MAPPING AND MODELING COTTON CROP, ITS PRODUCTION AND FUTURE PATTERNS USING GIS AND MACHINE LEARNING

FINAL YEAR PROJECT UG 2019

By

(Saba Fatima - 283612)

(Muhammad Uzair - 283922)

(Abdul Rehman Rahimi - 322364)

(Laiba Asif - 295212)

Institute of Geographical Information System

School of Civil and Environmental Engineering

National University of Science and Technology, Islamabad, Pakistan

YEAR 2023

This is to certify that

Final Year Project Titled

MAPPING AND MODELING COTTON CROP, ITS PRODUCTION AND FUTURE PATTERNS USING GIS AND MACHINE LEARNING

submitted by

(Saba Fatima - 283612)

(Muhammad Uzair - 283922)

(Abdul Rehman Rahimi - 322364)

(Laiba Asif - 295212)

has been accepted towards the requirements for the undergraduate degree

in

(BE Geoinformatics)

Quratulain Shafi

Lecturer

Institute of Geographical Information System

School of Civil and Environmental Engineering

National University of Sciences and Technology,
Islamabad, Pakistan

FORM UGP-OBE-1

INSTITUTE OF GEOGRAPHICAL INFORMATION SYSTEMS (IGIS)
UNDERGRADUATE FINAL PROJECT
(Formulation of Project Group and Advisor)

Date: _____

Project Title:
Project Advisor:
Name: Mrs Quratulain Shafi
Dept: IGIS
Project Co - Advisor:
Name: Mr. Arif Goheer
Dept: GCISC
Project Members

1. **Name (Group Leader):** Saba Fatima **CGPA:** 3.84
   NUST Regn No: 283612 **Signature:_____**
2. **Name:** Muhammad Uzair **CGPA:** 3.41
   NUST Regn No: 283922 **Signature:_____**
3. **Name:** Abdul Rahman Rahimi **CGPA:** 3.05
   NUST Regn No: 322364 **Signature:_____**
4. **Name:** Laiba Asif **CGPA:** 2.98
   NUST Regn No: 295212 **Signature:_____**

Note: Group cannot be more than 4 students

_____ _____

Signature of Advisor **Signature of Head of Department**

APPROVAL

_____

Signature of Associate Dean

ABSTRACT

Agriculture is the biggest contributor to Pakistan’s GDP that makes around 22.67 percent to the total GDP. Cotton production alone accounts for 4.1 percent of Pakistan agriculture, 0.8% of GDP and for roughly 60% of Pakistan's international earnings. The fact that crop yield per hectare in Pakistan is less than its competitors call for methods like integration of GIS with Machine Learning to assist farmers and policymakers in making decisions for sustainable cotton growth. This study aims to integrate GIS with Remote Sensing data and Machine Learning algorithms to: (i) Map the cotton covered region. (ii) Predict cotton yield. (iii) Project the long-term impacts of climate factors (maximum & minimum temperature, and precipitation) on cotton yield. Sentinel 2A Imagery was imported in GEE to extract specific ranges of five vegetation indices in cotton field and apply these ranges to other time periods to delineate cotton covered region from all other. Then 11 vegetation indices and climate factors were used to model and estimate cotton yield. ALM, GLM, RF, GBT and SVM models were used to compare their results on the study area and provided data. Lastly, CMIP6 future projections of temperature and precipitations were used to correlate them with the vegetation indices and find out the pattern of crop yield from 2023 until 2099. The result not only showed the use of GIS, Remote Sensing and Machine Learning to map cotton fields, model cotton yield, but also emphasized the need for mitigation and adaptation for climate change to save cotton crops for better crop management practices.

CONTRIBUTION TOWARDS SDGs

This project aligns with multiple Sustainable Development Goals (SDGs) and contributes to their objectives. To begin, it intends to develop a robust method for predicting crop output by leveraging Remote Sensing (RS), Geographic Information System (GIS), and Machine Learning (ML) techniques in relation to SDG 8. Accurate crop yield forecast enables policymakers to apply policies and make educated economic decisions, promoting sustainable agricultural practices.

Furthermore, it targets SDG 12 by focusing on cotton-field identification and mapping using RS and GIS. This selection technique lays the path for additional research and exploration of sustainable cotton crop methodologies. Hence contribute to the larger goal of developing sustainable agriculture in Pakistan by encouraging sustainable practices in the cotton sector. It recognizes the significant influence of long-term climate change on cropping patterns, especially cotton production, in relation to SDG 13 which can provide policymakers with vital insights into the environment's influence on agriculture by linking crop productivity with critical climate parameters such as temperature, humidity, and precipitation. This understanding of climatic impacts enables the application of adaptive measures, allowing policymakers to devise plans to reduce the consequences of climate change and ensure the resilience of Pakistan's cotton industry and agricultural sector as a whole.

Overall, it contributes to the overarching sustainable development agenda by providing a methodology for yield prediction, encouraging research into sustainable cotton production, and assisting in climate-related decision-making for agricultural practices by addressing SDG 8, SDG 12, and SDG 13.

ACKNOWLEGMENTS

We would like to convey our profound gratitude to Miss Quratulain Shafi, our supervisor, for her vision, unfailing support, guidance, and encouragement throughout the entire process of this project. Their experience, insights, and constructive input have been crucial in shaping this work. Her consistent encouragement and motivation have encouraged us to strive to be the best and never give up, even during the most difficult circumstances.

We are also grateful to Mr. Arif Goheer and Mr. Sher Shah Hassan from Global Change Impact Studies Centre (GCISC) for contributing their time and expertise to provide insightful comments and recommendations, and that helped us enhance the final product of this project. They continually gave us motivation and encouragement and guide us towards our goals. Their contributions have been crucial in influencing the path of this research, for which we are thankful.

We are thankful to all the other teachers for sharing their knowledge and excellent advice. We are grateful to everyone who has contributed, directly or indirectly, to our endeavor.

Last but not least, our heartfelt thanks go to our adoring parents for their unending encouragement, support, and care.

TABLE OF CONTENT

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ALM	Automatic Linear Modeling
ANOVA	Analysis of Variance
AOI	Area of Interest
CMHyd	Common Model for Hydrologic Simulation
CMIP6	Coupled Model Intercomparison Project Phase 6
DVI	Difference Vegetation Index
ESA	European Space Agency
EVI	Enhanced Vegetation Index
GBT	Gradient Boosting Trees
GEE	Google Earth Engine
GLM	Generalized Linear Model
MAE	Mean Absolute Error
MBE	Mean Bias Error
NDVI	Normalized Difference Vegetation Index
NetCDF	Network Common Data Form
NIR	Near-Infrared
OSAVI	Optimized Soil-Adjusted Vegetation Index
RDVI	Renormalized Difference Vegetation Index
RF	Random Forest
RMSE	Root Mean Squared Error
RVI	Ratio Vegetation Index
SARVI	Soil-Adjusted Ratio Vegetation Index
SAVI	Soil-Adjusted Vegetation Index
SPSS	Statistical Package for the Social Sciences
SSP	Shared Socioeconomic Pathway
STVI 01	Stress Related Vegetation Index
SVM	Support Vector Machine
TVI	Transformed Vegetation Index
USGS	United States Geological Survey
WDRVI	Wide Dynamic Range Vegetation Index

CHAPTER 1

INTRODUCTION

1.1 Background Information

Agriculture plays a vital role in helping life survive on earth, as it is the primary source of food for humans and animals. Other than food, it also supports many industries, such as textiles and biofuels, and offers employment opportunities for millions of people worldwide, hence becoming crucial for the economy. Since Pakistan is an agricultural state with development potential, agricultural advances are more important to its economy than any other industry. Agriculture is important because it provides food for people, raw materials for numerous businesses, and serves as a foundation for foreign exchange. Pakistan was ranked fifth in the world for cotton production (ICAC, 2021). 0.8% of GDP and 4.1% of the value added in agricultural GDP are accounted for by cotton production. Cotton provides the unprocessed supplies for textile commerce, the biggest agro-industrial segment of country's economy, retains 17% of the workforce and generates 60% of the country's foreign exchange (IFPRI, 2022). Cotton agriculture along the Indus River irrigation structure spans almost 3 million hectares which is crucial for the country's economic strength. Pakistan's cotton belt stretches about 1200 kilometers along the Indus River, between latitudes of 27° N and 35° N and heights of 27 m to 155 m. The soil transitions from clay loam to sandy, with clay dominating to the south (ADB, 2008). Cotton is a kind of crop that is highly sensitive to the higher temperature range in semi-arid regions and shows a similar kind of trend when it comes to rainfall (Centin et al., 2010). Cotton farming spans around 2.79 million hectares in total. Upland cotton is mostly farmed in two provinces of Pakistan: Sindh and Punjab. The Punjab province has the greatest explicit cotton agriculture area, but Sindh is also well recognized for cotton cultivation. Cotton yield in Sindh province is 855 kilograms per hectare, whereas cotton production in Punjab province is 695 kg/ha on average (Pakistan Bureau of Statistics, 2020). When compared to global cotton production, these two figures are underwatered in areas where cotton

production is low. Normally, the maximum temperature for cotton growth is 28.5° C to 35°C (Singh et al., 2007), but in Pakistan, the temperature breadths to 35° C in summers when cotton is being grown, and as summers proceed it can go to 41° C to 47°C, sometimes even 50°C, which is too high for human and animal survival. Many other issues affecting cotton yield and production in Pakistan, such as heat stress and high input prices, are also affecting cotton yield and production. Cotton Leaf Curl Virus infections CLCV illness, a lack of water availability, and seed adulteration are all major issues. Crop insurance and a cotton marketing challenges system are also key issues affecting cotton output.

1.2 GIS and RS in Agriculture

Agriculture is a significant trading industry for a country with a strong economy. Agricultural proficiency can be successfully improved by applying information technology tools such as Geographical Information System (GIS) and Remote Sensing. Food production at a low cost is the prime goal of all cultivators, large-scale farm management, and regional agricultural agency (Priya & Shibasaki, 2001). The application of remote sensing and GIS to assess and display agricultural terrain has proven to be extremely beneficial to farmers and industry. It helps farmers increase production, cut costs, and manage their land more effectively, having a big impact on agriculture all across the world. New methods for processing and utilizing geographical information for evaluation, planning, and monitoring have emerged as a result of advances in computer technology. Precision Agriculture, also known as Precision Farming, allows for the automation and simplification of data collection and analysis using GIS and other technology (Hazarika et al., 2001).

Geospatial Imagery and GIS are critical for understanding crop health, pest extent, potential yield, and soil conditions (Halder et al., 2013). It is used to look into agricultural applications like crop identification, area estimation, crop condition assessment, yield estimation, farm water management, and agrometeorological forecasting, among others. GIS-based mapping tools can help with monitoring crop health, locating crops that are growing across the nation, adjusting various variables, calculating yields from a specific farm, and increasing crop production.

1.3 Sentinel 2A

Sentinel 2A satellite was launched by the European Space Agency late in 2014. It is an essential tool for remote sensing in agriculture due to its high spatial and temporal resolutions. Sentinel 2A operates in visible, near infrared, and shortwave infrared regions of the electromagnetic spectrum which enables identification of a wide range of crop types, including cotton crops. Its high spatial and temporal resolutions with a pixel size of 10X10 meters makes it suitable for the detection of small-scale features such as individual crops. Moreover, its frequent revisits every five days enable the monitoring of crop growth and development over time.

1.4 Google Earth Engine

Google Earth Engine (GEE) is a free online cloud-based platform from Google that grants users access to an immense archive of satellite images from sources such as Sentinel, MODIS, Landsat, and others, as well as geospatial and metrological datasets. GEE possesses a wide range of tools for analyzing, processing, demonstrating, and manipulating both geospatial and non-geospatial data. With GEE, users can easily retrieve satellite imagery, meteorological and geospatial datasets and perform complex analyses and generate dynamic visualizations without downloading data or purchasing licenses. In addition to its vast archive of satellite imagery and other datasets, Google Earth Engine (GEE) provides strong processing capabilities that allow users to execute complicated analyses on massive datasets. Users may utilize GEE's cloud-based architecture and Google's servers to perform extensive computations on massive geospatial datasets, such as machine learning algorithms and statistical analysis. This enables researchers and analysts to easily examine and understand enormous volumes of data, generating insights and predictions that would otherwise be difficult or impossible to achieve with traditional desktop-based software.

1.5 Cotton Cycle

Cotton plants follow an interesting pattern throughout their cycle. From sowing to its harvesting, there are different crucial stages that come along, and each stage is closely monitored by the agronomists and farmers to get insights into how cotton plants are behaving this season. Cotton is generally grown in summers along the equator region. But its cycle varies from one region to another. The cotton cycle depends on different factors, weather conditions being one of them. The season generally starts in April-May and ends in September-November. This must be noticed that these patterns change over time thanks to the change in the climatic conditions (AARI, 2021).

In Pakistan, first stage is apparently Sowing or Plantation stage when the cotton seeds are planted into the ground. In Southern Punjab, the sowing period starts from mid-April until mid-May. Some varieties of cotton are sown earlier than others to account for their unique features. Then it takes about 35 days to move towards the next stage that is known as Emergence. At this critical stage, the cotton seeds start to emerge out of ground after germination. This means that now the photosynthesis and active growth of cotton plants will start.

After this stage, a square starts to appear on the cotton plant. Square is the fruiting bud that forms on the branches of cotton plant branches (Cotton Plant Development and Plant Mapping, n.d.). There are different kinds of square, first one is called First Square which includes the initial fruiting buds that appear on the plant, then comes Pinhead Squares where new squares can be classified, and then the Match Head Square which is subsequent stage of Pinhead Square. What follows next is the Flowering stage. It takes about 66-70 days after the sowing for flowers to sprout. Flowering is the blooming period of this crop. When squares are matured enough, they translate into little flowers and then pollination occurs. Then ovules develop into cotton fruits or Bolls which is next stage after Flowering, and it takes 20 more days.

Boll goes through three phases to reach the stage of maturation. The first step is enlargement where boll grows in size, followed by filling, and then maturation. Once the boll is fully mature, it is then harvested almost 163-170 days after the very first stage. After

harvesting, another step is critical raw cotton consisting of seeds and leaf trash are packed into modules, and then cotton fibers are separated from cotton seeds.

1.6 Vegetation Indices

Based on extensive literature review and previous research studies, this study opted 11 different vegetation indices that have been used for similar goals before. All these vegetation indices have been made to pick on particular features of vegetation areas and this combination was thought to be helpful for this study. Some of them are related to one and other in some way but their slight differences can help in better forecasting of crop yields. Out of these 11 bands, Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) comes along with the MOD13A1 product as stated earlier. They were used given their enormous utilities in agriculture related studies. (Morelli-Ferreira, 2021)

Table 01 lists the formulas and details of all indices used throughout this study. Difference Vegetation Index or DVI was calculated by subtracting near the red band reflectance from infrared band reflectance. It provides a measure of health and density of vegetation in a region that is less responsive to atmospheric interruptions. Then comes Ratio Vegetation Index or RVI is like DVI except it takes ration of NIR and Red bands. As efforts to improve indices results for different conditions increased, modified versions erupted to fill the gaps for different types of conditions. Soil Adjusted Vegetation Index or SAVI was developed to improve the NDVI performance by reducing the impacts of soil brightness. It includes a soil adjustment factor to account for the soil brightness. It measures the vegetation density with more accuracy when there is high soil brightness. But a difficulty occurred when soil background effect was adjusted for NDVI, the atmospheric variations got higher. This called for the making of a new and improved version of index called Soil Adjusted and Atmospherically Resistant Vegetation Index or SARVI. (Leprieur et al., 2000).

Similarly, Optimized Soil Adjusted Vegetation Index or OSAVI is adjusted version of SAVI to perform better in areas with sparse vegetation cover. Randomized Difference Vegetation Index or RDVI is an improved version of Difference Vegetation Index to give

off more accurate results in the areas with high levels of noise. Transformed Vegetation Index or TVI was another shot at improving the performance of NDVI in regards with vegetation density and health. TVI establishes normal distributions to improve results. Wide Dynamic Range Vegetation Index or WDRVI includes a weighting coefficient of 0.1-0.2 and increase association with vegetation section to improve performance of NDVI. Stress related Vegetation Index STVI-1 has shown a better relationship with vegetation cover and has proven useful for vegetation mapping. Online Index Database "Index Database. (n.d.)." can be searched to find other relevant information about any kind of index. Table 1 shows the vegetation indices included for this part of the project.

Table 1: Derived Vegetation Indices.

Index	Formula	Full Name	References
DVI	$\rho_{NIR} - \rho_{R}$	Difference Vegetation Index	(Wu, 2014)
RVI	P_{NIR} / ρ_{R}	Ratio Vegetation Index	(Wu, 2014)
NDVI	$(\rho_{NIR} - \rho_{R}) / (\rho_{NIR} + \rho_{R})$	Normalized Difference Vegetation Index	(Huete et al., 2010)
EVI	$2.5 * (\rho_{NIR} * \rho_{R}) / (\rho_{NIR} + C1 * \rho_{R} - C2 * \rho_{B} + L)$ $L = 1; C1 = 6, C2 = 7.5$	Enhanced Vegetation Index	(Ihuoma et al., 2019 & Wu, 2014)
SAVI	$(1+L) (\rho_{NIR} - \rho_{R}) / (\rho_{NIR} + \rho_{R} + L)$ $L = 0.5$	Soil Adjusted Vegetation Index	(Panda et al., 2010)
OSAVI	$(\rho_{NIR} - \rho_{R}) / (\rho_{NIR} - \rho_{R} + 0.16)$	Optimized Soil Adjusted Vegetation Index	(Steven, 1998 & Rondeaux et al., 1996)
RDVI	$(\rho_{NIR} - \rho_{R}) / (\rho_{NIR} + \rho_{R})^{1/2}$	Randomized Difference Vegetation Index	(Ihuoma et al., 2019)
SARVI	$(1 + L) (\rho_{NIR} - \rho_{RB}) / (\rho_{NIR} + \rho_{RB} + L)$ $\rho_{RB} = \rho_{R} - g * (\rho_{R} - \rho_{B})$ $\gamma = 1, \Lambda = 0.5$	Soil Adjusted and Atmospherically Resistant Vegetation Index	(Wu, 2014)
TVI	$(\rho_{NDVI} + 0.5)^{1/2}$	Transformed Vegetation Index	(Bannari et al., n.d.)
WDRVI	$(0.1 * \rho_{NIR} - \rho_{R}) / (0.1 * \rho_{NIR} + \rho_{R})$	Wide Dynamic Range Vegetation Index	(Gitelson, 2004)
STVI01	$(\rho_{MIR} * \rho_{R}) / \rho_{NIR}$	Stress related Vegetation Index	(Jafari et al., 2007)

Where R = Reflectance of Red Band, NIR = Reflectance of Near Infra-Red Band,
MIR = Reflectance of Middle Infra-Red Band

1.7 Objectives

This study was focused on mapping the cotton crop area, modelling the crop yield, and analyzing the future trends of cotton yield. Hence this study area is divided into three objectives:

(i) Identify cotton-fields using Remote Sensing and classification algorithms.

(ii) Develop a Machine Learning Model to predict cotton yield.

(iii) Analyze the impacts of climatic factors (temperature range and precipitation) on cotton growth in future.

1.8 Beneficiaries

- Cotton covered area identification from all agricultural land present in selected area with the help of remote sensing.
- Cotton production prediction to assist farmers and agronomists in making better management decisions.
- Assessment of the effects of rainfall, temperature, and humidity on cotton growth will result in improved crop management, better prediction of future trends, increased resilience to climate change and establishment of baseline for more accurate future research.

