type and the growth stage. The determination of K_c values followed expressions documented by (Cemek et al., 2023). Therefore, the combination of ET_a and ET_o is used to determine the correlation between K_c and NDVI within the study area. According to FAO, K_c values ranges from 0.2 to 1.2 with the following periodic breakdown: initial stage, crop development stage, mid-development stage, and late stage.

3.7 Crop Water Productivity

CWP then estimated using crop determined yield (Equation 9) and crop evapotranspiration (Equation 16). Yield units are in kilograms per meter squared (kg/m^2) and evapotranspiration units are m^3/m^2 . Water productivity maps were an output of dividing agricultural yield production results (Figure 4.5 and 4.6) by water use results (ET results in Figure 4.8 and Figure 4.9). The expression for CWP is as shown in equation 19.

$$\text{CWP} = (\text{Yield (Y)}) / (10 \times \text{ET}_{\text{mean}}) \quad (19)$$

where ET_{mean} is the mean actual evapotranspiration over the growing period when the estimation is performed

3.8 Machine Learning

3.8.1 Machine learning models used to estimate Yield, ET_a , and CWP

This section utilizes three machine learning models: Random Forest Regressor, Support vector machine, and Extreme Gradient Boosting.

Random Forest (RF) leverages the power of multiple decision trees by training them on diverse subsets of the training data through bootstrapping, ultimately creating a robust and integrated learning algorithm (Elbeltagi et al., 2022). The outcome of Random Forest (RF) approximations consists of the averages from each tree. As a result, RF can reduce variance and achieve more accurate predictions compared to typical tree-based algorithms. Nonetheless, in cases of predicting extreme observations, it can introduce bias. In particular, random forest predictions tend to be overly optimistic when there are limited observations available. In this

research, bias correction followed the RF bias correction flow applied by (Zhao et al., 2022). The scikit-learn package (Pedregosa et al., 2011) was utilized to train Random Forest regressors independently on each dataset.

Support Vector Regression (SVR) is a machine learning approach grounded in Vapnik-Chervonenkis (VC) theory and the principle of structural risk minimization (SRM). SVM is the machine learning algorithm that comes closest to deep learning. A two-layer neural network is equivalent to a nonlinear SVM. A multi-layer neural network may be simulated by adding extra kernel functions to nonlinear SVM (Mountrakis et al., 2011). SVR extends its applicability to nonlinear regression tasks by utilizing kernel functions, which help transform input data into a higher-dimensional feature space. SVR incorporates the concept of slack variables during training, allowing the model to accommodate and tolerate specific errors. Including slack variables enhances the model's capacity for generalization (Xu et al., 2023). In estimating ET, Yield, and CWP, the non-linear radial basis function kernel function is used. This kernel function performs better than other kernels for the SVM model. The optimal hyperparameters are determined where C for the kernel function is determined through trial and error.

Extreme Gradient Boosting (XGBoost) is a distributed gradient boosting library built to be efficient, adaptable, and portable. It employs machine learning methods using the Gradient Boosting framework. It utilizes a second-order Taylor expansion of the target function and the second derivative to enhance the speed of model convergence during training. Furthermore, a regularization component is incorporated into the target function to manage the complexity of the tree, resulting in a more straightforward model and guarding against overfitting (Geng et al., 2021). XGBoost introduces a novel sparsity-aware algorithm for handling sparse data and a weighted quantile sketch for approximate tree learning. The sparsity-aware algorithm is designed to efficiently handle sparse data, which is common in many real-world applications. It optimizes the tree construction process by only considering non-zero values, reducing the computational cost and memory usage. The weighted quantile sketch is a technique introduced in Xgboost to handle weighted data. It allows Xgboost to find quantiles on weighted data, essential for accurate tree learning. This technique is the first method to solve the problem of finding quantiles on weighted data. In this study, some of the parameters of importance used for XGBoost include

n_estimators, learning_rate, max_depth, subsample, and colsample_bytree. More information about XGBoost can be found in the documentation.

3.8.2 Performance Metrics and Evaluation

Calculated ETa, CWP, and Yield data are hereby compared to the modeled data. This is achieved through model performance evaluation. The dataset is split into 70% Training and 30% validation. To curb model overfitting, the K-fold cross-validation method is employed. To perform cross-validation (CV) in this study, the dataset was split into k subdivisions, explicitly using a 5-fold CV. In each of the 5 iterations, the model was trained, and during each iteration, a different fold was held out from the training set and used as the validation set. This approach ensures that each fold has a turn at being the validation set while the model is trained multiple times.

Model performance is assessed using mean absolute error (MAE) Equation (20), coefficient of determination (R^2) Equation (21), mean square error (MSE) Equation (22), and root mean square error (RMSE) Equation (23) (Elbeltagi et al., 2022).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (21)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (22)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (23)$$

Where n is the sum of all points used, y_i is the observed calculated value of the respective values of ETa, CWP, or Yield that are the target or response variables. Represent the modelled values, and \bar{y} represents the mean value of the respective reference value of the target variables.

4. Results

4.1 Crop Phenology

Study regarding crop growth stages is determined based on the cover characteristics within the study area. Using the NDVI and EVI, the analyzed average growth period and the resultant curve show each stage's dates. Figure 4.1 and Figure 4.2 visually present the vegetation growth characteristics for the specific day of the year. The continuous alignment of NDVI and EVI curves throughout reveals their significant association, emphasizing their importance in crop phenology assessment. This agreement highlights the durability and dependability of these vegetative indicators, reaffirming their importance.

There are two main growth seasons: the long and the short growing seasons. The short season ranges between DOY -220/270 and 355 (August/September to December). The long season happens DOY-410 to 605 (late January to August). The dates agree with planting periods as outlined in other studies within the study area (Muigai David et al., 2019). A 1 to 2-week window is applied to these growth stages to ensure it captures the differences that may arise within the Irrigation scheme.

The sowing period shows some discrepancy from what could happen in areas with complete reliance on rainfed agriculture. The study area being an irrigation scheme, the sowing period starts partially from zero NDVI or EVI index value as it could be in rainfed areas. This is a contribution by factors such as crop rotation and continuous crop alternation within the crop growth stages. This implies that both the sowing and maturity of maize crops happen concurrently within the study between DOY – 345 (Maturity) and DOY – 405 (sowing).

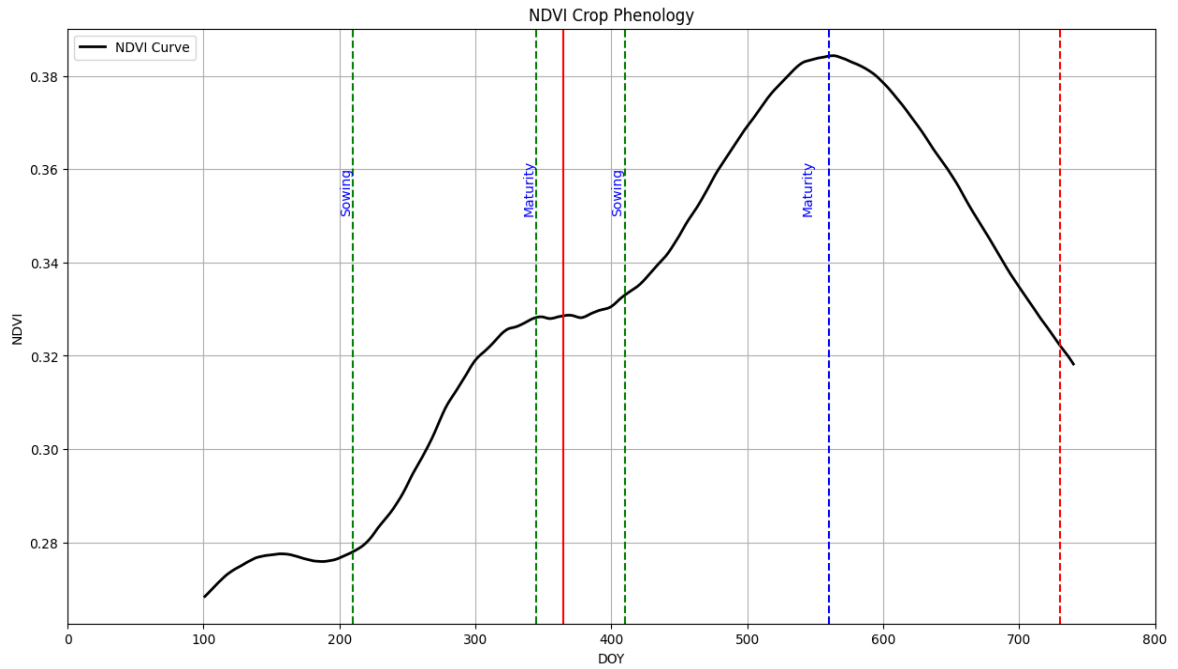


Figure 4. 1 Crop growth curve - Representation of crop phenology per growing period, using NDVI

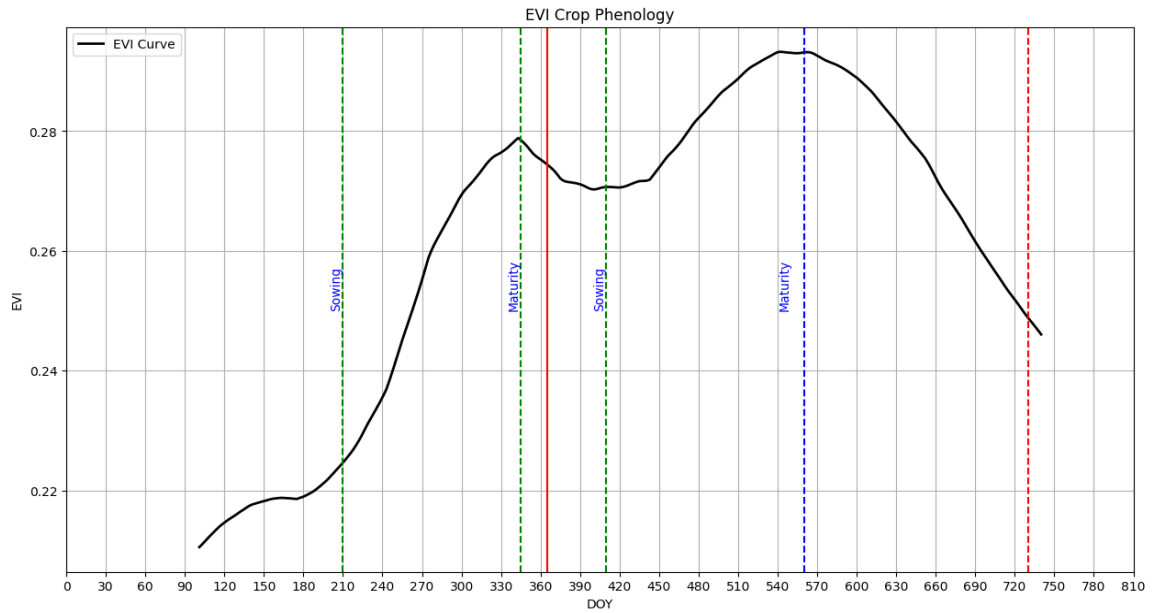


Figure 4. 2 Crop growth curve - Representing crop phenology per growth period using EVI

4.2 Yield Estimation

4.2.1 Classification

The land cover classification, accomplished through Random Forest analysis, successfully identified six distinct classes: bare land, tree cover, grassland, maize, other crops, and water bodies. The classification results exhibit high accuracy, enabling precise delineation and monitoring of land use patterns. Table 4.1 shows the error matrix for classification using random forest for the year.

Table 4.1 Classification error matrix obtained from classification of land use and land cover of Bura Irrigation Scheme using random forest, 2022

Classification	Reference							
	Grassland	Forest	Maize	Other Crop	Bare land	Water	Row Total	User's Accuracy
Grassland	0.188	0.009	0.153	0.162	0.012	0.017	0.542	35.60%
Forest	0.024	83.243	1.546	0.044	0.175	0.035	83.496	97.40%
Maize	0.201	0.339	11.005	0.023	0.141	0	11.71	94.00%
Other Crop	0.024	0.005	0.05	0.456	0.02	0.023	0.578	78.90%
Bare land	0.001	0.089	0.149	0.113	0.568	0.001	0.92	61.70%
Water	0.021	0.072	0.023	0.014	0.031	3.980	2.354	95.35%
Column Total	0.459	83.653	11.82	0.902	0.942	4.056	Overall Accuracy: 95.8%	
Producer's Accuracy	44.50%	97.40%	84.57%	57.43%	60.26%	95.12%	Kappa: 0.916	

Cropland delineation is achieved through classification with the primary goal of using the maize cropland for masking other layers of land pattern. The supervised classification using random forest performed well in mapping tree cover, maize, and water. The three classes have an accuracy of above 85, both user and producer accuracy.

This classification was applied every year between 2018 and 2022 to determine cropland patterns in the study area. Figure 4.3 below shows the classification map for the year 2022.

