Politecnico di Torino

Laurea magistrale in Ingegneria Per L'Ambiente E Il Territorio
A.a. 2023/2024
Sessione di Laurea Ottobre 2024

Development of a Data-Driven Model for the Estimate of Drought Risk for Agriculture

A Case Study on Maize Crop Yield in Northern Italy

Relatori:

Prof. Jost-Diedrich G. Von Hardenberg

PhD. Bartolo Albanese, Eoliann S.r.l.

Candidata:

Alice Ajassa

Abstract

Climate change is driving significant alterations in weather patterns, particularly in southern Europe and Italy, where the frequency and intensity of droughts are projected to increase [1], [2]. Under the RCP 8.5 scenario, Italy is expected to experience temperature rises of 3°C to 4.5°C by 2100, along with shifts in precipitation patterns that will lead to drier summers and increased drought severity [2], [3]. This thesis focuses on maize cultivation in Italy's highest-producing provinces, aiming to develop a methodology for estimating agricultural damage due to drought. Maize is a crucial crop for food, livestock feed, and industrial uses [4], but it is vulnerable to climate stressors such as drought and heatwaves [3], [5].

The analysis uses simple climatic features, temperature and precipitation, aggregated from high-resolution datasets (VHR-REA_IT for historical data and VHR-PRO_IT for future projections) at a monthly temporal and provincial spatial level. The climatic features span from March to October, covering the growing season of maize. Preprocessing strategies like normalization and standardization were applied, and the model was evaluated using raw, standardized, and normalized features.

The target variable, maize crop yield, sourced from ISTAT, was detrended provincially to remove long-term growth trends and transformed into a yield anomaly to assess deviations due to short-term drought.

Different regression models were evaluated: first of all, hyperparameter tuning was performed using grid search, guided by a custom evaluation metric to minimize overfitting and maximize accuracy. Models with optimal hyperparameters were trained using cross-validation, and the final model was selected based on the best R2 score, provided it achieved a Mean Absolute Scaled Error (MASE) of less than 1.

The model that produced the best results was the k-NN model, applied uniformly across all provinces. Non-preprocessed features led to better performance, suggesting that the preprocessing strategies were not optimal for this dataset. However, despite yielding the best results, the overall level of accuracy was not particularly satisfactory.

The comparison between model-predicted and observed yield anomalies revealed that the predicted distribution was more concentrated around the mean, failing to capture extremes critical for evaluating anomalies. The model performed well in distinguishing between gain and loss, correctly predicting nearly 76% of cases, though it struggled with extremes.

Under the RCP 8.5 scenario, the model's predictions aligned with the literature, showing that several provinces could experience average annual yield losses, with some provinces seeing peaks of up to 4%. However, as with historical data, the predicted distributions were narrower compared to observed values, indicating that the model underestimated yield variability.

The methodology of this study has several important limitations. The main ones include the relatively short time series for maize yield data, which covers only 17 years, limiting the robustness of the results. Additionally, the features used for the analysis may not be the most optimal for accurately predicting maize yield anomalies, limiting the model's performance. External factors, such as market conditions, other climatic events, and irrigation, were not considered in the model, though these factors can have significant impacts on yield.

In conclusion, while the model performed reasonably in distinguishing gain and loss, its overall accuracy was limited, especially in capturing extreme yield variations. A potential improvement would be to transition from a regression model to multiple binary classification models, each based on a threshold of yield anomaly.

Table of Contents

Abstract 2

List of Figures 7

List of Tables 10

1 Introduction 11

1.1 Climate and Drought Projections: Impacts and Modelling Challenges 11

1.1.1 Climate Projections and Scenarios 11

1.1.2 Drought Projections 13

1.1.3 Challenges in Modelling Drought 14

1.2 Maize: Cultivation, Characteristics, and Climate Vulnerabilities 17

1.2.1 Maize Production and Applications 18

1.2.2 Maize General Characteristics 19

1.2.3 Climate Change Impacts on Maize Production 22

1.3 Research Objectives 23

1.4 Contribution and Structure of This Work 23

2 Methodology 25

2.1 Data description 25

2.1.1 Historical Climatic Data: overview and bias assessment 26

2.1.2 Future Projections Climatic Data 30

2.1.3 Maize Cultivation Data 32

2.2 Data preparation and preprocessing 35

2.2.1 Data Engineering and Features Preparation 35

2.2.2 Feature selection 37

2.2.3 Quantitative and Qualitative Data Analysis 38
2.2.4 Features preprocessing 39
2.2.5 Target preparation 40
2.3 Regression Models 43
2.3.1 Linear Regression 43
2.3.2 Random Forest Regressor 44
2.3.3 Extra Trees (Extremely Randomized Trees) 45
2.3.4 k-Nearest Neighbours (KNN) 45
2.3.5 XGBoost Regressor 46
2.4 Models' calibration and validation 46
2.4.1 Hyperparameter Tuning Process 47
2.4.2 Models Evaluation Method and Metrics 48
2.4.3 SHAP Feature Importance Analysis 49
2.5 Inference Procedure: Historical and RCP8.5 Scenarios 50
2.6 Methodology limitations 52
3 Results 54
3.1 Datasets analysis and comparison 54
3.1.1 Maize production data analysis 55
3.1.2 Climatic Variables in the Selected provinces 61
3.1.3 Variables and Target Correlation 63
3.2 Model Evaluation Results 65
3.2.1 Models performance: Metrics 65
3.2.2 KNN Model: Predicted vs. Observed Data 67
3.2.3 Feature Importance in KNN Model 73
3.3 Inference Results on RCP 8.5 Projection Using KNN 75

4 Results Discussion and Conclusions 78

4.1 Recommendations for Future Work 79

4.2 Conclusions..... 81

Reference List 83

List of Figures

Figure 1:Yearly maize production (1961-2022) in Southern Europe (orange) and Italy (blue), linear trends show a production increase. Source: FAOSTAT 18
Figure 2: Maize schematic growth periods. Source: FAO, Crop Information, Maize..... 20
Figure 3 Seasonal spatial distribution of 2 m temperature (period 1989–2020) of E-OBS, ERA5, and VHR-REA_IT on their relative grids (first three columns). In the last two columns: seasonal spatial distribution of 2 m temperature bias with ERA5, and VHR-REA_IT. Source: Figure 2 from Raffa et al. [35] 28
Figure 4: Seasonal spatial distribution of precipitations (period 1989–2020) of E-OBS, ERA5, and VHR-REA_IT on their relative grids (first three columns). In the last two columns: seasonal spatial distribution of 2 m temperature bias with ERA5, and VHR-REA_IT. Source: Figure 3 from Raffa et al. [35] 29
Figure 5: Maize cultivated area and production in Turin province. 32
Figure 6: Maize cultivated area and production in Treviso province. 33
Figure 7: Maize cultivated area and production in Perugia province. 33
Figure 8: Maize cultivated area and production in Foggia province. Notice vertical axes scale change compared to previous provinces (Figure 5, Figure 6, Figure 7). Some data are missing. 34
Figure 9: Maize cultivated area and production in Foggia province. Notice vertical axes scale change compared to previous provinces (Figure 5, Figure 6, Figure 7). Some data are missing. 34
Figure 10: average provincial maize production records per macroregions (NUTS1)..... 55
Figure 11: average provincial maize cultivated area per macroregion (NUTS1)..... 55
Figure 12: Italy map with provinces selected for the final dataset (green) 57
Figure 13: Crop yield trend (slope and p-value reported) for Alessandria, Cuneo, Reggio Emilia, Lodi, Trieste, Vercelli provinces..... 58
Figure 14: Significant Crop Yield trends comparison with p-value and confidence interval 59
Figure 15: Detrended crop yield anomaly distribution for selected provinces 60

Figure 16: Monthly precipitation comparison, Historical (1982-2022) vs Projection (2030-2070) mean and standard deviation 61
Figure 17: Monthly temperature comparison, Historical (1982-2022) vs Projection (2030-2070) mean and standard deviation 62
Figure 18: Variables-Target Pearson correlation..... 63
Figure 19: Variables-Target Spearman correlation 64
Figure 20: Frequency distribution of Model Predictions with density curve. The histogram and density curve illustrate the predicted values from the KNN model, with most values clustered near 0%..... 68
Figure 21: Cumulative Frequency of Model Predictions. The plot shows the cumulative frequency of predicted values from the KNN model, with most predictions concentrated around the centre 68
Figure 22: Frequency distribution comparison: Predictions (blue) vs Observed Values (red). The predicted values are tightly centred, while the observed values show more spread. 69
Figure 23: Cumulative Frequency Comparison: Predictions (blue) vs Observed Values (red). The predicted values are concentrated around 0%, while the observed values display greater variability..... 70
Figure 24: Predicted vs. Observed Yield Anomalies: the scatter plot compares predicted yield anomalies with observed values, with the red dashed line representing a perfect fit. While a weak correlation is evident, the model underestimates extreme values and struggles to capture the full variability of the observed data..... 71
Figure 25: Confusion Matrix [%] of Predicted vs Observed Yield Anomalies. The matrix shows the model's accuracy in predicting positive (gain) and negative (loss) anomalies, with more errors in false positives 72
Figure 26: KNN Feature Importance (SHAP values)..... 73
Figure 27: SHAP Summary Plot for KNN, showing the impact of each feature on yield anomaly predictions. Higher precipitation values (magenta) in June and July have the greatest influence. 74

Figure 28: Frequency distribution of predicted yield anomalies for the periods 2030-2050 and 2050-2070 under the RCP 8.5 scenario 76
Figure 29: Average annual yield anomaly for selected provinces under the RCP 8.5 scenario, centered around 2040 and 2060 77

List of Tables

Table 1: Models' hyperparameters involved in tuning process 47

Table 2: Summary of key statistical measures for the detrended crop yield anomaly distribution, including standard deviation, interquartile range (IQR), range, and 95% confidence interval, providing insights into the spread and variability of the data. 60

Table 3: Comparison of R^2 and RMSE values across different regression models using original data preprocessing. The R^2 column includes the 90% confidence intervals, highlighting the predictive accuracy of each model. 66

Table 4: Hyperparameter tuning results for Random Forest, Extra Trees, and K-Nearest Neighbours (K-NN) models that achieved MASE < 1. The selected hyperparameters and corresponding 66

Table 5: Summary of Distribution Measures for Model Predictions. The table shows the standard deviation, interquartile range, range, and 95% confidence interval of the predicted values from the KNN model 67

Table 6: Summary of key statistical measures for RCP8.5 scenario; 2030-2050 and 2050-2070 periods comparison 76

1 Introduction

The increasing frequency and severity of droughts, exacerbated by climate change, pose significant challenges to agricultural systems worldwide. As global temperatures rise and precipitation patterns shift, the agricultural sector must grapple with the implications of these changes, particularly in regions where crops are vulnerable to water scarcity. In this context, the ability to predict and quantify agricultural damage resulting from drought becomes crucial. Understanding the potential impacts of climate change on crop yields is essential for developing effective adaptation strategies that can mitigate the economic and food security risks associated with such climatic events [3].

1.1 Climate and Drought Projections: Impacts and Modelling Challenges

This section explores the critical implications of climate change for agriculture, with a particular focus on how shifting climate patterns, especially in southern Europe and Italy, are expected to intensify droughts. It reviews the projected changes in climate scenarios and the complexities involved in modelling agricultural drought, providing a foundation for understanding the climatic challenges that agricultural systems will face in the coming decades.

1.1.1 Climate Projections and Scenarios

Climate change is having widespread impacts on global weather patterns, including temperature increases, shifts in precipitation, and an increase in the frequency of extreme weather events such as droughts and heatwaves. These changes are driven by human activities, particularly the emission of greenhouse gases, which have contributed to a measurable rise in global temperatures [6]. According to the IPCC's Sixth Assessment Report [7], global surface temperatures have risen by approximately 1.1°C above pre-industrial levels, and continued warming is expected unless significant mitigation efforts are undertaken [7].

To project future climate outcomes, the Representative Concentration Pathways (RCPs) are commonly used. These scenarios represent different trajectories of greenhouse gas concentrations and their resulting radiative forcing on the climate system. The RCP 2.6 scenario represents an ambitious mitigation pathway, where global temperatures are projected to rise by less than 2°C by 2100, assuming aggressive cuts in emissions. At the other end of the spectrum, RCP 8.5 represents a "business-as-usual" scenario, where emissions continue to rise, potentially leading to global temperature increases exceeding 4°C by the end of the century [1], [3]. Other intermediate scenarios, such as RCP 4.5 and RCP 6.0, represent more moderate increases, where emissions peak mid-century and gradually decline [7].

Italy, positioned in the Mediterranean Basin, is especially vulnerable to climate change. The Mediterranean is a recognized climate change "hotspot" because it is projected to experience higher-than-average warming compared to other global regions [2]. Under the RCP 8.5 scenario, temperatures in Italy are expected to rise by 3°C to 4.5°C by 2100, with northern regions such as the Po Valley being particularly affected. Even under more moderate scenarios like RCP 4.5, Italy could experience temperature increases of between 2°C and 3°C [2]. These rising temperatures are expected to intensify the frequency of extreme heat events, leading to challenges for both ecosystems and human infrastructure [1], [2].

The Po Valley is highly vulnerable to changing precipitation patterns. This region may experience increased winter rainfall, especially in the areas near the Alps, leading to a heightened risk of flooding. In contrast, summers are expected to become much drier. Studies indicate that in the Po Valley, precipitation patterns could shift toward wetter winters and much drier summers, intensifying seasonal water shortages and placing additional stress on agricultural systems dependent on summer rainfall [8], [9], [10]. This pattern of drying summers and wetter winters in northern Italy aligns with projections for the broader Mediterranean region, where RCP 8.5 scenarios predict a decrease in annual rainfall of up to 20% in southern regions. For the Po Valley, while winter precipitation may

slightly increase, the drier summers could still result in overall water scarcity during the growing season [11].

The combination of higher temperatures and altered precipitation patterns in the Po Valley will have profound implications for Italian agriculture. This region produces a significant portion of the country's crops, including water-intensive varieties such as maize and rice. The projected changes in seasonal water availability will require adaptation strategies, particularly in water management and irrigation, to mitigate the impacts of water shortages during critical growth periods [8], [11], [12].

1.1.2 Drought Projections

Projections of drought frequency and intensity show that climate change will lead to an increase in both the duration and severity of drought events across Europe, with the Mediterranean region, including Italy, bearing the brunt of these impacts. Studies project that, under the *RCP 8.5 scenario*, southern Europe is likely to experience an increase in agricultural drought frequency by as much as 30% by the mid-21st century [13]. Drought risks are not solely driven by changes in precipitation; rising temperatures contribute to increased evaporation, further straining water resources, particularly during the summer growing season [1], [10].