CHAPTER 2

LITERATURE REVIEW

Feng et al. (2019) discusses the significance of remote sensing data for crop classification, which is required for large-scale agricultural remote sensing monitoring, agricultural monitoring types, and government decision-making. There are two key approaches used for crop classification using Remote Sensing: one uses spectral features of high spatial resolution data combined with multitemporal features, and the other studies crop growth patterns and phenological characteristics. The article also discusses crop classification using Landsat, MODIS, and Sentinel data, as well as the benefits of using machine learning models like SVMs and random forests for crop multiclassification problems in comparison to the maximum likelihood model for classifying crops in the city of Yushu, China using Sentinel-2A images. Sentinel-2A images from the year 2017 were successfully used to extract spectral reflectance of 12 bands, 96 texture parameters, 7 vegetation indices, and 11 phenological parameters.

The following are the primary outcomes of this paper:

I. The results show the combination of 13 features results in 88.96% accuracy for traditional classification and 98% for machine learning classification.
II. The shortwave infrared band has a significant effect on classifying rice, corn, and soybean.
III. The water vapor band distinguishes corn and rice.
IV. GCVI assists in discriminating corn and soybean, while coastal band does so for other crops from dry fields.
V. Rice identification is easier than maize and soybean identification, and machine learning methods perform significantly better than traditional classification procedures for recognition of multifeatured crops.

It is also mentioned that the amount and accuracy of information extracted from Sentinel-2A data can be affected by factors like atmospheric factors. Furthermore, the number of

available data periods influences the ability to extract phenological features from time series data. However, to fully benefit from the red edge band's high temporal and spatial resolution and additional information, methods for de-clouding and denoising the data must be developed.

Bargiel, 2017 discusses the significance of accurate and complete crop type classification for assessing the effect of agrarian landuse on ecosystems. The study presents a modern multitemporal based categorization method that uses information about phenological shifts in crop lands to identify crop type phenological sequence patterns (PSP). The PSP approach's performance was evaluated over two vegetation terms using Sentinel-1 data and over 200 ground truth fields in northern Germany. The outcomes favored PSP over standard order strategies for meadows, maize, canola, sugar beets, and potatoes, with the PSP approach beating standard grouping techniques for grain harvests like spring grain, oat, winter grain, and rye. The PSP approach is likewise stronger to contrasts in cultivating the board and development conditions, as well as more delicate to unobtrusive changes, like weed extents inside a field. The strategy is reasonable for huge scope arrangement and can be assessed further with different multitemporal input information, for example, polarimetric highlights, optical sensor information, or imaging radar information at different frequencies.

As indicated by research by Kwak et al. (2019), automated airborne vehicle (UAV) pictures can possibly be utilized in crop order in light of their high spatial and temporal goal. The utilization of GLCM-based surface data for crop distinction with time-series UAV pictures and machine learning algorithms calculations is examined in this exploration paper. For cases with at least one UAV picture as information, the effect of consolidating surface and otherworldly data on grouping execution is assessed. The impact of surface data on precision was determined by the GLCM. The extraction of GLCM-based surface elements requires cautious kernel size determination. The review found that multi-transient UAV pictures joined with GLCM-based surface elements accomplished the most precision, while surface data further developed proficient execution for a single August UAV picture. These discoveries infer that surface data can be helpful for crop order when just a set number of UAV pictures are accessible. Be that as it may, while utilizing multi-fleeting

pictures, the blend of surface highlights and unearthly data didn't fundamentally further develop better accuracy.

Mapping and monitoring cropland and crop type distribution is critical for assisting policymakers and international organizations in mitigating food security risks, particularly those caused by climate change as mentioned in the article by Tariq et al. (2022). Remote sensing has grown in popularity as a tool for these purposes. However, because of the spectral similarity of crop types and cropping patterns, it is difficult to identify specific types and patterns using satellite data.

In Pakistan's Gujranwala District, researchers looked for crop types like tobacco, wheat, barley, and gramme, and patterns like wheat-tobacco, wheat-gram, wheat-barley, and wheat-maize. Sentinel-2 and Landsat-8 data was combined with Machine Learning algorithms, including Decision Tree Classifier and Random Forest, for the study.

The study used machine learning algorithms to link NDVI-based time-series from Sentinel-2 and Landsat 8 with phenological parameters to identify most suitable time-periods for distinguishing cropland from other landuse. Landsat with crop data from 2020 and 2021, and ground data on patterns were used to evaluate the methodology. The study also tested temporal changes in cropping patterns and types, as well as a comparison of the spatial and temporal resolution of medium-resolution imagery.

Following results were derived using 184 crop samples:

I. The landscape configuration influenced cropland mapping accuracy.
II. R2 values were high in 2021 at the sub-district level between crop-statistical and Sentinel-2 derived cropland data.
III. Sentinel-2-derived zones for various crop categories matched well with crop-statistical data.
IV. Using Sentinel-2 derived data, the paper extracted the cropland on a large scale with high accuracy. These results vouched for accurate and precise results that can be considered at lower costs than other methods.
V. Among all crop types, tobacco had the best estimation results.

VI. Wheat was the most widely cultivated crop in study region, covering 85% of total agricultural land, while barley had the least, covering only 0.07%.

In the end, the importance of large-area studies that cover wide ecological gradients to reveal the benefits and drawbacks of using optical and radar data for crop type mapping in a variety of ecological conditions and data sources were highlighted.

Fang et al. (2020) discusses the importance of wheat production in Henan Province, China, and the use of remote sensing technology for crop mapping. According to the article, remote sensing technology has been used for crop biomass, leaf area index, and yield mapping, as well as crop identification using machine learning algorithms such as SVMs, RF and NNs. However, the crop identification process is time-consuming and inefficient. Traditional remote sensing methods' limitations, such as low- and medium-resolution images and susceptibility to mixed pixels, are also highlighted. Then the Google Earth Engine (GEE) platform as a solution to these constraints is introduced, offering efficient computing capabilities as well as access to public geospatial datasets such as Landsat, MODIS, and Sentinel. GEE includes a number of machine learning algorithms that can be used for vegetation monitoring, land use/cover analysis, water change analysis, and drought analysis. The paper presents the use of machine learning algorithms, specifically support vector machine (SVM), random forest (RF), and classification and regression tree (CART), to identify and map winter wheat using Sentinel-2 images. The algorithms' hyperparameters were tuned using grid search and cross-validation, and their classification performances were compared. Finally, the effect of MODIS mixed pixel with medium resolution on crop mapping accuracy is investigated.

With an overall accuracy (OA) of 0.95, a user accuracy (UA) of 0.95, a producer accuracy (PA) of 0.95, and a kappa coefficient of 0.92, the SVM algorithm performed the best in terms of classification. Investigations into the sensitivity of the algorithms to the hyperparameters revealed that SVM was more sensitive to C and gamma, RF was less susceptible to tree and split, and CART was more sensitive to maxD and minSP. According to the study, SVM and RF perform better at classifying data than CART. While RF is not sensitive to algorithm settings, SVM and CART are. The study's recommendations for

further research include looking into other machines and deep learning algorithms for this purpose.

The approach for county-scale cotton mapping proposed in this research uses a random forest (RF) feature selection algorithm and classifier to choose multi-features like spectral, vegetation indices, and texture characteristics. SVM, ANN, and RF were the three machine learning methods used by Fei et al. (2022) to classify cotton using spectral characteristics, vegetation index, and texture features. The random forest technique was used to choose these features, and the classification outcomes were examined and contrasted based on the image date, feature selection, and classifier choice. The study found that adding more features can considerably increase classification accuracy, with RF exhibiting the most stability and effectiveness. Along with spectral data and the vegetation index, the study assesses the impact of texture features on cotton categorization accuracy. The feature-based classification improves on the pixel-based classification by including texture features. Characteristics of crops at various times taken into account. The study also determined the importance of various features in classification, with NIR ranking first among spectral features and GLCM ranking first among texture features.

The findings demonstrated:

1. The grey level co-occurrence matrix (GLCM) texture feature, which ranked second in contribution among all examined spectral, VI, and texture features, is useful for enhancing classification accuracy.

II. The RF classifier surpassed SVM and ANN in terms of accuracy and stability.

III. The average OA of the classification incorporating multiple features was 93.36%, which was 7.33% higher than the average OA of the single-time spectrum and 2.05% higher than the average OA of the multi-time spectrum.

IV. The classification accuracy can be 92.12% after feature selection using RF, indicating outstanding accuracy and efficiency.

This technique has the potential to be an effective county-scale method for classifying cotton. A method reference for precision cotton management at the county level was also offered by the study.

To better understand the techniques and features utilized in crop production prediction studies, Klompenburg et al. (2020) conducted a thorough literature evaluation of 567 pertinent articles. After reviewing 50 papers, it was found that the most frequently utilized features were temperature, rainfall, and soil type, with Artificial Neural Networks (ANN) being the most frequently employed algorithm in these models. The study found that the scope of the study and the availability of data influence the choice of features. Not usually did models with more features outperform those with fewer characteristics. Random forest, neural networks, linear regression, and gradient boosting trees were the most frequently employed models. Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), and Deep Neural Networks (DNN) were found to be the most often utilized deep learning algorithms after a second search to find papers based on deep learning. These results have significant ramifications for future agricultural yield prediction studies based on machine learning.

The article discusses the importance of early crop yield estimation in countries like Pakistan, where agriculture is a major source of income. The study looks into the feasibility of using MODIS-derived vegetation indices and remote sensing to predict wheat yield in Pakistan's Potohar region. It also shows how multiple linear regression (MLR) models can be used in agricultural decision support, specifically yield forecasting. The study develops a statistical model using two MODIS products, MOD15A2H and MOD13A1, and wheat yield data from each district in Pakistan from 2009 to 2018. The results show that using geospatial techniques in conjunction with the statistical modelling approach, accurate wheat yield prediction can be made almost 2 months before harvesting, with an average difference of -1.986% between the actual and predicted yield (Hassan et al., 2020).

The article does, however, acknowledge the MLR model's limitations, which include the potential overestimation of yield due to environmental factors that affect the crop after the forecasting date.

The capacity of conventional techniques based on vegetation indices (VI) to evaluate environmental stress-related disorders that cannot be assessed using Vegetation Indices is constrained. These techniques can estimate agricultural yields. This paper suggests a new approach that uses a random forest regression algorithm and extra environmental factors

to estimate US corn and soybean yields. To increase the precision of a vegetation index-based agricultural production estimation approach, the study used additional environmental factors and a random forest regression machine learning algorithm. According to the study, the RF approach delivers the most precise estimations, particularly in irrigated areas. The study supports the RF method's ability to forecast crop loss due to drought accurately (Sakamoto, 2019).

The study quantified drought episodes from 2000 to 2018 utilizing indices such as STVI 01, NDVI, EVI, and SAVI, as well as climatic data, with the goal of assisting decision-making for drought monitoring and yield prediction. The research found a significant inverse relationship between wheat yield and temperature and a significant inverse relationship between wheat production and rainfall. The findings indicate that throughout this time, the region had comparatively higher winter mean temperatures and significant seasonal rainfall changes, which led to consistently low soil moisture and frequent drought occurrences. Vegetation indices identified two more drought events, and STVI 01 identified three moderate and two light drought events. The study also found that, after temperature and rainfall, soil moisture had the most effect on wheat yield (Ijaz et al., 2021).

Azmat et al. (2021) aimed to utilize two crop models and regional climate models to assess the effects of climate change on winter wheat in Pakistan's rainfed and irrigated regions. Different adaptation strategies were investigated, and experimental data for wheat phenology, biomass, and yield were obtained. Climate change had a significant impact on wheat phenology, biomass, and yield, with stronger impacts under RCP8.5. Wheat biomass and yield were improved by adaptation strategies, particularly sowing-2 under RCP4.5 in rainfed regions and irrigation-2 and the combination of sowing-1 + irrigation-2 in irrigated regions.

STICS and APSIM crop models are used to simulate the effects of climate change and different adaptation strategies. According to the study, temperature and precipitation changes will have a significant impact on wheat growth and yield, with the APSIM model showing a more significant response to temperature changes. According to the STICS model, the sowing-2 adaptation strategy in rainfed regions and the combination of S1 and I2 strategies in irrigated regions can provide the best yield recovery. The study provides

policymakers and researchers with a credible range of potential outcomes regarding the potential impacts of climate change and adaptation strategies on wheat yield in South Asia in general, and Pakistan in particular.

The article by Arunrat et al. (2021) aims to assess the impact of climate change on crop yields and water footprint (WF) in Thailand's lower north. The study employs five global circulation models to forecast crop yields and WF changes in the future under Shared Socioeconomic Pathways (SSPs) scenarios.

i. Precipitation, maximum and lowest temperatures, and all time periods were predicted to increase under the SSP245 and SSP585 scenarios.
ii. Under the RRR cropping method, it was anticipated that rice yields for all three crops would increase progressively.
iii. According to the SSP585 scenario, the yields of the three rice crops would only modestly grow in the near future while decreasing in the mid- and long-term.
iv. The first and second rice crop yields under the RR cropping system (in the rain-fed area) were both decreased by the SSP585 scenario (6.0-14.4% and 7.4-17.7%, respectively).
v. The first and second rice crop yields improved by 3.0% and 4.3%, respectively, in the near future, according to the SSP245 scenario.
vi. Future climate change had less impact on the production of maize, soybeans, and mung beans as opposed to a second rice crop, especially because the yield of mung beans was expected to modestly rise in all time periods under the SSP245 and SSP585 scenarios.

Future WF variations were connected to future crop production variations; hence, a decline in WFs was brought on by the anticipated rise in crop yield, and vice versa.

CHAPTER 3

MATERIAL & METHODOLOGY

3.1 Study Area

The focal point of study was Dera Ghazi Khan. It contains four districts named Dera Ghazi Khan, Layyah, Rajanpur, and Muzaffargarh with geographic extent of 30◦02’56” N, 70◦38’43” E. It is a flat and agricultural area and the most prominent cotton producing region after Multan (Lodhran, Khanewal) ang Bahawalpur (RYK). Figure 1 shows the map of Dera Ghazi Khan with all districts. The city's overall climate is dry, with minimal rain. The winters are pleasant and dry, while the summers are extremely scorching. The mean high temperature in the summer is around 107 °F (42 °C), while the average low temperature in the winter is 40 °F (4 °C). Summer temperatures in Pakistan are typically among the hottest.

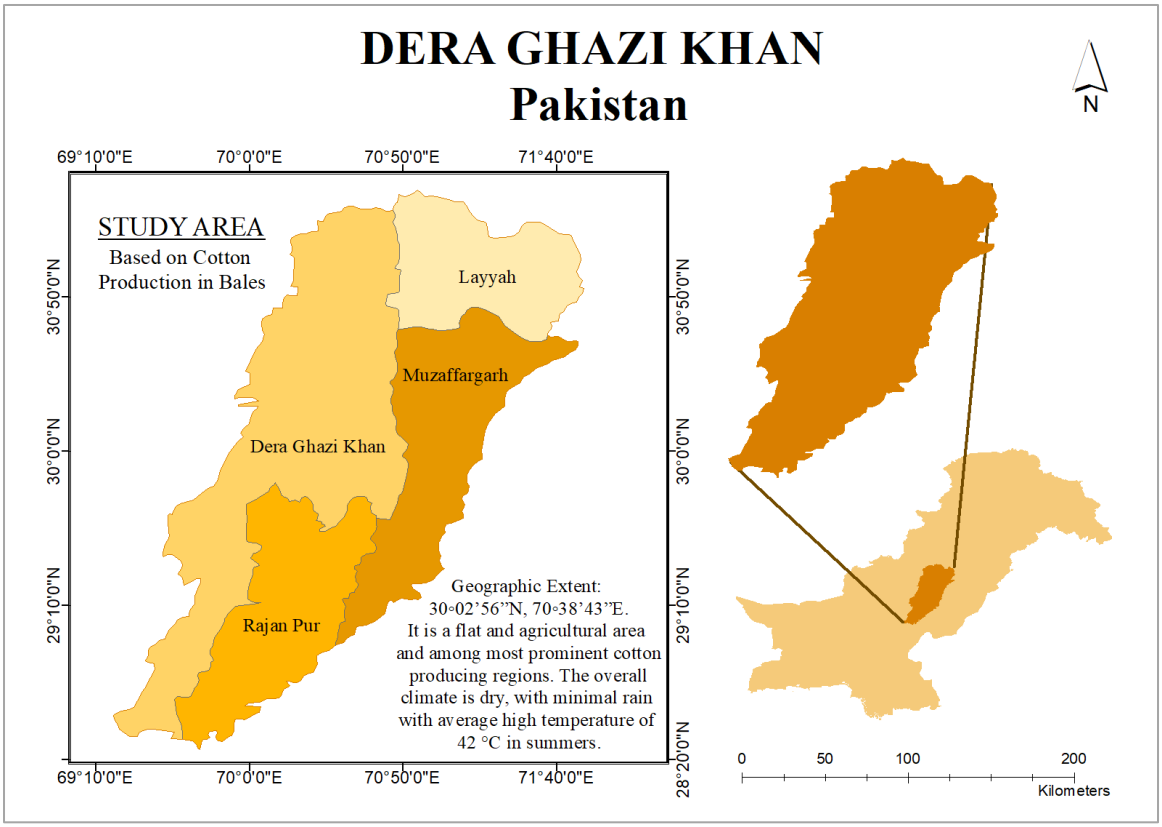


Figure 1: Map of study area

3.2 Data Acquisition and Preprocessing

3.2.1 Sentinel 2A Imagery

Analyzing satellite images to identify specific crops like cotton can still be a challenging task as it requires advanced data processing and extensive analysis techniques. Therefore, the development of an accurate and reliable method to detect and classify cotton crops in Sentinel 2A imagery using different remote sensing techniques and various programming platforms have the potential to revolutionize the way cotton crops are mapped. Accurate mapping of cotton crops using RS techniques can help in predicting crop yield, monitoring crop health, and providing early warnings of crop stress which ultimately lead to increased productivity and improved food security. (Moumni et al., 2021)

In order to extract cotton mask, the Sentinel 2A images were extracted through Google Earth Engine from ESA archive for the months of cotton cycle for one year to analyze the

images later and apply the classification algorithm and cotton mask extraction. Once the Sentinel-2 image from the COPERNICUS/S2_HARMONIZED dataset was selected using ee.ImageCollection function, it was filtered by geographic bounds and date range, sorted by cloud cover, and the first image with least cloud percentage was selected for further analysis. After that, the processed images were created by mosaicking ten separate images together and then clipping the resulting mosaic to a specific area of interest (AOI). The final image is displayed as a three-band composite using the Red, Green, and Blue (RGB) bands ('B4', 'B3', 'B2') with a specified minimum and maximum value range (min: 0, max:3000) and gamma correction (gamma: 1.4). These steps were used to preprocess other various remote sensing to analyze the required combining multiple images into a single image for further analysis or visualization. The code for the above image processing and extraction is provided in Appendix A (Data extraction and mosaicking).

3.2.2 MOD13A1 Imagery

Given the use of MODIS or Moderate Resolution Image Spectroradiometer in previous research studies regarding crop yield forecasts and other crop related studies, it was considered a suitable choice. Also, MODIS has a high radiometric sensitivity (12-bit) with 36 spectral bands having wavelength extending from 0.4 µm to 14.4 µm. MOD13A1 product of Terra MODIS was selected for its wide use in agricultural and environmental studies. It has per pixel based spatial resolution of 500 meters with a temporal resolution of just 16 days. This enables users who need to monitor changes of a region with short intervals. This frequency of data was a prominent factor for choosing MODIS product over other options. Data from the year 2008 until 2022 was readily available.

MOD13A1 provides a Quality Assurance band as well, and the algorithm selects a pixel that gives off the best result in a 16-day period. Other bands used in this study were Red Reflectance band or Band 01, Near Infrared Reflectance band or Band 02, Blue Reflectance band or Band 03, and Mid Infrared Reflectance band or Band 04. Red reflectance band is sensitive to chlorophyll absorption and is used to estimate vegetation density. Near Infrared band can also be used for vegetation health and density, but it is more sensitive to the moisture content of vegetation. While Blue reflectance band mainly focus of atmospheric aerosols, the Mid Infrared band is sensitive to moisture content and atmospheric temperature and is also used to study vegetation stress (Myneni et al., 2007).