Supervised Classification - Random Forest

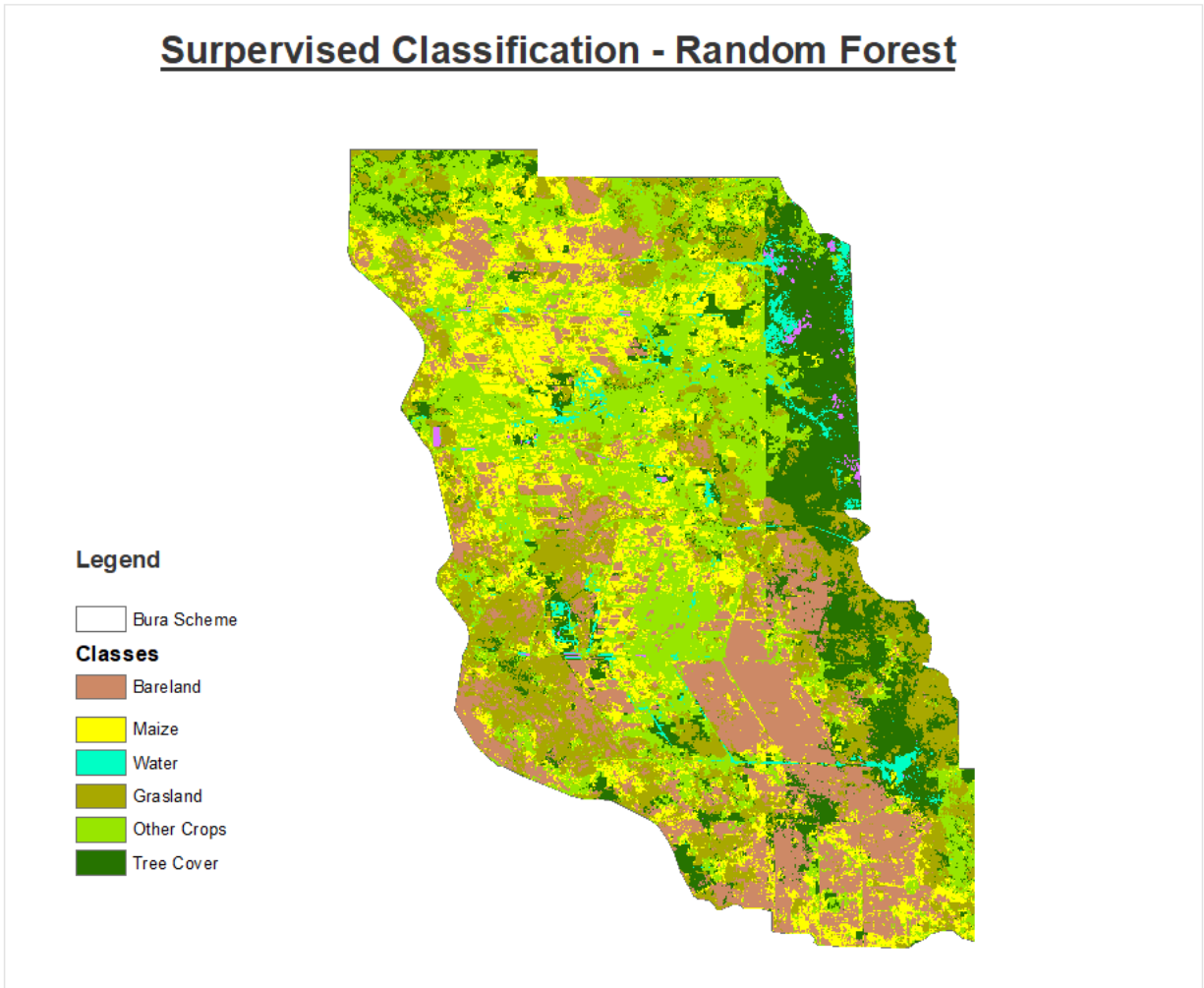


Figure 4. 3 Land Use Land Cover map

4.2.2 Vegetation distribution

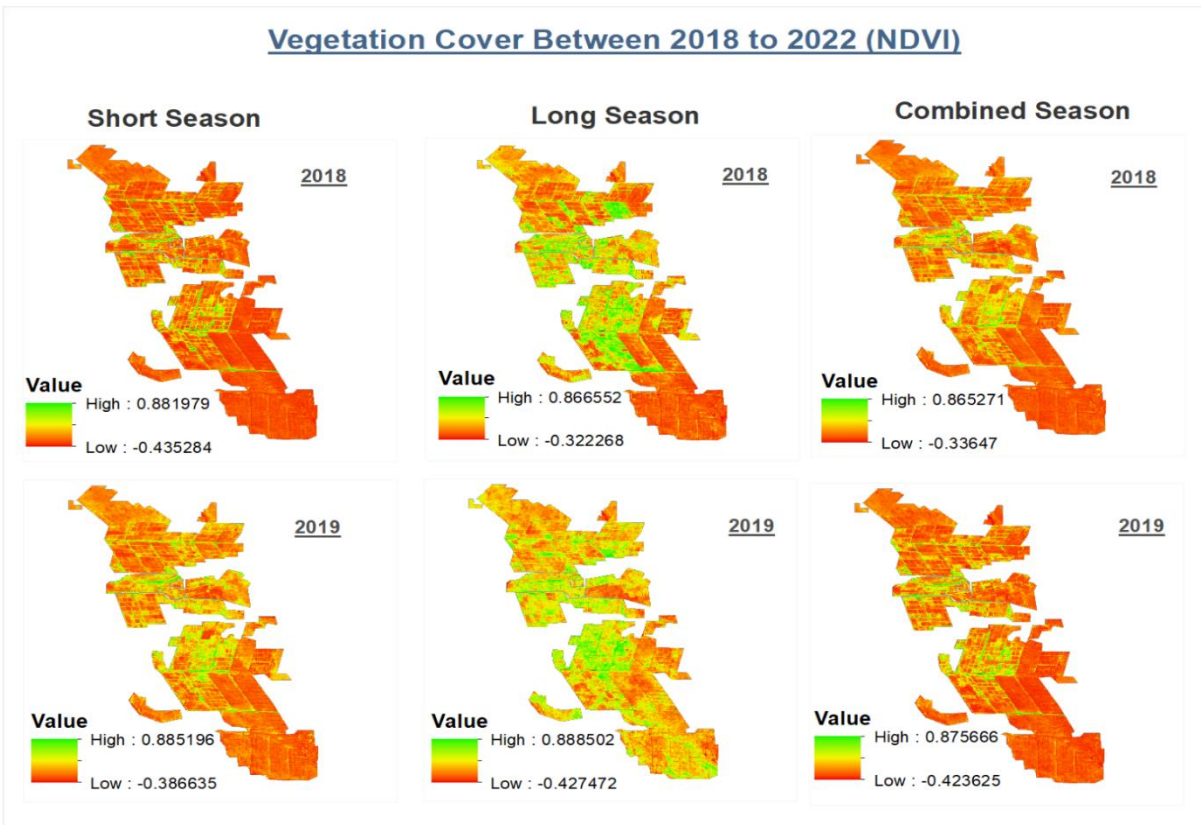


Figure 4. 4 Visualization of vegetation cover throughout the study period (2018 - 2022).

The Normalized Difference Vegetation Index (NDVI) maps in Figure 4.4 above were generated for both short and long growing periods and combined periods over four years (2018, 2019, 2021, and 2022) as part of this study on crop water productivity estimation for resilient agriculture. NDVI is calculated from remote sensing reflectance data and indicates vegetation greenness and health.

The NDVI maps allow for spatial and temporal analysis of vegetation conditions. Comparing NDVI values across the different growing periods and years provides insights into crop productivity and crop water use efficiency. Higher NDVI values generally indicate healthier, greener vegetation and higher photosynthetic activity. Lower NDVI values may indicate water stress, poor fertility, or other factors limiting plant growth.

The NDVI values vary between the short and long growing periods each year due to differences in crop phenology and water availability. During short growing periods, crops may

not reach full canopy cover before the end of the season, resulting in lower overall NDVI. In long growing periods, crops have more time to mature and develop dense, green canopies, leading to higher NDVI values. Finally, the combined season provides a preview of all the season's images merged into one.

4.2.3 Seasonal Yield Distribution

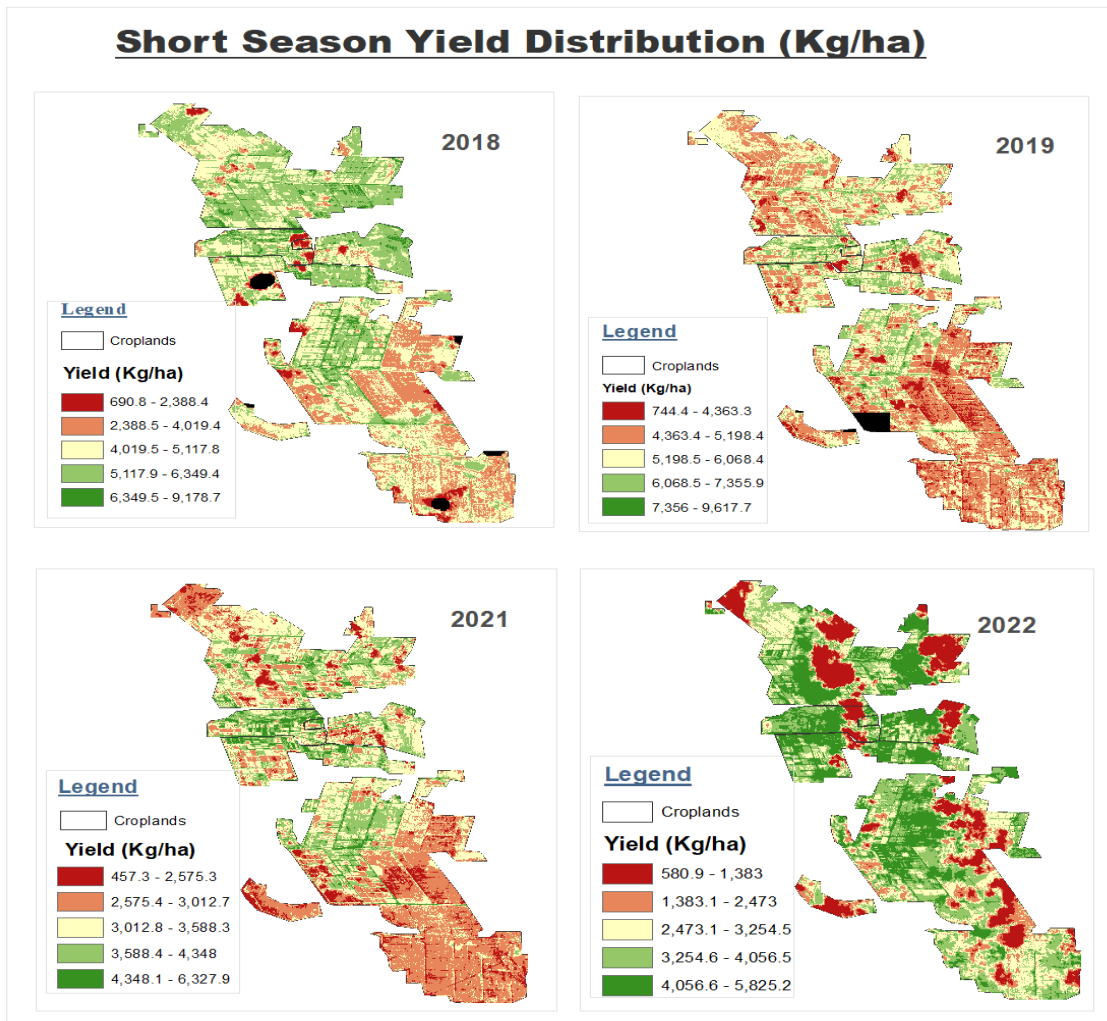


Figure 4. 5a Spatial Temporal Yield distribution during the short period between 2018 to 2022.

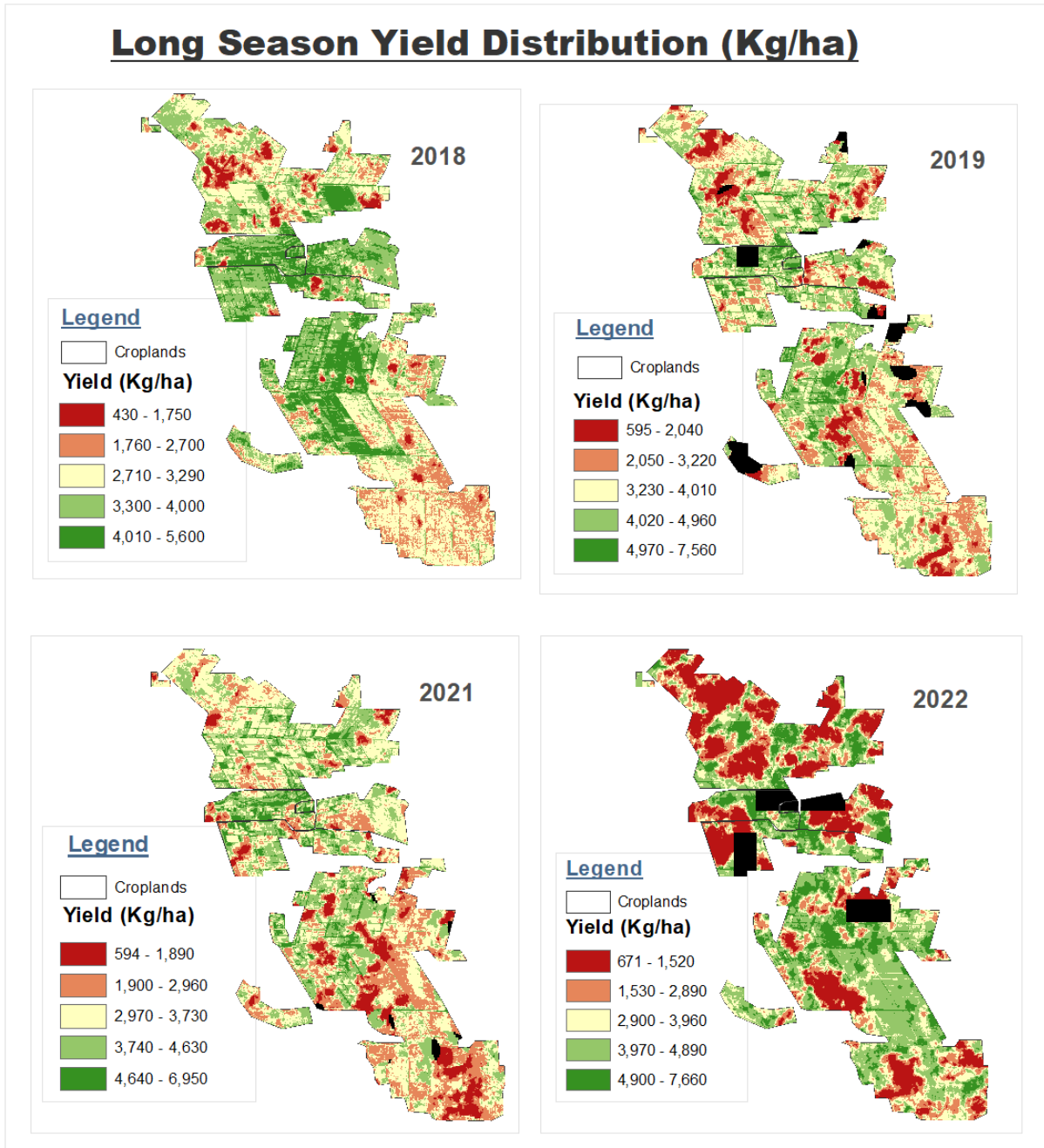


Figure 4. 6 Spatial Temporal Yield distribution during the long growing period between 2018 to 2022.

The crop yield mapping from 2018-2022 revealed clear spatial and temporal patterns in crop production across the Bura Irrigation Scheme. The yield estimates were derived by

combining remotely sensed NDVI data with a light-use efficiency model and APAR to estimate biomass, which was converted to yield using a crop-specific harvest index (HI).

For the short growing seasons, yield ranges showed high variability, with minimum yields as low as 457 kg/ha in 2021 and maximum maize yields up to 9178 kg/ha in 2018 (Fig. 4.4a). The lowest yields tended to be concentrated in the northern and southern parts of the scheme, and the 2019 season had the highest overall yields, while 2021 had the poorest yields, likely due to weather fluctuations.

The long growing seasons showed less variability, with minimum yields between 429-671 kg/ha and maximum yields of 5586-7630 kg/ha (Fig. 4.4b). The lowest yields occurred primarily in the southern areas of the scheme. The highest yields were found in the northeast region. Overall, the long-season yields were higher and more stable than the short seasons.

The maps reveal substantial spatial heterogeneity in crop yields, highlighting low-crop water productivity areas. While irrigation supports cropping in this ASAL region, water distribution, and drainage issues may limit yields in certain scheme parts. The temporal variability highlights the impacts of changing weather patterns and differences in cropping patterns and management.

In addition to the spatial yield maps, average crop yield was calculated for the entire Bura Irrigation Scheme region for each growing season and year. For the short seasons, the average yield ranged from 3.12 t/ha in 2022 to 5.05 t/ha in 2020 (Table 4.1) below. The long-season average regional yields were lower, varying from 3.22 t/ha in 2022 to 3.76 t/ha in 2019. The combined average yield for both seasons fluctuated between 6.34 t/ha in 2022 and 8.68 t/ha in 2020. The combined yield was obtained as a sum of all the season's yields in that year. A bar graph was generated for the same results to visualize the trend of yield production at the regional level within the Bura scheme (Figure 4.7) below.