For Italy, the impacts of climate change on drought are already visible. The droughts of 2022 and 2023 affected large parts of Italy, with the agricultural sector suffering significant yield reductions in key crops like maize [14], [15]. According to the *European Drought Risk Atlas* [3], agricultural droughts are expected to increase in severity, particularly for non-irrigated crops. In the Po Valley the reduction in soil moisture due to reduced rainfall and increased evaporation will likely lead to yield losses of up to 10%, with southern regions facing even steeper declines in agricultural productivity [1], [3], [13]. Projections from the *European Drought Risk Atlas (2023)* show that Italy's agricultural drought risk under the RCP 8.5 scenario will significantly increase if no adaptive measures are taken [10].

The complexities of these projections are further highlighted by the fact that agricultural drought risk is not uniform across Italy. For instance, Po basin agriculture depends heavily

on water from the Po River, which is fed by both rainfall and snowmelt from the Alps [16]. As temperatures rise, the reduction in snow accumulation, along with increased evaporation, threatens to diminish this critical water source, exacerbating drought conditions during the critical growing months [3], [13], [14]. This highlights the importance of integrated water management strategies that account for both local and regional climatic conditions [3].

The *European Drought Risk Atlas* emphasizes that addressing drought risk requires a holistic approach that takes into account not only the direct effects of water scarcity on crops but also the broader socio-economic impacts, including the competition for water resources between agriculture, industry, and urban areas. In Italy, where competition for water is already high, particularly in the Po Valley, the increasing frequency of droughts could exacerbate tensions between sectors, further complicating the management of water resources [3], [9]. Moreover, this competition may limit the effectiveness of irrigation as an adaptation strategy, underscoring the need for more sustainable water use practices and innovations in drought-resistant crop varieties [3], [9].

Finally, drought projections indicate that, under climate change, the time between drought events is shrinking, reducing the recovery period for both agricultural systems and water resources [1]. This “shrinking recovery window” poses a significant threat to Italy’s agricultural sector, particularly for crops that are highly sensitive to water stress during critical growth stages [3], [14], [15].

1.1.3 Challenges in Modelling Drought

Modelling drought presents numerous challenges due to the complexity and variability of drought events. One of the primary difficulties arises from the infrequency of extreme droughts, making it harder to gather comprehensive data and establish reliable patterns. Droughts are inherently multifaceted phenomena, classified into various types based on their primary effects: meteorological drought (precipitation deficit), agricultural drought (soil moisture deficit impacting crops), and hydrological drought (deficiency in water resources such as rivers, lakes, and groundwater). Each type emphasizes different aspects of water scarcity, adding layers of complexity to the modelling process [3], [17].

The inherent variability of drought duration and severity further complicates the task of drought prediction and management. For instance, droughts can last anywhere from a few weeks to several years, and their onset is typically slow and difficult to predict [18]. This makes the precise identification of when a drought begins or ends a challenge, often dependent on which drought indicator or index is used. The lack of a universally accepted metric for drought adds to the difficulty of producing consistent predictions across different regions [10], [19].

In terms of agricultural drought, which this research focuses on, the challenge lies in accurately predicting crop yield losses resulting from moisture deficits [17], [18]. Agricultural drought is often the result of a complex interplay between weather conditions, soil properties, and crop-specific vulnerabilities [11]. Factors such as soil moisture retention capacity, farming practices, and the stage of crop growth all influence how drought affects agricultural output [3], [11]. This complexity requires a highly tailored modelling approach, as general drought models may not adequately capture the specific risks to different crops or regions [18], [19].

Furthermore, agricultural production losses are influenced by other climatic factors beyond soil moisture deficits, such as waterlogging and heat waves. For example, excessive rainfall leading to waterlogging can reduce crop yields, especially when soil is saturated, impeding root function and leading to oxygen deprivation [20]. Similarly, heat waves, which are increasing in frequency due to climate change, can exacerbate water stress and cause heat damage to crops. These factors illustrate the complexity of drought modelling, where multiple climatic stressors intersect and influence agricultural outcomes [21].

Moreover, the difficulty in defining agricultural drought is compounded by the use of different drought indices, such as the Standardized Precipitation Evapotranspiration Index (SPEI) and Soil Moisture Deficit Index (SMDI), each of which measures drought impacts from different perspectives [21]. The choice of index significantly influences the outcomes of the model, and even small changes in index thresholds can result in vastly different predictions for the severity of drought impacts [11].

Finally, the integration of climate projections into drought models introduces further complexity. As climate change alters precipitation patterns and temperature dynamics, predicting future drought risks becomes more uncertain. The European Drought Risk Atlas (2023) stresses that while general trends can be predicted, such as an increase in drought severity in southern Europe, regional variations in climate change impacts make it difficult to produce precise, localized forecasts. This uncertainty emphasizes the need for models that can flexibly adapt to new data as climate conditions evolve, allowing for ongoing recalibration of predictions [3].

1.1.3.1 Influence of Irrigation in Agricultural Drought Modelling

In the context of agricultural drought, irrigation plays a pivotal role in mitigating the impacts of water scarcity on crop production. However, incorporating irrigation into agricultural drought models presents unique challenges, largely due to the complexity of irrigation practices, the variability in data availability, and the need to account for the interaction between human and natural systems.

One of the major challenges in modelling agricultural drought with irrigation is the limited availability of reliable data. Unlike meteorological variables such as precipitation and temperature, which are widely monitored and recorded, data on irrigation practices are often sparse or inconsistent. In many regions, including Europe, irrigation data is not systematically collected at the necessary spatial and temporal resolutions. Even when such data exists, it is often aggregated at a regional or national level, obscuring important variations in irrigation intensity, frequency, and water sources used by individual farms [3], [22].

In Italy, where irrigation is a critical adaptation tool for maintaining maize yields, the lack of detailed, farm-level irrigation data complicates efforts to model the full extent of drought resilience. Irrigation systems vary widely in efficiency, from traditional surface irrigation to more advanced drip irrigation techniques. Drip irrigation, for example, significantly reduces water losses and improves water use efficiency compared to flood irrigation, but data on the adoption rates and effectiveness of these systems are often unavailable [3], [22]. As a result, models that fail to differentiate between irrigation

methods may either overestimate or underestimate the extent to which irrigation can mitigate drought impacts.

The timing of irrigation is another crucial aspect to consider. Water stress during sensitive growth stages, such as pollination and grain filling, can cause significant yield reductions, even if irrigation is applied later in the season. Modelling the effectiveness of irrigation thus requires accurate data not only on the amount of water used but also on when it is applied in relation to the crop's phenological stages. The variability in irrigation timing and its impact on crop growth further complicates drought modelling [3], [23].

Another important factor in modelling irrigation is the dynamic relationship between irrigation and water resources. While irrigation can provide short-term relief during droughts, it also contributes to long-term water scarcity, especially in regions where groundwater extraction is unsustainable. In the Mediterranean region, including Italy, where water resources are already stressed, increased irrigation during prolonged dry periods can exacerbate water shortages. This creates a feedback loop that can further deplete water resources and increase vulnerability to future droughts [3], [11].

Finally, irrigation modelling must also consider the future viability of irrigation as an adaptation strategy in the face of climate change. Projections from the *European Drought Risk Atlas* suggest that, under the RCP 8.5 scenario, water availability in southern Europe, including Italy, will decrease significantly, which could limit the ability of farmers to rely on irrigation during future droughts. As temperatures rise and rainfall becomes more uneven, competition for water resources between agriculture, industry, and urban areas is expected to increase, further complicating the use of irrigation as a drought mitigation tool [3].

1.2 Maize: Cultivation, Characteristics, and Climate Vulnerabilities

This section reviews the specific vulnerabilities of maize cultivation in the context of climate change. It examines the crop's sensitivity to environmental stressors such as drought and heatwaves, detailing how these factors affect maize yields. Additionally, it

highlights the broader socio-economic implications of yield variability and discusses potential adaptation strategies to mitigate climate risks.

1.2.1 Maize Production and Applications

Maize (*Zea mays*) stands as one of the world's most significant crops [4], with a wide range of applications spanning from food production to industrial uses. It holds a central position in global agriculture due to its versatility and adaptability [24]. As a staple food crop, maize serves as a crucial source of calories and nutrients for millions of people, particularly in developing countries. In many regions, it forms the foundation of traditional diets, consumed in various forms such as flour, cornmeal, and oil [24].

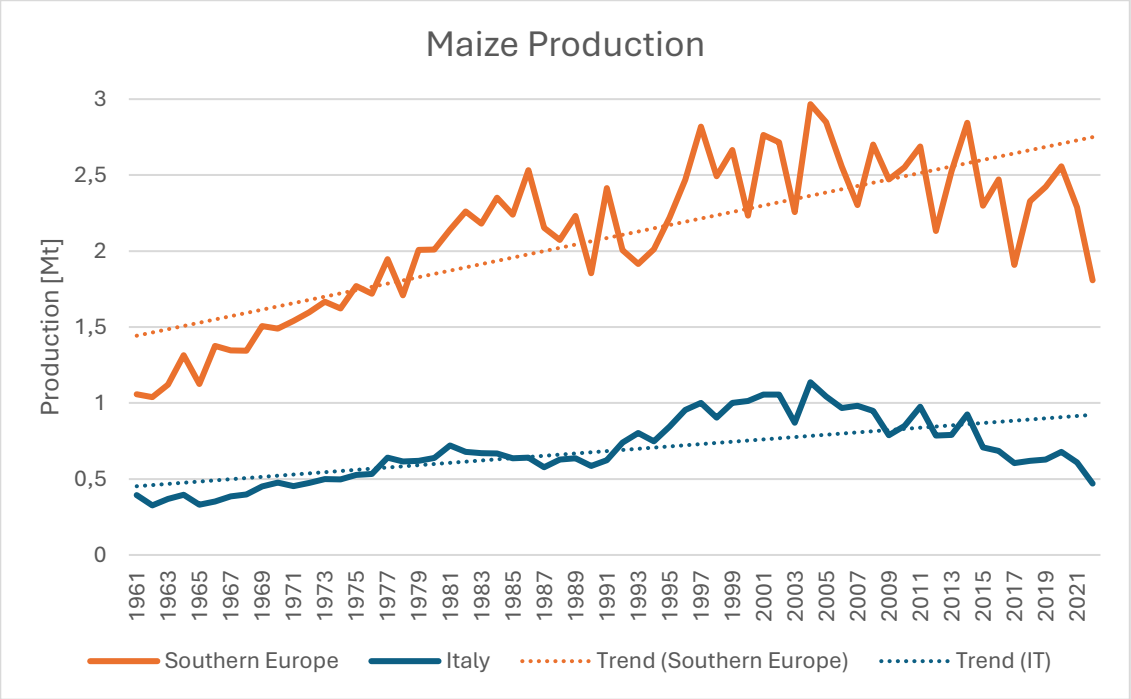


Figure 1:Yearly maize production (1961-2022) in Southern Europe (orange) and Italy (blue), linear trends show a production increase. Source: FAOSTAT

Globally, maize production has been on a steady rise, driven by increasing demand for food, animal feed, and industrial uses. According to a report by the FAO, maize production has grown significantly over the past few decades and continues to expand, particularly in developing countries where it plays a vital role in food security and economic

development [25]. This growth is fuelled by advancements in agricultural technology, improved seed varieties, and the expanding demand for biofuels and livestock feed, making maize a cornerstone of both traditional and modern agricultural systems [25].

Beyond its role in human nutrition, maize is vital in animal husbandry, where it is commonly used as feed for livestock. Its high-energy content makes it an efficient feedstock, particularly for cattle, swine, and poultry. The crop's ability to grow in diverse climates and its adaptability to different soil conditions have made it a preferred choice for forage production globally, further solidifying its importance in the agricultural sector [4], [24]. According to recent studies, maize ranks as one of the most important crops worldwide, both in terms of acreage and economic value. Its contributions to global food security and agricultural sustainability cannot be overstated [4], [24].

In addition to its uses in food and feed, maize is increasingly recognized for its potential in industrial applications. It serves as a primary feedstock for biofuel production, especially ethanol. The rise in demand for renewable energy sources has accelerated the expansion of maize cultivation dedicated to biofuels, particularly in regions like the United States and parts of Europe. As concerns about climate change and energy security grow, maize-based biofuels offer an alternative to fossil fuels, contributing to the global shift towards more sustainable energy solutions [24].

1.2.2 Maize General Characteristics

Maize is a versatile crop, known for its ability to adapt to a wide range of climatic and soil conditions. Maize utilizes a specialized photosynthesis process called C4 photosynthesis, which allows it to efficiently capture carbon dioxide and convert it into energy even under high temperatures and low water availability [25]. This adaptation makes maize highly productive in environments with sufficient warmth and sunlight. The crop thrives in temperatures ranging from 20°C to 30°C and can grow in a variety of soils, though it performs best in fertile, well-drained soils with a neutral pH [25]. Maize is sensitive to frost, which limits its cultivation in colder regions, particularly during early growth stages. Its broad adaptability has contributed to its global expansion, where it is cultivated extensively for both human consumption and industrial uses [25].

The life cycle of maize is divided into distinct growth phases that determine the final yield. Germination and emergence occur shortly after sowing, provided soil temperatures are above 10°C. During this phase, the seed absorbs water, and the root system begins to establish. As the plant enters the vegetative growth stage, rapid leaf and stem development occurs, allowing the plant to form a canopy that captures sunlight. This stage is critical for maximizing photosynthesis, which supports grain development later on. The flowering phase, also known as the reproductive stage, marks the formation of tassels (male flowers) and silks (female flowers). This period is highly sensitive to water stress, as pollination occurs, and any deficit in water availability can significantly reduce kernel formation. Following successful pollination, maize enters the grain-filling stage, where the kernels develop and accumulate starch, directly influencing yield. Finally, the plant reaches physiological maturity when the kernels have accumulated maximum dry matter, at which point the plant ceases growth [25].