MOD13A1 data is uploaded to Land Processes Distributed Active Archive Center and this data is Level 03 gridded product. Using the code attached in Appendix A (MODIS imagery acquisition and mosaicking). MOD13A1 data has already undergone atmospheric correction, geometric correction, and calibration. This means generally there is no need to do further preprocessing, unless required for the task. MOD13A1 provides two primary vegetation layers as well, named Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) which are handy for any study associated with biophysical parameters. NDVI is a commonly used measure which is calculated using the ratio of near infrared reflectance and red band's reflection. It is used to monitor vegetation dynamics, as its value ranges from -1 to +1. Value closer to +1 indicates denser vegetation. EVI is another vegetation index which is calculated using near infrared, red and blue bands

of electromagnetic spectrum. It also indicates vegetation dynamics but does it better than NDVI for regions that have higher density of atmospheric aerosols as it has better sensitivity for biomass.

3.2.3 Meteorological Data

It was clear from the previous research studies that just having phenological parameters associated with a certain crop cannot be enough to predict crop yield. There are plenty of other factors that have direct or indirect effect on the crop biophysical properties. A few of them could be soil related properties like soil pH, soil salinity etc.; others could be Cotton seed varieties related factors, and some could be about farmer practices and so on. It is near impossible to include all those factors at this stage of study. But climate related factors are on the top of list amongst these factors. Climate change is having an alarming impact on different aspects of life in this region, agriculture being one of most prominent amongst them being a source of income for major part of the residents there. (Akbar et al., 2020)

For this study, three climate related factors were shortlisted being Maximum Temperature, Minimum Temperature, and Precipitation. The data for downloaded from the NASA Power online portal where historical data has been available for users to download free of cost. National Aeronautics and Space Administration or NASA maintains the data of NASA Power that provides solar and meteorological data for globe using combinations of satellite measurements, global climate models and ground weather stations. The data is available for the past 40 years starting from 1980. It is mainly focused on communities working for renewable energy, sustainable buildings, or agroclimatology. The data was downloaded separately for all the four districts in our study area, named Layyah, Dera Ghazi Khan, Muzaffargarh, and Rajanpur, from year 2008 until the year 2022.

3.2.4 Annual Yield Data

Cotton yield per hectare over the study period was downloaded to model cotton yield over it. Yield data was downloaded from Crop Reporting Services of Punjab “Crop Reporting

Service Punjab (2021)", the data from 2008 until 2022 was downloaded for all the four districts in the study area.

3.2.7 Climatic Projections Data

To project the long-term impact of climatic factor on yield according to two scenarios i.e., SSP2-4.5 and SSP5-8.5 of CMIP6, first step was the collection of the climatic factors data under two major categorized, named as Observed data, and SSPs data. The SSPs data was later divided into Historical and Future data after data cleaning and processing. The observed data was extracted from NASA POWER, whereas the SSPs data was downloaded from CDS Climate - Copernicus Climate Change Service website. SSP2-4.5 and SSP5-8.5 are two of the Shared Socioeconomic Pathways (SSPs) used to project future climate scenarios in the Coupled Model Intercomparison Project phase 6 (CMIP6). Based on the level of effort done to counteract climate change, these two scenarios illustrate different possible futures for the globe. SSP2-4.5 assumes moderate greenhouse gas emission reduction efforts, whereas SSP5-8.5 assumes no emissions reduction efforts and a high rate of world economic development. The collecting and analysis of these data sets is vital for understanding the possible implications of climate change on crop yields and creating mitigation strategies.

Observed data consisted of daily minimum temperature, maximum temperature, and precipitation data from year 1981 till 2022 (Figure 2). To get this data, go to data access viewer portal of power.larc.nasa.gov and choose the temporal variation average (Daily), latitude and longitude of the target area i.e., coordinates of each district, time extent i.e., January 1981 – December 2022, file format(.csv) and selection parameters (Temperature at 2 meters maximum, Temperature at 2 meters minimum and precipitation). Once all the options are chosen, submit the request which provides the required data file after processing.

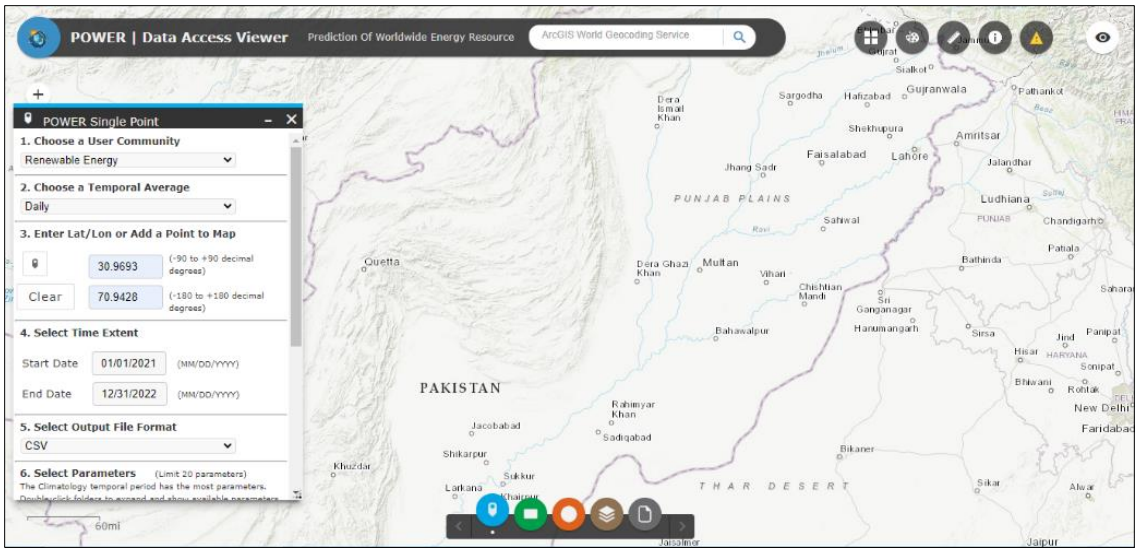


Figure 2: NASA Power viewer portal

Both the Historical (January 2015 – December 2022) and Future data (January 2023 – December 2099) was obtained in form of the SSPs scenarios data of the three variable, Daily minimum near-surface air temperature, Daily maximum near-surface air temperature and Daily Precipitation, downloaded from online climate data store. In order to get this data, go to website and search CMIP6 climate projections from Datasets section. This leads to a webpage which provides an overview and documentation related to data that is available on the website and a download section. In the download section, some options need to be specified to get desired data such as Temporal resolution (Daily), Experiments (SSP2-4.5 and SSP5-8.5), variables (Temperature and Precipitation), Model (MIROC6(Japan)), Years (2015 – 2099) and file format. Once the requirements are specified, submit the request to get desired data in NetCDF format. To get the data for each scenario, this process was repeated individually for each variable.

3.2.8 SSPs Data Preparation

The data in NetCDF format cannot be visualized and used to proceed, so it was processed using ArcGIS. The raster tile of the downloaded data, obtained using NetCDF to Raster tool, displayed Global level view of data which was then processed to narrow down to get

data of pixels that lie in the study area by using NetCDF to Table tool. The NetCDF to Table tool provided data of each specified pixel for specified year i.e., 2015-2099. Later the pixels level data was edited in excel district wise based on which pixels lie in which district. (Figure 3 and 4)

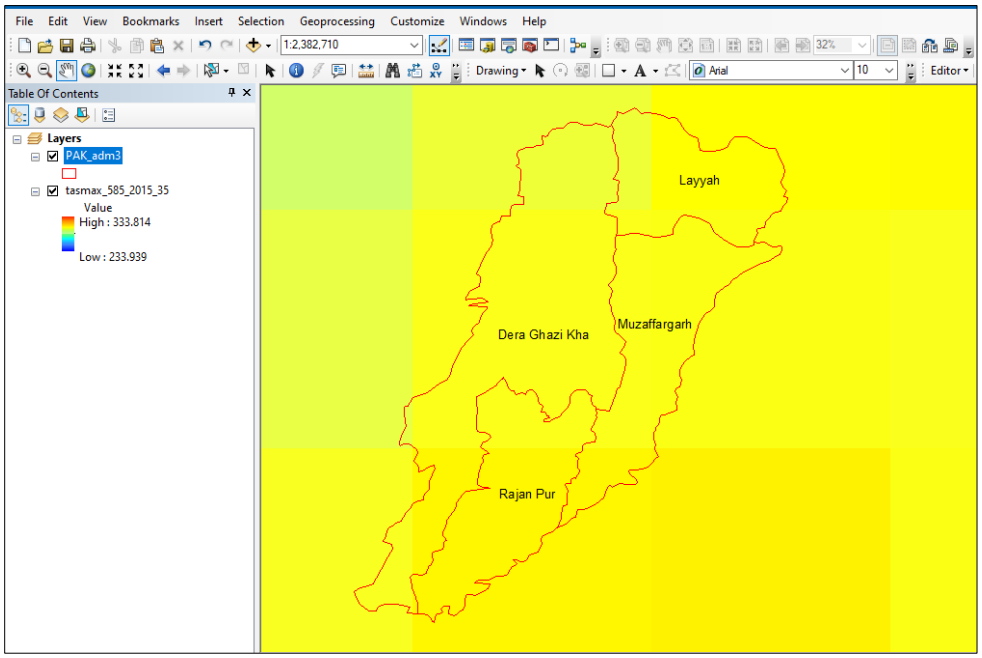


Figure 3: Visualization of data in ArcGIS

Table 2: Latitude, Longitude and Number of pixels in each district

District	Lat	Long	No. of Pixels
Layyah	31.052617	71.353702	4
Dera Ghazi Khan	30.298565	70.480589	3
Muzaffargarh	30.2362	71.211963	3
Rajanpur	29.017244	70.163094	2

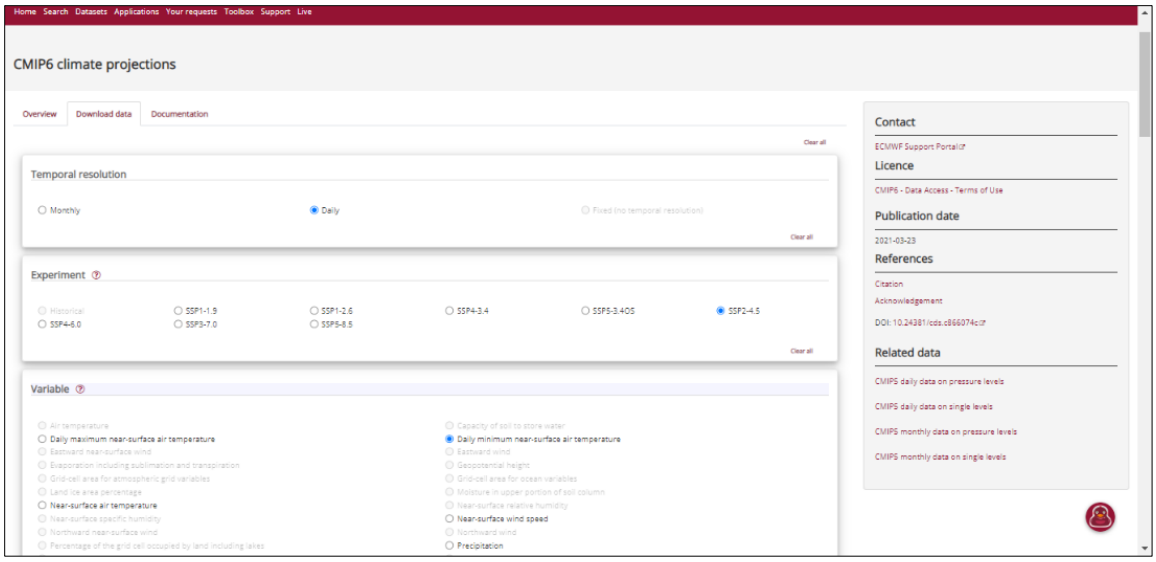


Figure 4: CMIP6 climate projections download portal.

3.3 Cotton crop area identification using phenology-based Approach

The study of the schedule of recurring natural phenomenon in plants and animals, such as blooming, fruiting, and migration, is known as phenology. Phenology can be used in remote sensing to distinguish various crop varieties based on their distinct patterns of growth and development. Each crop has its own phenological cycle, which relates to the progression and development phases during growing season, such as emergence, vegetative growth, blooming, and senescence.

Remote sensing data, such as satellite imagery can be used to detect changes in vegetation patterns over time and to monitor crop phenology. In Addition, Remote Sensing data, for example, can be utilized to measure the timing of leaf emergence, peak leaf area, and leaf senescence, which can ultimately help in distinguishing various crop types. For instance, analyzing vegetation indices patterns in particular Normalized Different Vegetation Index NDVI over time can help identify crop types based on their unique phenological signature. For the cotton crop area identification, two different platforms were used: Python and Google Earth Engine.

3.3.1 Python Platform

The primary goal of this part was to create a method that could accurately map cotton crops using Sentinel 2A imagery in python programming language. Python is an open-source popular programming language for data analysis and image processing in remote sensing applications. The proposed method involved using various Python programming language libraries such as NumPy, Pandas, Geopandas, Rasterio, Matplotlib and others, to read, analyze and process satellite downloaded imagery from European Space Agency's (ESA) official website. Initially, Normalized Difference Vegetation Index (NDVI) was computed for Rajanpur district. 5Subsequently, a unique NDVI range for cotton crops during a specific month was identified with the help of google earth pro, Extract by Mask and Random Points functions. A mask function was defined to convert the NDVI image to an array and then classify NDVI image pixels into two classes. Those pixels falling within the range were classified as class 1 (representing cotton crops) and the remaining pixels were classified as class 0 (representing additional features in the imagery). The end result was a new image with two distinct classes: Cotton Crops and Other features.

3.3.2 Google Earth Engine Platform

Two major methods were used to determine and map the Cotton Crop Area, one of which was carried out using GEE. Furthermore, two different approaches were used to determine which classification method more accurately determines the cotton area. The code to implement both these approaches is attached in Appendix A and explained in detail as follow:

A. The suitability of GEE was evaluated and chosen as an appropriate platform to perform the analysis. The first step was to load the verified Ground Truth Cotton Fields shapefile from CRS (Crop Reporting Service) of Punjab and determine the unique boundary values of cotton crop NDVI reflectance using ee.Reducer.max(), ee.Reducer.min() functions for NDVI image that lies within the boundary of Reference cotton fields. Once the range was determined, this classification method of using NDVI range was applied on the preprocessed mosaicked images of study area as discussed in Section 3.2.1 consisting of

four districts of Deri Ghazi. The resulting classified image consisted of two classes: cotton crops and other features. The threshold range for NDVI that was unique to cotton only was between 0.49681705 and 0.629582268, whereas any pixel value falling within this range was classified as "1", while any pixel outside the range was classified as "2". The resulting cotton classified image was then converted to a shapefile to obtain the final Cotton mask for the study area.

B. Since the resulting mask from the first approach had less accuracy, the Second approach was adopted which included four other indices as well with NDVI which enhanced the accuracy of the output cotton mask. Subsequently, five vegetation indices, included were Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Difference Vegetation Index (DVI), Wide Dynamic Range Vegetation Index (WDRVI), and Stress Related Vegetation Index (STVI). A range for each indices were then determined and used in a mask function to derive a final classified image which consisted of two classes: Cotton Crops and Other Features. The process was employed three times, and the findings from each year were combined with the prior year's results to generate a time-series dataset. This method could be used to investigate the spatial and temporal patterns of any target vegetation, cotton in this research.

3.4 Indices Derivation and Data Organization

After the eleven indices were selected and their mathematical derivations were lined up, the next step was to go about the ways to implement them and drive their values. There is the usual way of using available GIS software to calculate vegetation indices in batch. ArcMap could be one option where the option of Raster Calculator is available to write mathematical equations to generate customized indices. While this seems quick, the fact is that to start with this method, a lot of data would need to be downloaded.

To get a glimpse, the first step would be to go to LPDAAC online portal, provide them with dates of cotton cycle and shapefile of boundaries of study area selected for this thesis, and download the specific bands required for this study. There would be more than 100 data files per year and combining them with a 14-year study period would elevate this

number significantly. Now each of these files would need to be imported in ArcMap software, which is heavy by its nature, and even importing data would consume huge amounts of time. Next step would be to calculate separate raster indices for each year that is again a humongous task. Finally, when 11 indices for each date of all 14 years are calculated, there would still be needed to mask these hundreds of files to the specific cotton region extracted in the earlier phase of this thesis. This is an extremely long task that demands a lot of attention and time as a small mistake could prove fatal for research.

The alternate path is to use Google Earth Engine. GEE is a cloud-based platform to perform analysis related to geospatial data. It is also scalable for complex datasets and provides a wide range of satellite driven and climate related data. It has immense programming capabilities that can be used to write JavaScript code and automate processes that could otherwise be a hassle. Since the GEE provides access to huge satellite data, so downloading was not needed for this part of methodology as data can directly be imported in Earth Engine Code Editor. Secondly, instead of going for GIS software toolboxes, code could be written to automate the process of calculating vegetation indices.

After signing up to GEE using Google account, code editor needs to be opened. Then there are options for Scripts, Documents and Assets. In the Asset section, the study area boundary shapefile and cotton shapefile extracted in earlier phase of this thesis needs to be uploaded. These Assets are related to the research study and can help to focus the data on study area and extract features of that region. Next there are documents or Docs where there are different methods and algorithms already available to help the users. Then in the Script section, new script must be created. This is where code needs to be written. The first step in Google Earth Engine Code Editor was to import the MOD13A1 product of Terra MODIS into the environment. Image Collection is a constructor in GEE where asset ID needs to be pasted to import required satellite data into the environment. The approach for this study was to import MOD13A1 data separately for all four districts of the study area to get some extra values of vegetation indices. The product was focused on specified districts by clipping and dates were also specified for required time. Once one district was done, the next district would be entered, and data would be clipped on that one.

Another crucial step was to write code for all the indices. As indices formulas were already known, the necessary bands had to be selected for each vegetative index function. All the indices' values were calculated in GEE using the code provided in Appendix A (Indices Calculation). Once all indices were calculated, the next step was to calculate the mean values of each month of the whole cotton cycle (Code given in Appendix A). This was mandatory so that different phenological parameters can be gathered to estimate how the cotton pattern changes on monthly basis, so pattern can be identified for estimation of yield instead of just the values. Just to make sure that all the formulas have been correctly implemented and issues that arise while writing code have been addressed properly, a recheck was advised using traditional methods. MOD13A1 data from LPDAAC was downloaded but only for just a few days amongst 14 years. The purpose of this activity was to confirm that the mean values calculated using the GEE code are in agreement with the mean values calculated using Raster Calculator values of ArcMap. This was not otherwise necessary, but the circumstances demanded this verification. The values came out to be synonymous.

Once means were calculated and verified, an excel file was generated inside the code which contained all the mean values of eleven vegetation indices. Then that excel file was exported as CSV to Google Drive from where it was manually downloaded. Keeping the cotton cycle explained above, 8 months were picked for each year from 2008 until 2022. The starting month was April as it is the month for cotton seed plantation, and then the ending month was November was harvesting is done in this month. There were not just mean vegetation indices values of these 08 months in final excel file, but also three climate related factors explained above including maximum temperature, minimum temperature, and sum precipitation.

After data was organized monthly, the next target was to average them over the years, because yield per hectare was available on a yearly basis. The annual average of all the eleven indices was taken separately for each year, starting from 2008 until 2022. Similarly, the annual average was taken for Maximum Temperature and Minimum Temperature. While in the case of Precipitation, sum of all monthly values was taken for each year. This is because temperature is a continuous variable that changes from day to night each day.

To get an accurate idea of temperature, there is need to average it over period to see its overall range. But in case of precipitation that is a discrete variable and happens over individual events throughout the time, it needs to be summed over that period. This helps to have a precise idea of how the climate has behaved in a region.