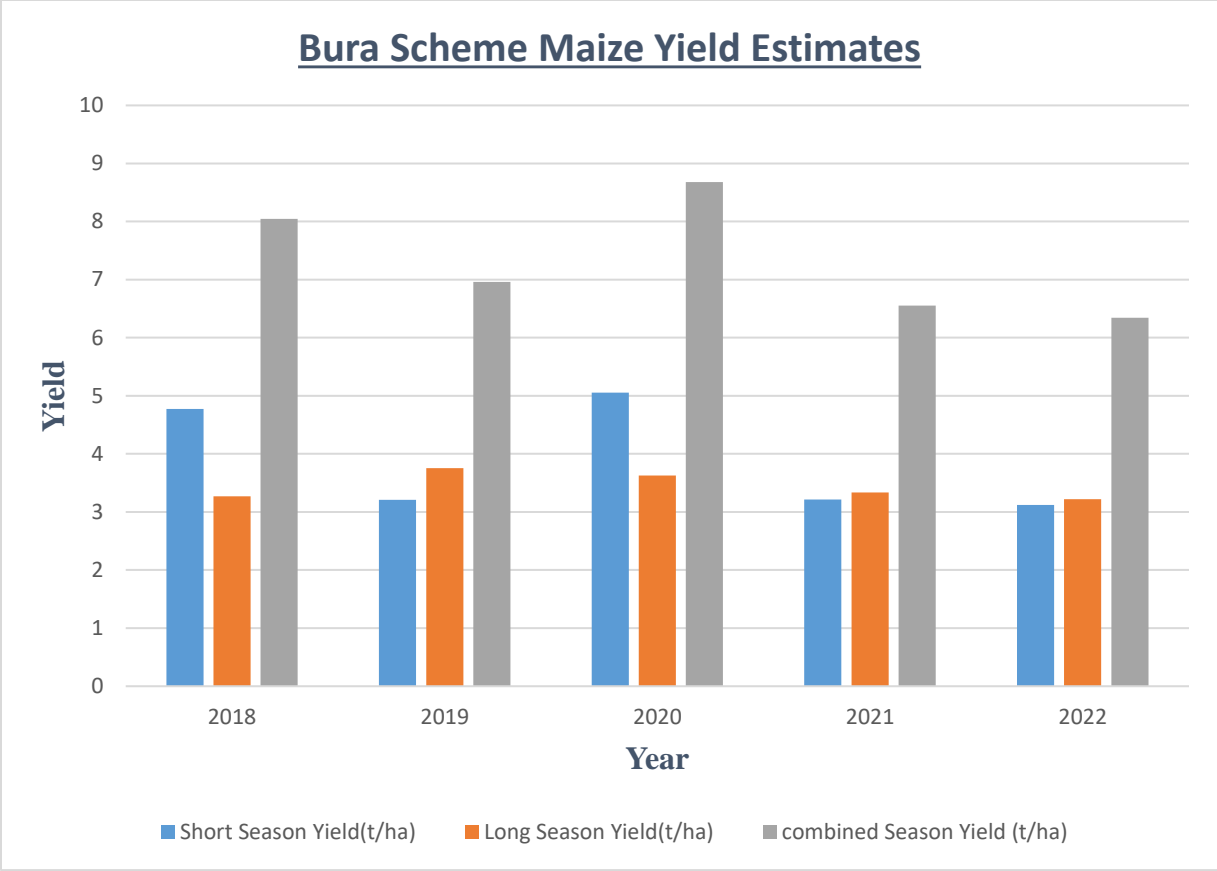


Figure 4. 7 Yearly maize yield regional based average.

Table 4. 1 Yield zonal statistics at season level and on yearly basis

Year	Short Season Yield(t/ha)	Long Season Yield(t/ha)	Combined Season Yield (t/ha)
2018	4.776366357	3.271972696	8.048339053
2019	3.207342922	3.7555	6.962842922
2020	5.052628943	3.626553298	8.679182241
2021	3.215566055	3.336088949	6.551655004
2022	3.121359342	3.222519736	6.343879078

4.3 Regional Evapotranspiration Estimates

Evapotranspiration (ET) was estimated for the Bura Irrigation Scheme study area from 2018 to 2022 using the SEBAL algorithm and Penman-Monteith equation. During the short growing season (Figure 4.8), ET ranged from a minimum of 9.3 mm to a maximum of 117.0 mm

in 2018, averaging 63.2 mm across the study period. In 2019, the range was wider, from 24 mm to 143.5 mm, with an average of 80.8 mm. The minimum ET stayed consistent in 2021 and 2022 at 9.3 mm, while maximum values were 119.0 mm and 112.0 mm, respectively. Average ET for 2021 was 64.2 mm, and 61.0 mm for 2022.

For the long growing season (Figure 4.9), the minimum ET was 9.3 mm from 2018 to 2022. Maximum ET reached 118.0 mm in 2018, 76.1 mm in 2019, 114.5 mm in 2021, and 143.9 mm in 2022. The average ET for the long season was 73.3 mm in 2018, 42.7 mm in 2019, 62.4 mm in 2021, and 78.6 mm in 2022. Overall, ET was lower and had a narrower range during the short growing season compared to the long season.

The year with the highest ET estimates in the short season was 2019, with an average of 80.8 mm and a maximum of 143.5 mm. For the long season, 2022 had the highest ET, with an average of 78.6 mm and a maximum of 143.9 mm. The lowest ET for both seasons occurred in 2019, with an average of 42.7 mm in the long season.

Higher ET in a particular season and year indicates increased crop water use and demand. Years with higher ET likely experienced better-growing conditions, less water stress, and higher potential crop yields. Lower ET suggests crops were more water-limited, leading to reduced productivity. For example, the high ET in 2019's short season means crops had adequate moisture, while the low ET in the long season indicates water stress

ET estimates from the two models were relatively consistent, indicating reliable results. Some variability between years can be attributed to changing weather patterns influencing factors like rainfall, temperature, and wind speed, which affect ET. These ET maps and estimates provide insights into crop water use in the Bura Irrigation Scheme over time. The data can help inform water management and agricultural decisions to optimize crop productivity.

Figures 4.8 and 4.9 below show ET's spatial variability across the scheme.

Short Season Actual ET Distribution

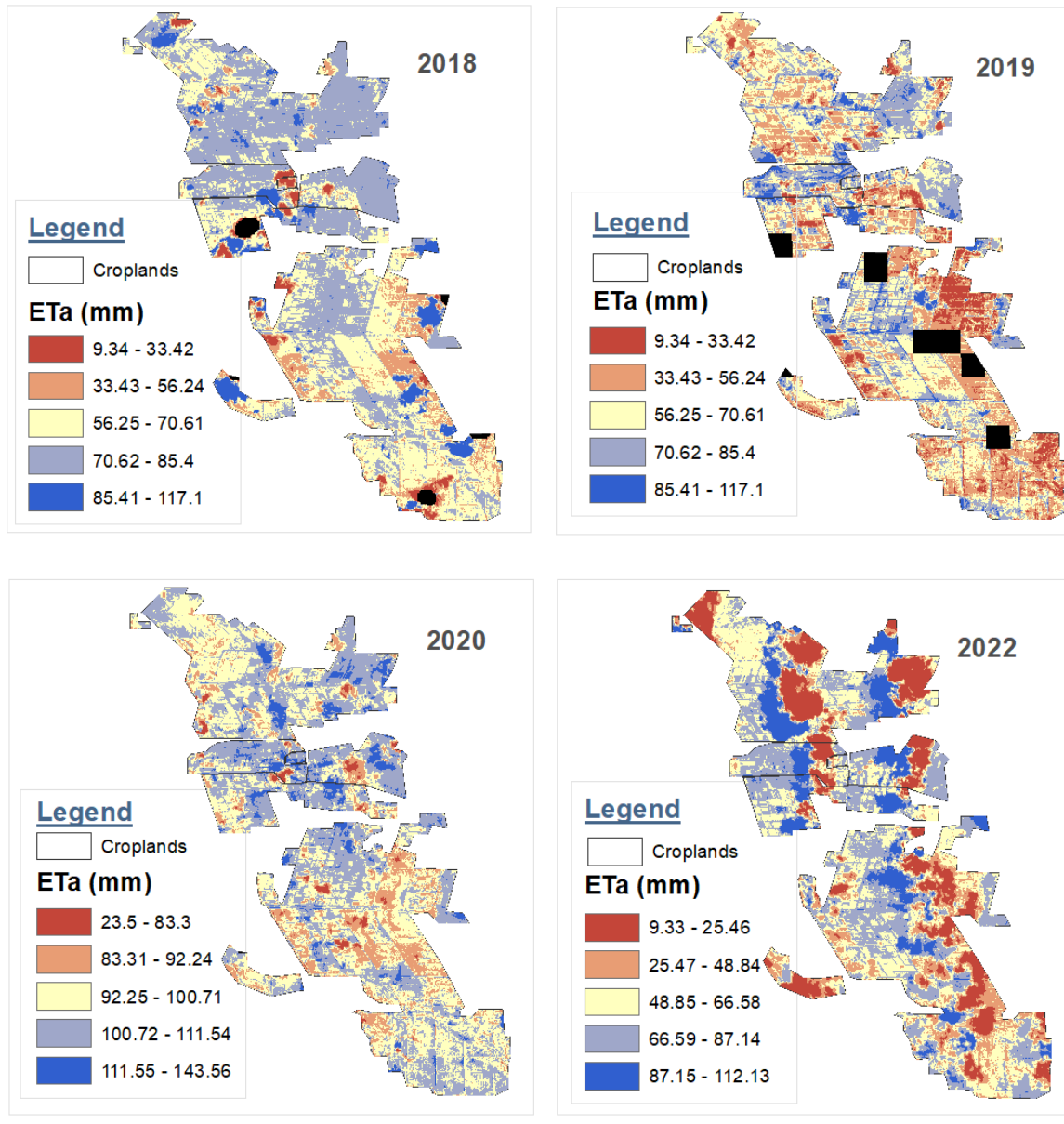
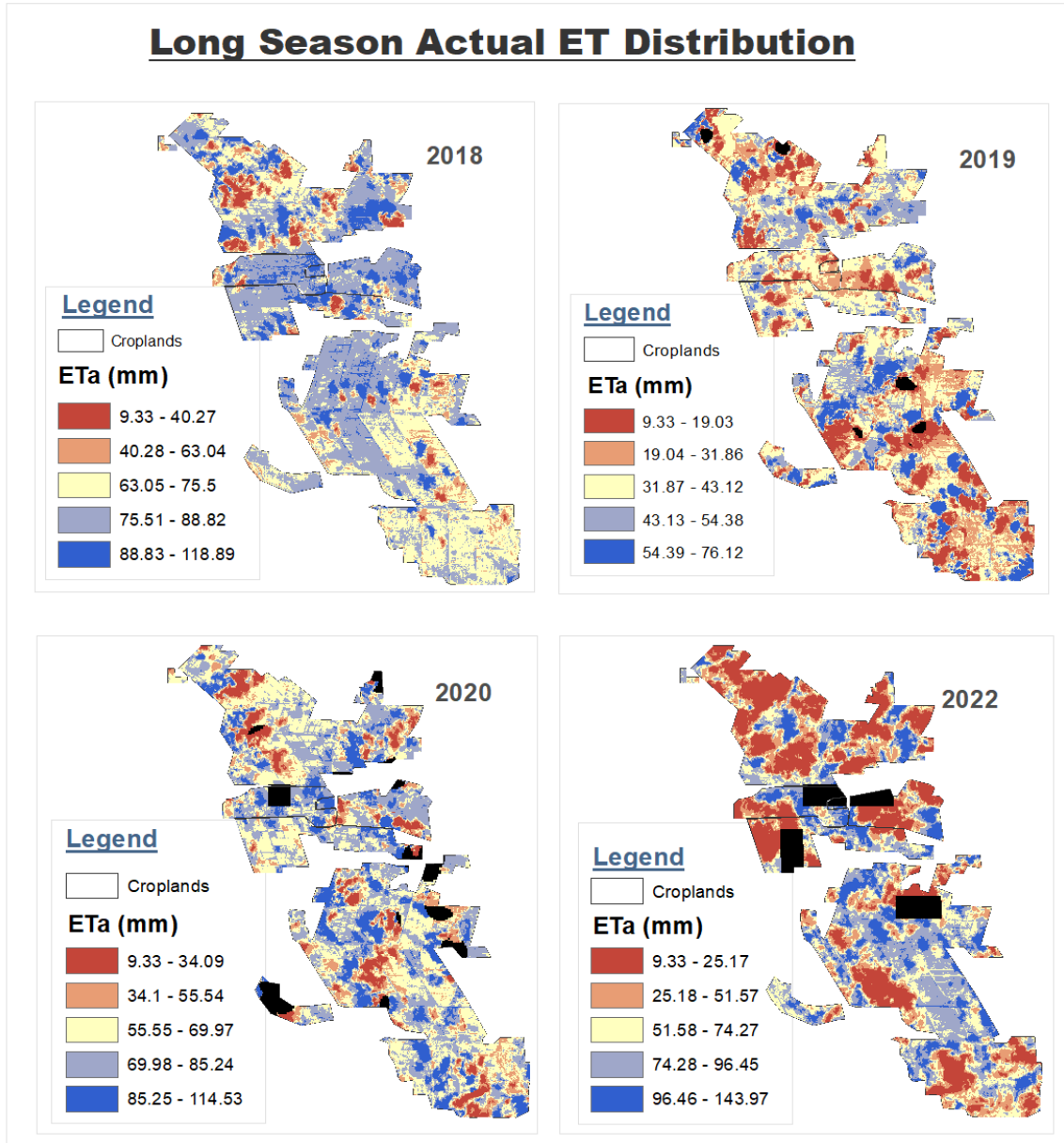