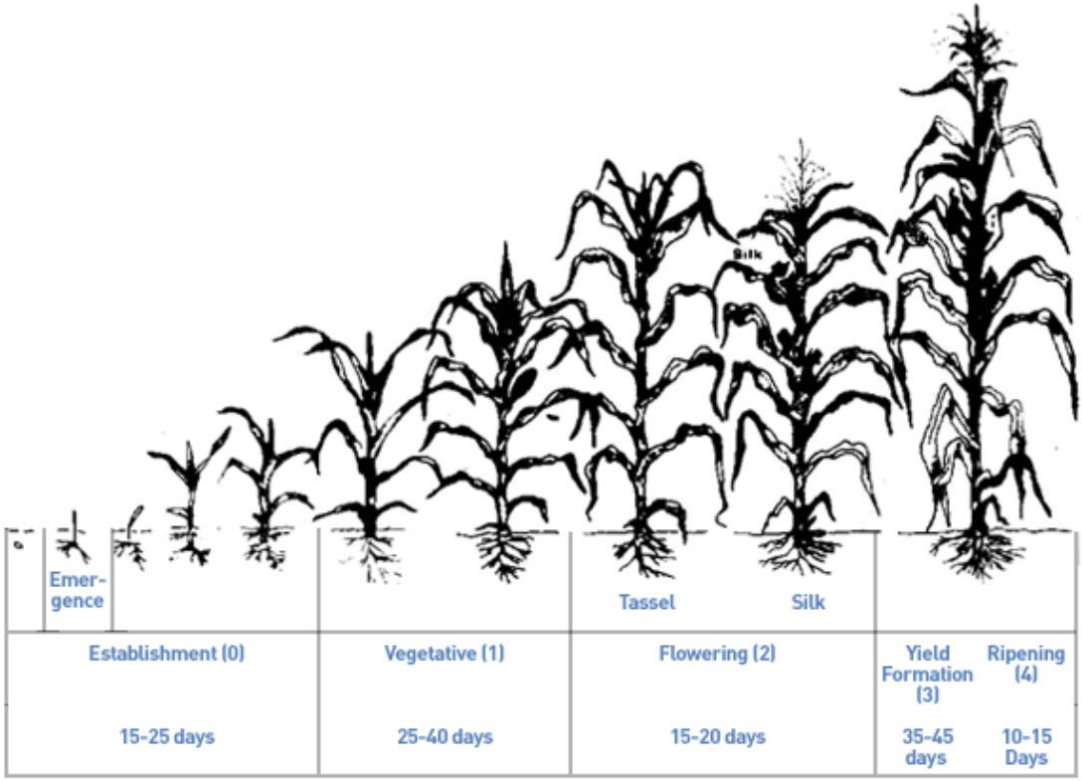


Figure 2: Maize schematic growth periods. Source: FAO, Crop Information, Maize

The duration of maize's growth cycle is influenced by temperature, with faster growth occurring in warmer climates. When average daily temperatures during the growing season exceed 20°C, early maturing varieties typically complete their growth cycle in 80 to 110 days, while medium-maturing varieties may take between 110 and 140 days. For varieties grown as a vegetable, such as baby corn, the maturation period is shorter by approximately 15 to 20 days. The duration of each phase and the overall growing season is determined by various factors, including environmental conditions and the maize variety used. Appropriate water management, particularly during the flowering and grain-filling stages, is essential to achieving optimal yields [25].

Maize's water requirements are substantial, with the crop needing between 500 and 800 mm of water throughout its growing cycle. Water stress during the flowering and grain-filling stages is particularly detrimental, as it can prevent kernel formation and reduce grain size, leading to lower yields. In contrast, excessive water, particularly in poorly drained soils, can restrict root oxygen, further impacting plant growth. Efficient water management, including timely irrigation, is crucial for achieving optimal yields, especially in regions with variable rainfall patterns [25].

Sowing dates play a critical role in the success of maize cultivation, aligning the growth cycle with favourable climatic conditions. Maize sowing can be categorized into early, normal, and late sowing periods, each with distinct impacts on crop development and yield [26]. Early sowing is typically done when soil temperatures have reached the minimum threshold (above 10°C) to encourage rapid germination and ensure that the plant avoids heat stress during flowering. This method can increase the yield potential, but it carries a risk of exposure to frost. Normal sowing is usually carried out when the risk of frost has passed, providing the crop with favourable growing conditions for optimal leaf and root development. Late sowing, on the other hand, occurs after the optimal window and often results in lower yields due to heat stress during pollination and a shorter grain-filling period. In Southern Europe, including Italy, maize is generally sown between April and May to take advantage of early summer rainfall, while in tropical regions, staggered sowing throughout the year allows for multiple harvests [26].

1.2.3 Climate Change Impacts on Maize Production

Climate change presents a complex set of challenges for maize production, with the effects varying depending on specific climate stressors such as water excess, heat waves, and drought. Each of these factors has the potential to significantly disrupt the growth cycle of maize and affect yield outcomes. The increasing frequency of intense rainfall and flooding due to climate change can result in waterlogging, a condition where excess water saturates the soil, preventing roots from receiving sufficient oxygen. Waterlogged conditions impede the plant's ability to absorb essential nutrients, leading to stunted growth, poor kernel development, and reduced overall productivity. Moreover, water excess during critical growth periods, such as the grain-filling stage, can significantly reduce yield potential by restricting photosynthesis and increasing the prevalence of diseases that thrive in overly moist environments [3], [20], [27].

Simultaneously, heat waves are becoming more frequent and severe, placing additional strain on maize cultivation. High temperatures, particularly during the flowering and grain-filling stages, can exacerbate water loss through increased evapotranspiration, resulting in heightened water stress. Maize is particularly sensitive to heat stress during pollination, when temperatures exceeding 30°C can impair kernel formation and reduce yield. Studies have shown that prolonged heat waves not only decrease maize yield but also reduce the crop's quality, as extreme heat accelerates the plant's life cycle, shortening critical stages like grain filling [3], [5]. This accelerated development often leads to smaller, less dense kernels, which diminishes both the quantity and nutritional value of the crop [5], [28].

Drought, or water stress, represents one of the most significant threats to maize production under climate change. Maize requires substantial water throughout its growing season, particularly during the reproductive phase [9]. In drought-prone areas, such as the Mediterranean region, including northern Italy [29] [30], irrigation is often used to mitigate the effects of insufficient rainfall [31]. However, the sustainability of relying on irrigation is increasingly in question. Irrigation helps maintain productivity in dry conditions, but as competition for water intensifies—especially in agriculture-dependent regions—this approach may not provide a long-term solution [31]. As already presented in

section 1.1.2, according to studies on Mediterranean agriculture, future projections suggest that water availability will decline significantly, particularly in southern Europe, placing further strain on irrigation systems. This is reflected in the *European Drought Risk Atlas* [3], which highlights that while irrigation is essential for maize production, it cannot fully compensate for the expected reductions in water resources under climate change scenarios [12], [31]. Research has shown that while irrigation can improve yields in the short term, over-reliance on it risks depleting groundwater and exacerbating water scarcity. In fact, some studies predict that the region could save up to 35% of water by implementing more efficient irrigation methods, yet the long-term sustainability of these practices depends heavily on addressing broader water management issues [12], [31], [32]. As water becomes scarcer, the pressure to balance the needs of various sectors is expected to grow, potentially leading to more frequent conflicts over resource allocation [3], [33].

1.3 Research Objectives

This thesis aims to develop an initial methodology for the preliminary assessment of agricultural damage caused by drought, focusing specifically on maize cultivation with a provincial spatial resolution over Italy. The framework established in this research is designed to predict yield anomalies under the RCP 8.5 climate scenario, using basic climatic variables as foundational elements. Although the process is developed for maize, the ultimate aim is to create a methodology that can be easily adapted to other crops present in the ISTAT database, supporting broader applications across different agricultural sectors.

1.4 Contribution and Structure of This Work

The contribution of this work lies in the development of a methodology to assess agricultural damage caused by drought, focusing specifically on maize cultivation within the context of climate change. This methodology provides an initial approach using basic

climatic variables to predict yield anomalies under the RCP 8.5 scenario. The research contributes to the field by offering a simple model that integrates climate projections and agricultural risk assessments at a more localized scale, concentrating on the provincial level and focusing on high-productivity zones in Italy. Moreover, no similar study has been conducted on a provincial scale for Italy, making this a valuable addition to understanding regional vulnerabilities to drought.

In section 1, a comprehensive literature review of key elements relevant to the thesis is presented, followed by the presentation of research objectives. The review explores existing research on climate projections, the impact of drought on agriculture, challenges in modelling agricultural drought, and the role of irrigation. These topics provide a foundation for the methodology and research focus, highlighting the most pertinent aspects already established in the literature.

In section 2, the study methodology is presented, in particular the data sources employed in this study are outlined; the methodology for modelling agricultural drought is explained, with specific focus on how climate features are used to estimate yield anomalies at the provincial scale. This section also discusses the limitations of the model, particularly regarding the simplification of climatic variables and the challenge of applying local-level data to broader climate projections.

In section 3, the results of the study are presented, including projected yield anomalies for maize under future climate scenarios. These results are analysed at the provincial level, highlighting the regions most vulnerable to drought within the specific zone studied.

In section 4, the findings are discussed in relation to existing literature. Finally, suggestions for future research are provided, highlighting the need for more detailed irrigation data and more sophisticated models that incorporate additional variables affecting drought resilience.

2 Methodology

This chapter outlines the methodology employed for conducting the analyses in this thesis. The goal of the methodology is twofold: first, to establish a clear, structured process for building a predictive model that evaluates the impact of agricultural drought on maize production, and second, to highlight both the limitations and areas for future improvement. By providing a coherent and reproducible framework, this study aims to lay the groundwork for further refinement and enhancement of the model.

The methodology emphasizes the development of a model capable of quantifying the damage or variation in agricultural yield due to agricultural drought conditions. The focus is on identifying yield anomalies—deviations from expected production levels—potentially driven by climatic factors, such as changes in temperature and precipitation. By applying statistical and regression-based approaches, the study primarily aims to capture the short-term impacts of drought. While local trends were analysed and accounted for through detrending, the underlying causes of these long-term trends were not the focus of this thesis.

Additionally, this chapter details the steps taken for data preparation, preprocessing, feature engineering, and the selection of regression models, with a particular focus on hyperparameter tuning to maximize performance and minimize overfitting. Each step is explained in detail to provide transparency and ensure that the methodology can be replicated in future improvements of the pipeline. Throughout, the limitations encountered during the study are highlighted, offering a roadmap for future enhancements and refinements in modelling agricultural yield anomalies under climate stress.

2.1 Data description

This section provides a comprehensive overview of the data sources used in this study. The data encompass two key components: climatic data (both historical reanalysis and future projections) and maize cultivation data. The climatic data includes high-resolution temperature and precipitation records, which serve as the primary variables for assessing

yield anomalies. The historical data were sourced from the VHR-REA_IT dataset [34], produced through a detailed downscaling process from ERA5 reanalysis to ensure accuracy across regional scales. The future projections data were sourced from the VHR-PRO_IT dataset, which offers high-resolution climate projections under different greenhouse gas concentration scenarios [35].

The VHR-REA_IT and VHR-PRO_IT datasets were selected because they represent the most comprehensive and high-resolution climate datasets available for Italy. There are no other comparable sources for both climate and agricultural production data for the country. Agricultural production data from Eurostat, for instance, are derived from ISTAT [36], which is the primary source for provincial-level records of cultivated area and production. As such, the ISTAT Agriculture Database [29] was chosen for maize cultivation data, providing annual provincial-level records that allow for a detailed analysis of maize yields across Italy.

Although other climate datasets were considered, such as the Copernicus SIS Hydrology Meteorology Derived Projections database [37], it was not selected due to its lower spatial resolution (5x5 km), which is less suitable for the provincial-scale analysis required in this study.

2.1.1 Historical Climatic Data: overview and bias assessment

The selection of temperature and precipitation as the key variables for this study was primarily driven by their fundamental role in agricultural processes and the availability of high-quality data. Given the objective of this study to develop a model for predicting yield anomalies, focusing on these primary climate variables provided a clear and manageable starting point for analysing the relationship between climatic conditions and agricultural performance.

The climatic data used for this thesis were extracted from the VHR-REA_IT dataset, which stands for Very High Resolution Dynamical Downscaling of ERA5 Reanalysis over Italy [38], provided by the CMCC Delivery System [34]. VHR-REA_IT was developed as part of the HIGHLANDER European project [39] and was produced using the COSMO-CLM

regional climate model. This model employs dynamic downscaling, which involves using a high-resolution regional climate model, COSMO-CLM in this case, to refine the coarser output from a global reanalysis, such as ERA5, by simulating atmospheric processes at a finer scale. This process is essential for improving regional-level climate predictions that account for local topography and atmospheric dynamics [38]. The ERA5 reanalysis, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), provides a global dataset of weather data from 1950 to the present. It assimilates historical weather observations into a numerical model to provide a consistent reconstruction of past atmospheric conditions. For more details on the ERA5 dataset, refer to Hersbach et al. (2020), which outlines the reanalysis method and its applications [40].

The COSMO-CLM model (COSMO model in CLimate Mode), used for the downscaling process, is a well-established regional climate model based on non-hydrostatic atmospheric dynamics, making it ideal for high-resolution simulations in complex terrains like Italy. For further details on the setup and production of the VHR-REA_IT dataset, reference can be made to the work of Raffa et al. (2021), which discusses different nesting strategies in the downscaling process [41].

The dataset spans from 1981 to 2023 and provides hourly values for temperature and precipitation, stored in NetCDF format, which facilitates efficient management and analysis of multi-dimensional climate data. The dataset follows the WGS84 (EPSG 4326) coordinate system, ensuring compatibility with global geospatial standards. The data are retrieved through the CMCC Foundation's Data Delivery System (DDS), which offers streamlined access to high-resolution climate datasets.

Although the accuracy assessment of the VHR-REA_IT dataset was not conducted as part of this thesis, such evaluations were carried out in the original study that developed VHR-REA_IT. The dataset's performance was compared to observed data, with bias and variance analysed for both temperature and precipitation [38].

For temperature, the analysis revealed a notable seasonal variation in bias. Across Italy as a whole, the largest positive biases were observed during the summer months (JJA), with

VHR-REA_IT showing a bias of up to 1.9°C in some areas, particularly in the Northeast and Insular Italy. In contrast, bias was much smaller during the winter (DJF) and autumn (SON) seasons. The Northwest and Northeast regions exhibited minimal bias during the winter months, while Central and Southern Italy displayed relatively high bias in summer, reaching around 2.1°C in some regions (Figure 3). This suggests that while VHR-REA_IT captures winter temperatures accurately, it tends to overestimate temperatures during the summer, particularly in southern and insular regions.