3.5 Model Selection and Development

After organizing all the data, the input data was ready to be fed into machine learning models to predict the cotton yield. There were 14 features (independent variables) and one dependent variable i.e., yield. Out of 14 independent variables, 11 were vegetation indices including DVI, RVI, NDVI, EVI, SAVI, OSAVI, RDVI, SARVI, TVI, WDRVI and STVI01 (Table 1); while the rest 03 were climate factors including maximum temperature, minimum temperature, and sum precipitation. All There were five models selected based on extensive literature review that have been used frequently for crop modelling.

3.5.1 Automatic Linear Modelling

Automatic Linear Modeling or ALM is a systematic model that includes variable selection, model fitting, and diagnostic checking to construct linear regression models. ALM is an expansion of stepwise regression, which picks variables iteratively based on their statistical significance, but with extra features that enhance the effectiveness of the resulting model. All predictor variables are first included in the model by the ALM algorithm, and then those that have a small impact on the prediction of the response variable are gradually removed. This is accomplished by utilizing hypothesis testing. It is especially helpful when there are a lot of factors to consider and the relationship between the predictors and the response variable is complex. (Panneerselvam et al., 2021)

3.5.2 Generalized Linear Model

Generalized Linear Model or GLMs performs the statistical modelling of relationships between a response variable and one or more predictor variables. By relaxing the constraint of constant variance and allowing the response variable to have a non-normal distribution,

the GLM generalizes the conventional linear regression model. A link function, which converts the predicted value of the response variable to a linear combination of the predictor variables, is used in a GLM to express the relationship between the response variable and the predictor variables. The relationship between the mean value of the response variable and the predictor factors can be modelled using the link function. This makes GLMs approach and applications different from those of ALM even though both are used for linear regression analysis. (Faramiñan, 2022)

3.5.3 Random Forest

Random Forest or RF is another machine learning which comprises of several decision trees that use random subsets of data to train these trees, hence the name. These decision trees then come together to create an output that is most accurate and robust amongst all. They are used for their accuracy and being less susceptible to overfitting. Like Linear Models, they can be used for regression as well as classification problems.(Paul et al., 2018) and Charoen-Ung et al., 2018) have preferred the utility of Random Forest in crop modelling.

3.5.4 Gradient Boosted Trees

Gradient Boosted Trees or GBT are a refined and better version of Random Forest. They contain decision trees like RF but GBT functions by constructing a series of decision trees that are trained using the leftovers from the prior tree. Each subsequent tree in the sequence is created to fix the flaws of the preceding trees. GBTs are known for their accuracy and precision. However, they could need a lot of training data and may be computationally taxing (Krauss et al., 2017).

3.5.5 Support Vector Machine

Support Vector Machine or SVM is a machine learning model that operates by locating the hyperplane in the feature space that maximally separates the data points. The hyperplane

is selected so that the margin between the two groups of data points is maximized. SVM may be used with a number of kernel functions to handle various sorts of data and is quite effective at handling high-dimensional data. They can also be used for regression as well as classification problems. (Derek A. Pisner et al., 2020) & (Su Ying-xue et al., 2017).

To develop Automatic Linear Model or ALM, a statistical software called IBM Statistical Package for the Social Sciences or SPSS was used. Through its graphical user interface, SPSS enables users to do different statistical analyses without having need to be familiar with any programming languages. Excel, CSV, and SQL databases are a few of the sources from which users can import data. Additionally, SPSS comes with a data editor that gives users the ability to work with data, clean it up, handle missing numbers, and change variables. The data visualization tools in SPSS include histograms, scatterplots, bar charts, and pie charts. The software also allows users to design their own graphs and charts. By using scripts and syntax commands, SPSS users can increase the software's functionality beyond its primary features. This can be very helpful for automating routine chores or running intricate analyses.

GLM, RF, GBT and SVM were developed using an open-source platform for predictive analytics, machine learning, and data science called RapidMiner. This software presents numerous machine learning techniques, including decision trees, clustering, regression, neural networks, and deep learning, are offered by RapidMiner. A variety of tools for model evaluation and performance evaluation are also available. RapidMiner's visual interface, which enables users to construct data pipelines using drag-and-drop components, is one of its primary advantages. These components can be used for modelling and assessing prediction models as well as for several data preprocessing activities, including cleaning, filtering, and transforming data.

3.6 Model Statistical Evaluation

The performance of machine learning models for different applications is measured through many statistical methods available. This part of the study utilizes five frequently used machine learning models so to assess and validate the model's processing, three

evaluation metrices were used. The yields predicted by each model were compared with the actual yield recorded in input data to measure the number of errors in each. The Root Means Square Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE) were also used to assess the model's efficacy (Kahimba et al., 2009). Equation 1 was used for the approximation of the weighted error difference between the predicted and actual yield:

$$RMSE = \sqrt{\frac{\sum(P_j - A_j)^2}{n}} \qquad (1)$$

where A_j is the recorded yield, P_j is the projected yield and n is the rows in data that was 100 in this study. MAE measures the average absolute differentiation between the values of actual yield and predicted yield. Equation 2 was used for the computation of the weighted average of the absolute error:

$$MAE = \frac{\sum|P_j - A_j|}{n} \qquad (2)$$

To evaluate the consistency of the error division and identify whether the model is under- or over-predicting, MBE was calculated. Positive or negative sign denote underprediction and overprediction, respectively. Positive and negative values are equally distributed when the value is zero. MBE was calculated utilizing Equation 3:

$$MBE = \frac{\sum(P_j - A_j)}{n} \qquad (3)$$

3.7 Future cotton yield patterns

3.7.1 Bias Correction and Downscaling

After converting the collected data into CSV format, the next step was to remove the biasness in SSPs data using CMhyd tool. Climatic models are important instruments for understanding and predicting future climatic conditions, but their outputs are prone to biases that can impair the accuracy and dependability of their projections. To address this issue, bias correction approaches for adjusting climate model outputs using observable data

have been developed, minimizing systematic errors, and improving the depiction of current and future climate conditions.

The CMhyd tool, which is created exclusively for hydrological variables such as precipitation and temperature, is one such bias correcting method. The CMhyd tool is a statistical downscaling strategy that uses a high-resolution collection of actual data to adjust for biases in climate model outputs. On a daily resolution, bias correction processes were employed to condense the variance amid observed and simulated climate variables, ensuring that models based on corrected simulated climate data reasonably match simulations based on observed climate data. Generally assuming stationarity, bias correction techniques used in climate science use the same methodology and parameterization for both present and future climate circumstances. Regarding their capacity to continue putting in strong work in the face of shifting circumstances in the future, there is some doubt. The statistical characteristics of climate data and the connections between climatic variables may change in the future, causing bias correction approaches to lose some of their efficiency even though they performed well during an evaluation period. In order to reduce the possible influence of future changes on the performance of these technologies, it is critical to test their performance under various future climatic scenarios, including extreme occurrences.

1. To perform Bias correction and downscaling, an important step was to select which bias-correction method to employ as there are many options available on CMhyd including Linear scaling, Delta-change correction, power transformation, variance scaling and Distribution Mapping. The method used as Delta-change correction as it assumes that the changes in the model output between the historical period and the future period are consistent with the changes observed in the observations. The method works by first calculating the difference between the historical model output and the observed data for a given time period (the delta-change). This delta-change is then applied to the future model output to correct any biases. Additionally, it is a relatively simple and practical method to apply that takes into account variations in the mean and variability of the data, making it a robust method for correcting biases in climate model outputs. (Figure 5)

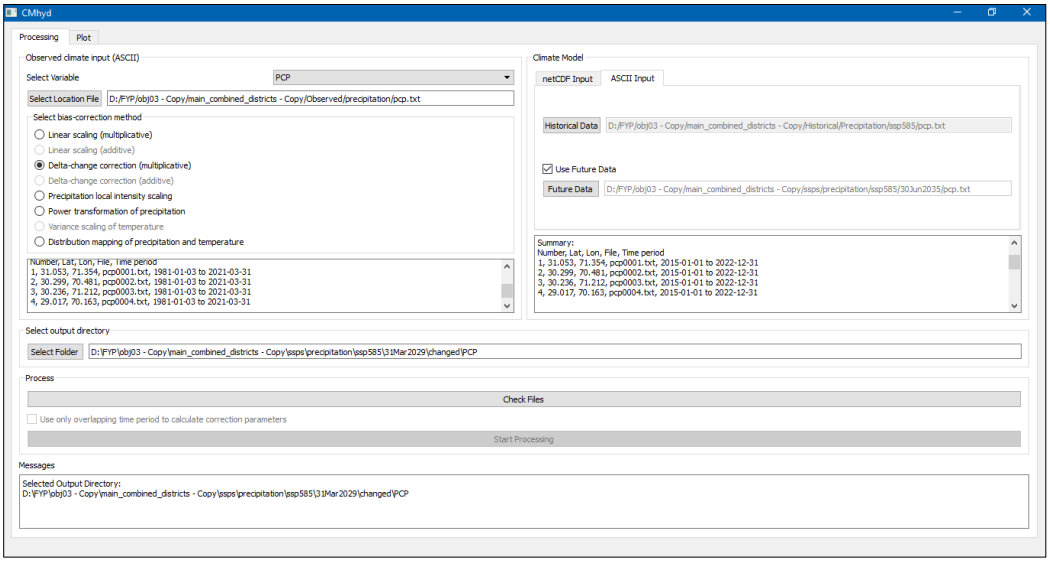


Figure 5: Selecting Bias Correction method on CMhyd

2. Alongside, the SSPs data was split into Historical and Future datasets, 2015-2022 and ten sets of future data i.e., 2023-30, 2031-38, 2039-46, 2047-54, 2055-62, 2063-70, 2071-78, 2079-86, 2087-94, and 2095-99 respectively with an eight-year interval and then the datasets were converted into .txt format.
3. Finally, the prepared data was input into the CMHyd as follows:
   i. Observed data into Observed climate input section.
   ii. Under Climate Model section, choose Historical data into historical data tab and,
   iii. Each Future dataset one after another into future data tab.
4. Output data directory was then chosen to get the bias corrected data.
5. The output included corrected data and several Graphs in Figure 6 that showed different aspects of Temperature and precipitation under both scenarios of CMIP6 bias correction.
   i. Mean Monthly Precipitation: This graph shows the average amount of variable (i.e., Temperature and precipitation) that falls in each month of

the year in the respective scenario. This information can be useful for understanding seasonal patterns of precipitation.

ii. Monthly Standard Deviation: This graph shows how much variation there is in variable from month to month in the respective scenario. A high standard deviation indicates that some months are much wetter or drier than others, while a low standard deviation indicates more consistent precipitation patterns.

iii. 90th Percentile: This graph shows the amount of precipitation that exceeds the 90th percentile in the SSP285 scenario. This information can be useful for understanding extreme precipitation events.

iv. Coefficient of Variation: This graph shows the ratio of the standard deviation to the mean precipitation in the SSP285 scenario. Bigger coefficient of variation indicates that there is a lot of variation in precipitation comparative to the mean, while a low coefficient of variation designates more consistent precipitation patterns.

v. Wet Day Probability: This graph shows the probability of a day having precipitation in the SSP285 scenario. This information can be useful for understanding the frequency of precipitation events.

vi. Precipitation Intensity: This graph shows the intensity of precipitation events in the SSP285 scenario. This information can be useful for understanding the severity of precipitation events.

vii. All these steps were repeated for each future dataset under each SSP scenario for both variables Temperature and Precipitation.

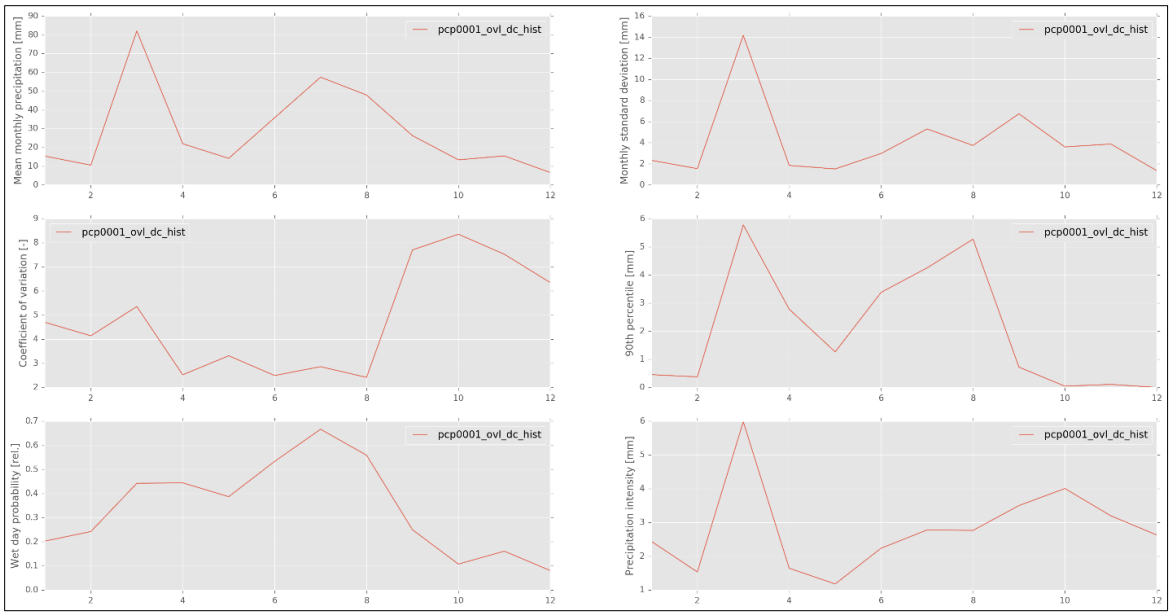


Figure 6: Graphs showing different aspects of Temperature and Precipitation under both scenarios of CMIP6 bias correction.

The bias corrected daily data was then combined into three files named Short, Mid and Long-term.

1. Short-term projection: 2023-2040 (about 17 years)

   Short-term projection was focused on the near future, where the impact of climate change may not yet be fully realized. However, changes in weather patterns, water availability, and temperature can still affect crop yield.

2. Mid-term projections: 2041-2060 (about 20 years)

   Mid-term projections could be focused on a period where the influence of climate variables on crop yield is expected to become more significant. This could be a period where adaptation measures are likely to be taken, and new technologies may be developed to cope with the changing climate.

3. Long-term projections: 2061-2100 (about 39 years)

   Long-term projections could be focused on a period where the effect of climate change on crop yield is expected to be substantial, and adaptation measures may have reached their limits.

The data in these three files was then shrunk into Monthly resolution to be later used for predicting the future monthly values of vegetation indices that had relatively high correlation with cotton yield.

3.7.2 Indices Prediction Model

The next step was to make Models for highly correlating vegetation indices from Objective 02 model prediction results, in order to get the future yield for all these three longs, mid, and short terms to later predict and analyze the impact of change in climatic factor on cotton yield under each SSP scenario. Those indices included NDVI, OSAVI, SARVI, STVI, and WDRVI. All the models were made on Google Collaboratory that is an online python code editing tool available freely for use by anyone.

First, the past monthly Minimum temperature, Maximum temperature, Precipitation, and respective Indices data from 2015-2022 was loaded into the colab and then split into train and test dataset. Later, a function was made that consisted of various Machine learning models to determine which one performs best on the dataset provided and the data was then fed into that function. The models incorporated as options in the function included Linear Regression, Lasso, Ridge, ElasticNet, Decision Tree, Random Forest, gradient Boosting and SVM. The one that performed best for these Indices prediction was Random Forest with the least RMSE error and greatest Accuracy score for all the Indices.

After Models were compared, the Optimized Parameters for the Random Forest were determined using GridSearchCV function and again model was trained and tested with the optimized parameters. Then the monthly three future data files were fed one after another into the trained model to predict future values for these Vegetation Indices. Once the future indices values were obtained on monthly basis, this indices data and the future SSPs climatic data was converted into even more general temporal resolution i.e., Yearly basis, so later can be used for future yield prediction.

3.7.3 Future Yield Prediction Model

To predict the future yield for the years 2023-2099, the past yearly data of years 2008-2022 was considered for training the Prediction model. To achieve this, a linear regression model was built using advanced statistical methods, specifically the SPSS software. The model was developed using various climatic factors such as minimum and maximum temperature, precipitation, and Vegetation Indices such as NDVI, OSAVI, SARVI, STVI, and WDRVI.

The first step in the process was to collect and preprocess the dataset, which involved removing missing values, duplicates, and outliers. Then, the dependent variable (cotton yield) and independent variables were defined, and the data created was divided into a training set and testing set. The Linear Regression tool in SPSS was used to build the model, and its performance was evaluated using metrics such as R-squared, adjusted R-squared, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). To visualize the variety in data values, a scatterplot was made using ZRESID on the y-axis and ZPRED on the x-axis. Table 3 shows the mean and standard deviation of data. Table 4 includes the suggested indices for the model.

Table 3: Descriptive Statistics of Input Factors

Descriptive Statistics

	Mean	Std. Deviation	N
Yield	305.2206875	139.2160144	100
tasmax	42.83080163	1.292229469	100
tasmin	20.72152364	1.109164535	100
prec	199.3514750	176.7621900	100
NDVI	.3887500194	.0198651433	100
OSAVI	.3887396019	.0198607833	100
SARVI	.8869521993	.0221435861	100
STVI	774.9514375	60.69119610	100
WDRVI	-.614083446	.0166805615	100

Table 4: Variables used in model

Model	Variables Entered	Variables Removed	Method
1	WDRVI, tasmin, prec, STVI, tasmax, SARVI, NDVI[b]	.	Enter

a. Dependent Variable: Yield

b. Tolerance = .000 limit reached.

The developed model was then used to predict future cotton yield for the years 2023-2099 based on the values of independent variables such as temperature, precipitation, and vegetation indices. The predicted values were validated against actual values to assess the accuracy of the model and make any necessary adjustments to improve its performance.

The final step involved using the developed equation to forecast future cotton output based on the values of independent variables, and this information can be used to inform decision-making in areas such as crop management, production planning, and risk assessment.

Because the training data was small, advanced statistical methods such as SPSS software were utilized to build a linear regression model. The overall accuracy of the model was found to be moderate, exceeding 60%. However, to identify the most important factors influencing agricultural productivity, a Pearson correlation analysis was conducted. The results of the analysis revealed a significant correlation between cotton yield and maximum temperature. This features the effectiveness of studying maximum temperature as a key variable when developing climate change mitigation and adaptation strategies for the cotton sector. Figure 7 summarizes this whole methodology into a flowchart starting from the left and heading towards right after the achieving result from each dataset included.