Figure 4. 8: Short season spatial temporal distribution of Evapotranspiration

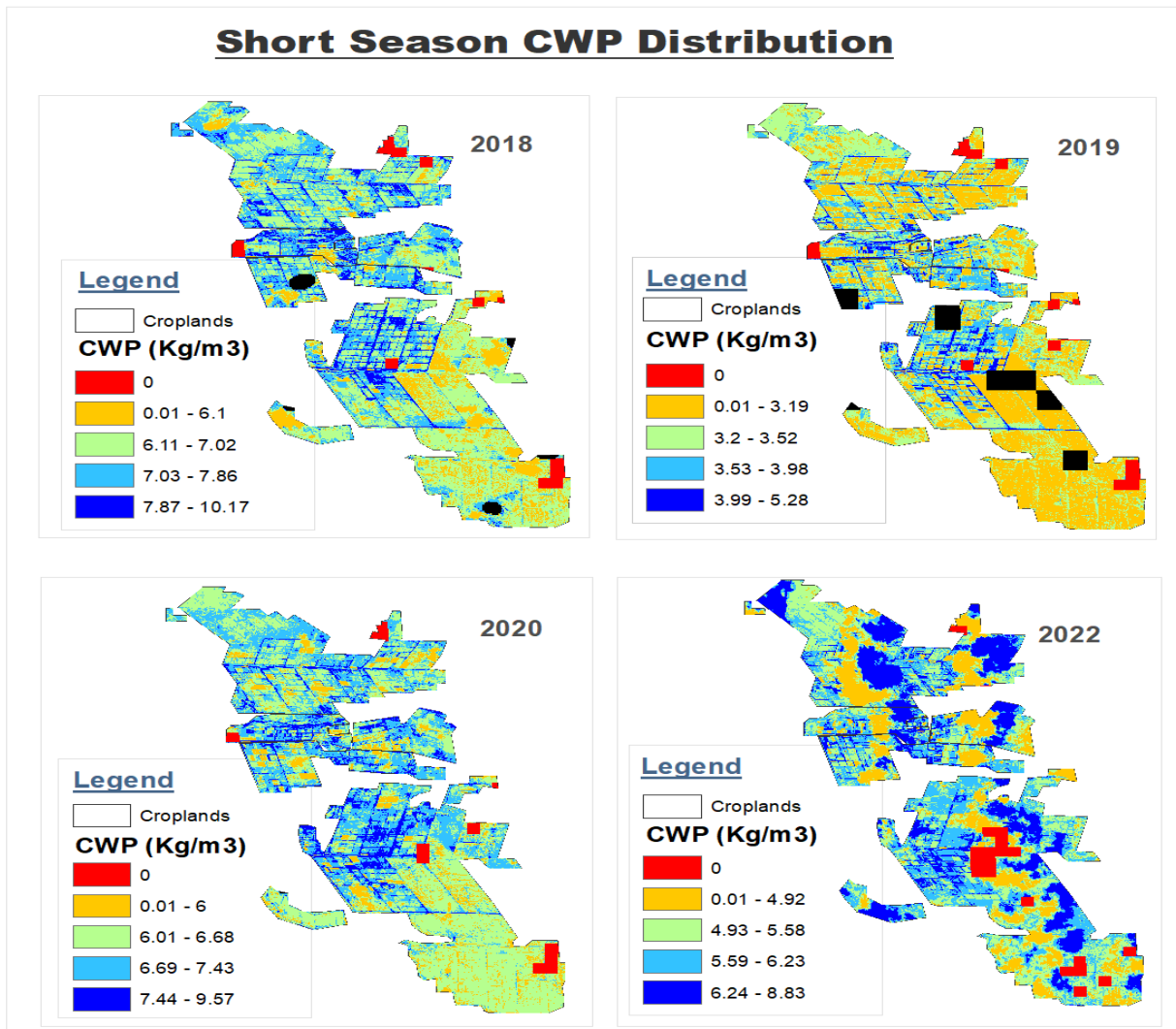
Long Season Actual ET Distribution



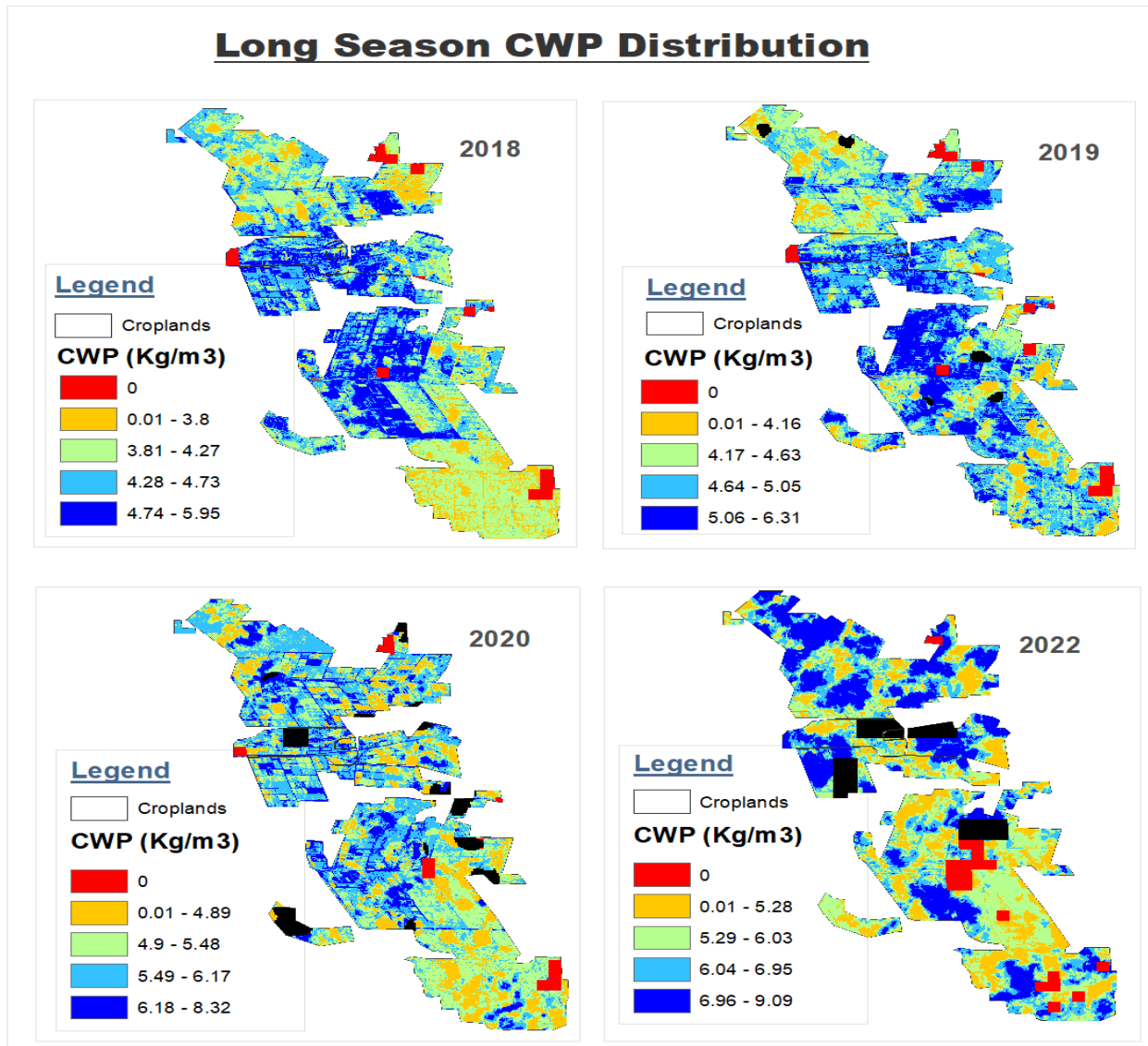
4.4 CWP Spatial Distribution

CWP maps are an output based on ET and yield. It was achieved using Equation 19. The exact timeframe and seasons were processed, and the spatial-temporal distribution of CWP maps was generated. High values of CWP represent higher productivity, while lower values represent lower productivity. Figures 4.8 and 4.9 below show the spatial distribution of CWP between 2018 and 2022 for short and long growing seasons.

In a short growing season, CWP ranges between a minimum value of 0 kg/m³ and 10.5 kg/m³ for the given timeframe. Generally, the southern part of the scheme has low crop water productivity. The central part of the study area has a higher CWP, which is the main part where irrigation happens. On the other hand, the long-growing season CWP was found to range between 0 kg/m³ to 9.5 kg/m³. Spatially, CWP was lower in 2022 in most parts of the study area for both seasons. Similar observations can be made in the 2021 long season and 2019 short growing season. These low CWP results for the seasons contribute to high ET with low yields in the maize-growing regions.



The crop water productivity maps reveal several important spatial and temporal patterns across the Bura Irrigation Scheme. While the southern areas consistently showed lower CWP, the central and north-western irrigated zones displayed higher productivity that aligned with the intensive agricultural activity in this region. The maps indicate that water management and agronomic practices may need re-evaluation in the southern scheme to improve CWP. See Figure 4.10 above and 4.11 below.



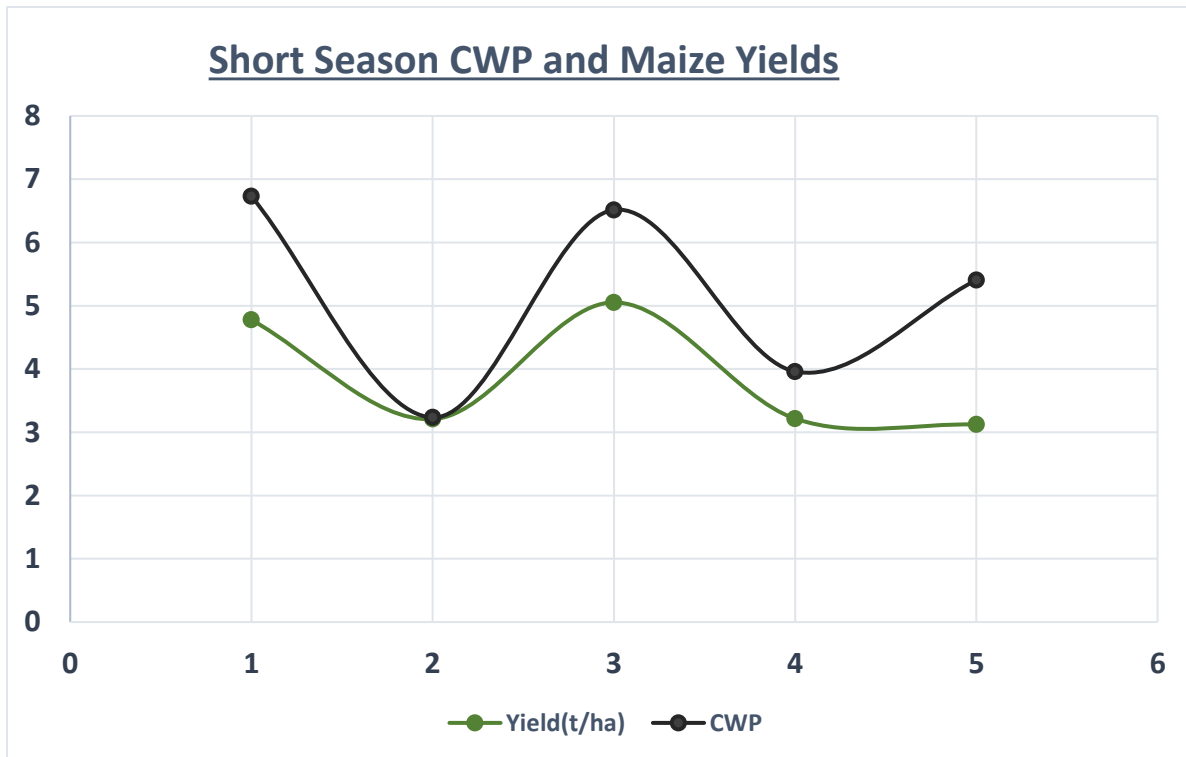


Figure 4. 9: Trend analysis of the changes between yield and CWP for the whole period of study.

4.5 CWP, Yield and ET Trend View

In addition to the spatial maps, zonal statistics were calculated for each season and year to quantify crop water productivity (Table 4.2). The reference evapotranspiration (ET₀), actual evapotranspiration (ET_a), ET_a standard deviation, yield, and CWP were summarized. This provides the temporal variability in critical water use and productivity metrics.

CWP showed differences between the short and long growing seasons. In the short season, CWP ranged from 3.23 kg/m³ in 2019 to 6.73 kg/m³ in 2018. For the long season, CWP was lower, varying between 4.21 kg/m³ in 2018 to 5.88 kg/m³ in 2022. Yield also fluctuated year-to-year, with the lowest values in the 2019 short season (3.21 t/ha) and the 2021 long season (3.34 t/ha).

Line graphs and bar plots were created to analyze further the relationships between water use, yield, and CWP. Figure 4.13 shows the connected trends in CWP and yield over time for both seasons. CWP and yields decreased in the 2019 and 2021 short seasons compared to 2018 while remaining more stable in the long season.

Meanwhile, Figures 4.13 and 4.14 illustrate the comparative bars of CWP, yield, and ETa standard deviation by season. The highest CWP and yields aligned with lower ETa variability in the 2018 short season and 2022 long season. ETa deviation was more significant in years with reduced productivity.

Overall, the temporal graphs coupled with the spatial CWP maps provide insights into the factors influencing crop water productivity over the study period. The statistics and visualizations can guide water and yield optimization.

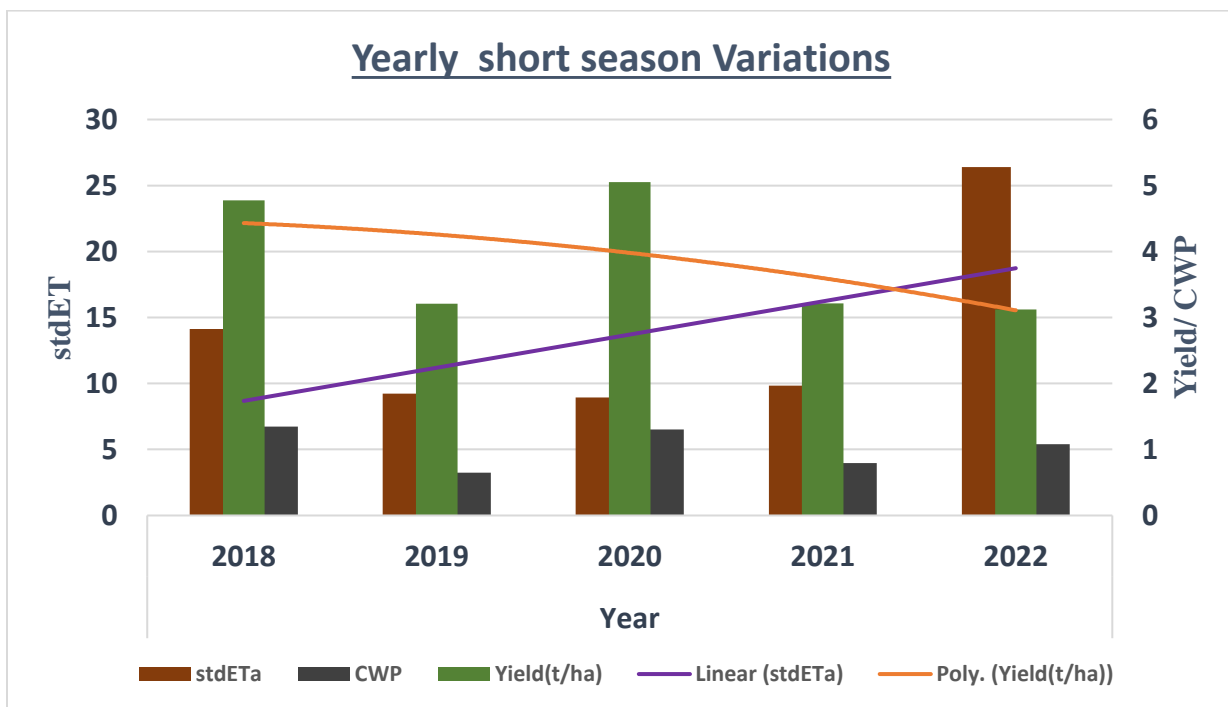


Figure 4. 10: Multivariable comparison and trend analysis of CWP, Yield and Evapotranspiration during the short growing season.