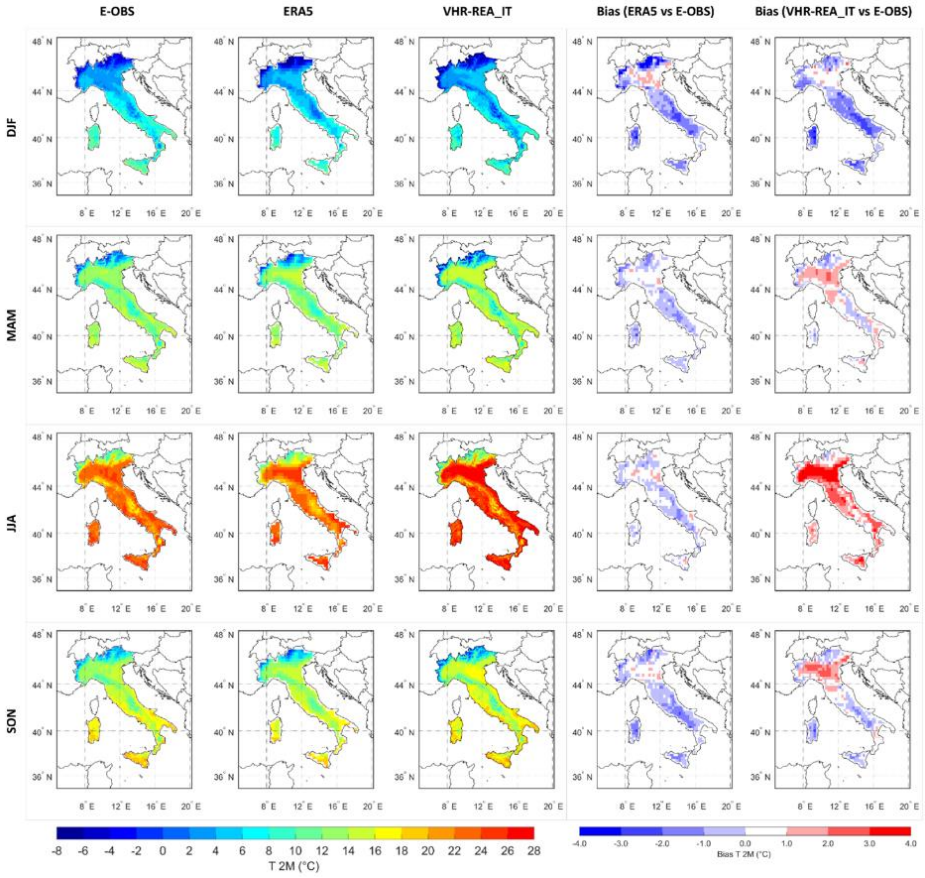


Figure 3 Seasonal spatial distribution of 2 m temperature (period 1989–2020) of E-OBS, ERA5, and VHR-REA_IT on their relative grids (first three columns). In the last two columns: seasonal spatial distribution of 2 m temperature bias with ERA5, and VHR-REA_IT. Source: Figure 2 from Raffa et al. [35]

For precipitation, the dataset exhibited more pronounced bias and variability. As shown in Figure 4**,** ERA5 generally underestimated precipitation, with a stronger negative bias during the summer (JJA) and autumn (SON) seasons. For instance, in Northeast Italy, ERA5 underestimated summer precipitation by around 70%, while VHR-REA_IT provided a closer approximation, though still overestimating precipitation in certain regions. In Southern Italy and Insular Italy, VHR-REA_IT displayed a positive bias, particularly during the summer months, with overestimations reaching 86% in some cases. Overall, VHR-REA_IT tended to show less bias than ERA5, particularly in the winter and spring months, but the magnitude of overestimation during the summer, especially in southern regions, should be considered.

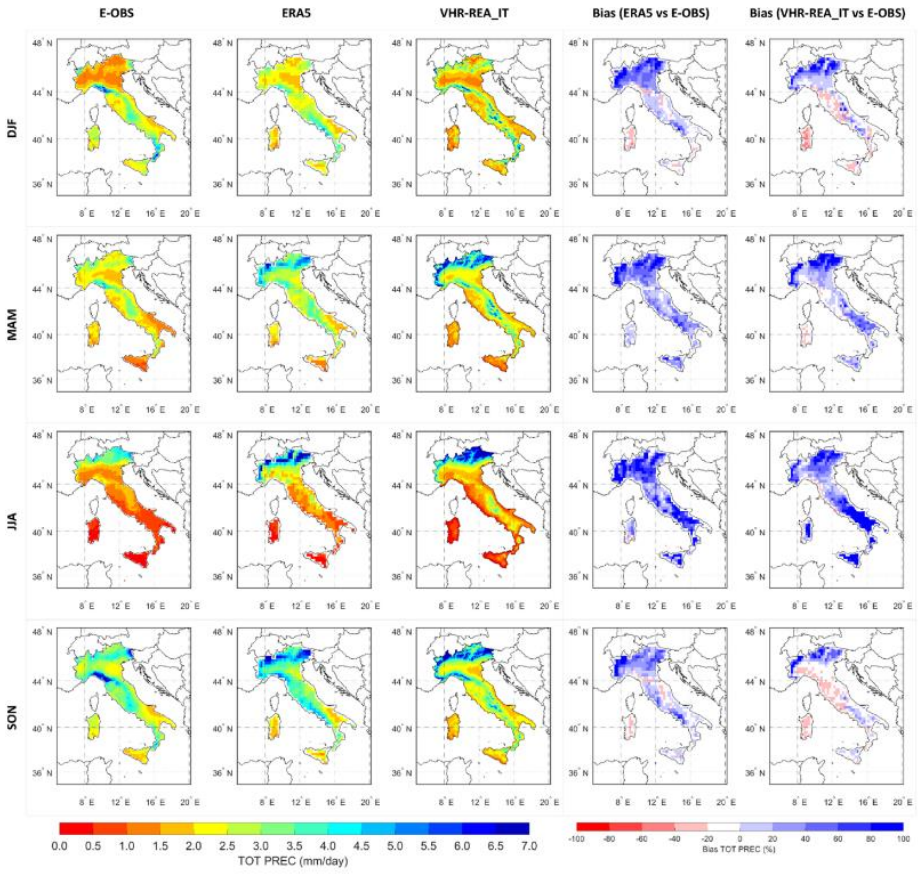


Figure 4: Seasonal spatial distribution of precipitations (period 1989–2020) of E-OBS, ERA5, and VHR-REA_IT on their relative grids (first three columns). In the last two columns: seasonal spatial distribution of 2 m temperature bias with ERA5, and VHR-REA_IT. Source: Figure 3 from Raffa et al. [35]

This analysis highlights the regional variability in both temperature and precipitation biases across the dataset, with VHR-REA_IT performing better than ERA5 in capturing precipitation patterns, though still facing challenges in accurately estimating summer conditions, particularly in the southern and insular regions of Italy. These regional differences are crucial for understanding the limitations and applicability of the dataset in climate impact studies, especially for agricultural analyses like those conducted in this thesis.

2.1.2 Future Projections Climatic Data

The future climatic data used in this study were sourced from the VHR-PRO_IT dataset (Very High-Resolution PROjections for ITaly) [42], also developed as part of the HIGHLANDER [39] and it is available through the CMCC Foundation's Data Delivery System (DDS) [42]. The project data provide hourly projections of temperature and precipitation under the RCP 8.5 scenario for the period from 2030 to 2070.

The VHR-PRO_IT dataset was selected for this study due to its high spatial resolution of 2.2 km [35], which aligns with the resolution of the historical VHR-REA_IT dataset, ensuring spatial consistency between the two time periods. Additionally, its structure—stored in NetCDF format and using the WGS84 (EPSG 4326) coordinate system—is identical to that of the reanalysis data. This consistency allows for seamless integration and processing using the same tools developed for managing historical data, without the implication of further data engineering, simplifying data handling and analysis across both datasets.

The downscaling process was applied to projections from the Italy8km-CM climate projection (CMIP5 global model), using the same COSMO-CLM regional climate model as in VHR-REA_IT (2.1.1) to improve the resolution of the projections. The model configuration was based on the COSMO-DE setup, originally developed by the Deutscher Wetterdienst (DWD) for numerical weather prediction, ensuring high reliability in simulating regional climate dynamics [35]. For detailed information on the model setup, reference can be made to the relevant study. The downscaling was processed by dividing

Italy into Northern, Central, and Southern regions, allowing for more detailed regional analysis of climate impacts [35].

It is important to note that the historical data used in this study extend only up to 2005. The global forcing is driven by the observed natural and anthropogenic atmospheric composition for the period 1989–2005, and the RCP4.5 and RCP8.5 greenhouse gas concentration trajectories for the years 2006–2050. Therefore, projection data for the period 2006–2018 are derived from the RCP8.5 scenario and not from historical observations [35].

The dataset underwent extensive performance and consistency validation on a daily basis for both temperature and precipitation. The VHR-PRO_IT dataset shows a significant improvement in bias reduction when compared to other models, particularly Italy8km-CM, for both 2m-temperature and total precipitation. However, the dataset still exhibits some biases, with a wet bias for precipitation in Northern Italy and a dry bias in Central and Southern Italy. These biases align with the Euro-CORDEX ensemble for the northern regions but differ in central areas, where most Euro-CORDEX models report a wet bias [35].

In terms of model consistency, VHR-PRO_IT demonstrates strong alignment with the Italy8km-CM projections and the Euro-CORDEX ensemble mean for climate changes in temperature and precipitation. The primary differences occur in Northern Italy under RCP4.5 and Central Italy under RCP8.5, primarily due to variations in precipitation. The model falls within the Euro-CORDEX envelope, highlighting its reliability in projecting climate changes at a regional scale [35].

Additionally, at the daily scale, VHR-PRO_IT has been validated for 2021–2050 against historical data (1989–2018), showing consistency with projected temperature and precipitation changes. The dataset's performance is particularly strong in Northern Italy, although there are some discrepancies in precipitation in central regions, indicating areas for further refinement [35].

2.1.3 Maize Cultivation Data

The crop chosen for this study is maize, selected due to its representativeness within the available data and the opportunity to compare the results with existing studies in the literature. The data on maize cultivation were sourced from the ISTAT (Italian National Institute of Statistics) Agriculture Database, specifically from the section on crop statistics [29]. The data include two key variables: maize cultivated area and the total maize production. These variables were collected on an annual basis at the provincial level, covering the period from 2006 to 2022, which was the most recent data available at the time of this study.

The data on cultivated area and production are derived through estimative methods, as reported in database metadata. These estimates are made based on evaluations conducted by experts in the agricultural sector, with information provided by local authorities. The experts' estimates may include data obtained from direct inspections in the field, as well as information from external sources such as professional organizations, producer associations, and administrative data. Additionally, auxiliary data sources related to maize cultivation may also be considered in the estimation process [29].

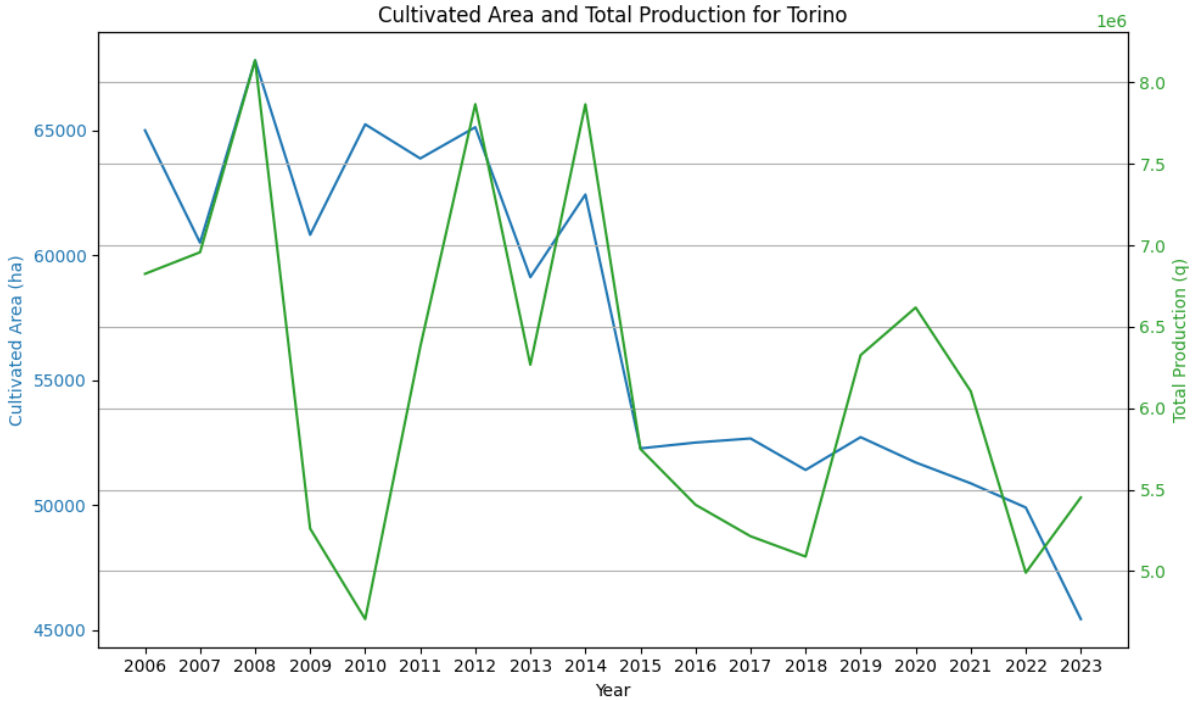
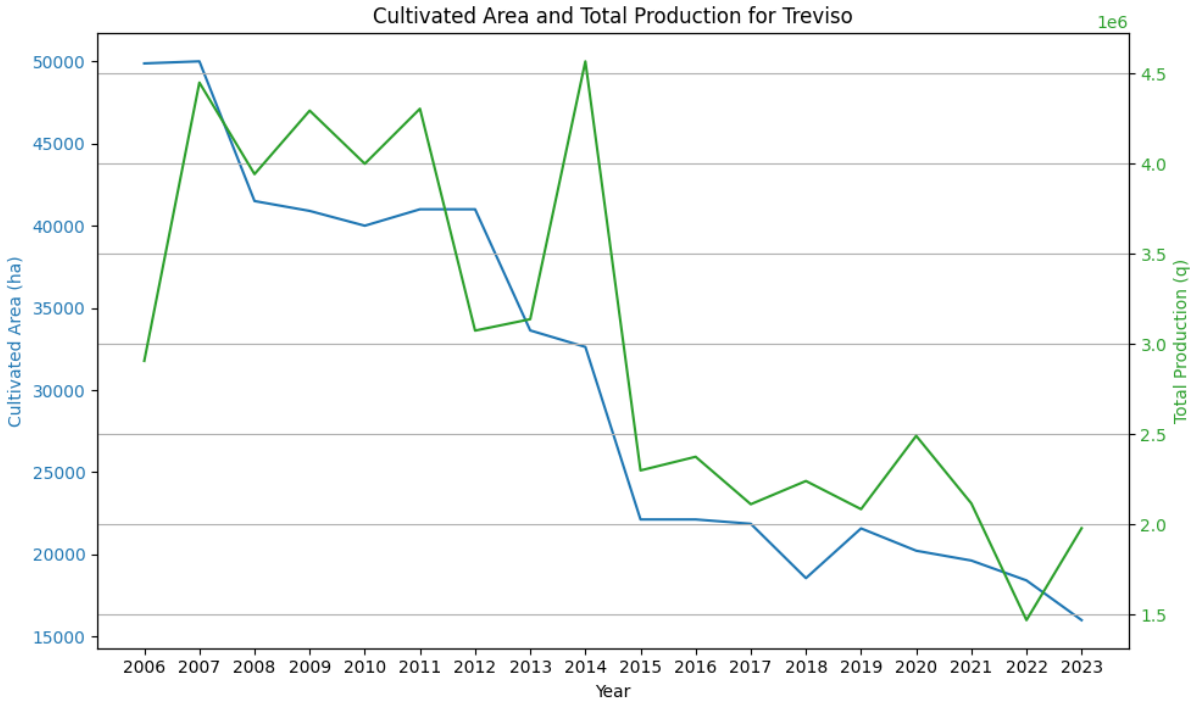


Figure 5: Maize cultivated area and production in Turin province.