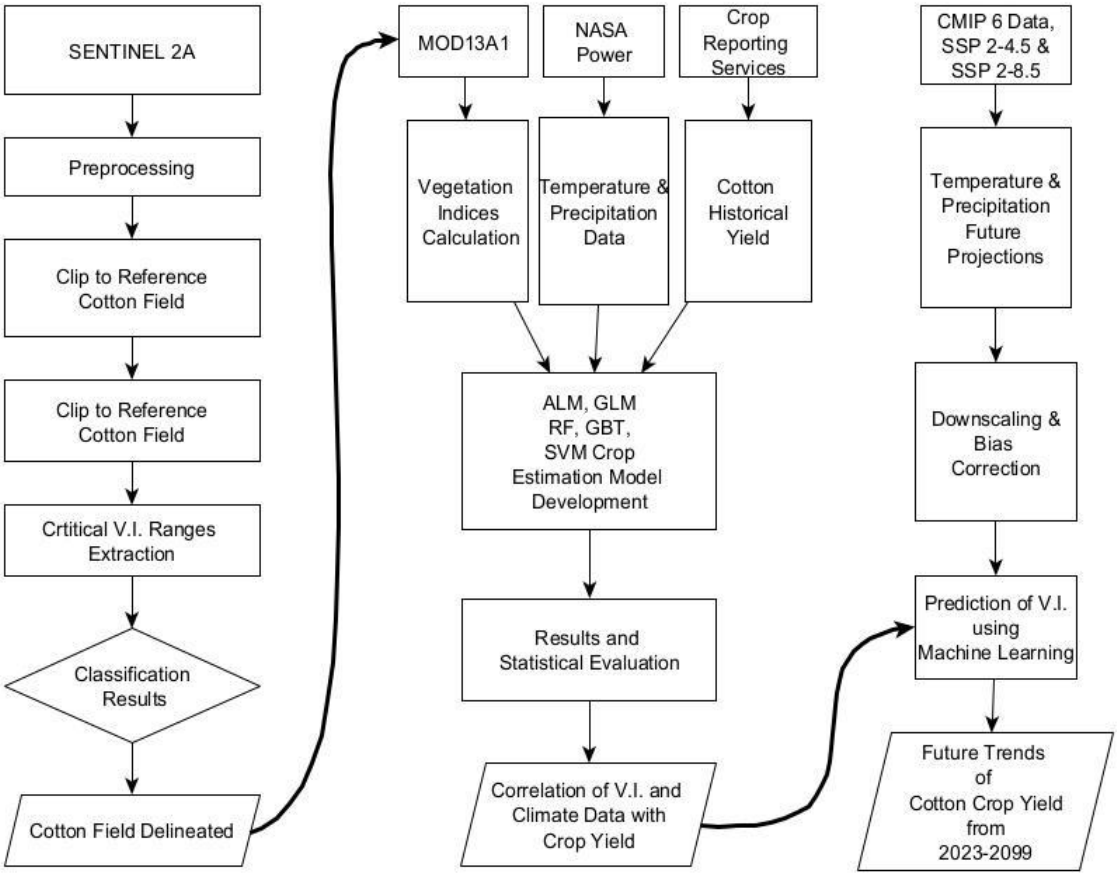


Figure 7: Methodology Flowchart

CHAPTER 4

RESULT & DISCUSSION

4.1 Cotton Covered Area

4.1.1 Python Results

The accuracy of the classified image with python programming language was found to be unsatisfactory when assessed with user-producer accuracy assessment. As result, most of additional features were misclassified as cotton crops. The overall accuracy of the study, as measured by user-producer accuracy assessment was below 50 percent and kappa coefficient were around 40 percent. Hence, that is an unacceptable percentage for accuracy. Additionally, attempts to assign a specific coordinate system to the classified image using python libraries were unsuccessful. Assigning a coordinate system to an image is critical for many applications, including Geographic Information Systems (GIS) and Remote Sensing and leads to help in interpretation and analysis within spatial context.

Moreover, in order to read, process and analyze an image in python it is necessary to download satellite imagery from a source which can be time-consuming and require strong internet connection. Python is unable to load imagery covering a large area and requires reducing imagery scale and extent (Figure 8).

In comparison with the python approach which involved analyzing a single sentinel image, follow by Normalized Vegetation Index (NDVI) calculation and then range determination, the image classified with the google earth engine (GEE) which involved only Normalized Vegetation index (NDVI) range cotton crops identification appeared to be more accurate. However, a user-producer accuracy assessment was conducted to validate the result against ground truth and a shapefile provided by Global Change Impact Studies Center (GCISC). The assessment revealed an overall accuracy below 60 percent, indicating that the classification result was not adequate for the intended purpose.

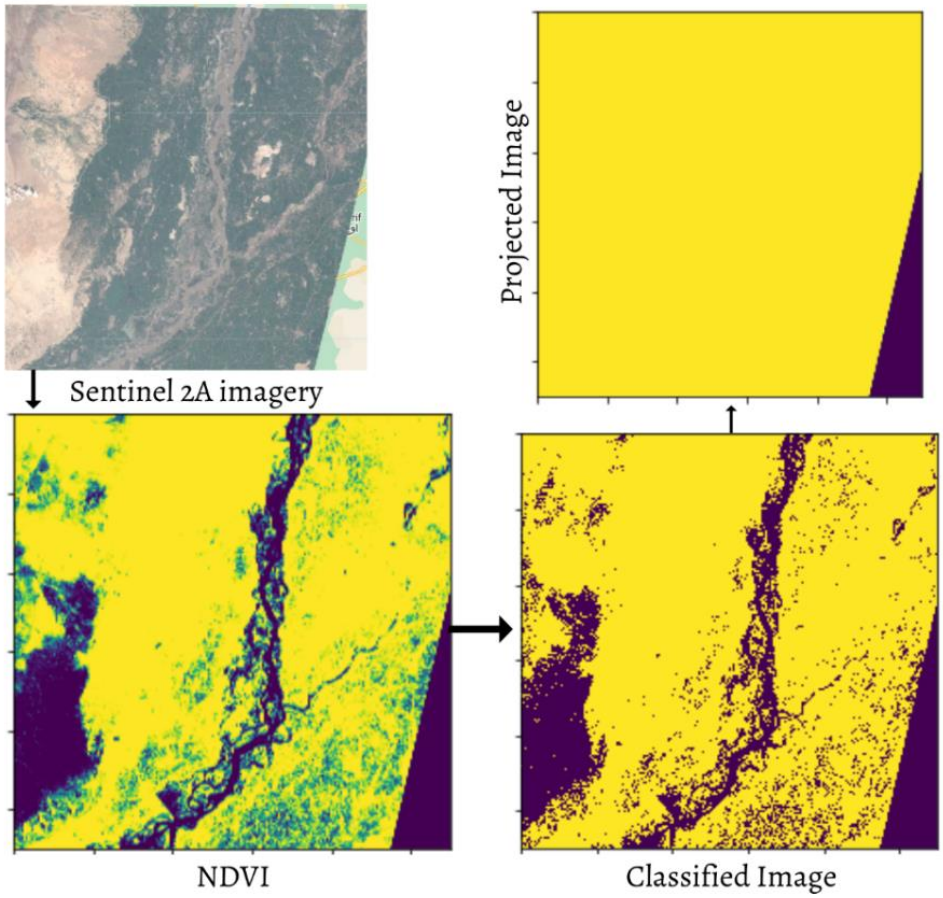


Figure 8: Result of Python approach to extract cotton mask

4.1.2 Google Earth Engine Results

A. In consideration of the accuracy of the previous method, a new approach was devised. Initially, satellite imagery from different dates throughout a particular year were collected. Subsequently, the imagery was mosaicked and clipped to the area of interest shapefile to obtain the final imagery of the intended location. Cotton crop fields were digitized manually in Google Earth Pro as polygons using a reference shapefile and visual observation. Using ArcGIS, a new shapefile of digitized cotton crop fields was created. The final image was then clipped to the cotton crop fields in Google Earth Engine, and five vegetation indices (insert names) were computed. Subsequently, the ranges for each index were determined for cotton crops. Example index of NDVI is shown in Figure 9.

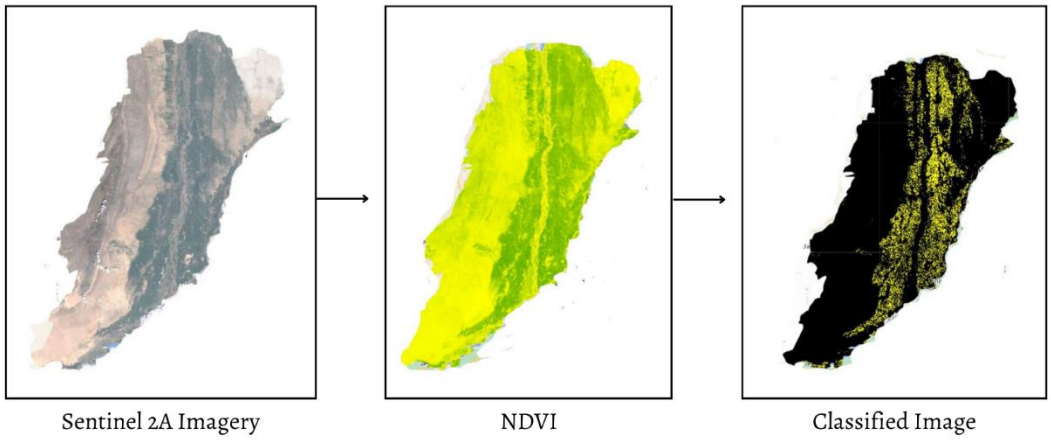


Figure 9: Classified Image

User producer accuracy was employed to evaluate the correctness of this study's classification results. We found that the overall accuracy of our classification was 87%, with a Kappa coefficient of 83.7%. These outcomes show that the given methodology is helpful in mapping cotton crops using Sentinel 2A imagery and five vegetation indices. (Figure 10)

To validate our results, we compared them to a reference cotton shapefile that was created based on field survey. The reference shapefile had an accuracy of over 95%, which is considered to be highly accurate. Our results showed satisfactory agreement with the reference shapefile, indicating that our approach is reliable and robust.

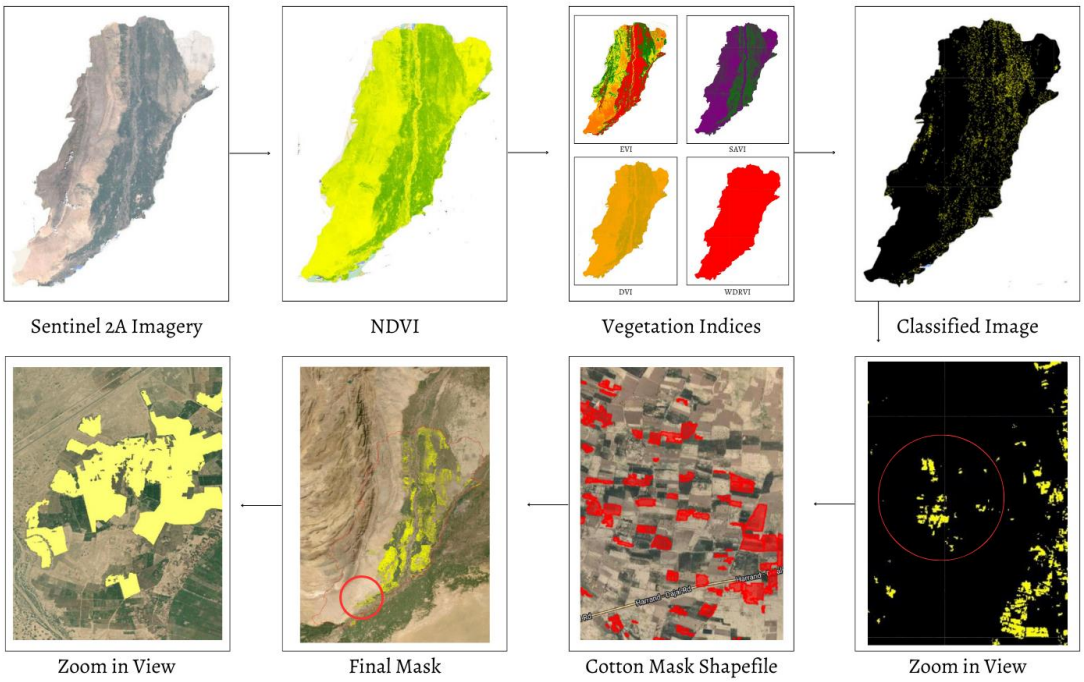


Figure 10: Cotton mask extraction process using GEE

4.2 Yield Prediction Model Results

4.2.1 Automatic Linear Modelling

Automatic Linear Model was developed with the objective of enhancing the accuracy of the model by using boosting techniques, instead of just using a standardized model. Forward stepwise was chosen as model method. The results given in Figure 11 depict the scatter plot between the predicted yield by the model on Y-axis versus the actual yield recorded on X-axis.

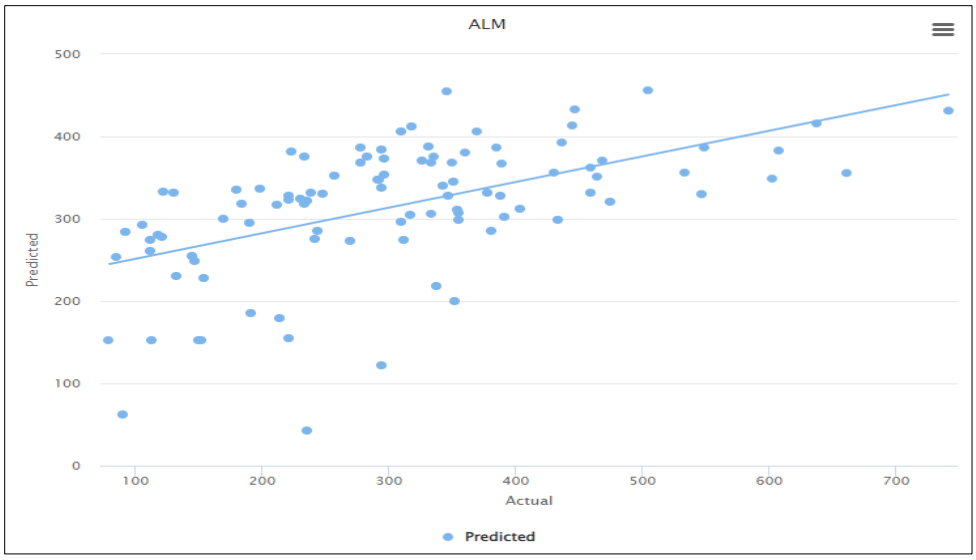


Figure 11: Scatter Plot of ALM Yield Prediction Result

Table 5 shows the importance of different independent variables to predict yield using bar plot. It shows that the model gave most importance to maximum temperature, followed by precipitation, then SARVI and OSAVI. Also, it gave the least importance to the minimum temperature. Since cotton yield in this region is mostly impacted by maximum temperatures so this observation is valid.

Table 5: ALM Weights Assigned to Independent Variables

Attribute	Weight
tasmax	0.4
perc	0.37
SARVI	0.09
OSAVI	0.06
WDRVI	0.04

4.2.2 Generalized Linear Model

The results of Generalized Linear Production Model are given in Figure 12 which depicts the scatter plot between the predicted yield by the model on Y-axis versus the actual yield recorded on X-axis. The points are more confined than the ALM model as can be seen.

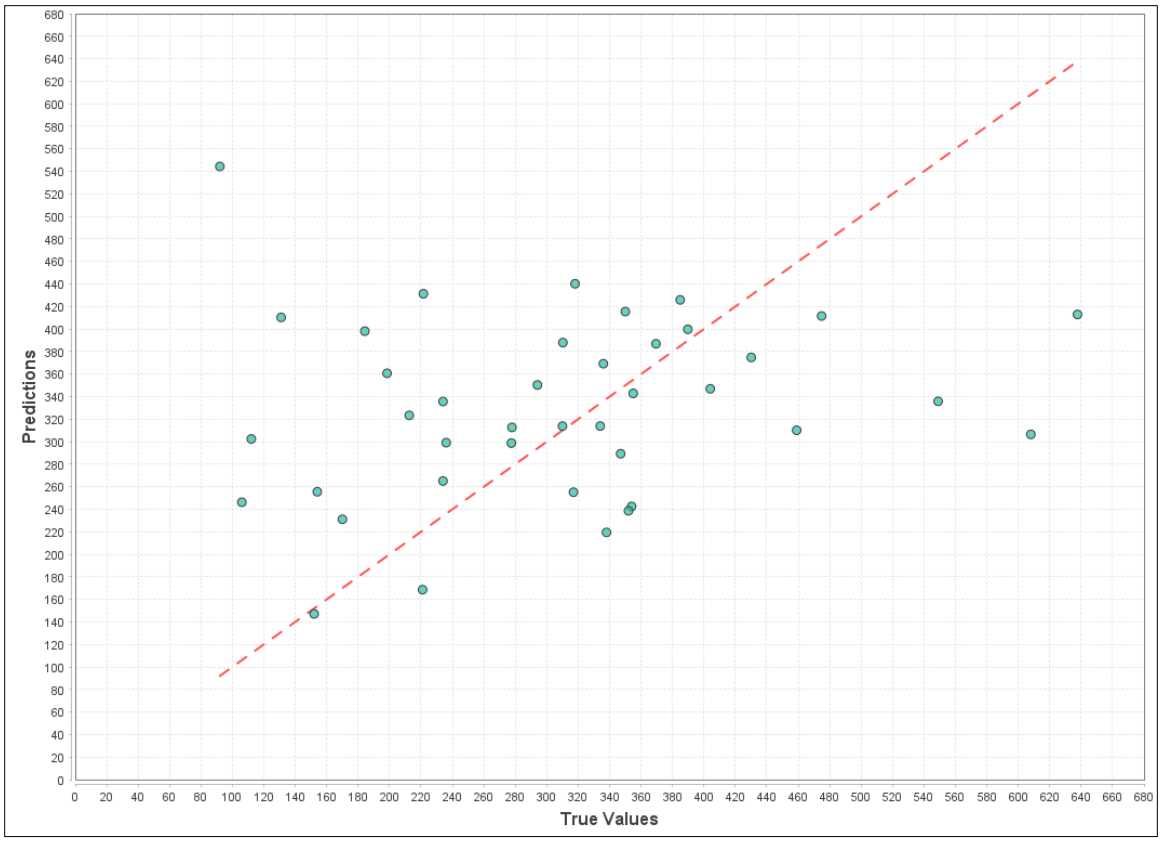


Figure 12: Scatter Plot of GLM Yield Prediction Result

Table 6 shows that the model gave most importance to NDVI, STVI and WDRVI. While this GLM did not give importance to any other parameter. The most logical explanation leads to the fact that GLM does not consider non-linear relationship between the input variables and outcome.

Table 6: GLM Weights Assigned to Independent Variables

Generalized Linear Model - Weights

Attribute	Weight
NDVI	0.597
STVI	0.423
WDRVI	0.133
tasmax	0
tasmin	0
prec	0
OSAVI	0
SARVI	0

4.2.3 Random Forest

The results of Random Forest Model are given in Figure 13 which depicts the scatter plot between the predicted yield by the model on Y-axis versus the actual yield recorded on X-axis.

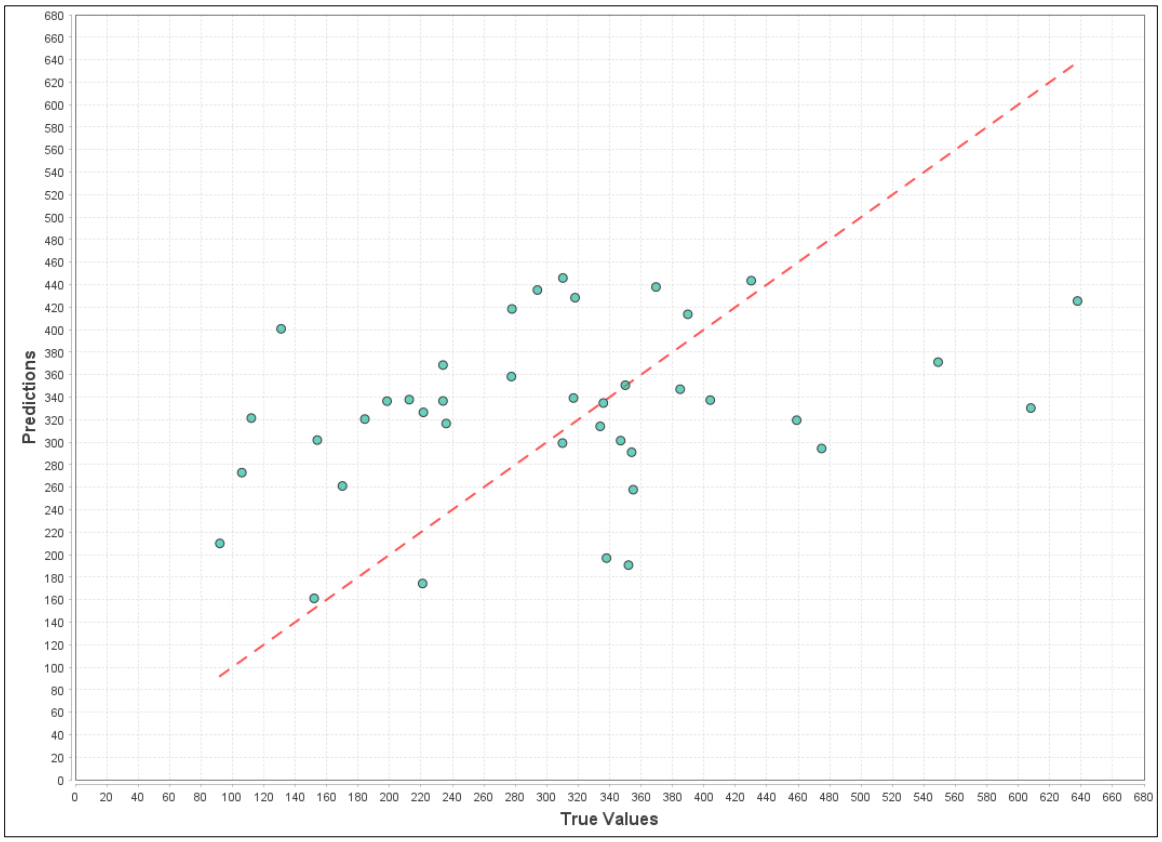


Figure 13: Scatter Plot of RF Yield Prediction Result

Table 7 shows the importance of different independent variables to predict yield using bar plot. It shows that the model gave most importance to maximum temperature, STVI and then minimum temperature, followed by vegetation indices like WDRVI and OSAVI. The minimum temperature could have popped up due to complex computation of random forest instead of simple decision trees.