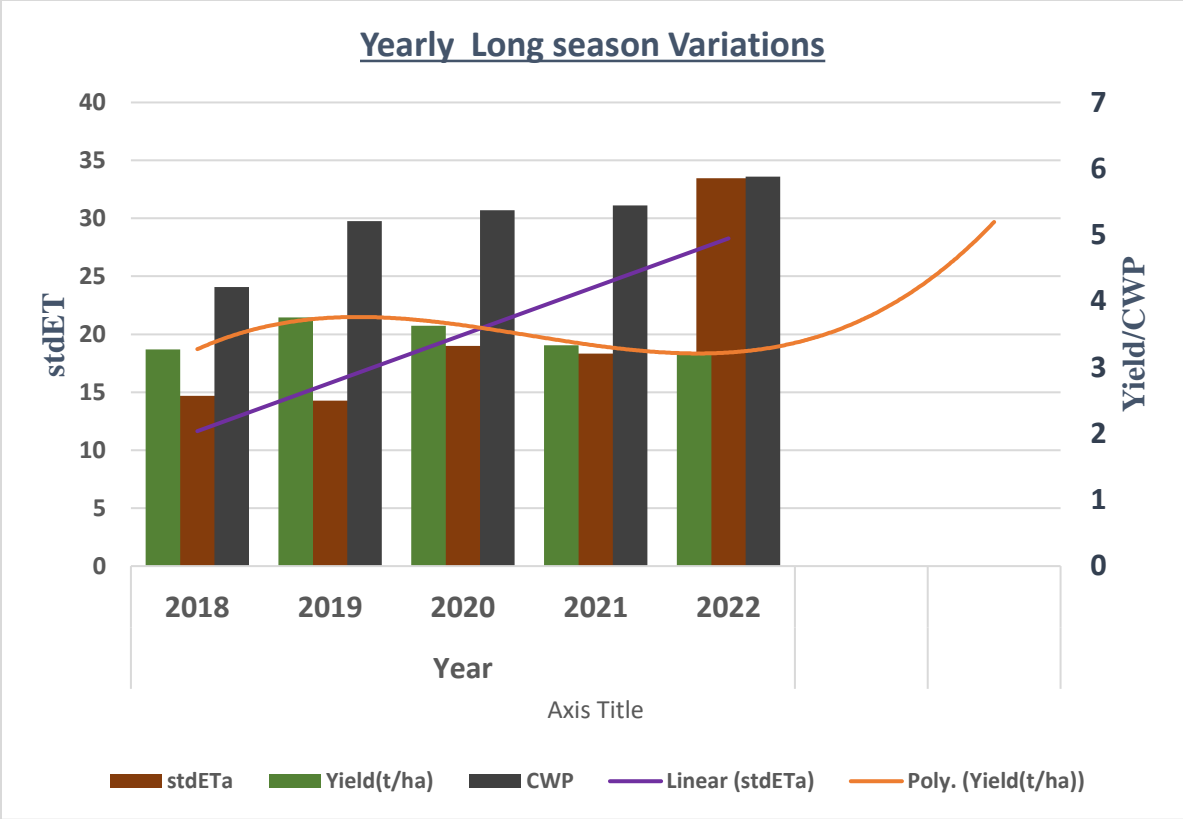


Figure 4. 11: Multivariable comparison and trend analysis of CWP, Yield and Evapotranspiration during the long season.

Table 4.2 Regional Zonal statistics for CWP, Eta, ETo, and Yield were obtained yearly and seasonal.

Short Season Statistics						
		ETo	Eta	stdETa	Yield(t/ha)	CWP
Year	2018	237.0842	69.47594	14.12928	4.776366357	6.730136
	2019	258.0159	96.215	9.227823	3.207342922	3.233628
	2020	237.2859	76.05709	8.942988	5.052628943	6.515204
	2021	261.6644	79.7649	9.837985	3.215566055	3.960766
	2022	249.1607	58.34623	26.39416	3.121359342	5.406843
Long Season Statistics						
		ETo	Eta	stdETa	Yield(t/ha)	CWP
Year	2018	215.4942	75.8447	14.69787	3.271972696	4.213657
	2019	227.6036	72.08	14.28246	3.7555	5.21018
	2020	243.2024	67.29996	19.00252	3.626553298	5.372413
	2021	249.0766	61.38575	18.32356	3.336088949	5.443546
	2022	252.2586	57.67675	33.46675	3.222519736	5.881332

4.6 Relative Importance of CWP Estimation Parameters

A correlation matrix was generated to examine the relationships between the vegetation indices, remote sensing metrics, and the target variables of crop water productivity (CWP), evapotranspiration (ET), and yield, whose results are as in Figure 4.15 below.

The vegetation indices of EVI, GNDVI, and SAVI showed strong positive correlations with each other, with coefficients ranging from 0.83 to 1.0. These three indices also demonstrated positive correlations with the target variables, with GNDVI having the strongest correlations of 0.19 with yield, 0.07 with CWP, and 0.05 with ET.

The remote sensing metrics of SR, EDI, TCI, and albedo had strong inter-correlations, with coefficients from 0.48 to 0.79. EDI and TCI showed the highest correlations with yield at 0.27 and 0.29, respectively. Albedo and land surface temperature (LST) were most strongly correlated with CWP at -0.64 and -0.49.

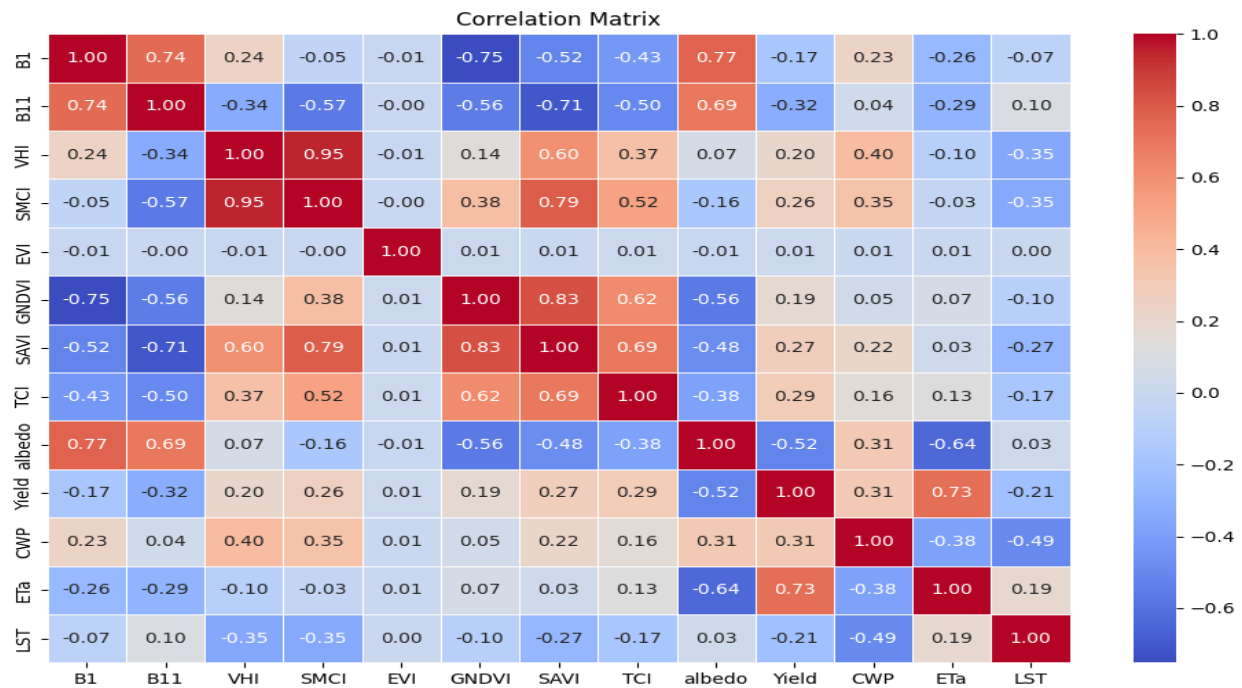


Figure 4. 12: Results of the Correlation analysis for the variables used in the machine learning modelling.

An XGBoost model was developed to predict crop water productivity (CWP), yield, and actual evapotranspiration (ETa). The model's performance was evaluated using the metrics of mean absolute error (MAE), R-squared (R²), root mean squared error (RMSE), and mean squared error (MSE).

For CWP, the model achieved an R² of 0.79, indicating that the model explains 79% of the variability in the actual CWP values. The RMSE and MSE were 0.58 and 0.34, respectively, showing the errors in predicting CWP. The model performed very well for yield prediction, with an R² of 0.88, a low RMSE of 0.45, and an MSE of 0.20. This demonstrates the model's strong ability to predict crop yield. For ETa, the model had an R² of 0.91, so over 90% of the ETa variation is explained. However, the errors were higher than CWP and yield, with RMSE of 7.27 and MSE of 52.87. So, while still good, the model's predictions of ETa were less accurate than for the other targets.

Overall, the XGBoost model strongly predicted CWP, yield, and ETa from the given data. The highest accuracy was achieved for yield, followed by ETa and CWP. These evaluation metrics quantify the model's ability to generalize and accurately estimate the target variables for this crop system.

The trained XGBoost model was used to generate predictions for crop water productivity (CWP), yield, and actual evapotranspiration (ETa) across the Bura Irrigation Scheme. Statistical summaries were calculated on the predictions and compared to the summaries of the actual field data. The model predictions had a mean of 6.12, close to the actual mean of 5.85 for CWP. The standard deviation of the predictions was 0.92 compared to 0.88 for the actual data.

The predicted yield had a mean of 3.14 and a standard deviation 1.55. The actual yield statistics were a mean of 3.10 and a standard deviation of 1.41. So, the yield predictions aligned well with the actual yield distribution. ETa prediction, on the other hand, showed the most prominent difference from the field data. The predicted ETa mean was 56.29, and the standard deviation was 32.92. However, the actual ETa mean was 56.92, with a standard deviation 29.99. So, the model slightly overestimated the ETa mean.

The scatter plot in Figure 4.16 allows visualization of the relationship between the model's predicted values and actual field-measured values for crop water productivity (CWP),

yield, and actual evapotranspiration (ETa). Each point on the plot represents one pair of predicted and observed values. The distribution and clustering of these points provide insight into how well the predictions correspond to the actual data.

A linear fit line is added to the scatter plot with an equation of $y=0.98x + 0.41$. The slope of 0.98 indicates a nearly 1:1 agreement between predicted and actual values across the variables. This slope close to 1 demonstrates the strong correlation between model predictions and ground truth data. The intercept of 0.41 suggests a slight bias where the model slightly overestimates the field values. The fit line quantifies the close linear relationship between predictions and measurements.

4.7 Xgboost Model results

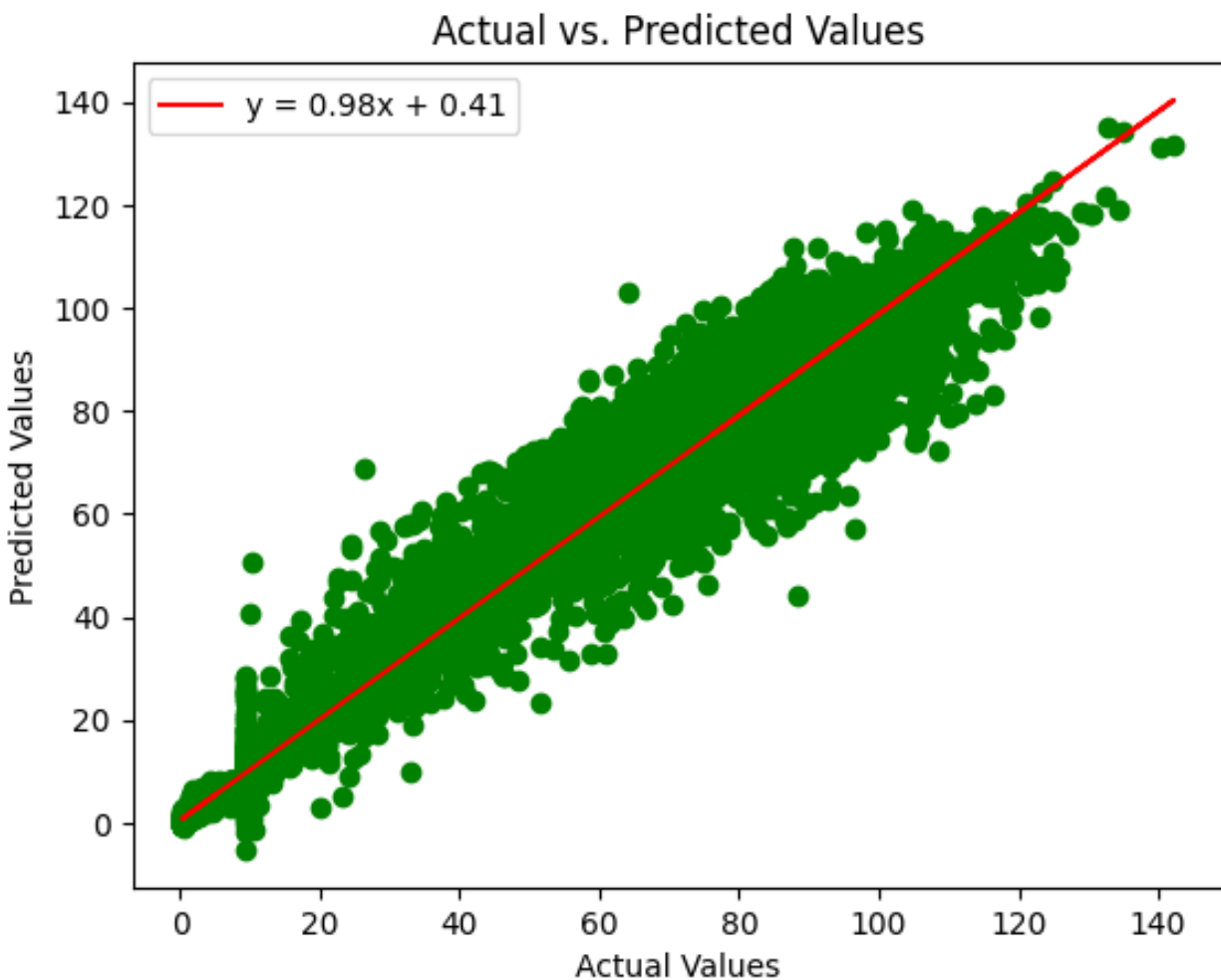


Figure 4. 13: XGBoost model results analysis for CWP, Yield and ET; Scatter plot.