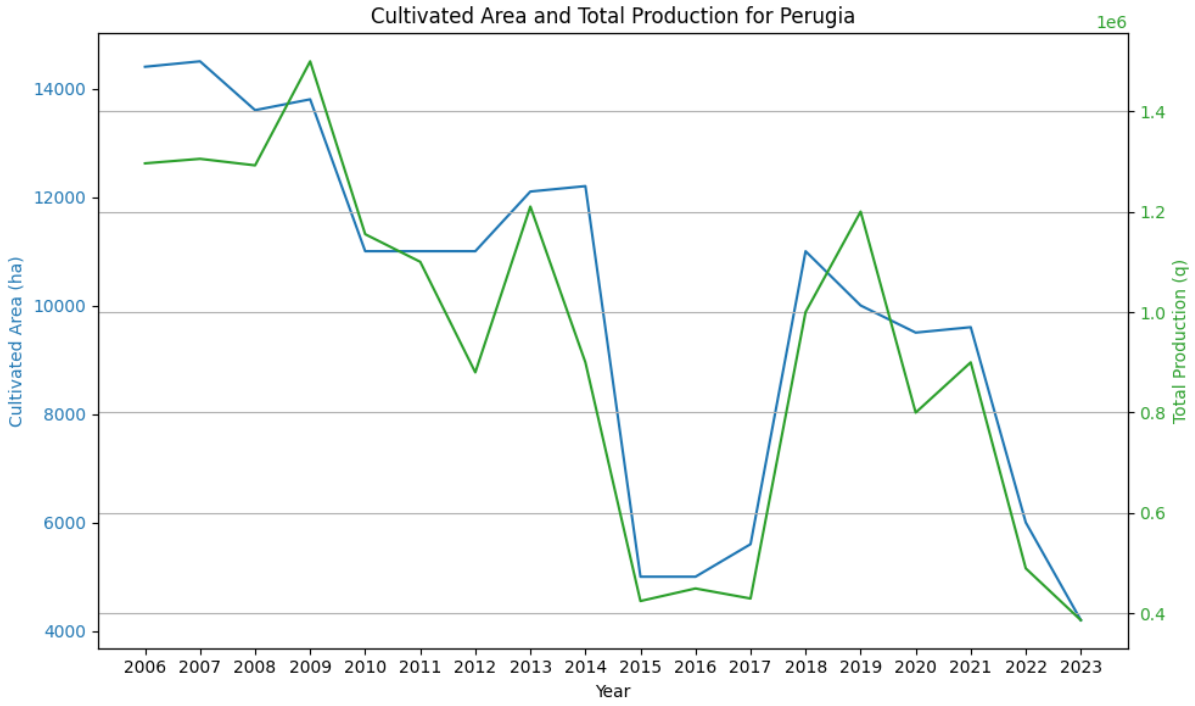


Figure 7: Maize cultivated area and production in Perugia province.

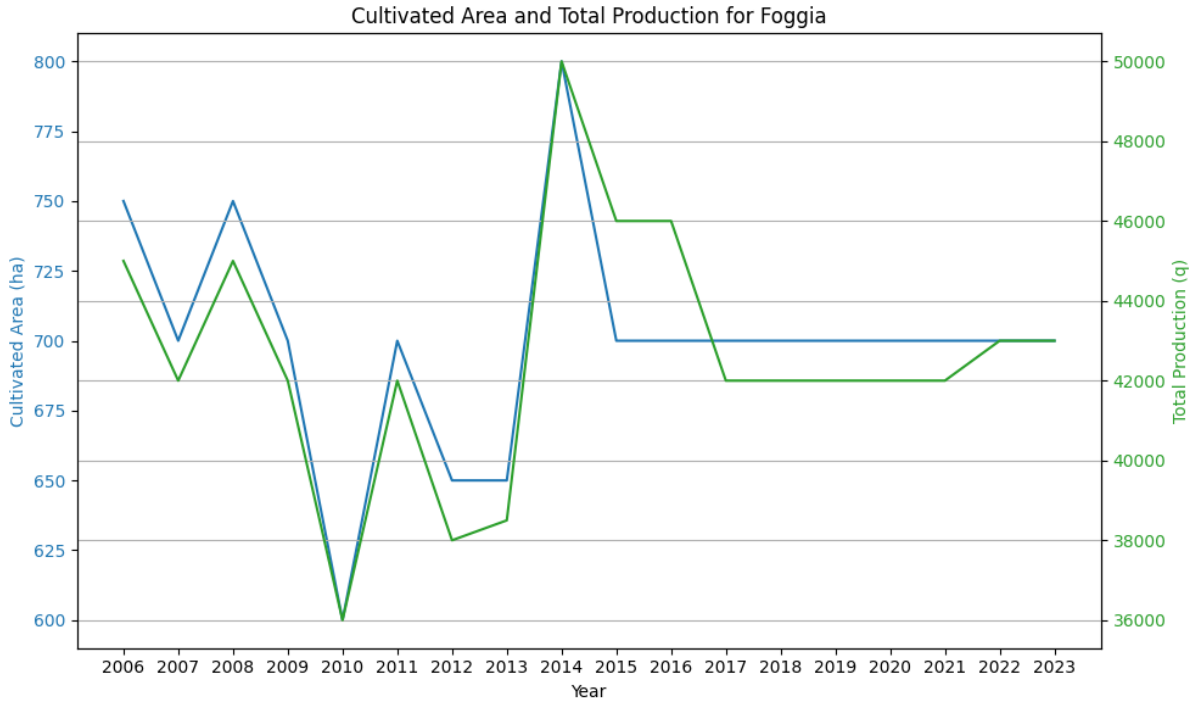


Figure 8: Maize cultivated area and production in Foggia province. Notice vertical axes scale change compared to previous provinces (Figure 5, Figure 6, Figure 7). Some data are missing.

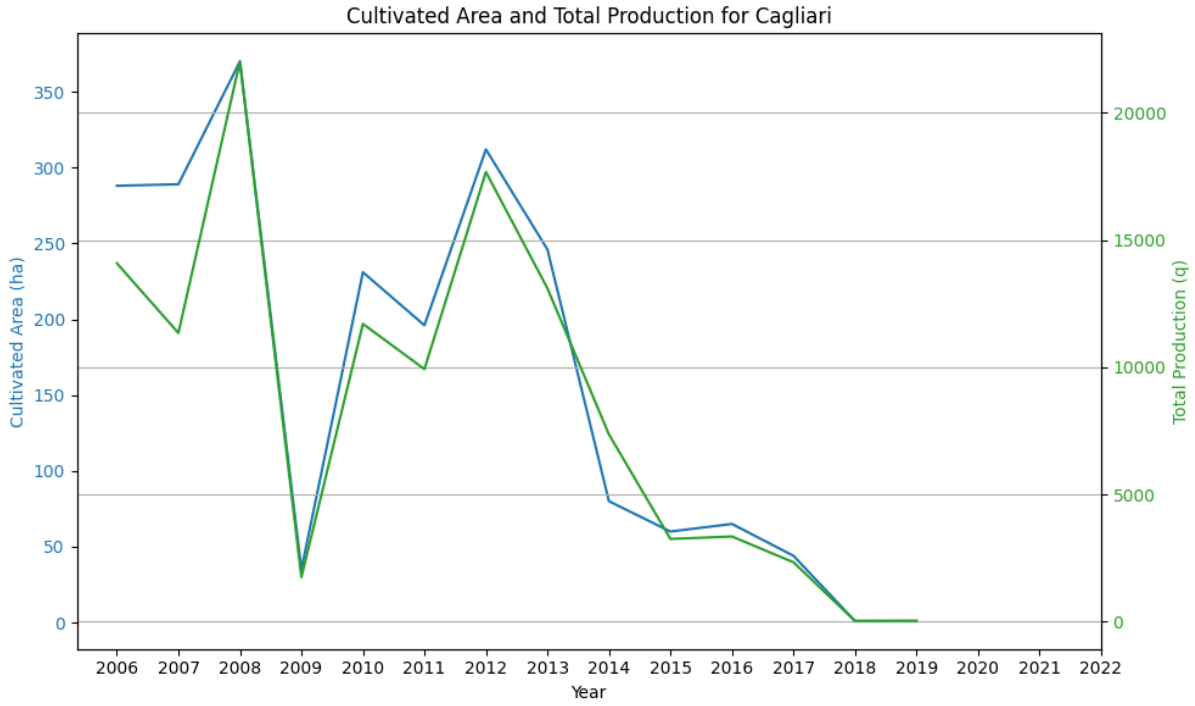


Figure 9: Maize cultivated area and production in Foggia province. Notice vertical axes scale change compared to previous provinces (Figure 5, Figure 6, Figure 7). Some data are missing.

2.2 Data preparation and preprocessing

2.2.1 Data Engineering and Features Preparation

The precipitation and temperature data retrieved from the VHR-REA_IT dataset were subsequently processed to meet the specific requirements of this study using custom Python tools. These tools were developed to ensure efficient data handling, focusing on both monthly temporal and provincial spatial aggregation. The data were processed in two main phases: the creation of meteorological temporal aggregated georasters and the calculation of spatial averages at the provincial level. The spatial boundaries were derived from georeferenced polygons with a 1-meter resolution, provided by Eurostat. The nomenclature for territorial units at levels 0 (NUTS 0) and 3 (NUTS 3), updated to 2021, was employed as it was the most recent available at the time of the study [43].

Phase 1: Creation of temporal aggregated meteorological Georasters

In this phase, the temperature and precipitation data were first downloaded as hourly NetCDF files. For each year and month in the selected time period, the script iterated through the following steps:

1. Data Download:

   Using a custom Python function, hourly temperature and precipitation data were downloaded from the CMCC database in NetCDF format for each month. This step involved looping over each year and month, saving the hourly data in a temporary directory.

2. Temporal Aggregation:

   The downloaded hourly data were aggregated to a monthly resolution. For temperature, the monthly average was calculated, while for precipitation, the monthly sum was computed. These operations were done using Python's xarray library to resample and perform temporal aggregation.

3. Merging NetCDF Files:

   The monthly NetCDF files were then merged to create a continuous dataset for each year. This step reduced the complexity of handling individual monthly files, streamlining subsequent spatial processing.

4. Conversion to GeoTIFF:

   The merged NetCDF files were converted into GeoTIFF format, a widely used geospatial raster data format. This format facilitated the subsequent spatial averaging process and ensured compatibility with geographical information systems (GIS).

5. Cropping by Region (NUTS 0):

   After generating the GeoTIFFs, the data were spatially cropped to fit within the boundaries of Italy using the NUTS 0 level (country-level) georeferenced polygons from Eurostat. This step was necessary because the original download covered Europe, and cropping ensured that only data relevant to Italy were stored in fixed memory, resulting in lighter and more manageable files.

Phase 2: Spatial Averaging

In the second phase, the processed georasters were spatially aggregated to calculate provincial-level averages. This step involved:

1. Province Selection:

   The NUTS 3 boundaries for Italy were used to define the provinces. A list of all provinces was generated from the Eurostat dataset.

2. Spatial Cropping:

   For each province, the geospatial data were cropped to match the specific provincial boundaries. This ensured that the calculated averages were representative of the spatial extent of each province.

3. Calculation of Averages:

Once the data were cropped to provincial boundaries, the script calculated the spatial average of mean temperature and cumulative precipitation for each province. This was achieved by masking the data outside the provincial polygons and computing the mean of the remaining values.

4. Transformation into a Time-Series Dataset:

   The calculated provincial monthly aggregated (average temperature and cumulative precipitation) data were transformed into a structured time series, where each row represented a province, and each column corresponded to a specific year and month. This monthly dataset was then further transformed into an annual dataset, where monthly data were stacked, and both year and month were extracted from the date column. The data was pivoted to display 12 columns, each representing one of the months of the year, with each row corresponding to a province for a specific year.

5. Merging Temperature and Precipitation Data:

   Given that separate georasters, and consequently time-series datasets, were created for temperature and precipitation, the final step involved merging the two datasets on a provincial basis. This unified dataset included both monthly temperature and precipitation values.

2.2.2 Feature selection

As previously mentioned in the preceding section, the selected climate features for this analysis are monthly average temperature and cumulative monthly precipitation. This choice was made to facilitate the development of the prototype pipeline for model production and interpretation, which represents the focus of this thesis. These features were easily accessible to produce via data engineering and to work with, which sufficed for the purposes of this study. Furthermore, for a first approximation, temperature and precipitation are considered significant enough to assess their influence on maize yield anomalies [3], [27].