Table 7: RF Weights Assigned to Independent Variables

Random Forest - Weights

Attribute	Weight
tasmax	0.325
STVI	0.079
tasmin	0.075
WDRVI	0.048
OSAVI	0.024
NDVI	0.013
prec	0.008
SARVI	0.002

4.2.4 Gradient Boosted Trees

The results of GBT Model are given in Figure 14 which illustrates the scatter plot between the predicted yield by the model on Y-axis versus the actual yield recorded on X-axis. The results of Gradient Boosted Trees are more scattered than the three models presented before, including ALM, GLM and RF.

Table 08 shows the importance of different independent variables to predict yield using bar plot. It shows that the model gave most importance to WDRVI, followed by minimum temperature, STVI and then maximum temperature. This gives reasoning to the hypothesis made earlier that the minimum temperature has popped up above others due to even more complex computation of gradient boosted trees instead of simple decision trees. This means that the feature importance given by each model to the input parameters can be seen changing with each model. The most related output features would come from the model that shows better results.

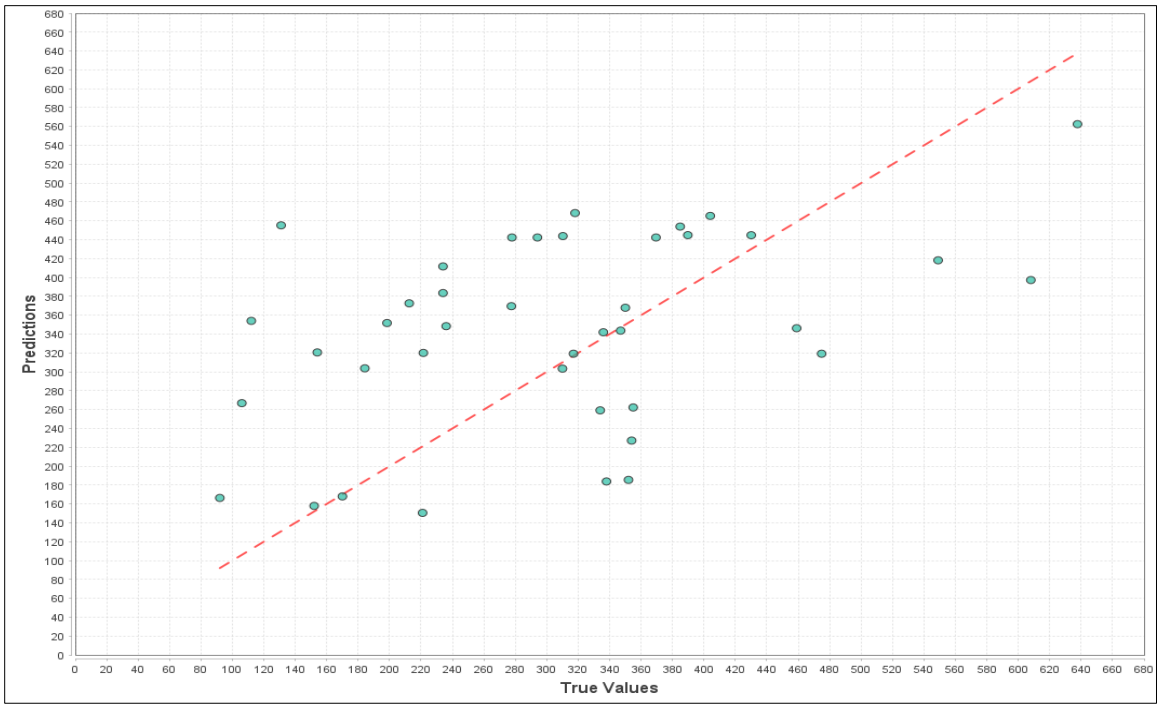


Figure 14: Scatter Plot of GBT Yield Prediction Result

Table 8: GBT Weights Assigned to Independent Variables

Gradient Boosted Trees - Weights

Attribute	Weight
WDRVI	0.143
tasmin	0.083
STVI	0.069
tasmax	0.066
OSAVI	0.030
NDVI	0.004
prec	0.003
SARVI	0.000

4.2.5 Support Vector Machine

The results of Support Vector Machine prediction model are given in Figure 15 which depicts the scatter plot between the predicted yield by the model on Y-axis versus the actual yield recorded on X-axis. The points are more confined in this chart apparently than others.

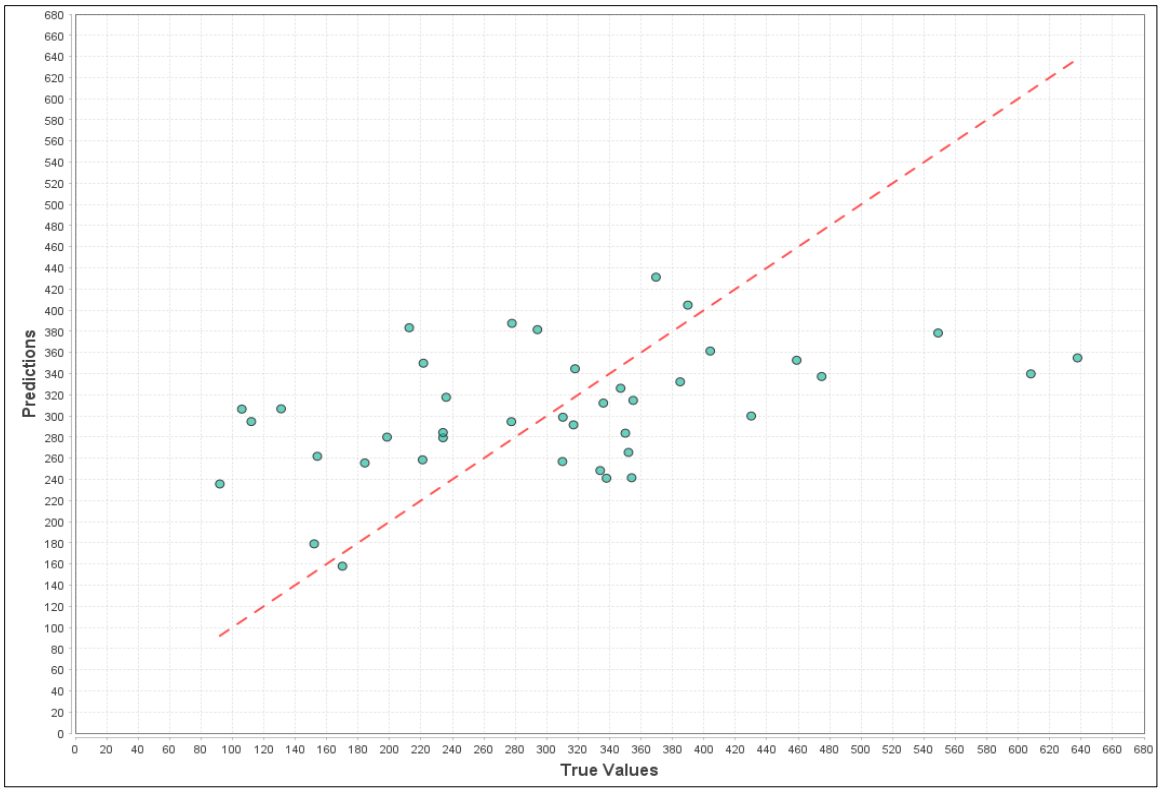
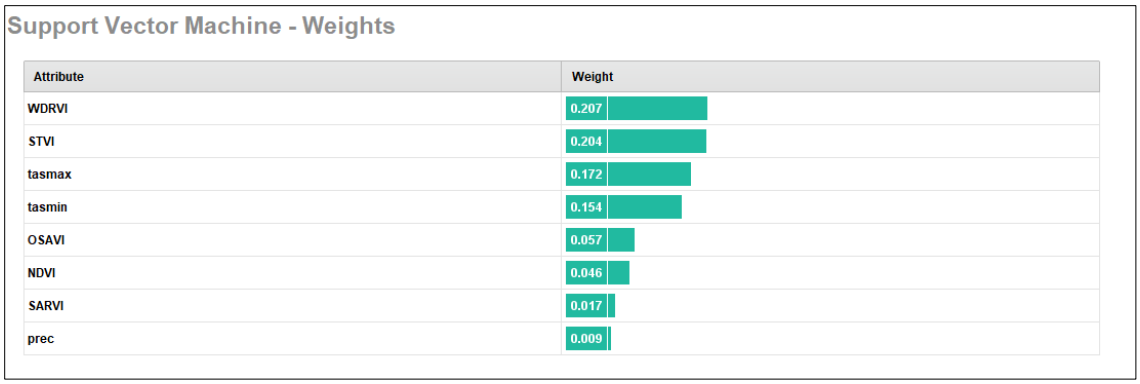


Figure 15: Scatter Plot of SVM Yield Prediction Result

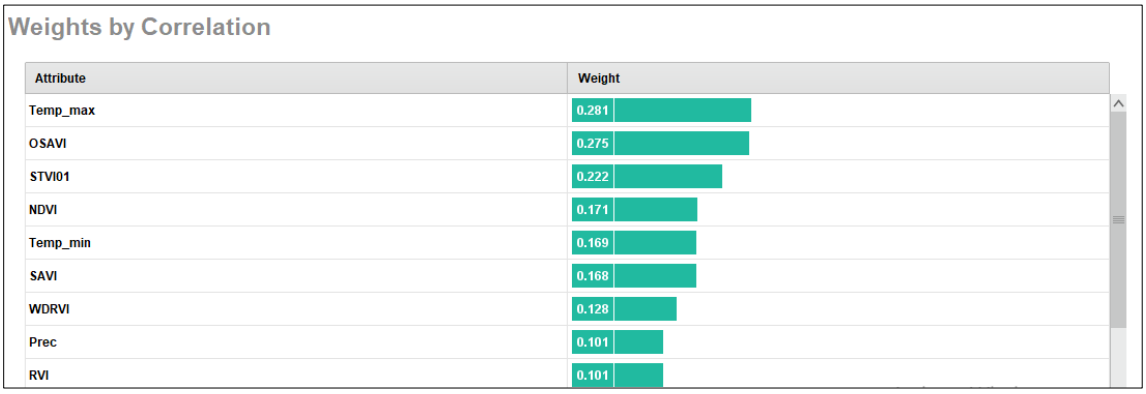
Table 9 shows the importance of different independent variables to predict yield using bar plot. It shows that the model gave most importance to WDRVI and STVI and followed by maximum temperature and minimum temperature. The precipitation has been put down as the study region is irrigated one.

Table 9: SVM Weights Assigned to Independent Variables



The bar graph shown in Table 10 displays the weights assigned to the input variables as per their correlation with the yield. Maximum temperature was given highest weight as cotton yield needs on optimum temperature for its growth over which the flower starts to suffer. OSAVI, STVI and NDVI followed the maximum temperature in their correlation with yield. Precipitation can be seen at the lower end as the study area in this research is an irrigated region which does not depend much on rainwater for its crops.

Table 10: Weights Assigned per Correlation of Independent Variables with Yield



From above scattered plots, there can be seen that the points are highly clustered together near the line in case of Support Vector Machine model, while they are a little less clustered in case of Generalized Linear Model. Automatic Linear Model results follow the GLM in this perspective while in case of Random Forest and Gradient Boosted Trees results, points are loosely packed.

The graph in Figure 16 shows the comparison between results of Automatic Linear Model, Generalized Linear Model, Random Forest model, Support Vector Machine and Gradient

Boosted Trees. It can be observed from the graph that SVM performed much better than all the other models this study experimented with. It has an RMSE of 28.34, with MAE of 22.70 and MBE of just -2.59. It shows it has the least variation from the actual cotton yield values. GLM comes second in this evaluation as it has an RMSE of 29.10, MAE of 23.76 and MBE of -10.0. These values further corroborate the fitness of SVM model with provided features in the given area. Subsequently, ALM undermines the yield prediction by -22.31, GBT by -24.60 and RF does it by -38.73. This shows that those models performed better which do not require huge amounts of features and number of samples.

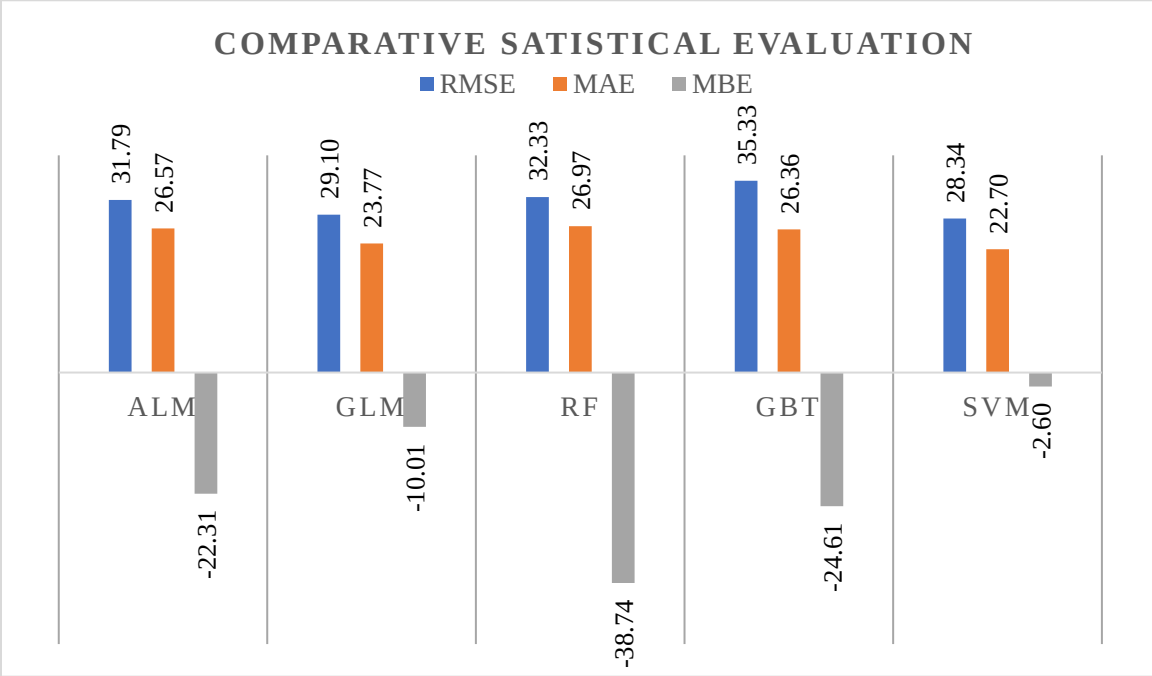


Figure 16:RMSE, MAE and MBE results of ALM, GLM, RF, GBT and SVM

4.3 Future yield estimation using climatic projections and crop yield patterns.

The descriptive statistic of the model shows that the Mean yield was around 305 units while Maximum Temperature was around 42 degrees Celsius. While taking yield as a dependent variable and various climatic factors and vegetation indices as independent variable to

make predictions, the resulting trained model had an R square error of 0.266 and R value of 0.516 which demonstrates that the climatic and vegetation variables used in the model explain only 51% of the variation in cotton yield (Table 11). This suggests that there can be other factors as well that might be impacting cotton output that are outside the purview of this model, and more research can be done to better predict the crop's productivity. The correlation table emphasized the positive and significant correlation between Yield and Maximum Temperature with unstandardized B value of 18.830. Whereas it showed a significant negative correlation of Yield with Precipitation with a standardized coefficients Beta value of -0.182 and a t value of -1.054.

Meanwhile, the ANOVA in Table 12 showed an alpha value or p-value less than 0.05 which indicates that the results obtained from the analysis are significant and trustworthy, and that the model is valuable in explaining the variation in the dependent variable supported by $F(7,92) = 47.64$ and $p = 0.000$. Table 13 shows beta statistics and t-values that shows the weights assigned to predictor variables and significance of coefficients respectively. Furthermore, the mean of residuals is 0.00 in Table 14, indicating that the linear regression model's predicted values are, on average, very close to the actual values. A mean residual of 0 shows that the model is unbiased, with errors distributed at random around zero. This checks that the model fits the data well and accurately captures the link between the dependent variable (yield) and the model's independent variables (climatic conditions and vegetation indices).

The frequency vs regression standardized residual plot also displays a normal distribution, indicating that the residuals are normally distributed, and that the linear regression model fits the data well. Figures 17, 18 and 19 demonstrate that the regression model's errors are random and not biased towards any certain value or direction. It also implies that the model has caught the underlying patterns in the data, and that any remaining variability is due to random fluctuations that the model cannot predict. This fact is supported by the scatter plot between regression standardized residual vs regression standardized predicted values, if this were to be following some pattern it would have mean that the results are biased and not well distributed.