4.8 Validity of Machine Learning Model (XGBoost)

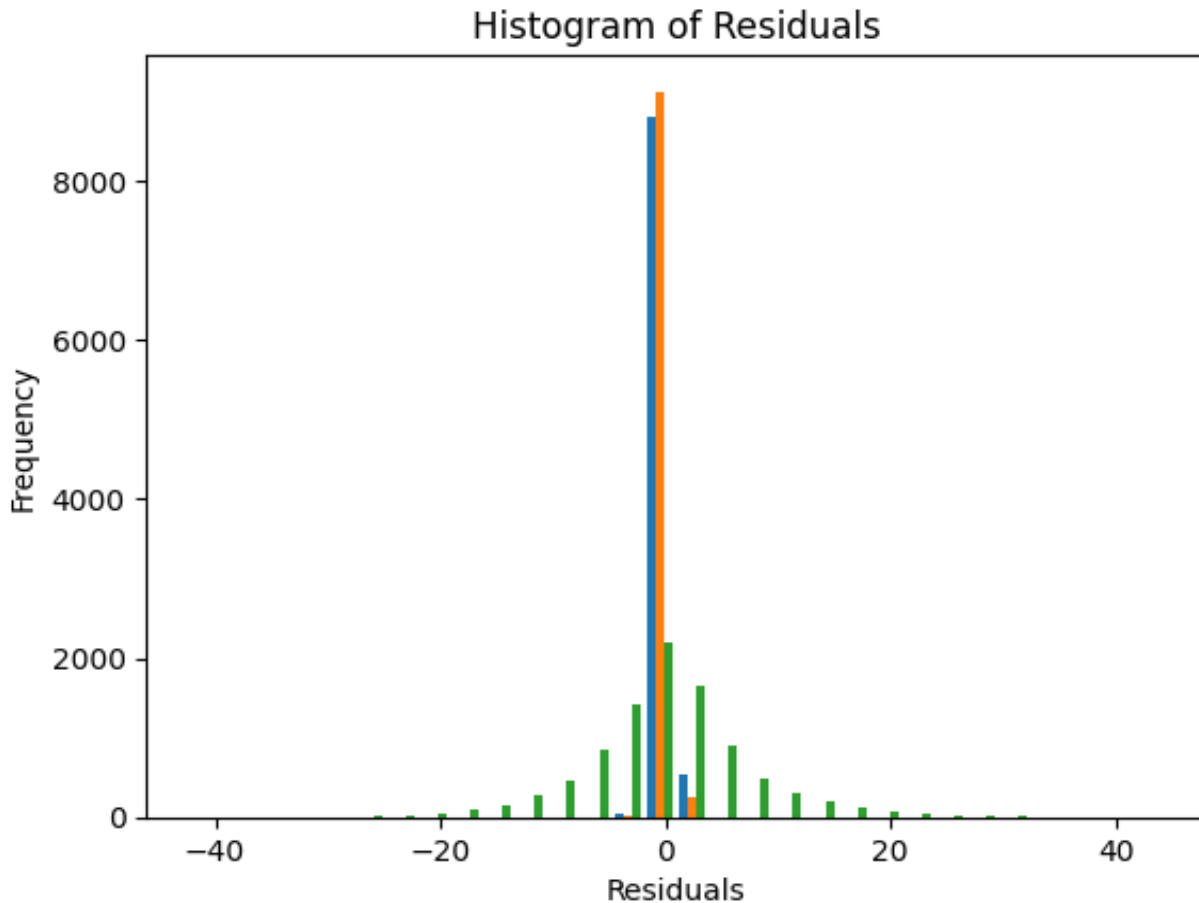


Figure 4. 14: Machine learning model (XGBoost) prediction residuals.

A residual plot (Figure 4.17) was generated to evaluate the trained XGBoost model by plotting a histogram of the prediction residuals (errors). The residual is the difference between the model's predicted value and the actual observed value for each data point. The residual plot shows the highest frequency of residuals near 0, with decreasing frequencies towards the left and right extremes. The concentration of residuals near 0 indicates the model is making more minor errors overall, predicting values close to the actual observations. The tapering residuals further from 0 suggest more significant prediction errors occur less often. The symmetrical, bell-curve-like shape shows that the errors are approximately normally distributed, which is ideal.

No strong skewness or multiple peaks are apparent, signifying the lack of systemic biases in the model's errors. The residual plot demonstrates that this XGBoost model accurately predicts the target variables, with the majority of errors small and centered on 0. Therefore, the residual distribution validates the model's strong performance and ability to generalize predictions across the dataset with minimal bias.

The trained XGBoost model was then directly applied to a 2022 satellite image (daily collection for the specified short and long maize growing season) covering the Bura Irrigation Scheme area. Predictions were generated for crop water productivity (CWP), yield, and actual evapotranspiration (ETa) across the scheme using various spectral bands and vegetation indices as model inputs.

The zonal statistics of the predicted CWP had a mean of 5.80 and standard deviation of 0.25 across the study area. Compared to the actual field-measured CWP statistics of a mean of 5.85 and a standard deviation of 0.88, the predicted mean is very close to the actual value. However, the model underestimates the variability in CWP, as seen in the lower standard deviation. The model estimated a mean of 3.27 and a standard deviation of 0.50 for predicted yield. The actual yield measurements had a mean of 3.10 and a higher standard deviation of 1.41. So, the model overestimates the actual yield mean slightly but underestimates the yield variability across the scheme.

The most considerable difference between predicted and measured statistics is seen for ETa. The model predicted an ETa mean of 59.25 with a standard deviation 12.78. However, the ground measurements showed a lower ETa mean of 56.92 and a much higher variability of 29.99. This indicates that the model is overestimating ETa while also underestimating the spatial differences in ETa across the study region.

5. Discussion

5.1 Crop Phenology Dynamics

Analysis of vegetation indices over time provided insights into crop growth stages and phenology patterns in the Bura Irrigation Scheme. The NDVI and EVI curves showed two distinct maize growing periods aligned with the short and long rainy seasons.

Sowing typically occurs around late January and in the short season, sowing is done in late August to early September. Peak greenness and canopy cover were reached at tasseling or silking around late April to May in the long season and around late September to October in the short season. Finally, maturity and senescence phases occurred around late July to early August for long-season maize and mid-December for short-season crops.

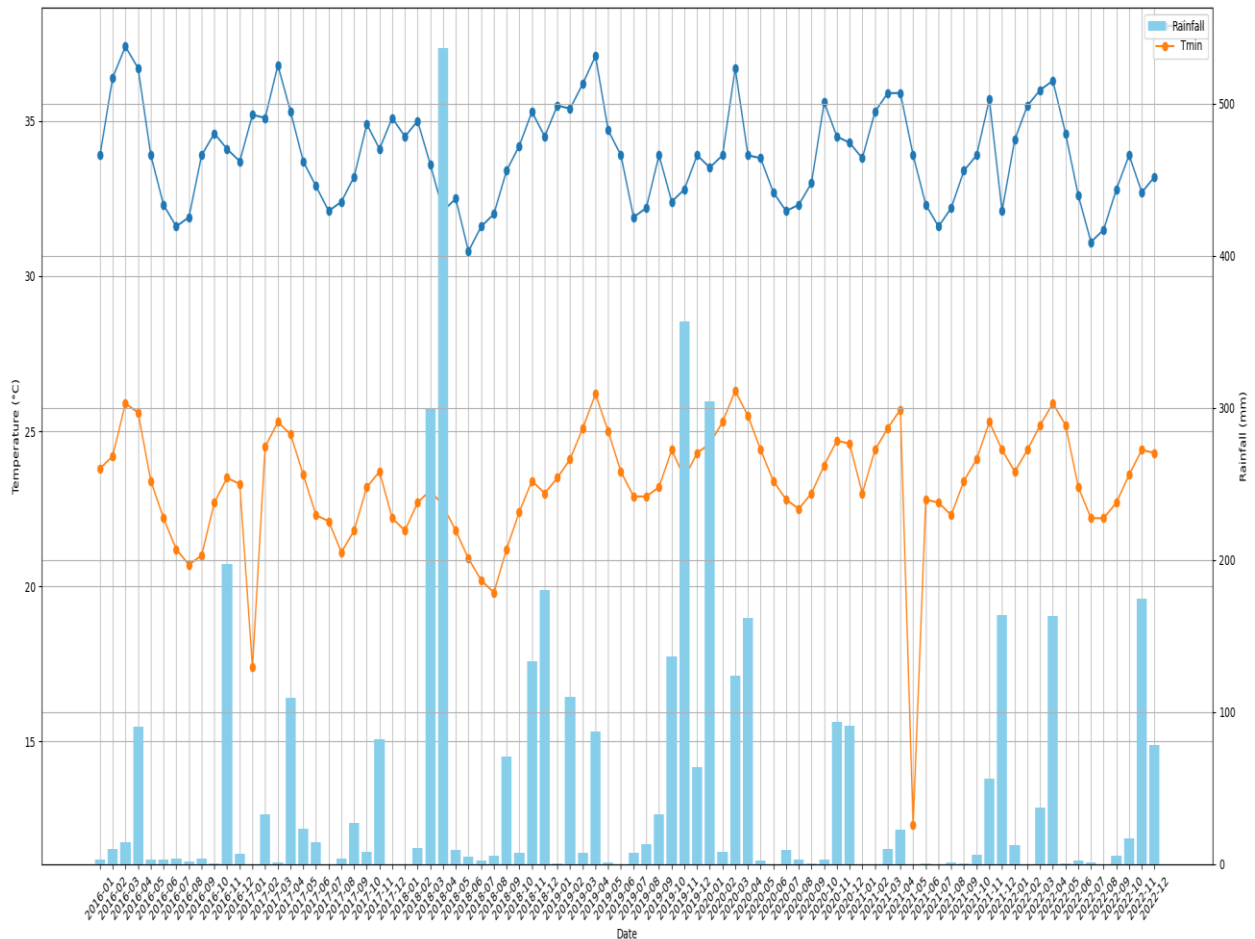


Figure 5. 1: Main Local weather dataset showing the seasonal variation in rainfall and temperature between 2016 to 2022 on monthly basis

These patterns agree well with crop calendars for the region reported by KMD and local agricultural agencies (Mugai et al., 2019; KALRO, 2022). Figure 5.1 shows the actual local weather pattern based on the KMD dataset.

The bimodal cycles enable year-round production but can complicate irrigation and other management. In particular, the short-season growth stage is vulnerable to terminal moisture stress due to cessation of rains (KMD, 2022).

Phenology monitoring showed that sowing periods were less discrete than typical rainfed systems, likely due to year-round irrigation availability allowing continuous cropping. Overlaps in sowing and maturity phases indicate farmers' capacity to stagger planting dates to manage water demand and respond to climate variability.

Capturing crop growth dynamics is critical for indexing vegetation health and parameterizing CWP models. For example, relating ETa patterns to different crop stages helps diagnose likely water stress causes (e.g., emergence vs. grain filling). Similarly, the timing of NDVI peaks and yield estimation is keyed to specific developmental milestones like flowering and maturity (Bastiaanssen & Ali, 2003).

5.2 Analysis of Yield and ET Results

The light-use efficiency modeling approach enabled the mapping of maize yield across the study area. Estimated yields showed high spatio-temporal variability, ranging from 429 kg/ha to 9178 kg/ha, as evidenced in the results. Research has found that factors like soil fertility, crop genetics, and management practices can significantly influence productivity in this region (Muli et al., 2015). While the model provides valuable yield estimations, ground-truthing and calibration using field measurements could help refine parameterization and improve accuracy over time (Bastiaanssen & Ali, 2003). Field collection of yield data through surveys or on-farm trials supports the continued advancement of the remote sensing-based yield model.

Evapotranspiration (ET) was estimated using the SEBAL energy balance model, incorporating satellite imagery and weather data. Estimated ET fluctuated between 42.7 mm and 117 mm over the study period, reflecting changing crop water demand. Prior research found that maize ET in similar environments varies depending on climate, soils, and irrigation. Tracking ET

provides insights into crop water usage dynamics and factors influencing productivity like water stress (Allen et al., 1998). However, uncertainties remain in the ET modeling, requiring further validation using lysimetry or other field techniques. Improved characterization of soil hydraulic properties, crop development, and micrometeorology can aid ET estimation.

5.3 Spatio-temporal Patterns in Crop Water Productivity

The crop water productivity (CWP) analysis for the Bura Irrigation Scheme revealed distinct spatial and temporal patterns. CWP showed high variability both between seasons and across years. In the short growing season, CWP was lowest in 2019 at 3.23 kg/m³ and highest in 2018 at 6.73 kg/m³. For the long season, CWP ranged from 4.21 kg/m³ in 2018 to 5.88 kg/m³ in 2022. This aligns with findings by Muigai et al. (2019), who reported CWP between 3.9-6.5 kg/m³ for maize in the Bura scheme, depending on the planting date.

This variability highlights the impacts of changing weather patterns, differences in cropping systems, and water management across the scheme (Blatchford et al., 2018). Years with higher CWP, like the 2018 short season, suggest better moisture availability, while low CWP in the 2019 short growing period indicates water limitation reduced productivity despite irrigation potential (Sarshad et al., 2021). As CWP integrates crop yield and actual evapotranspiration, it provides an index of plant growth, vigor, and the efficiency of water use (Hellegers et al., 2009; Booker & Trees, 2020).

The temporal CWP fluctuations demonstrate the need for flexible, responsive water management to account for inter-annual weather variability and associated crop water requirements in this semi-arid region (Hommadi & Almasraf, 2019). For instance, the higher CWP in the 2018 short season resulted from sufficient moisture availability to meet crop demands. Nevertheless 2019, CWP declined as crops faced water limitations during the critical growth stages, reducing yields (Talpur et al., 2023). Adapting irrigation scheduling and volume based on real-time crop water use and plant water stress monitoring could stabilize productivity over variable seasons.

CWP responds to agronomic practices like irrigation, planting density, and fertilizer application. Thus, the CWP assessment provides an integrated measure of crop performance and water use efficiency to guide management. Monitoring of CWP using remote sensing offers a

scalable approach to benchmarking and improving agricultural water productivity in the area. However, several sources of uncertainty persist and require ongoing research. Soil variability, microclimate fluctuations, and complex crop-water interactions may not be fully captured (Gibson et al., 2018). Field instrumentation like soil moisture probes, weather stations, and plant water status sensors can provide data to refine parameterization and improve estimation performance (Dalla Marta et al., 2018). Higher resolution inputs from satellite, drone, or airplane remote sensing could better resolve within-field heterogeneity.


5.4 Machine Learning for Enhanced CWP Estimation

Machine learning techniques have emerged as powerful tools for enhancing crop water productivity (CWP) estimation using remote sensing and meteorological data. As demonstrated in this study, models like XGBoost can effectively integrate and analyze different data sources to generate reliable predictions of critical parameters, including CWP, yield, and evapotranspiration (Elbeltagi et al., 2022).

Implementing XGBoost allowed the exploitation of complex interactions and patterns within the dataset that would be difficult to model using traditional statistical approaches. Virnodkar et al. (2020) discussed that machine learning algorithms like XGBoost have distinct advantages for handling large, multi-dimensional agricultural datasets comprising various soil, climate, spectral, and crop-specific variables. By capturing non-linear relationships, high-order interactions, and latent data structures, the XGBoost model provided robust generalization capability beyond the training data.

The model evaluation confirmed the strong predictive performance of XGBoost for the target output variables. Metrics like the coefficient of determination (R^2) and residual analysis demonstrated good model fit with minimal bias. As Patel et al. (2021) noted, machine-learning approaches require careful validation to assess real-world applicability. The residual errors centered around zero provided evidence that XGBoost could produce reliable CWP, yield, and ET estimates from unseen data in the operational setting.

Another benefit of using XGBoost was the intrinsic feature selection, ranking the most informative input variables through each iteration of tree splits during model training (Islam et al., 2023). This allowed automated optimization of the many spectral, soil, vegetation and



climate variables available to determine an optimal subset for accurate CWP and ET prediction. Such data-driven variable selection removes subjective biases common in manual techniques.

Overall, the machine learning methods implemented significantly advanced the crop water productivity assessment, providing a scalable approach to integrating large datasets and extracting actionable insights.

6. Conclusions

In conclusion, this research used remote sensing and machine learning techniques to estimate crop water productivity (CWP) for maize in the Bura Irrigation Scheme from 2018-2022. The goal was to support improved water management and agricultural resilience in this semi-arid region prone to drought and water scarcity.