An a priori feature selection was performed based on the lifecycle of maize, which typically spans from March (when the first sowing takes place) to October (the latest harvest). As reported in sections 1.2.2 and 1.2.3, maize is particularly sensitive to temperature, requiring minimum temperatures of at least 12°C during its early growth phase to prevent damage, as growth halts if temperatures fall below 10°C. The crop's growth cycle typically lasts between 100 and 140 days, and multiple sowings occur during the agricultural year, depending on local conditions and practices [26], [44]. Therefore, for each year, the final selection of features focused on the monthly average temperature and monthly cumulative precipitation during this period, spanning from March to October.

2.2.3 Quantitative and Qualitative Data Analysis

Once the dataset was processed, both quantitative and qualitative analyses were conducted to better understand the structure and relationships within the data:

- Correlation Analysis (Features): Pearson and Spearman correlations were computed to measure the relationships between the features themselves. Pearson's correlation assesses the linear relationships between two continuous variables, while Spearman's correlation captures ranked and non-linear relationships. Evaluating the correlation between features is crucial to understanding how they interact and ensuring that highly correlated features are not redundant. This helps prevent multicollinearity, which could skew the model's ability to identify the unique contributions of each feature to yield anomalies.
- Correlation Analysis (Features and Target): Additionally, Pearson and Spearman correlations were used to evaluate the relationship between the features (temperature, precipitation) and the target variable (crop yield anomaly). These correlations help assess which features have the strongest linear and ranked relationships with the target, guiding feature selection and model development.

Furthermore, potential correlations between the target variable and features corresponding to months outside the lifecycle of maize (before March and after October)

may reflect broader intra-seasonal climatic patterns. These correlations could stem from climatic conditions in off-season months that influence the subsequent growing season, indirectly affecting yield anomalies. While this hypothesis was not examined in detail in this study, future research could explore these relationships to understand how climatic interactions between seasons contribute to variations in crop yield.

2.2.4 Features preprocessing

Several preprocessing strategies were employed to ensure that the data was appropriately transformed for analysis, all applied at the provincial level:

- Standardization:

  Standardization was performed using StandardScaler from sklearn.preprocessing. This method rescales the features so that they have a mean of 0 and a standard deviation of 1, ensuring that all features are on the same scale. The formula used for standardization is:

$$X' = \frac{X - \mu}{\sigma}$$

  where X is the original feature value, μ is the mean of the feature, and σ is the standard deviation.

- Normalization:

  For normalization, Normalizer(norm='l2') from sklearn.preprocessing was applied. This method scales each sample individually to have unit norm, making it particularly useful for sparse data or datasets with high variance. The L2 normalization formula is:

$$X' = \frac{X}{\left|\left|X\right|\right|_2} = \frac{X}{\sqrt{\sum_{i=1}^{n} X_i^2}}$$

  where ‖X‖ represents the L2 norm (Euclidean distance) of the feature vector.

Initially, a log transformation was considered for precipitation data to reduce skewness and stabilize variance, addressing the inherent variability often found in precipitation distributions. The transformation was applied using the formula:

$$X' = \log(X + 1)$$

where 1 is added to avoid the issue of taking the logarithm of zero values. However, given that the precipitation data was standardized prior to the log transformation, it was recognized that the addition of 1 might significantly affect smaller values. In such cases, the selection of an appropriate constant (smaller than 1) should be carefully evaluated to avoid distorting the data. After considering these factors, it was decided to exclude the log transformation in the final analysis to maintain consistency in the data processing pipeline. This approach, however, could be revisited and further explored in future steps, particularly in relation to selecting a more suitable constant or considering alternative transformations.

Following these preprocessing strategies, three different datasets were utilized in the analysis: one with original features, one with standardized features, and one with normalized features.

2.2.5 Target preparation

First of all, the spatial distribution of the available data of agricultural production and cultivated area data, sourced from ISTAT [29], was assessed by grouping the provinces into four Italian macro areas based on similar climates: Islands, North, Central, and South. To ensure the inclusion of provinces with significant maize production but potentially sparse data collection, a criterion was necessary to select provinces for the study. The chosen criterion was that at least in one year, from 2006 to 2022, the area cultivated with maize in the province exceeded 10,000 hectares, and that each province had at least three records of maize production data. This dual-threshold approach ensured that only provinces with significant maize cultivation and at least three records of maize production data were included, providing a robust dataset for temporal analysis. This strategy not only guards against the inclusion of provinces with sporadic data that

could undermine the study's robustness and reliability but also guarantees a statistically meaningful time series for each selected province.

Following the selection of provinces based on the criteria of significant maize cultivation and sufficient data records, the agricultural production and cultivated area data, sourced from ISTAT [29], were reprocessed to compute the crop yield according to the formula:

$$Yield_{crop}\left[\frac{kg}{ha}\right] = \frac{Production_{crop}\,[kg]}{Area_{crop}[ha]}$$

This focus on yield rather than total production is crucial because yield normalizes production relative to the size of the cultivated area. Maize production alone does not provide a full picture, as it varies greatly depending on the size of the area planted with the crop in any given year. By calculating yield, the study captures agricultural efficiency, how much production is obtained per hectare, making regions with different land areas under cultivation comparable. This allows for a more accurate comparison across provinces and years, as it removes the variability introduced by differing or fluctuating cultivated areas.

Next, trends in crop yields were analysed using linear regression to determine whether there was a consistent increase or decrease in yields over time. This trend analysis is essential because the study focuses on yield anomalies—the deviation from what is considered a normal or expected production level. While long-term changes in yields fall outside the scope of this study, instead, the primary focus is on understanding short-term deviations from the established trends, which are hypothesized to be influenced by seasonal climatic factors, such as fluctuations in temperature and changes in precipitation patterns.

The central hypothesis being tested in this study is that these yield anomalies are primarily driven by climate variability. By analysing the deviations from long-term trends, the study aims to isolate the short-term climatic events on maize production.

The significance of the observed trends was determined by examining the p-values and confidence intervals from the regression analysis. A p-value is a statistical measure that helps determine the probability of obtaining the observed results of a test under the

assumption that the null hypothesis is true. In this context, A p-value less than 0.05 indicates a statistically significant trend.

Additionally, the confidence intervals were analysed to assess the precision of the estimated trends. A confidence interval provides a range of values which is believed to contain the true value of the parameter with a certain level of confidence. Significant trends were identified when the confidence intervals did not include zero, indicating that the observed trend was different from no change.

After identifying significant trends within the crop yield data for each province, necessary adjustments were made. For those provinces with established trends, the crop yield data were detrended by subtracting the trend component using the formula:

$$Yield_{p,y_{detrended}} = Yield_{p,y} - S * (y - y_{start}) \qquad where\ S\text{: } trend\ slope, p\text{: } province, y\text{: } year$$

Provinces without detectable trends retained their original crop yield data unchanged.

To derive the target variable for the analysis, the crop yield anomaly, an average crop yield was first calculated for each province. This average was based on simple crop yield data for provinces without trends and on detrended crop yield data for those with trends. Subsequently, the crop yield anomaly was computed annually using the formula:

$$Yield\ Anomaly_{p,y}\ [\%] = \frac{Yield_{p,y} - Yield_{avg_p}}{Yield_{avg_p}} * 100 \qquad where\ \ p\text{: } province, y\text{: } year$$

This formula normalizes the crop yield anomaly, rendering it as a ratio that indicates how much each year's yield deviates from the average in percentage. This normalization allows for comparisons across different regions and scales by accounting for variations in absolute yield figures, thus establishing the crop yield anomaly as the key metric for subsequent analysis in this study.

2.3 Regression Models

To predict the crop yield anomaly based on climatic variables, a regression analysis was conducted. Five different types of regression models were employed and compared: Linear Regression, Random Forest Regressor, Extra Trees Regressor, k-Nearest Neighbors (KNN) Regressor, and XGBoost Regressor. These five regression models were selected to capture a wide range of relationships in the data, from simple linear patterns to complex non-linear interactions. For more detailed insights beyond what is covered in the following paragraphs, refer to *scikit-learn* library documentation [45], [46], [47], Dobson et al. [48], Avila [49], Liaw et al. [50], Geurts et al. [51], Taunk et al. [52], *XGBoost* library documentation [53] and Chen et al. [54].

2.3.1 Linear Regression

Linear regression is one of the most fundamental and widely used models, valued for its simplicity and interpretability. It assumes a linear relationship between the input variables (features) and the target variable (yield anomaly). The general form of the linear regression model is expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:

- y represents the crop yield anomaly (target variable),
- $X_1, X_2, \ldots, X_n$ represent the features (temperature, precipitation),
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients or weights,
- ϵ is the error term.

While linear regression provides insight into the direct relationships between the features and the target variable, its limitation lies in its inability to model non-linear interactions between variables, which can be common in complex phenomena. However, non-linear

interactions could be captured by introducing interaction terms (e.g., $X_1 \times X_2$) within the linear model. Additionally, a broader class of models, such as Generalized Linear Models (GLM), can introduce a non-linear link function, providing a more flexible framework for capturing non-linear relationships [48].

In this thesis, however, the analysis was performed using only the simple linear model without the inclusion of interaction terms or more advanced models.

2.3.2 Random Forest Regressor

Random Forest [49], [50] is an ensemble model built on multiple decision trees. By averaging predictions from multiple trees, Random Forest can account for non-linear relationships and reduce overfitting. It is particularly suited for complex datasets where interactions between variables may not follow a linear path. Each tree in the forest is constructed from a random subset of both features and training samples, enhancing the model's robustness against noise [45].

The model's prediction is:

$$y = \frac{1}{T} \sum_{t=1}^{T} h_t(X)$$

Where:

- $h_t(X)$ is the prediction from the t-th tree,
- T is the total number of trees.

Random Forest was selected due to its resilience to overfitting, high-dimensional data handling, and its effectiveness in capturing intricate, non-linear dependencies.

2.3.3 Extra Trees (Extremely Randomized Trees)

Extra Trees is another ensemble method similar to Random Forest, but with one key difference: it introduces more randomness by selecting split points for trees at random, rather than choosing the best split based on a criterion or information gain. This added randomness helps the model generalize better, particularly in noisy datasets, as it reduces variance and avoids overfitting [49], [51].

The Extra Trees Regressor generates multiple trees and aggregates their outputs, but with more randomness in the selection of splits. The prediction for a new instance X is the average of the predictions from all the trees. Like Random Forest, Extra Trees generates multiple trees and aggregates their predictions. However, the random nature of split selection means that it might discover unique patterns within the data that other models may overlook. The decision to include Extra Trees was motivated by its potential to perform well with diverse climate data and capture subtleties in yield variation [46].

2.3.4 k-Nearest Neighbours (KNN)

k-Nearest Neighbours is a non-parametric model that makes predictions based on the values of the k-closest neighbours in the training data. It operates by computing the distance (Euclidean in this case) between the input and the points in the dataset, selecting the closest neighbours, and averaging their outcomes to predict the yield anomaly [47], [49].

The formula for KNN regression is:

$$y = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}$$

Where y_i represents the yield anomaly values of the k-nearest neighbours, and w_i represents the weight assigned to each neighbour. The weight w_i can be uniform (if all neighbours are treated equally) or distance-based (where closer neighbours have a higher

influence on the prediction). In the case of distance-based weighting, $wi = \frac{1}{d_i}$, where d_i is the distance between the input and the i-th nearest neighbour.

While KNN does not assume any prior relationships between features, its simplicity comes with limitations: it can struggle with high-dimensional data and may be sensitive to the choice of k [52].

2.3.5 XGBoost Regressor

XGBoost is an optimized implementation of the Gradient Boosting algorithm, designed to be both efficient and accurate. It builds trees sequentially, with each new tree correcting the errors made by the previous ones, thereby improving accuracy over time. Additionally, XGBoost includes regularization mechanisms that prevent overfitting, a common issue in highly flexible models [53], [54].

The prediction for XGBoost is based on the sum of the outputs from all trees:

$$y = \sum_{t=1}^{T} f_t(X)$$

where $f_t(X)$ represents the prediction from the t-th tree. XGBoost was chosen due to its strong track record in machine learning tasks, especially when dealing with large datasets and non-linear interactions between features.

2.4 Models' calibration and validation

In this section, the process of calibrating and validating the models is outlined, focusing on the steps taken to ensure optimal performance and generalizability. The goal of this phase was to fine-tune the models by selecting the best hyperparameters and validating their performance on unseen data, ensuring a balance between accuracy and the risk of overfitting. Additionally, feature importance was evaluated both through correlation

analysis and SHAP values [55], [56], which are later batter explained, providing deeper insights into the relationships between the input features and the model predictions.

2.4.1 Hyperparameter Tuning Process

For the algorithms requiring hyperparameter tuning (Random Forest Regressor, Extra Trees Regressor, k-Nearest Neighbours Regressor, and XGBoost Regressor), a grid search was employed to identify the optimal combination of hyperparameters. The key hyperparameters involved in the analysis are listed in Table 1. The grid search was performed using a custom metric, D, specifically designed to balance the R2 scores of both the training and test sets, ensuring the model generalizes well without overfitting. The definition of the R2 metric will be provided in the following paragraph (2.4.2).

Table 1: Models' hyperparameters involved in tuning process

Regression model	**Hyperparameters**
Random Forest Regressor	n_estimators, max_depth
Extra Trees Regressor	n_estimators, max_depth
K-Nearest Neighbours	n_neighbours, weights
XGBoost Regressor	n_estimators, max_depth

The custom metric D is a Euclidean distance measure that reflects how close the model's performance is to the ideal case, where the training and test R2 scores are equal. The formula for the D metric is:

$$D_{R2} = \frac{\sqrt{2} - \sqrt{(R2_{test} - 1)^2 + \left(\frac{R2_{test}}{R2_{train}} - 1\right)^2}}{\sqrt{2}}$$

In this equation, $D_{R2} = 1$ represents the optimal case where $\mathrm{R2_{train}} = \mathrm{R2_{test}}$, indicating no overfitting and perfect generalization. By maximizing D, the hyperparameters are

selected to ensure a balance between training and test performance, reducing the risk of overfitting (which would be indicated by a high $R2_{train}$ but a lower $R2_{test}$).

The grid search was repeated 30 times, each time using a fixed test set size, where 20% of the dataset was reserved for testing. Reiterating the grid search over multiple data splits helps to mitigate the impact of a single, potentially unrepresentative train-test split on the final hyperparameter selection. Averaging the results across these 30 iterations reduces the influence of outliers or random variations in the data, thus providing more reliable hyperparameter estimates.

The final set of hyperparameters was selected based on the combination that achieved the highest mean D score across iterations, ensuring that the chosen model configuration consistently balanced performance on both the training and test sets across multiple data splits.