Table 11: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.516[a]	.266	.210	123.7224480

Table 12: Regression and Residual values for ANOVA analysis

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	510462.309	7	72923.187	4.764	.000[b]
	Residual	1408266.461	92	15307.244		
	Total	1918728.769	99			

a. Dependent Variable: Yield

b. Predictors: (Constant), WDRVI, tasmin, prec, STVI, tasmax, SARVI, NDVI

Table 13: Coefficients analysis of all factors

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4675.645	5935.100		.788	.433
	tasmax	18.830	19.184	.175	.982	.329
	tasmin	-17.148	14.451	-.137	-1.187	.238
	prec	-.143	.136	-.182	-1.054	.295
	NDVI	4769.949	7128.085	.681	.669	.505
	SARVI	-5919.018	1905.102	-.941	-3.107	.003
	STVI	.617	.971	.269	.635	.527
	WDRVI	3054.505	5027.834	.366	.608	.545

Table 14: Residual Statistics of input data

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	76.67642212	473.3541565	305.2206875	71.80658002	100
Residual	-375.928986	333.4325562	.0000000000	119.2682428	100
Std. Predicted Value	-3.183	2.341	.000	1.000	100
Std. Residual	-3.038	2.695	.000	.964	100

a. Dependent Variable: Yield

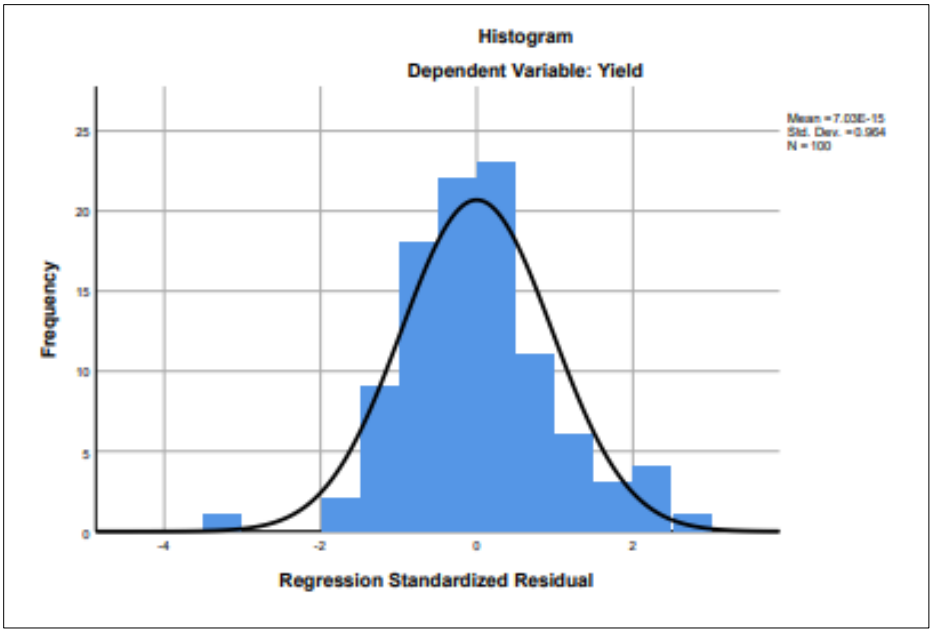


Figure 17:Graph between Regression Standardized Residual and Frequency

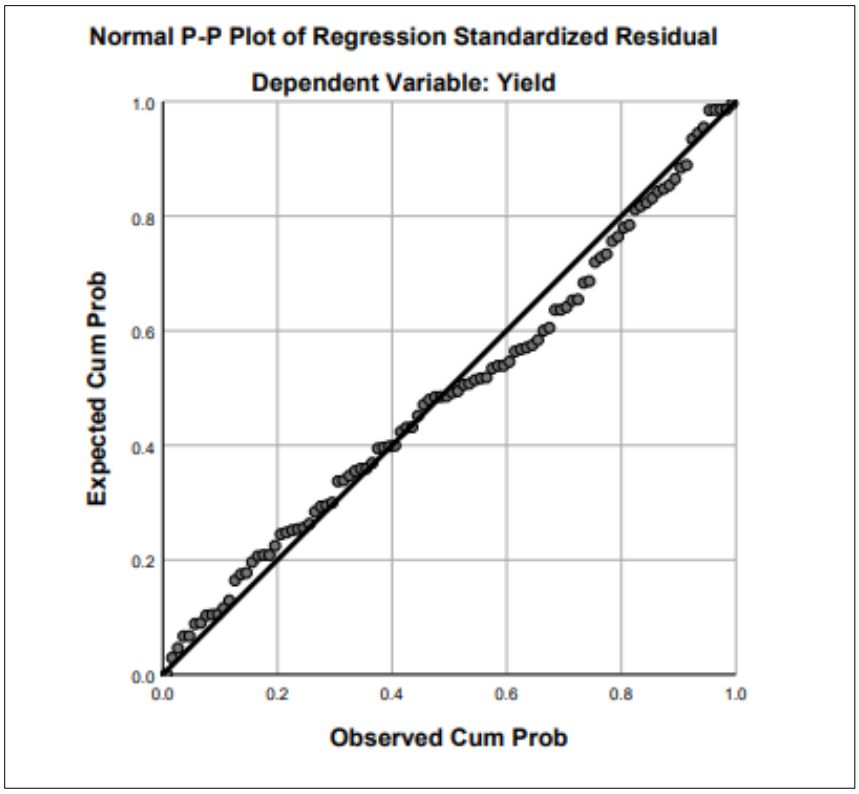


Figure 18: Scatterplot between Observed vs Expected Cumulative Probability

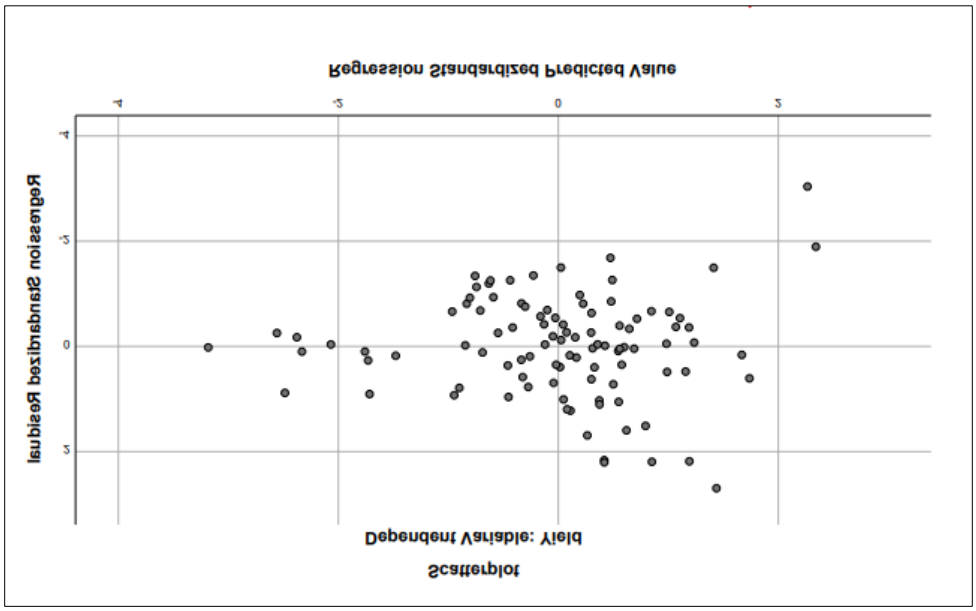


Figure 19: Scatterplot between Regression Standardized Predicted & Residuals

Getting the results from yield prediction model, the yield for Short, Mid, and Long-term periods was plotted under two distinct climatic scenarios, SSP2 4.5 and SSP5 8.5. The graphs revealed interesting ups and downs for both scenarios. It was observed that the yield was higher under SSP2 4.5 in the Short and Mid-term periods in Figure 20 and 21 respectively, which is the setting where Greenhouse Gas productions peak around mid-century and then fall steadily. This scenario results in a 2.4°C addition in global temperature by the last of the century. Conversely, the trends for the SSP5 8.5 yield line were relatively lower, which envisions a society resulting in a 4.8°C surge in global temperature by the end of the century, with severe and potentially disastrous repercussions for the world and its population.

However, the Long-term graph in Figure 22 showed the opposite trend, with the yield being higher under SSP5 8.5. For all three time periods, the general graph showed a wavy trend, with an increase in yield observed in specific years for both scenarios. For example, the yield showed an increasing trend in the years 2028, 2032, 2034, 2036, 2047, 2049, 2053, 2057, 2060, 2069, 2077, and 2087, under both scenarios. However, there were also instances where the curve increased for SSP2 4.5 while decreasing for SSP5 8.5, such as in the years 2028, 2039, 2045, 2054, 2063, and 2095. Moreover, extreme peaks were observed in the years 2035, 2047, 2053, and 2077.

These results emphasize the importance of contemplating the effects of different climatic scenarios on crop yield, as the yield varies significantly depending on the climatic conditions. The findings suggest that reducing greenhouse gas emissions could lead to higher yields in the Short and Mid-term periods, while in the Long-term, the impact could be different.

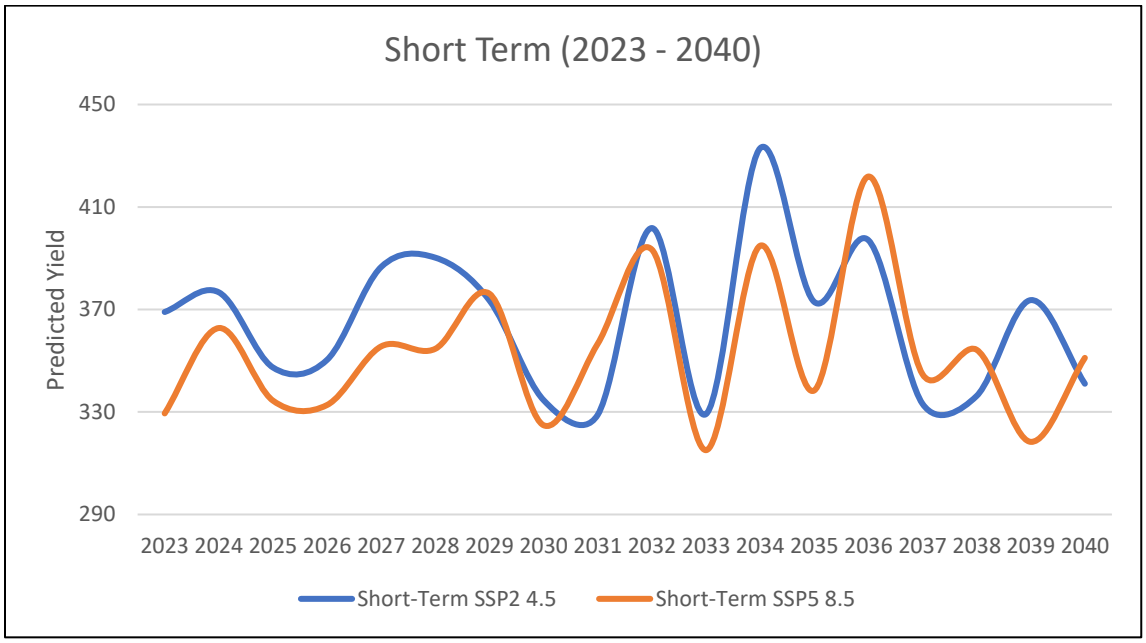


Figure 20: Short Term predicted yield

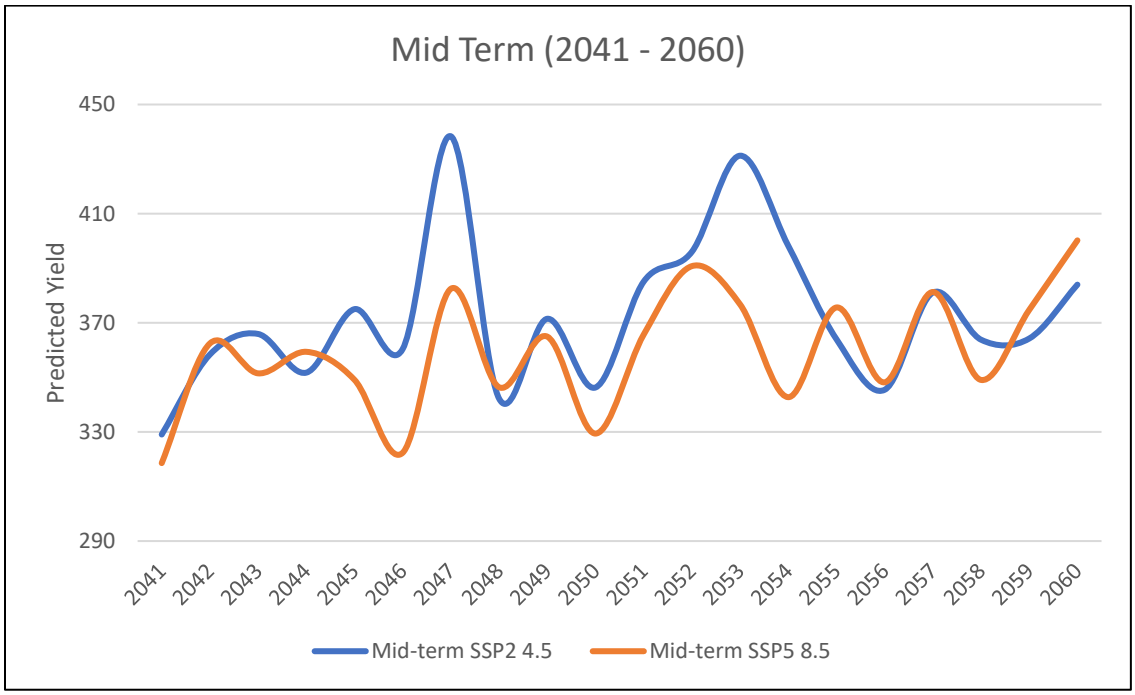


Figure 21: Mid Term predicted yield

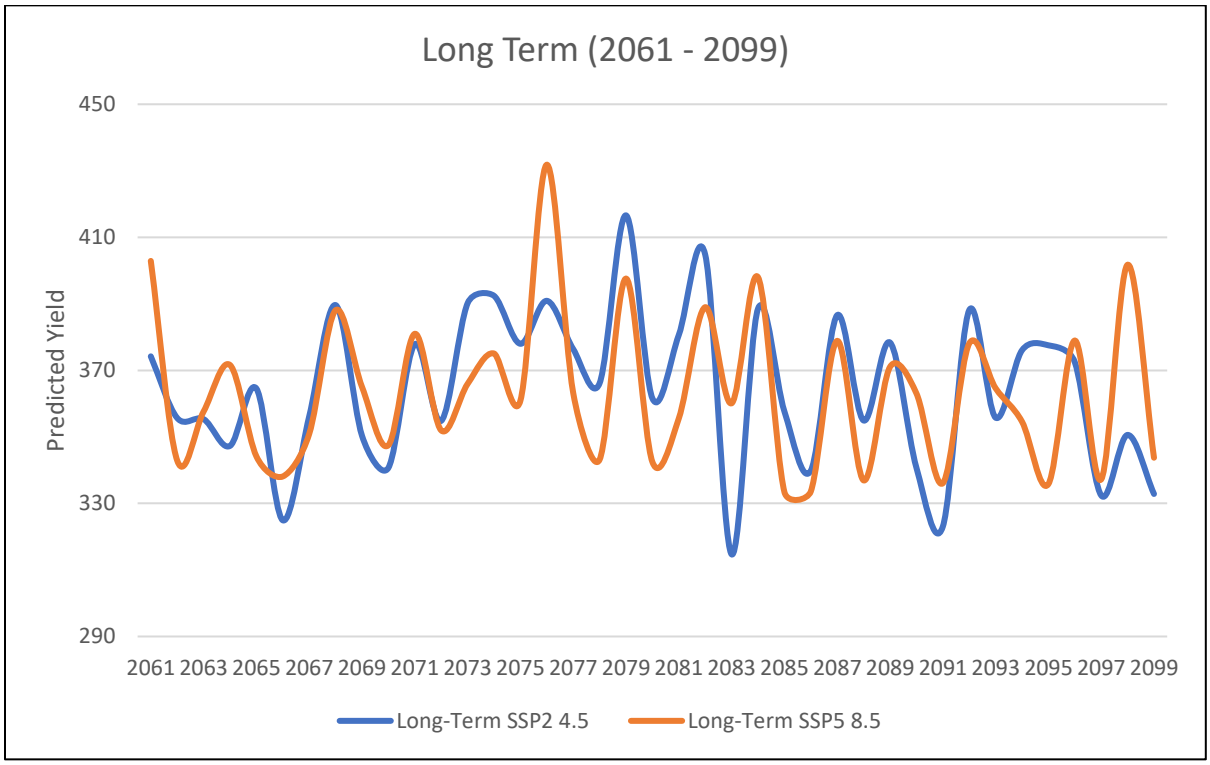


Figure 22: Long Term predicted yield

Chapter 5

CONCLUSION & RECOMMENDATION

5.1 Conclusions

In conclusion, the study demonstrates the practicality of satellite imageries for cotton crops mapping. The use of Sentinel 2A imagery and five vegetation indices ranges enables the mapping of cotton crops with a high degree of accuracy. The approach can be extended to other crops and regions, providing valuable information for agricultural management and decision-making. In addition to crops mapping, the study emphasizes the adaptability and usefulness of Remote Sensing technologies in agriculture. This can solve the problem of accurate and reliable data for crop mapping at lower costs.

The second part of study explains a descriptive approach of integrating Geographic Information System with Remote Sensing and Machine Learning to predict the crop yield. Different machine learning models were applied to the eleven vegetation indices and climate factors including temperature and precipitation to model crop yield. RMSE, MSE and MAE were used to prove that SVM showed the best results for the study area in given data, followed by GLM and then ALM. The study results can be improved by the addition of more features like soil data, additional climate data and other related variables. Farmers and agricultural managers may improve crop yields, cut expenses, and make well-informed decisions regarding sustainable agricultural practices by harnessing GIS integration in agriculture.

The yield prediction model displayed interesting ups and downs under various climatic scenarios, emphasizing the importance of taking climate change into account when predicting crop yield. These findings underline the significance of considering the influence of different climatic scenarios on crop yield, as the yield varies significantly depending on the climatic conditions. The findings suggest that reducing greenhouse gas emissions could lead to higher yields in the Short and Mid-term periods, while in the Long-

term, the impact could be different. This calls for suggestive adaptation strategies customized as per results of this and related studies.

5.2 Recommendations

Implementing the usage of cotton masks based on the Vegetation Index (VI), particularly during the critical month of September, is advised in order to optimize crop management and increase efficiency within the cotton business. This method can enable farmers to spot the locations where their cotton crops are growing poorly and then take the necessary corrective action. Farmers can improve their decision-making skills and produce better crops by using this proactive method.

Machine learning techniques, such as Support Vector Machines (SVM), can help improve the accuracy of cotton yield projections. These models can consider a variety of elements, including weather conditions, historical yield data, and vegetation indices. SVM models can provide more precise estimations of cotton yield by considering these various variables. Using advanced machine learning techniques allows farmers to anticipate probable output variations and adapt quickly, improving profitability and sustainability in the cotton business.

Given the expected effects of climate change on cotton productivity, policymakers and farmers must investigate alternate planting patterns that are more resilient to shifting climatic circumstances. It is advised to explore and promote the production of crops that demand less water and are temperature tolerant. The negative effects of climate change on cotton production can be reduced by diversifying crop choices and prioritizing resilient alternatives. Furthermore, this deliberate shift towards more sustainable cropping patterns benefits to agriculture's overall resilience and lifespan in Pakistan.