The study incorporated satellite imagery across multiple sensors (Landsat, Sentinel-2, MODIS) as well as meteorological data from ERA5 and local weather stations. Vegetation indices, including NDVI and EVI, enabled crop growth stage monitoring while ET was modelled using SEBAL energy balance and the FAO Penman-Monteith method. Crop yield was estimated using a light-use efficiency approach relating biomass production to absorbed photosynthetically active radiation and harvest index.

The generated CWP maps and statistics revealed distinct spatial patterns, with higher productivity aligning to intensely irrigated zones while southern areas showed poorer CWP. Temporal fluctuations occurred between seasons and years, highlighting the impacts of variable weather and water availability on crop-water dynamics. Estimated CWP ranged from 3.2-6.7 kg/m³ over the period, agreeing with previous local studies. The yield model performed well but could be further improved through calibration with field measurements.

Machine learning, specifically the XGBoost algorithm, was implemented to enhance CWP, yield, and ET estimation by exploiting complex data relationships. The model evaluation confirmed excellent predictive capability based on metrics like R², RMSE, and residual analysis. Applied to unseen satellite imagery, the trained model generated reasonable CWP, yield, and ET estimates across the scheme.

Finally, this study highlighted the capabilities of emerging digital techniques for water productivity monitoring to support agricultural sustainability. With further refinement, the developed methods could aid real-time irrigation decisions and benchmarking to improve yields, optimize water utilization, and build regional climate change resilience. Integrating such data-driven approaches with water governance and policy frameworks will be essential to translate information into action for sustainable agricultural water management.

Recommendations

Expand the ground data collection to additional locations within the irrigation scheme. While the current analysis relied primarily on remotely sensed data, gridded, and single station local weather datasets, collecting in situ data on crop yields, soil moisture, weather parameters, and water use at more sites would allow for improved validation and parameterization of the CWP model across the spatial extent of the scheme. This could be achieved through test plots and intensive measurement campaigns during critical crop growth stages.

Incorporate higher-resolution satellite imagery from sensors such as GeoEye, Pleiades, or drone and airborne data. The 10-30 m resolution data currently limits the characterization of fine-scale heterogeneity in factors driving field-level variations in CWP. Sub-meter satellites could better capture soil, topographic, and drainage variability to diagnose tightly localized yield constraints.

Test alternate deep learning algorithms and ensemble approaches to improve generalizability. While XGBoost performed well presently, evaluating other techniques like neural networks or combinations of statistical and deep learning models may lead to better out-of-sample prediction accuracy. Given the limited training dataset size, tuning model structural complexity could prevent overfitting. Augmenting input data through generative adversarial networks (GANs) is another exploration avenue.

Integrate findings into a user-friendly decision support system for local stakeholders. Packaging the results into an interactive web dashboard or mobile application is imperative to enable accessible adoption of the CWP insights by irrigation scheme managers and agricultural extension officers. The system could provide real-time visualization of the spatial CWP and yield patterns while summarizing trends and variability for the region. User input mechanisms to customize recommendations based on crop types, field locations and available budget should also be enabled for practical utility. Leveraging such ICT tools can bridge the research-implementation gap more effectively.

7. References

- Abou Zakhem, B., Al Ain, F., & Hafez, R. (2019). Assessment of field water budget components for increasing water productivity under drip irrigation in arid and semi-arid areas, Syria. *Irrigation and Drainage*, 68(3), 452–463. <https://doi.org/10.1002/ird.2286>
- Agilandeewari, L., Prabukumar, M., Radhesyam, V., Phaneendra, K. L., & Farhan, A. (2022). Crop classification for agricultural applications in hyperspectral remote sensing images. *Applied Sciences*, 12(3), 1670. <https://doi.org/10.3390/app12031670>
- Ahmadpour, A., Farhadi Bansouleh, B., & Azari, A. (2022). Proposing a combined method for the estimation of spatial and temporal variation of crop water productivity under deficit irrigation scenarios based on the AQUACROP model. *Applied Water Science*, 12(7). <https://doi.org/10.1007/s13201-022-01666-8>
- Ashwitha, A., & Latha, C. A. (2022). Crop recommendation and yield estimation using machine learning. *Journal of Mobile Multimedia*. <https://doi.org/10.13052/jmm1550-4646.18320>
- Bansouleh, B., Karimi, A., & Hesadi, H. (2015). Evaluation of sebal and SEBS algorithms in the estimation of maize evapotranspiration. *International Journal of Plant & Soil Science*, 6(6), 350–358. <https://doi.org/10.9734/ijpss/2015/15711>
- Bastiaanssen, W.G.M. and Ali, S. (2003). ‘A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan’, *Agriculture, Ecosystems & Environment*, 94(3), pp. 321–340. [https://doi.org/10.1016/s0167-8809\(02\)00034-8](https://doi.org/10.1016/s0167-8809(02)00034-8).
- Bekchanov, M., Lamers, J. P., Karimov, A., & Müller, M. (2011). Estimation of spatial and temporal variability of crop water productivity with incomplete data. *Cotton, Water, Salts and Soums*, 329–344. https://doi.org/10.1007/978-94-007-1963-7_20
- Blatchford, M., Karimi, P., Bastiaanssen, W. G. M., & Nouri, H. (2018). From global goals to local gains—a framework for crop water productivity. *ISPRS International Journal of Geo-Information*, 7(11), 414. <https://doi.org/10.3390/ijgi7110414>

Bolfe, É. L., Parreiras, T. C., Silva, L. A., Sano, E. E., Bettioli, G. M., Victoria, D. de, Sanches, I. D., & Vicente, L. E. (2023). Mapping agricultural intensification in the Brazilian savanna: A machine learning approach using harmonized data from Landsat Sentinel-2. *ISPRS International Journal of Geo-Information*, *12*(7), 263. <https://doi.org/10.3390/ijgi12070263>

Booker, J. F., & Trees, W. S. (2020). Implications of water scarcity for water productivity and Farm Labor. *Water*, *12*(1), 308. <https://doi.org/10.3390/w12010308>

Cemek, B., Tasan, S., Canturk, A., Tasan, M., & Simsek, H. (2023). Machine learning techniques in estimation of eggplant crop evapotranspiration. *Applied Water Science*, *13*(6). <https://doi.org/10.1007/s13201-023-01942-1>

Copernicus, E. (2018, June 14). *Copernicus Climate Data Store*. Copernicus Climate Data Store |. <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview>.

Cuculeanu, V., Tuinea, P., & Bălțeanu, D. (2002). Climate change impacts in Romania: Vulnerability and adaptation options. *GeoJournal*, *57*(3), 203–209. <https://doi.org/10.1023/b:gejo.0000003613.15101.d9>

Dalla Marta, A., Eitzinger, J., Kersebaum, K.-C., Todorovic, M., & Altobelli, F. (2018). Assessment and monitoring of crop water use and productivity in response to climate change. *The Journal of Agricultural Science*, *156*(5), 575–576. <https://doi.org/10.1017/s002185961800076x>

Darwish, M. A., Elkot, A. F., Elfanah, A. M., Selim, A. I., Yassin, M. M., Abomarzoka, E. A., El-Maghraby, M. A., Rebouh, N. Y., & Ali, A. M. (2023). Evaluation of wheat genotypes under water regimes using hyperspectral reflectance and agro-physiological parameters via genotype by yield*trait approaches in Sakha Station, Delta, Egypt. *Agriculture*, *13*(7), 1338. <https://doi.org/10.3390/agriculture13071338>

Dhau, I., Dube, T., & Mushore, T. D. (2019). Examining the prospects of sentinel-2 multispectral data in detecting and mapping maize streak virus severity in smallholder

Ofcolaco farms, South Africa. *Geocarto International*, 36(16), 1873–1883.
<https://doi.org/10.1080/10106049.2019.1669724>

Dhillon, M. S., Kübert-Flock, C., Dahms, T., Rummeler, T., Arnault, J., Steffan-Dewenter, I., & Ullmann, T. (2023). Evaluation of modis, Landsat 8 and sentinel-2 data for accurate crop yield predictions: A case study using STARFM NDVI in Bavaria, Germany. *Remote Sensing*, 15(7), 1830. <https://doi.org/10.3390/rs15071830>

Elbeltagi, A., Srivastava, A., Kushwaha, N. L., Juhász, C., Tamás, J., & Nagy, A. (2022). Meteorological data fusion approach for modeling crop water productivity based on ensemble machine learning. *Water*, 15(1), 30. <https://doi.org/10.3390/w15010030>

Emami, M., Ahmadi, A., Daccache, A., Nazif, S., Mousavi, S.-F., & Karami, H. (2022). County-level irrigation water demand estimation using Machine Learning: Case Study of California. *Water*, 14(12), 1937. <https://doi.org/10.3390/w14121937>

Fan, M., Shen, J., Yuan, L., Jiang, R., Chen, X., Davies, W. J., & Zhang, F. (2011). Improving crop productivity and resource use efficiency to ensure food security and environmental quality in China. *Journal of Experimental Botany*, 63(1), 13–24.
<https://doi.org/10.1093/jxb/err248>

Farrell, M., Gili, A., & Noellemeyer, E. (2018). Spectral indices from aerial images and their relationship with properties of a corn crop. *Precision Agriculture*, 19(6), 1127–1137.
<https://doi.org/10.1007/s11119-018-9570-9>

Gao, F., Hilker, T., Zhu, X., Anderson, M., Masek, J., Wang, P., & Yang, Y. (2015). Fusing Landsat and Modis data for vegetation monitoring. *IEEE Geoscience and Remote Sensing Magazine*, 3(3), 47–60. <https://doi.org/10.1109/mgrs.2015.2434351>

Gao, H., Zhang, X., Wang, X., & Zeng, Y. (2023a). Phenology-based remote sensing assessment of Crop Water Productivity. *Water*, 15(2), 329.
<https://doi.org/10.3390/w15020329>

Gebremedhin, M. A., Lubczynski, M. W., Maathuis, B. H. P., & Teka, D. (2022). Deriving potential evapotranspiration from satellite-based reference evapotranspiration, Upper

Tekeze Basin, northern Ethiopia. *Journal of Hydrology: Regional Studies*, 41, 101059. <https://doi.org/10.1016/j.ejrh.2022.101059>

Geng, L., Che, T., Ma, M., Tan, J., & Wang, H. (2021). Corn biomass estimation by integrating remote sensing and long-term observation data based on machine learning techniques. *Remote Sensing*, 13(12), 2352. <https://doi.org/10.3390/rs13122352>

Ghorbanpour, A. K., Kisekka, I., Afshar, A., Hessels, T., Taraghi, M., Hessari, B., Tourian, M. J., & Duan, Z. (2022). Crop water productivity mapping and benchmarking using remote sensing and Google Earth Engine Cloud Computing. *Remote Sensing*, 14(19), 4934. <https://doi.org/10.3390/rs14194934>

Gibson, L., Jarman, C., Su, Z., & Eckardt, F. (2013). Review: Estimating evapotranspiration using remote sensing and the Surface Energy Balance System – a South African perspective. *Water SA*, 39(4). <https://doi.org/10.4314/wsa.v39i4.5>

Gonçalves, I. Z., Ruhoff, A., Laipelt, L., Bispo, R. C., Hernandez, F. B. T., Neale, C. M. U., Teixeira, A. H. C., & Marin, F. R. (2022). Remote Sensing-based evapotranspiration modeling using geesebal for sugarcane irrigation management in Brazil. *Agricultural Water Management*, 274, 107965. <https://doi.org/10.1016/j.agwat.2022.107965>

Hassan, D., Abdalkadhum, A., Mohammed, R., & Shaban, A. (2022). Integration Remote Sensing and meteorological data to monitoring plant phenology and estimation crop coefficient and evapotranspiration. *Journal of Ecological Engineering*, 23(4), 325–335. <https://doi.org/10.12911/22998993/146267>

Hellegers, P. J., Soppe, R., Perry, C. J., & Bastiaanssen, W. G. (2008). Combining remote sensing and economic analysis to support decisions that affect water productivity. *Irrigation Science*, 27(3), 243–251. <https://doi.org/10.1007/s00271-008-0139-7>

Holzman, M. E., & Rivas, R. E. (2016). Early maize yield forecasting from remotely sensed temperature/vegetation index measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(1), 507–519. <https://doi.org/10.1109/jstars.2015.2504262>

Hommedi, A. H., & Almasraf, S. A. (2019). Water retention techniques under crop's root zone a tool to enhance water use efficiency and economic water productivity for Zucchini. *Journal of Engineering*, 25(6), 44–52. <https://doi.org/10.31026/j.eng.2019.06.04>

Huang, L. *et al.* (2022) 'A two-stage light-use efficiency model for improving gross primary production estimation in Agroecosystems', *Environmental Research Letters*, 17(10), p. 104021. <https://doi.org/10.1088/1748-9326/ac8b98>

International Committee of the Red Cross. (2022, December 5). *Kenya drought response: Facts & figures May - November 2022*. International Committee of the Red Cross. <https://www.icrc.org/en/document/kenya-drought-response-facts-figures-may-november-2022>

Islam, M. D., Di, L., Qamer, F. M., Shrestha, S., Guo, L., Lin, L., Mayer, T. J., & Phalke, A. R. (2023). Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning. *Remote Sensing*, 15(9), 2374. <https://doi.org/10.3390/rs15092374>

Jaafar, H. H., & Ahmad, F. A. (2015). Crop yield prediction from remotely sensed vegetation indices and primary productivity in arid and semi-arid lands. *International Journal of Remote Sensing*, 36(18), 4570–4589. <https://doi.org/10.1080/01431161.2015.1084434>

Jaafar, H. H., & Ahmad, F. A. (2020). Time series trends of landsat-based ET using automated calibration in metric and SEBAL: The Bekaa Valley, Lebanon. *Remote Sensing of Environment*, 238, 111034. <https://doi.org/10.1016/j.rse.2018.12.033>

Kamyab, A. D., Mokhtari, S., & Jafarinia, R. (2022). A comparative study in quantification of maize evapotranspiration for Iranian maize farm using Sebal and metric-1 EEFLUX algorithms. *Acta Geophysica*, 70(1), 319–332. <https://doi.org/10.1007/s11600-021-00704-4>

Khan, M. I., Liu, D., Fu, Q., & Faiz, M. A. (2017). Detecting the persistence of drying trends under changing climate conditions using four meteorological drought indices. *Meteorological Applications*, 25(2), 184–194. <https://doi.org/10.1002/met.1680>

Kisekka, I., Schlegel, A., Ma, L., Gowda, P. H., & Prasad, P. V. V. (2017). Optimizing preplant irrigation for maize under limited water in the High Plains. *Agricultural Water Management*, 187, 154–163. <https://doi.org/10.1016/j.agwat.2017.03.023>

Laipelt, L., Henrique Bloedow Kayser, R., Santos Fleischmann, A., Ruhoff, A., Bastiaanssen, W., Erickson, T. A., & Melton, F. (2021). Long-term monitoring of evapotranspiration using the SEBAL algorithm and Google Earth Engine Cloud Computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178, 81–96. <https://doi.org/10.1016/j.isprsjprs.2021.05.018>

Laonamsai, J., Julphunthong, P., Saprathet, T., Kimmany, B., Ganchanasuragit, T., Chomcheawchan, P., & Tomun, N. (2023). Utilizing NDWI, MNDWI, Savi, WRI, and Awei for estimating erosion and deposition in Ping River in Thailand. *Hydrology*, 10(3), 70. <https://doi.org/10.3390/hydrology10030070>

Li, H., Zheng, L., Lei, Y., Li, C., Liu, Z., & Zhang, S. (2008). Estimation of water consumption and crop water productivity of winter wheat in north China plain using Remote Sensing Technology. *Agricultural Water Management*, 95(11), 1271–1278. <https://doi.org/10.1016/j.agwat.2008.05.003>

Lim, C.-H., Kim, S., Choi, Y., Kafatos, M., & Lee, W.-K. (2017). Estimation of the virtual water content of main crops on the Korean peninsula using multiple regional climate models and evapotranspiration methods. *Sustainability*, 9(7), 1172. <https://doi.org/10.3390/su9071172>

Luo, Y., Guan, K., Peng, J., Wang, S., & Huang, Y. (2020). Stair 2.0: A generic and automatic algorithm to Fuse Modis, landsat, and sentinel-2 to generate 10 m, daily, and cloud-/gap-free surface reflectance product. *Remote Sensing*, 12(19), 3209. <https://doi.org/10.3390/rs12193209>

Mbayaki, C. W. (2021). Performance and water productivity of selected sweet potatoes (*Ipomoea Batatas* L) varieties intercropped with common beans in Katumani-Kenya. <https://doi.org/http://dx.doi.org/10.13140/RG.2.2.25977.60000>

Mohanasundaram, S., Kasiviswanathan, K. S., Purnanjali, C., Santikayasa, I. P., & Singh, S. (2022). Downscaling global gridded crop yield data products and crop water productivity mapping using remote sensing derived variables in the South Asia. *International Journal of Plant Production*, 17(1), 1–16. <https://doi.org/10.1007/s42106-022-00223-2>

Moriondo, M., Maselli, F. and Bindi, M. (2007) ‘A simple model of regional wheat yield based on NDVI data’, *European Journal of Agronomy*, 26(3), pp. 266–274. <https://doi.org/10.1016/j.eja.2006.10.007> .