2.4.2 Models Evaluation Method and Metrics

After determining the optimal hyperparameters through grid search, the final models were trained and evaluated using 5-fold cross-validation. This technique divides the dataset into five subsets (folds), with each fold serving as the validation set once while the remaining four are used for training. Cross-validation is particularly useful for reducing overfitting, as it ensures that the model is tested on different subsets of the data, providing a more reliable estimate of its generalization ability [57]. The evaluation metrics were computed for each fold, and the average scores were reported to provide a robust assessment of the model's general performance. The selected evaluation metrics are:

- **R^2 (Coefficient of Determination)**: This metric assesses how well the model's predictions correspond to the actual values, with values ranging from 0 to 1. A value of 1 indicates a perfect fit, so higher R^2 values suggest that the model captures more variance in the target variable.
- **RMSE (Root Mean Squared Error)**: RMSE measures the square root of the average squared differences between predicted and actual values, indicating how far

predictions deviate from reality. Lower RMSE values signify more accurate predictions, as large errors are penalized more heavily.

- **MASE (Mean Absolute Scaled Error)**: MASE is a scale-independent metric that compares the model's error to that of a naive baseline model. Specifically, it compares the Mean Absolute Error (MAE) of the model with the MAE of the naive model. The MAE is calculated using the formula:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | y_t - \widehat{y_t}|$$

where y_t is the actual value at time t, $\widehat{y_t}$ is the predicted value at time t, and n is the number of observations.

MASE is then calculated as:

$$MASE = \frac{MAE_{model}}{MAE_{naive}}$$

A MASE value lower than 1 means the model outperforms the baseline, while values higher than 1 indicate worse performance. The naive model used in this study generates predictions based on a normal distribution with mean and standard deviation equal to the target mean and standard deviation.

The final models were selected based on their R^2 scores, prioritizing the models that achieved the highest R^2 values across the cross-validation folds given the condition that they present $MASE < 1$. This ensured the selection of models with the best predictive accuracy while minimizing the risk of overfitting, thanks to the prior hyperparameter tuning process.

2.4.3 SHAP Feature Importance Analysis

To gain a deeper understanding of how each feature contributed to the model's predictions, a SHAP (SHapley Additive exPlanations) analysis was performed. SHAP is a method based on game theory that explains individual predictions by computing the contribution of each feature to the final prediction. Specifically, SHAP computes Shapley values, which fairly allocate the "payout" (the model's prediction) among the features by treating each feature as a "player" in a cooperative game. The Shapley values are the only

values that satisfy key properties such as local accuracy, consistency, and missingness, making SHAP a highly reliable method for interpreting model predictions [55], [56].

In addition to providing explanations for individual predictions, SHAP allows for global model interpretation through feature importance measures. SHAP feature importance is calculated as the mean absolute Shapley value of each feature across all instances, allowing to determine which features have the greatest overall influence on the model. Unlike permutation-based feature importance, which measures the decrease in model performance when a feature is shuffled, SHAP feature importance is based on the magnitude of feature attributions [55].

Moreover, SHAP summary plots combine feature importance with feature effects, showing the distribution of Shapley values for each feature across all data points. This helps in understanding both the overall importance of a feature and the direction of its impact (positive or negative) on the predictions. For example, a high SHAP value for a feature could indicate that increasing its value tends to increase the predicted target [55]. SHAP also enables the interpretation of models like k-Nearest Neighbours (KNN), which otherwise lack built-in feature interpretability.

This SHAP-based analysis was conducted to complement the a priori correlation analysis of features and their relationship with the target variable. By comparing SHAP values with the previously calculated Pearson and Spearman correlations, we could validate whether the initial feature assumptions of relevance were consistent with the model's actual behaviour during inference. This provided a more holistic view of how well the selected climatic variables explained the crop yield anomalies and whether further refinements could be made in future model iterations.

2.5 Inference Procedure: Historical and RCP8.5 Scenarios

With the processed data ready, inference was performed using the best-performing models that were identified through the model evaluation process. The model.predict function from Scikit-learn was used to estimate agricultural yield anomalies, and this was

applied to two datasets: the baseline dataset (2006–2023) and the future scenario dataset based on the RCP 8.5 projection. The baseline inference allowed for a comparison between predicted and actual yield anomalies over the training period, while the RCP 8.5 scenario provided predictions for future agricultural yield anomalies.

For the baseline dataset, predictions were compared with actual observations by calculating the Mean Absolute Error (MAE), which quantifies the average magnitude of errors between predicted and actual values, and it was also used to compute Mean Absolute Scaled Error (MASE). This provided insight into the model's performance on data it was trained on. Additionally, the frequency distribution of predicted anomalies was evaluated using histogram plots and Kernel Density Estimation, providing a visual representation of the spread and shape of predicted yield anomalies.

For both the baseline and RCP 8.5 datasets, the average yearly yield anomaly was calculated. In the case of the RCP 8.5 scenario, the average was computed over a 20-year period centred on 2040 and 2060, offering insights into how agricultural yield is expected to change under future climate conditions. For the baseline dataset, the average yearly yield anomaly was calculated over the full period of 17 years (2006–2023), centred on 2014, to compare the model’s historical predictions with the actual conditions.

Finally, provincial average yearly yield anomalies are visualized on maps, showing yield anomalies as percentage values. These maps enable a clear comparison between regions under both baseline and future climate conditions, providing an accessible interpretation of the regional impacts on agricultural yield.

2.6 Methodology limitations

Several limitations underpin this study, which may affect both the accuracy and generalizability of the results:

1. **Data Limitations**:
    - The historical climatic data used in this study were not bias-corrected, although bias was evaluated. Uncorrected biases may influence the predictions [38].
    - Additionally, regarding future projection data, from 1989 to 2005, the dataset consistency was evaluated using historical data, while for the period from 2006 to 2018, the projections from the RCP8.5 scenario were used. This transition creates a limitation, as the model relies on scenario data starting in 2006, effectively losing 20 years of historical control data for validation purposes [35].
    - The time series for maize yield data is relatively short, covering only 17 years. Moreover, the spatial coverage is limited, as there are few provinces with a sufficiently long and complete historical record, which constrains the robustness of the results.
2. **Model Limitations**:
    - Although a comparison of different models was conducted, the analysis was limited to a small set of models, which may not fully capture the complexity of maize yield anomalies. As a result, the models used might not be the most optimal for this specific task.
    - Based on existing literature [3], it is known that the features used in this study may not be the most optimal for yield prediction, and thus the results are not expected to be highly accurate.
    - There is also a risk of spatial overfitting in the model training process. Provinces that are geographically and climatologically similar may be

overrepresented in the training set compared to the validation set, which could skew the model's ability to generalize.

3. **External Factors**:
    - External influences, such as market conditions (e.g., shifts in demand), different climatic events (e.g. floods, hailstorms) or other non-climatic factors like irrigation, which can significantly impact agricultural production, were not included in the model. These unmodeled factors may have a considerable effect on yield, which was not accounted for in this study.
    - Additionally, the model may be sensitive to other climatic factors that were not explicitly considered, such as temperature extremes, storms, or humidity, which can also impact yield beyond the effects of drought alone.

3 Results

This chapter presents the results of the analyses performed on both the climate and maize datasets, as well as the evaluation of predictive models. The aim is to assess drought-related risks to maize cultivation, with a focus on both historical and future scenarios. The structure of this chapter is designed to guide the reader through the key stages of the analysis, from the data exploration phase to model evaluation and future inference.

The first section provides an analysis of the datasets used, including comparison between historical climate data and future projections. The maize dataset is examined through its geographic distribution, overall trends, and feature correlations. This section aims to establish a robust understanding of the data, laying the groundwork for the subsequent modelling and inference.

In the second section, the performance of various models is discussed, beginning with an initial filtering based on the Mean Absolute Scaled Error (MASE) metric. Models achieving a MASE below 1 are further analysed, with their R^2 and RMSE scores reported. The two most promising models are then compared in detail, evaluating their statistical distributions and performance in predicting true data. This comparison provides key insights into the models' accuracy and applicability for drought risk estimation.

Finally, the third section explores the future inference results, based on the best-performing models. Distributions and inference maps are presented to visualize the potential impact of drought on maize yields under future climate scenarios. These findings offer critical insights into the vulnerability of maize cultivation across different regions and the expected changes in yield anomalies under the RCP 8.5 scenario.

3.1 Datasets analysis and comparison

This section presents the results of the preliminary analyses that contributed to the definition of the final dataset used for model training. These basic evaluations were aimed

at ensuring that the dataset is sufficiently representative and valid for use as an initial structure in the model development pipeline.

3.1.1 Maize production data analysis

The preliminary analysis of maize production data across all provinces is presented first. Figure 10 and Figure 11 provide an overview of this analysis, which guided the selection of the provinces of interest for the study.

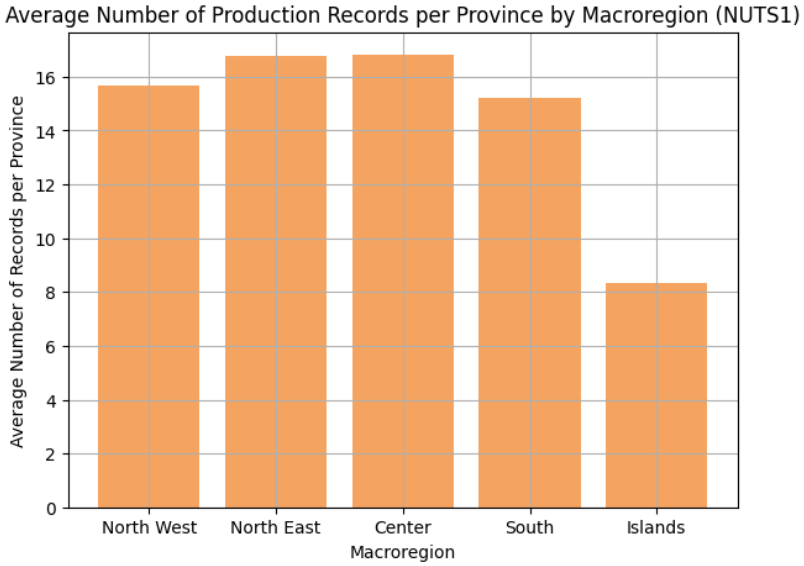


Figure 10: average provincial maize production records per macroregions (NUTS1)

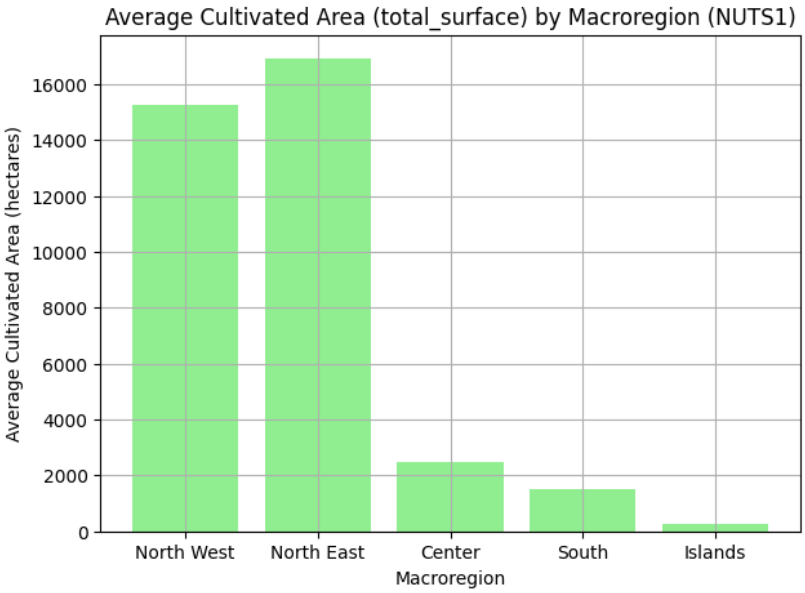


Figure 11: average provincial maize cultivated area per macroregion (NUTS1)

Figure 10 displays the average number of years with available data per province for each macroregion (North West, North East, Center, South, and Islands, NUTS1 classification major socio-economic regions [43]). On the surface, all macroregions appear to have, on average, a sufficient number of years with data per province to consider reliable time series. In other words, for most provinces, the 17-year period is well-represented with relatively few missing data points. However, Figure 11 highlights that the vast majority of cultivated land dedicated to maize is located in the North. Given the need to balance having sufficiently large production areas—areas that are less subject to very localized fluctuations in production—with having enough data on production and cultivated area (ensuring there are not too many missing records to form a reliable time series), provinces were selected based on the two criteria presented in the Methodology (2.2.5): they must have at least 10,000 hectares dedicated to maize cultivation in at least one record and at least three years of available records. The map in Figure 12 shows the provinces that meet these criteria.

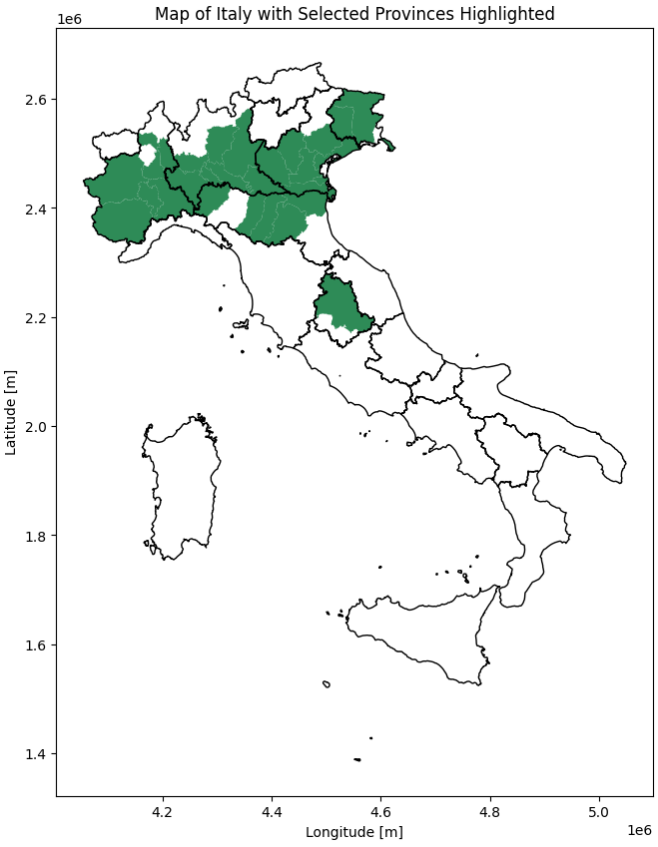


Figure 12: Italy map with provinces selected for the final dataset (green)

Following this, the presence of trends in crop yield was assessed for the provinces that met the selection criteria. In Figure 13 are presented the graphs for the provinces where trends were detected: the crop yield records (blue points) are compared with the regression curve (orange line), which illustrates the identified trend. The significance of these trends is ensured by a p-value less than 0.05, along with the examination of confidence intervals.