REFERENCES

1. Akbar, H., & Gheewala, S. H. (2020). Effect of climate change on cash crops yield in Pakistan. Arabian Journal of Geosciences, 13(11).
2. *Annual Analytical Report on External Trade Statistics of Pakistan FY 2020-21 | Pakistan Bureau of Statistics*. (2020). Pbs.gov.pk. https://www.pbs.gov.pk/publication/annual-analytical-report-external-trade-statistics-pakistan-fy-2020-21
3. Arunrat, N., Sereenonchai, S., Chaowiwat, W., & Wang, C. (2022). Climate change impact on major crop yield and water footprint under CMIP6 climate projections in repeated drought and flood areas in Thailand. Science of the Total Environment, 807, 150741.
4. Azmat, M., Ilyas, F., Sarwar, A., Huggel, C., Vaghefi, S. A., Hui, T., Qamar, M. T. U., Bilal, M., & Ahmed, Z. (2021). Impacts of climate change on wheat phenology and yield in Indus Basin, Pakistan. Science of the Total Environment, 790, 148221.
5. Bannari, A., Asalhi, H., & Teillet, P. M. (n.d.). Transformed difference vegetation index (TDVI) for vegetation cover mapping. *IEEE International Geoscience and Remote Sensing Symposium*. https://doi.org/10.1109/igarss.2002.1026867
6. Bargiel, D. (2017). A new method for crop classification combining time series of radar images and crop phenology information. Remote Sensing of Environment, 198, 369–383.
7. Cetin, O., & Basbag, S. (2010). Effects of climatic factors on cotton production in semi-arid regions—A review. Research on Crops, 11(3), 785-791.
8. Charoen-Ung, P., & Mittrapiyanuruk, P. (2018). Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques. In International Joint Conference on Computer Science and Software Engineering.
9. Charoen-Ung, P., & Mittrapiyanuruk, P. (2018b). Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques. In International Joint Conference on Computer Science and Software Engineering.
10. Cotton Plant Development and Plant Mapping. (n.d.). Extension.missouri.edu. https://extension.missouri.edu/publications/g4268
11. Cotton Plant Development and Plant Mapping. (n.d.). MU Extension.
12. Fang, P., Zhang, X., Wei, P., Wang, Y., Zhang, H., Liu, F., & Zhao, J. (2020). The Classification Performance and Mechanism of Machine Learning Algorithms in Winter Wheat Mapping Using Sentinel-2 10 m Resolution Imagery. Applied Sciences, 10(15), 5075.
13. Faramiñan, A., Rodriguez, P. O., Carmona, F., Holzman, M., Rivas, R., & Mancino, C. (2022). Estimation of actual evapotranspiration in barley crop

through a generalized linear model. *MethodsX*, *9*, 101665. https://doi.org/10.1016/j.mex.2022.101665
14. Fei, H., Fan, Z., Wang, C., Zhang, N., Wang, T., Chen, R., & Bai, T. (2022). "Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier." ("Remote Sensing | Free Full-Text | Cotton Classification Method ... - MDPI") Remote Sensing, 14(4), 829.
15. Feng, S., Zhao, J., Liu, T., Zhang, Z., & Guo, X. (2019). "Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data." ("Mapping corn dynamics using limited but representative samples with ...") IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(9), 3295–3306.
16. Gitelson, A. A. (2004). Wide Dynamic Range Vegetation Index for Remote Quantification of Biophysical Characteristics of Vegetation. *Journal of Plant Physiology*, *161*(2), 165–173. https://doi.org/10.1078/0176-1617-01176
17. Halder, J. C. (2013). Land suitability assessment for crop cultivation by using remote sensing and GIS. *Journal of geography and Geology*, *5*(3), 65.
18. Hassan, S. S., & Goheer, M. A. (2021). Modeling and Monitoring Wheat Crop Yield Using Geospatial Techniques: A Case Study of Potohar Region, Pakistan. Journal of the Indian Society of Remote Sensing, 49(6), 1331–1342.
19. Hazarika, M. K., & Honda, K. (2001). Estimation of soil erosion using remote sensing and GIS: Its valuation and economic implications on agricultural production. Sustaining the global farm, 1, 1090-1093.
20. Huete, A., Didan, K., van Leeuwen, W., Miura, T., & Glenn, E. (2010). MODIS Vegetation Indices. *Land Remote Sensing and Global Environmental Change*, 579–602. https://doi.org/10.1007/978-1-4419-6749-7_26
21. *IDB - Index DataBase*. (n.d.). Www.indexdatabase.de. Retrieved May 9, 2023, from https://www.indexdatabase.de/.
22. Ihuoma, S. O., & Madramootoo, C. A. (2019). Sensitivity of spectral vegetation indices for monitoring water stress in tomato plants. *Computers and Electronics in Agriculture*, *163*, 104860. https://doi.org/10.1016/j.compag.2019.104860
23. Ijaz, M., Zafar, Q., Khan, A. A., & Hassan, S. S. (2022). Assessing drought and its impacts on wheat yield using remotely sensed observations in rainfed Potohar region of Pakistan. Environment, Development and Sustainability. https://doi.org/10.1007/s10668-022-02200-1
24. Imran, A. (n.d.). *Cotton Crop Development in Central Punjab (Faisalabad, 2021)*. Retrieved May 9, 2023, from https://namc.pmd.gov.pk/assets/crop-reports/968773271Model-crop-report-Faisalabad-2021--Cotton.pdf.
25. International Food Policy Research Institute (2022). Ifpri.org. https://www.ifpri.org/publication/cotton-crop-situational-analysis-pakistan

26. *ISLAMIC REPUBLIC OF PAKISTAN COUNTRY ENVIRONMENT ANALYSIS*. (2008). https://www.adb.org/sites/default/files/institutional-document/32193/country-environment-analysis.pdf
27. Jafari, R., Lewis, M. M., & Ostendorf, B. (2007). Evaluation of vegetation indices for assessing vegetation cover in southern arid lands in South Australia. *The Rangeland Journal*, *29*(1), 39. https://doi.org/10.1071/rj06033
28. Kahimba, F., Bullock, P., Ranjan, R., & Cutforth, H. (2009). *Evaluation of the SolarCalc model for simulating hourly and daily incoming solar radiation in the Northern Great Plains of Canada*. Retrieved August 24, 2022, from https://library.csbe-scgab.ca/docs/journal/51/c0818.pdf
29. Krauss, C., Anh, X., DO, & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689–702.
30. Kwak, G., & Park, N. (2019). Impact of Texture Information on Crop Classification with Machine Learning and UAV Images. Applied Sciences, 9(4), 643.
31. Leprieur, C., Kerr, Y. H., Mastorchio, S., & Meunier, J. C. (2000). Monitoring vegetation cover across semi-arid regions: Comparison of remote observations from various scales. *International Journal of Remote Sensing*, *21*(2), 281–300. https://doi.org/10.1080/014311600210830
32. Leprieur, C., Verstraete, M. M., & Pinty, B. (1994). Evaluation of the performance of various vegetation indices to retrieve vegetation cover from AVHRR data. Remote Sensing Reviews, 10(4), 265–284.
33. LP DAAC - MOD13A1. (n.d.). https://lpdaac.usgs.gov/products/mod13a1v006/
34. Morelli-Ferreira, F., Maia, N. J. C., Tedesco, D., Kazama, E. H., Morlin Carneiro, F., Santos, L. B., Seben Junior, G. F., Rolim, G. S., Shiratsuchi, L. S., & Silva, R. P. (2021). Comparison of Machine Learning Techniques in Cotton Yield Prediction Using Satellite Remote Sensing.
35. Moumni, A., & Lahrouni, A. (2021). Machine Learning-Based Classification for Crop-Type Mapping Using the Fusion of High-Resolution Satellite Imagery in a Semiarid Area. Scientifica, 2021, 1–20.
36. Panda, S. S., Ames, D. P., & Panigrahi, S. (2010). Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sensing*, *2*(3), 673–696. https://doi.org/10.3390/rs2030673
37. Panneerselvam, B., Muniraj, K., Thomas, M., Ravichandran, N., & Bidorn, B. (2021). Identifying influencing groundwater parameter on human health associate with irrigation indices using the Automatic Linear Model (ALM) in a semi-arid region in India. Environmental Research, 202, 111778.

38. Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved Random Forest for Classification. IEEE Transactions on Image Processing, 27(8), 4012–4024.
39. Pisner, D., & Schnyer, D. M. (2020). Support vector machine. In Elsevier eBooks (pp. 101–121). Elsevier BV.
40. Priya, S., & Shibasaki, R. (2001). National spatial crop yield simulation using GIS-based crop production model. Ecological Modelling, 136(2-3), 113–129. https://doi.org/10.1016/s0304-3800(00)00364-1
41. Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, *55*(2), 95–107. https://doi.org/10.1016/0034-4257(95)00186-7
42. Sakamoto, T. (2020). Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. Isprs Journal of Photogrammetry and Remote Sensing, 160, 208–228. https://doi.org/10.1016/j.isprsjprs.2019.12.012
43. Singh, R. P., Prasad, P. V. V., Sunita, K., Giri, S. N., & Reddy, K. R. (2007, January 1). Influence of High Temperature and Breeding for Heat Tolerance in Cotton: A Review (D. L. Sparks, Ed.). ScienceDirect; Academic Press. https://www.sciencedirect.com/science/article/abs/pii/S0065211306930065
44. Steven, M. D. (1998). The Sensitivity of the OSAVI Vegetation Index to Observational Parameters. *Remote Sensing of Environment*, *63*(1), 49–60. https://doi.org/10.1016/S0034-4257(97)00114-4
45. Su, Y., Xu, H., & Yan, L. (2017). Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi Journal of Biological Sciences, 24(3), 537–547.
46. Tariq, A., Yan, J., Gagnon, A., Khan, M. R., & Mumtaz, F. (2022). Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. Geo-spatial Information Science, 1–19.
47. Van Klompenburg, T., Kassahun, A., & Rodriguez, D. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.
48. Wu, W. (2014). The Generalized Difference Vegetation Index (GDVI) for Dryland Characterization. *Remote Sensing*, *6*(2), 1211–1233. https://doi.org/10.3390/rs6021211

APPENDIX A

Sentinel 2A Data acquisition and mosaicking using GEE

```
var S1 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point1)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();
var S2 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point2)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();
var S3 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point3)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();
var S4 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point4)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();
var S5 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point5)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();

var S6 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point6)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();

var S7 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point7)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();

var S8 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point8)
    .filterDate('2017-09-15', '2017-09-30')
```

```
    .sort('CLOUD_COVER')
    .first();

var S9 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point9)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();

var S10 =ee.ImageCollection("COPERNICUS/S2_HARMONIZED")
    .filterBounds(point10)
    .filterDate('2017-09-15', '2017-09-30')
    .sort('CLOUD_COVER')
    .first();
var mosaic = ee.ImageCollection.fromImages([S1, S2, S3, S4, S5, S6, S7, S8,
S9, S10]).mosaic();

var FinalImage = mosaic.clip(AOI);
```

Google Earth Engine platform

```
// Loading Reference Cotton Fields (CRS data)
var cottonRef = ee.Table("projects/ee-FYP/assets/Cotton");
// Specify the style for the shapefile
var style = {
  color: 'red',
  width: 2,
  opacity: 0.5
};
// Add the shapefile to the map
Map.addLayer(CottonRef, style, 'cotton');
```

Approach A:

```
// ---------------- Calculate NDVI --------------------------

function getNDVI(image) {
  var evi = image.expression( '((NIR - RED) / (NIR + RED))',
    {'NIR': image.select('B8'),
    'RED': image.select('B4')
    }
  );
```

```
  return evi.divide(2.5);
}

var ndvi = getNDVI(FinalImage).rename('NDVI');

// Add the NDVI layer to the map
Map.addLayer(ndvi, {min: -1, max: 1, palette: ['blue', 'black', 'green']}, 'NDVI');

// NDVI range calculation
var maxNDVI = ndvi.reduceRegion({
  reducer: ee.Reducer.max(),
  geometry: cottonRef,
  scale: 250,
});
var minNDVI = ndvi.reduceRegion({
  reducer: ee.Reducer.min(),
  geometry: cottonRef,
  scale: 250,
});
print("Max NDVI: ", maxNDVI.get("ndvi"));
print("Min NDVI: ", minNDVI.get("ndvi"));


// ---------------------------- MASK ----------------------------
// Set the min and max values for the NDVI range
var minNDVI =  0.421478;
var maxNDVI =  0.51763;
// Create a binary mask where pixels within the specified range are set to 1 and all other
pixels are set to 0
var mask = ndvi.gte(minNDVI).and(ndvi.lte(maxNDVI));
// Create a new image with the mask applied
var classifiedImage = ndvi.updateMask(mask).unmask().where(mask, 1).where(mask.not(),
0).rename('Classified Image');
var studyAreaCottonMask = classifiedImage.clip(AOI);
// Add the classified image to the map
Map.addLayer(studyAreaCottonMask , {min: 0, max: 1, palette: ['black', 'yellow']},
'Classified Image');
```

Approach B:

```
var FinalImage = image.clip(CottonRef);
```

```
//------------------- Indices calculation Functions -------------------

// DVI
function getDVI(FinalImage) {
  return  FinalImage.expression('(NIR/10000)-(RED/10000)',
      {'NIR': FinalImage.select('B8'),
      'RED': FinalImage.select('B4')
      });
}

// NDVI

function getNDVI(image) {
  var evi = image.expression( '((NIR - RED) / (NIR + RED))',
    {'NIR': image.select('B8'),
    'RED': image.select('B4')
    }
  );
  return evi.divide(2.5);
}

// EVI

function getEVI(image) {
  var evi = image.expression( '2.5 * ((NIR - RED) / (NIR + 6 * RED - 7.5 * BLUE + 1))',
    {'NIR': image.select('B8'),
    'RED': image.select('B4'),
    'BLUE': image.select('B2')
    }
  );
  return evi.divide(2.5);
}

// WDRVI
function getWDRVI(image) {
  return image.expression( '(0.1 * NIR - RED) / (0.1 * NIR + RED)',
    {'NIR': image.select('B8'),
    'RED': image.select('B4')
    }
  );
}

// SAVI
function getSAVI(image) {
  return image.expression('1.5 * (NIR - RED) / (NIR + RED + 0.5)',
    {'NIR': image.select('B8'),
    'RED': image.select('B4')
    }
  );
```

```
}


//---------------- Indices Ranges Determination ----------
var dvi = getDVI(FinalImage).rename('DVI');
Map.addLayer(dvi, {min:-1, max:1, palette: ['blue', 'orange', 'green']}, 'DVI');
var savi = getSAVI(FinalImage).rename('savi');
Map.addLayer(savi, {min:-1, max:1, palette: ['blue', 'purple', 'green']}, 'savi');
…….  Repeat for all indices ……


var aoi = cotton;
var maxNDVI = dvi.reduceRegion({
  reducer: ee.Reducer.max(),
  geometry: aoi,
  scale: 250,
});
var minNDVI = dvi.reduceRegion({
  reducer: ee.Reducer.min(),
  geometry: aoi,
  scale: 250,
});
…….  Repeat for all indices ……


print("Max DVI: ", maxDVI.get("dvi"));
print("Min DVI: ", minDVI.get("dvi"));

print("Max NDVI: ", maxNDVI.get("dvi"));
print("Min NDVI: ", minNDVI.get("dvi"));
print("Max EVI: ", maxEVI.get("EVI"));
print("Min EVI: ", minEVI.get("EVI"));
print("Max WDRVI: ", maxWDRVI.get("wdvi"));
print("Min WDRVI: ", minWDRVI.get("wdrvi"));
print("Max SAVI: ", maxSAVI.get("savi"));
print("Min SAVI: ", minSAVI.get("savi"));



// Extracted Range values of each Indice for cotton crops
var dviMin = -0.139139;
var dviMax = 0.180406;
var ndviMin = 0.421478;
var ndviMax = 0.51763;
var wdviMin = 0.065255;
var wdviMax = 0.212198;
var saviMin = 0.321439;
var saviMax = 0.492743;
```

```
var eviMin = -0.240071;
var eviMax = 0.150266;


// Define a function to create a mask for cotton crops based on the vegetation indices
function createCottonMask(image) {
  var ndvi = image.select('NDVI');
  var dvi = image.select('DVI');
  var wdvi = image.select('WDRVI');
  var savi = image.select('SAVI');
  var evi = image.select('EVI');
  // Create a mask for each index based on the cotton crop range
  var ndviMask = ndvi.gte(ndviMin).and(ndvi.lte(ndviMax));
  var dviMask = dvi.gte(dviMin).and(dvi.lte(dviMax));
  var wdviMask = wdvi.gte(wdviMin).and(wdvi.lte(wdviMax));
  var saviMask = savi.gte(saviMin).and(savi.lte(saviMax));
  var eviMask = evi.gte(eviMin).and(evi.lte(eviMax));
  // Combine the masks for all indices using the "and" operator
  var combinedMask =
ndviMask.and(dviMask).and(wdviMask).and(saviMask).and(eviMask);
  // Return the final mask as a binary image
  return combinedMask;
}

// Apply the mask to an image and clip it to the region of interest
var maskedImage = image.updateMask(createCottonMask(image)).clip(regionOfInterest);
// Add the masked image to the map
Map.addLayer(maskedImage, {min: 0, max: 1, palette: ['black', 'white']},
'Cotton Masked Image');


// ---------------- Exporting Cotton Mask ----------------
// Set the parameters for the export
var exportParams = {
  image: studyAreaCottonMask,
  description: 'Classified Image',
  scale: 30,
  fileFormat: 'GeoTIFF',
  maxPixels: 1e13
};
// Export the MASK to Google Drive
Export.image.toDrive(exportParams);
```

MODIS Imagery acquisition and mosaicking

```
// Define function to get start and end dates from user
function getDates() {
  // Display dialog to get start and end dates
  var start = prompt('Enter start date (YYYY-MM-DD)');
  var end = prompt('Enter end date (YYYY-MM-DD)');

  // Check if both dates were entered
  if (start && end) {
    // Return start and end dates as an object
    return {
      start: ee.Date(start),
      end: ee.Date(end)
    };
  } else {
    // Return null if either date is missing
    return null;
  }
}
// Example usage
var selectedDates = getDates();
if (selectedDates) {
  print(selectedDates.start);
  print(selectedDates.end);
} else {
  print('Please enter both start and end dates.');
}
var modis =ee.ImageCollection('MODIS/061/MOD13A1')
      .filterBounds(first)
      .filterDate(selectedDates.start, selectedDates.end)
      .sort('CLOUD_COVER')
      .first();


// ---------------- Clip function ----------------
var image_1 = modis.clip(first);
var image_2 = modis.clip(scnd);
var image_3 = modis.clip(thrd);
var image_4 = modis.clip(frth);
var image_5 = modis.clip(fifth);
var image_6 = modis.clip(sixth);
var image_7 = modis.clip(svnth);
var image_8 = modis.clip(E_1_8);
var image_8_1 = modis.clip(E_2_8);
var image_9 = modis.clip(ninth);
```

```
// ---------------- Mosaicking function ----------------
var Rajanpor = ee.ImageCollection.fromImages([image_1, image_2, image_3]).mosaic();
var DGkhan = image_4;
var Muzzafragh = ee.ImageCollection.fromImages([image_5, image_6, image_7,image_8,
image_8_1]).mosaic();
var Layyah = image_9;


// ---------------- Function to get name based on input ----------------
function getName() {
  var input = parseInt(prompt("Enter 1 for Rajanpor, 2 for DGKhan, 3 for Muzzafragh and 4
for Layyah:"));
  if (input === 1) {
    return Rajanpor;
  } else if (input === 2) {
    return DGkhan;
  } else if (input === 3) {
    return Muzzafragh;
  } else if (input === 4) {
    return Layyah;
  } else {
    return 'Invalid input. Please enter a number between 1 and 4.';
  }
}


// Example usage
var image = getName(); // Get name based on user input
Map.centerObject(first, 8);
var visParams = {min: -0.2, max: 0.8, palette: ['red', 'yellow', 'green']};
```

Indices Calculation:

```
//----------------Compute NDVI----------------
var ndvi_1 = image.normalizedDifference(['B8', 'B4']);
var ndvi = ndvi_1.rename('NDVI');
Map.addLayer(ndvi, visParams, 'NDVI');

// ----------------Compute STVI01----------------
function getSTVI01(image) {
  return image.expression( '(MIR*RED) / NIR',
    {'NIR': image.select('sur_refl_b02'),
    'RED': image.select('sur_refl_b01'),
    'MIR': image.select('sur_refl_b07')
```

```
    }
  );
}
……. Repeat for all indices ……

var MeansOfNDVI = ndvi.reduceRegions({
  collection: AOI,
  reducer: ee.Reducer.median(),
  scale: 250,
});
print("NDVI")
print(ee.Feature(MeansOfNDVI.first()))
var MeansOfSTVI01 = stvio1.reduceRegions({
  collection: AOI,
  reducer: ee.Reducer.median(),
  scale: 250,
});
print("STVI01")
print(ee.Feature(MeansOfSTVI01.first()))
……. Repeat for all indices ……

// ----------------Create a table with the mean values----------------
var table = ee.FeatureCollection([
  ee.Feature(null, {'index': 'DVI', 'median':
MeansOfDVI.reduceColumns(ee.Reducer.median(), ['median']).get('median')}),
  ee.Feature(null, {'index': 'RVI', 'median':
ee.Feature(null, {'index': 'STVI01', 'median':
MeansOfSTVI01.reduceColumns(ee.Reducer.median(), ['median']).get('median')})
]);
……. Repeat for all indices ……

// Print the table to the console
print(table);

// Export the table to Google Drive as a CSV file
Export.table.toDrive({
  collection: table,
  description: 'index_medians',
  fileFormat: 'CSV'
});
```

FYP

ORIGINALITY REPORT

11%	7%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	www.mdpi.com Internet Source	1%
3	Noppol Arunrat, Sukanya Sereenonchai, Winai Chaowiwat, Can Wang. "Climate change impact on major crop yield and water footprint under CMIP6 climate projections in repeated drought and flood areas in Thailand", Science of The Total Environment, 2022 Publication	1%
4	link.springer.com Internet Source	1%
5	Aqil Tariq, Jianguo Yan, Alexandre S. Gagnon, Mobushir Riaz Khan, Faisal Mumtaz. "Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest", Geo-spatial Information Science, 2022	1%