Mostafa, M., Luo, W., Zou, J., & Salem, A. (2023). Optimizing rice irrigation strategies to maximize water productivity: A simulation study using AQUACROP model for the Yanyun Irrigation District, Yangzhou, China. *Earth*, 4(3), 445–460. <https://doi.org/10.3390/earth4030024>

Mountrakis, G., Im, J., & Ogole, C. (2011). Support Vector Machines in Remote Sensing: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001>

Muigai, D. K., Onwonga, R. N., Karuku, G. N., and Mohammed, A. Effect of irrigation

Muli. N., M., R. N., O., G. N., K., V. M., K., & M. O., N. (2015). Simulating soil moisture under different tillage practices, cropping systems and organic fertilizers using CropSyst model, in Matuu Division, Kenya. *Journal of Agricultural Science*, 7(2). <https://doi.org/10.5539/jas.v7n2p26>

Mulwa, F., Li, Z., & Fangninou, F. F. (2021). Water scarcity in Kenya: Current status, challenges and future solutions. *OALib*, 08(01), 1–15. <https://doi.org/10.4236/oalib.1107096>

PACHAC HUERTA, Y. C., & CHÁVARRI VELARDE, E. A. (2019). Spatial estimation of maize evapotranspiration (*Zea mays*) using the SEBAL algorithm in the valley of Huaylas, Huaraz -Peru, in 2016. *38th IAHR World Congress - “Water: Connecting the World.”* <https://doi.org/10.3850/38wc092019-0321>

Pandya, U., Mudaliar, A., & Gaikwad, A. (2023). Forecasting of banana crop productivity using geospatial approach: A case study of anand district. *ECWS-7 2023*. <https://doi.org/10.3390/ecws-7-14248>

Patel, A., Singh, P. K., & Tandon, S. (2021). Weather prediction using machine learning. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3836085>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pelosi, A., Belfiore, O. R., D’Urso, G., & Chirico, G. B. (2022). Assessing crop water requirement and yield by combining ERA5-land reanalysis data with CM-SAF satellite-based Radiation Data and sentinel-2 satellite imagery. *Remote Sensing*, 14(24), 6233. <https://doi.org/10.3390/rs14246233>

Prakash Mohan, M. M., Rajitha, K., & Varma, M. R. (2019). Integration of soil moisture as an auxiliary parameter for the anchor pixel selection process in Sebal using Landsat 8 and Sentinel - 1A Images. *International Journal of Remote Sensing*, 41(3), 1214–1231. <https://doi.org/10.1080/01431161.2019.1658239>

River County.

Ruhoff, A. L., Paz, A. R., Collischonn, W., Aragao, L. E. O. C., Rocha, H. R., & Malhi, Y. S. (2012). A Modis-based energy balance to estimate evapotranspiration for clear-sky days in Brazilian tropical savannas. *Remote Sensing*, 4(3), 703–725. <https://doi.org/10.3390/rs4030703>

Sadri, S., Famiglietti, J. S., Pan, M., Beck, H. E., Berg, A., & Wood, E. F. (2022). Farmcan: A physical, statistical, and machine learning model to forecast crop water deficit for farms. *Hydrology and Earth System Sciences*, 26(20), 5373–5390. <https://doi.org/10.5194/hess-26-5373-2022>

Saini, R., & Ghosh, S. K. (2018). Crop classification on single date sentinel-2 imagery using random forest and support Vector Machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII–5, 683–688. <https://doi.org/10.5194/isprs-archives-xlii-5-683-2018>

Sánchez, N., González-Zamora, Á., Piles, M., & Martínez-Fernández, J. (2016). A new soil moisture agricultural drought index (SMADI) integrating Modis and smos products: A case of study over the Iberian Peninsula. *Remote Sensing*, 8(4), 287. <https://doi.org/10.3390/rs8040287>

Sarshad, A., Talei, D., Torabi, M., Rafiei, F., & Nejatkhah, P. (2021). Morphological and biochemical responses of sorghum bicolor (L.) moench under drought stress. *SN Applied Sciences*, 3(1). <https://doi.org/10.1007/s42452-020-03977-4>

schedules on maize (*Zea mays* L.) growth and yield in Bura irrigation scheme, Tana

Scheierling, Susanne & Treguer, David. (2018). Beyond Crop per Drop: Assessing Agricultural Water Productivity and Efficiency in a Maturing Water Economy. <https://doi.org/10.1596/978-1-4648-1298-9>.

Sen, R., Zambreski, Z. T., & Sharda, V. (2023). Impact of spatial soil variability on rainfed maize yield in Kansas under a changing climate. *Agronomy*, 13(3), 906. <https://doi.org/10.3390/agronomy13030906>

Shan, Y., Li, G., Tan, S., Su, L., Sun, Y., Mu, W., & Wang, Q. (2023). Optimizing the maize irrigation strategy and yield prediction under future climate scenarios in the Yellow River Delta. *Agronomy*, 13(4), 960. <https://doi.org/10.3390/agronomy13040960>

Shoukat, M. R., Shafeeque, M., Sarwar, A., Mehmood, K., & Masud Cheema, M. J. (2021). Investigating effects of deficit irrigation levels and fertilizer rates on water use efficiency and productivity based on field observations and modeling approaches. *International Journal of Hydrology*, 5(5), 252–263. <https://doi.org/10.15406/ijh.2021.05.00287>

-
- Singh, P., Singh, A., & Kumar Upadhyay, R. (2021). A web based Google Earth engine approach for irrigation scheduling in Uttar Pradesh India using Crop Water Stress Index. *American Journal of Remote Sensing*, 9(1), 42. <https://doi.org/10.11648/j.ajrs.20210901.15>
- Souza, P. J., Silva, E. R., Silva, B. B., Ferreira, T. R., Sousa, D. de, Luz, D. B., Adami, M., Sousa, A. M., Nunes, H. G., Fernandes, G. S., Pinto, J. V., Farias, V. D., Oliveira, I. A., Silva, S. A., Costa, J. F., Rua, M. L., Costa, D. L., Moura, V. B., Lima, M. J., ... Ortega-Farias, S. (2023). Estimation of the evapotranspiration of irrigated açai (*Euterpe oleracea* M.), through the surface energy balance algorithm for land—SEBAL, in Eastern Amazonia. *Water*, 15(6), 1073. <https://doi.org/10.3390/w15061073>
- Spiliotopoulos, M., Alpanakis, N., Tziatzios, G. A., Faraslis, I., Sidiropoulos, P., Sakellariou, S., Karoutsos, G., Dalezios, N. R., & Dercas, N. (2023a). Estimation of remotely sensed actual evapotranspiration in water-limited Mediterranean agroecosystems for monitoring crop (cotton) water requirements. *ECWS-7 2023*. <https://doi.org/10.3390/ecws-7-14200>
- Talpur, Z., Zaidi, A. Z., Ahmed, S., Mengistu, T. D., Choi, S.-J., & Chung, I.-M. (2023). Estimation of crop water productivity using GIS and remote sensing techniques. *Sustainability*, 15(14), 11154. <https://doi.org/10.3390/su151411154>
- Tang, X., Li, H., Griffis, T., Xu, X., Ding, Z., & Liu, G. (2015). Tracking ecosystem water use efficiency of cropland by exclusive use of Modis Evi Data. *Remote Sensing*, 7(9), 11016–11035. <https://doi.org/10.3390/rs70911016>
- Tasumi, M., Allen, R. G., & Trezza, R. (2008). At-surface reflectance and albedo from satellite for operational calculation of Land Surface Energy Balance. *Journal of Hydrologic Engineering*, 13(2), 51–63. [https://doi.org/10.1061/\(asce\)1084-0699\(2008\)13:2\(51\)](https://doi.org/10.1061/(asce)1084-0699(2008)13:2(51))
- Thapa, B., Lovell, S., & Wilson, J. (2023). Remote Sensing and Machine learning applications for aboveground biomass estimation in agroforestry systems: A Review. *Agroforestry Systems*, 97(6), 1097–1111. <https://doi.org/10.1007/s10457-023-00850-2>
- United Nations. (2022). *Goal 6 | Department of Economic and Social Affairs*. United Nations. <https://sdgs.un.org/goals/goal6>

-
- Vergopolan, N., Xiong, S., Estes, L., Wanders, N., Chaney, N. W., Wood, E. F., Konar, M., Caylor, K., Beck, H. E., Gatti, N., Evans, T., & Sheffield, J. (2021). Field-scale soil moisture bridges the spatial-scale gap between drought monitoring and agricultural yields. *Hydrology and Earth System Sciences*, 25(4), 1827–1847. <https://doi.org/10.5194/hess-25-1827-2021>
- Virnodkar, S. S., Pachghare, V. K., Patil, V. C., & Jha, S. K. (2020). Remote Sensing and machine learning for crop water stress determination in various crops: A critical review. *Precision Agriculture*, 21(5), 1121–1155. <https://doi.org/10.1007/s11119-020-09711-9>
- Wang, Q.-Q., Geng, C.-X., Wang, L., Zheng, T.-T., Jiang, Q.-H., Yang, T., Liu, Y.-Q., & Wang, Z. (2023). Water conservation and ecological water requirement prediction of mining area in arid region based on RS-GIS and invest: A case study of bayan obo mine in Baotou, China. *Sustainability*, 15(5), 4238. <https://doi.org/10.3390/su15054238>
- Wang, Q.-Q., Geng, C.-X., Wang, L., Zheng, T.-T., Jiang, Q.-H., Yang, T., Liu, Y.-Q., & Wang, Z. (2023). Water conservation and ecological water requirement prediction of mining area in arid region based on RS-GIS and invest: A case study of bayan obo mine in Baotou, China. *Sustainability*, 15(5), 4238. <https://doi.org/10.3390/su15054238>
- Wang, Z., Shu, Y., Zhang, S., Li, H., & Lei, Y. (2009). Evaluating crop land productivity using modis derived time serious field greenness and water index in North China Plain. *SPIE Proceedings*. <https://doi.org/10.1117/12.830775>
- Wu, M., Huang, W., Niu, Z., Wang, C., Li, W., & Yu, B. (2018). Validation of Synthetic Daily Landsat NDVI time series data generated by the improved spatial and temporal data fusion approach. *Information Fusion*, 40, 34–44. <https://doi.org/10.1016/j.inffus.2017.06.005>
- Xu, L., Yu, H., Chen, Z., Du, W., Chen, N., & Huang, M. (2023). Hybrid deep learning and S2S model for improved sub-seasonal surface and root-zone soil moisture forecasting. *Remote Sensing*, 15(13), 3410. <https://doi.org/10.3390/rs15133410>

Yihun, Y. M., Haile, A. M., Schultz, B., & Erkossa, T. (2013). Crop water productivity of irrigated teff in a water stressed region. *Water Resources Management*, 27(8), 3115–3125. <https://doi.org/10.1007/s11269-013-0336-x>

Zhao, Y., Zhang, J., Bai, Y., Zhang, S., Yang, S., Henchiri, M., Seka, A. M., & Nanzad, L. (2022). Drought monitoring and performance evaluation based on machine learning fusion of multi-source remote sensing drought factors. *Remote Sensing*, 14(24), 6398. <https://doi.org/10.3390/rs14246398>

Appendix

1. Data

The validation data used in this project consists of weather data records provided by the Kenya Meteorological Department (KMD) and irrigation data from the Kenya Irrigation Authority. Due to data protection policies and security considerations, the complete datasets can only be accessed directly through requests submitted to KMD and the Irrigation Authority. However, presented below is a preview of the weather data records that informed the analysis undertaken in this project.

Table A1: Preview of Weather Dataset from KMD

Station Name	Year	Month	Tmax	Tmin	Rainfall
GARISSA METEOROLOGICAL STATION	2016	1		23.8	3.2
GARISSA METEOROLOGICAL STATION	2016	2	36.4	24.2	10.3
GARISSA METEOROLOGICAL STATION	2016	3	37.4	25.9	14.8
GARISSA METEOROLOGICAL STATION	2016	4	36.7	25.6	90.2
GARISSA METEOROLOGICAL STATION	2016	5		23.4	2.9
GARISSA METEOROLOGICAL STATION	2016	6	32.3	22.2	3.3
GARISSA METEOROLOGICAL STATION	2016	7	31.6	21.2	4
GARISSA METEOROLOGICAL STATION	2016	8	31.9	20.7	1.6
GARISSA METEOROLOGICAL STATION	2016	9		21.0	3.6
GARISSA METEOROLOGICAL STATION	2016	10	34.6	22.7	0.7
	...				
	...				
	Up to 2022				

2. Implementation Code and Models

The estimation of evapotranspiration in this project was adapted and inspired by the SEBAL algorithm available in the open-source [et-brazil](#) on the GitHub repository. This allowed leveraging proven code for efficient analysis using Google Earth Engine.

The core Python code detailing the workflows and logic I developed to implement this research is accessible via [Google Colab notebook](#) on request.

As the code spans several sections and is lengthy, it has not been appended to this report directly. However, researchers can conveniently review, run, and replicate the analysis using the Google Colab notebook. I will gladly share editable copies upon request.

Additionally, the optimized XGBoost machine learning model for crop yield prediction is available through [Google Drive link](#). The model accepts all Sentinel-2 bands and the following indices as input features: ['VHI', 'SMCI', 'EVI', 'GNDVI', 'SAVI', 'SR,' 'EDI,' 'TCI,' 'albedo,' 'LST']. This allows direct integration for prediction using new satellite imagery by researchers seeking to build on this work. Finally, all the datasets and code can be provided on request.