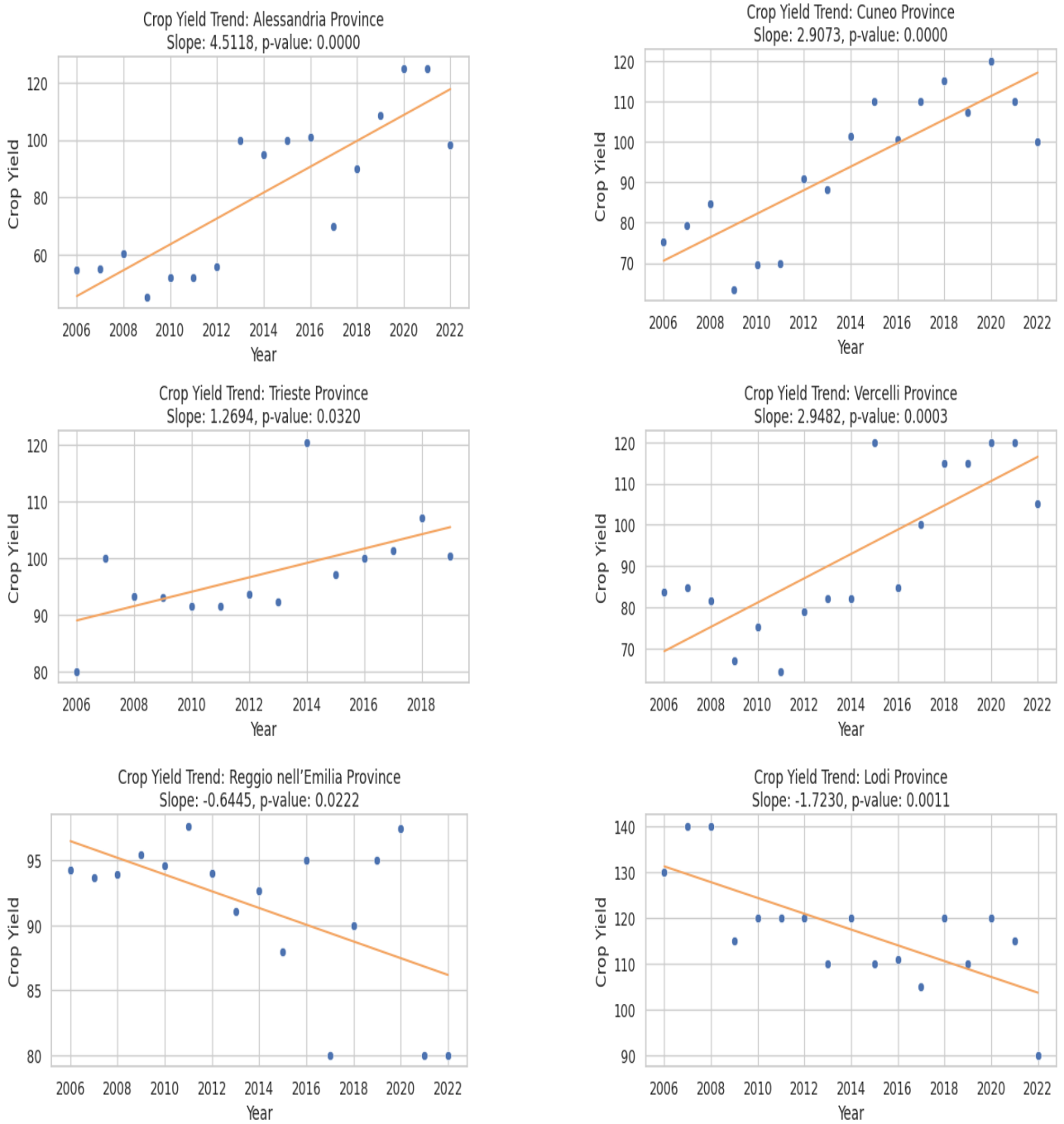


Figure 13: Crop yield trend (slope and p-value reported) for Alessandria, Cuneo, Reggio Emilia, Lodi, Trieste, Vercelli provinces

For negative slopes, a confidence interval entirely below zero confirms a significant decreasing trend, while for positive slopes, a confidence interval entirely above zero confirms a significant increasing trend. In both cases, a confidence interval that does not cross zero strengthens the reliability of the trend. Figure 14: Significant Crop Yield trends comparison with p-value and confidence interval illustrates all significant slope values, along with their respective p-values and confidence intervals.

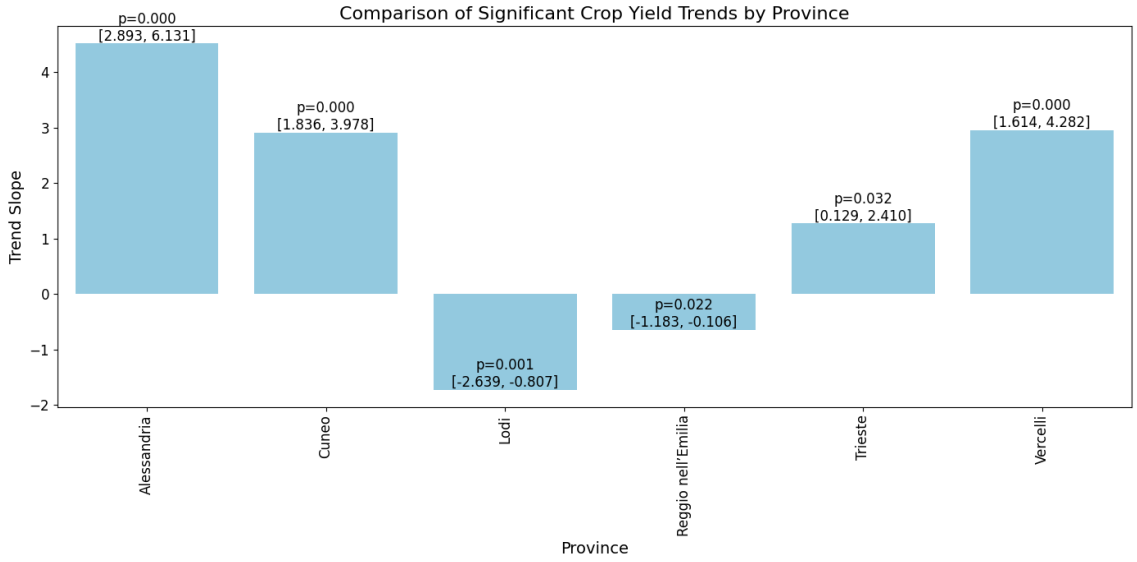


Figure 14: Significant Crop Yield trends comparison with p-value and confidence interval

The distribution of the detrended crop yield anomaly, as visualized in the histogram (Figure 14), does not conform to a Gaussian (normal) distribution. This conclusion is supported by the results of two formal statistical tests for normality: the Kolmogorov-Smirnov (K-S) test and the Anderson-Darling (A-D) test. The K-S test yielded a test statistic of 0.4448 and an extremely low p-value (7.728e-86), indicating a significant departure from normality. Similarly, the A-D test produced a test statistic of 2.1627, which exceeded the critical values for all significance levels (from 15% to 1%), further reinforcing the rejection of the null hypothesis that the data follows a normal distribution.

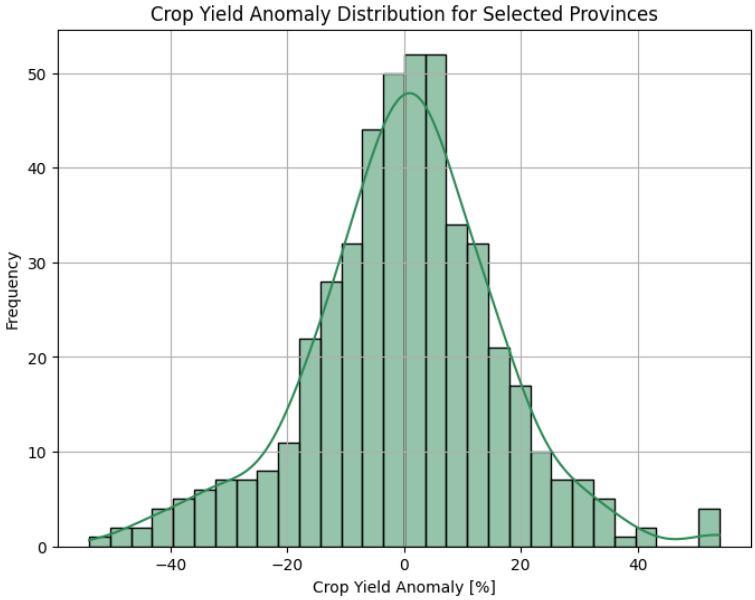


Figure 15: Detrended crop yield anomaly distribution for selected provinces

To better understand the spread of the detrended crop yield anomaly, the following statistical measures were calculated:

Table 2: Summary of key statistical measures for the detrended crop yield anomaly distribution, including standard deviation, interquartile range (IQR), range, and 95% confidence interval, providing insights into the spread and variability of the data.

Distribution measure	**Value**
Standard Deviation	16.4 %
Interquartile Range	18.91
Range	107.89
95% Confidence Interval	[-32.14, 32.14]

The standard deviation of 16.4% reflects the average variability in crop yield across provinces, while the Interquartile Range (18.91) shows the spread of the middle 50% of the data. Additionally, the 95% confidence interval indicates that most of the data lies within the range of [-32.14, 32.14], capturing the extent of the anomaly values. Given the nature of the data - characterized by regional variability and potential localized production shocks - this non-normality is not unexpected and does not hinder the analysis. The distribution, although non-Gaussian, remains informative for modelling purposes. Furthermore, the dataset's size and the variability between provinces likely contribute to

the observed distribution shape, which still provides a reliable basis for the subsequent steps of the model development.

3.1.2 Climatic Variables in the Selected provinces

This study does not focus on the validation of the dataset used, as the climatic data is derived from pre-existing, reputable sources, VHR-REA-IT [34] for historical data and VHR-PRO-IT [42] for future projections. Given the reliability of these sources, the dataset was taken as valid without further verification within this research.

The two graphs in Figure 16 and Figure 17 provide a comparison between historical (1981-2022) and projected future (2030-2070) climate data for the selected provinces. Specifically, they depict the mean and standard deviation of monthly temperatures (in Kelvin) and monthly cumulative precipitation (in kg/m^2) across the year.

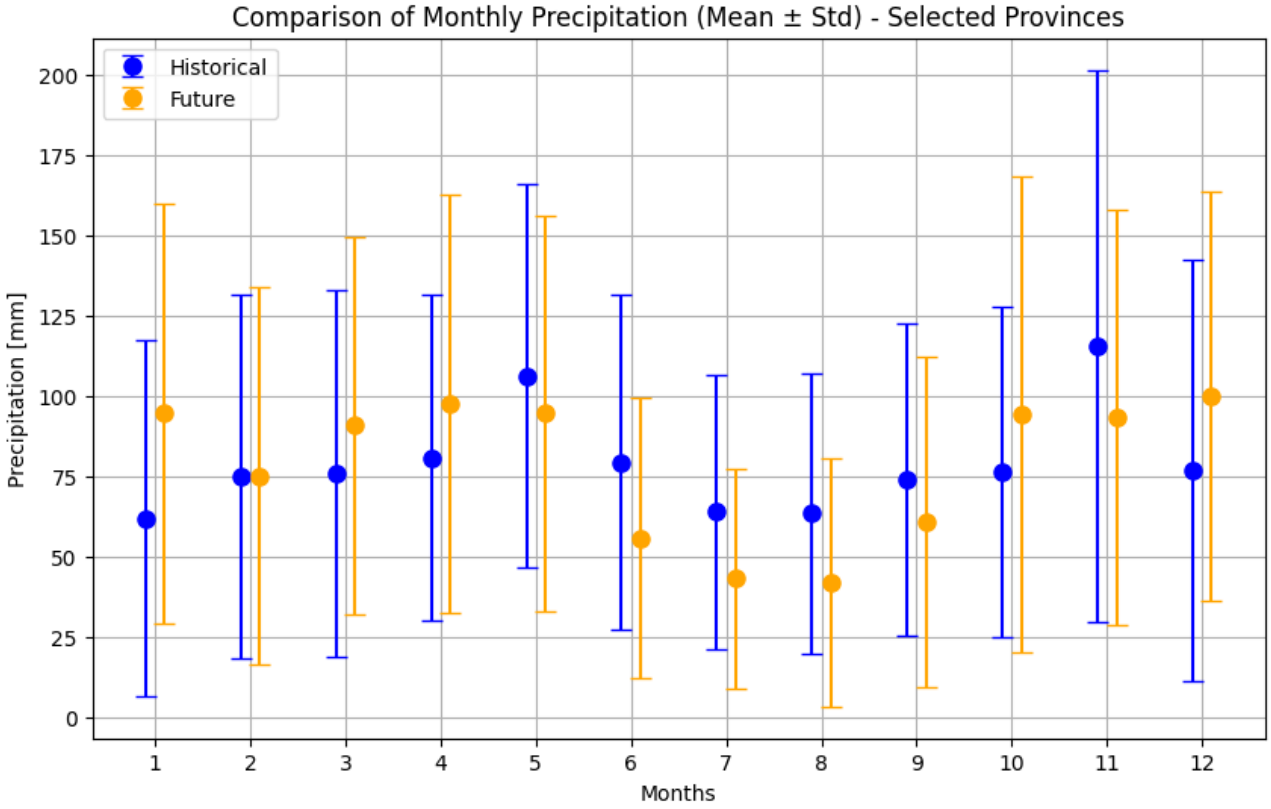


Figure 16: Monthly precipitation comparison, Historical (1982-2022) vs Projection (2030-2070) mean and standard deviation

The graph in Figure 16 illustrates the changes in monthly precipitation. Historical data is represented by blue dots, while future projections are shown in orange. The error bars reflect the standard deviation for each month. In the future projections, a general trend of reduced precipitation is observed, particularly in the summer months (June to August), where the projected values are significantly lower. The variability (standard deviation) is also more pronounced in certain months, such as January, indicating potential shifts in seasonal precipitation patterns.

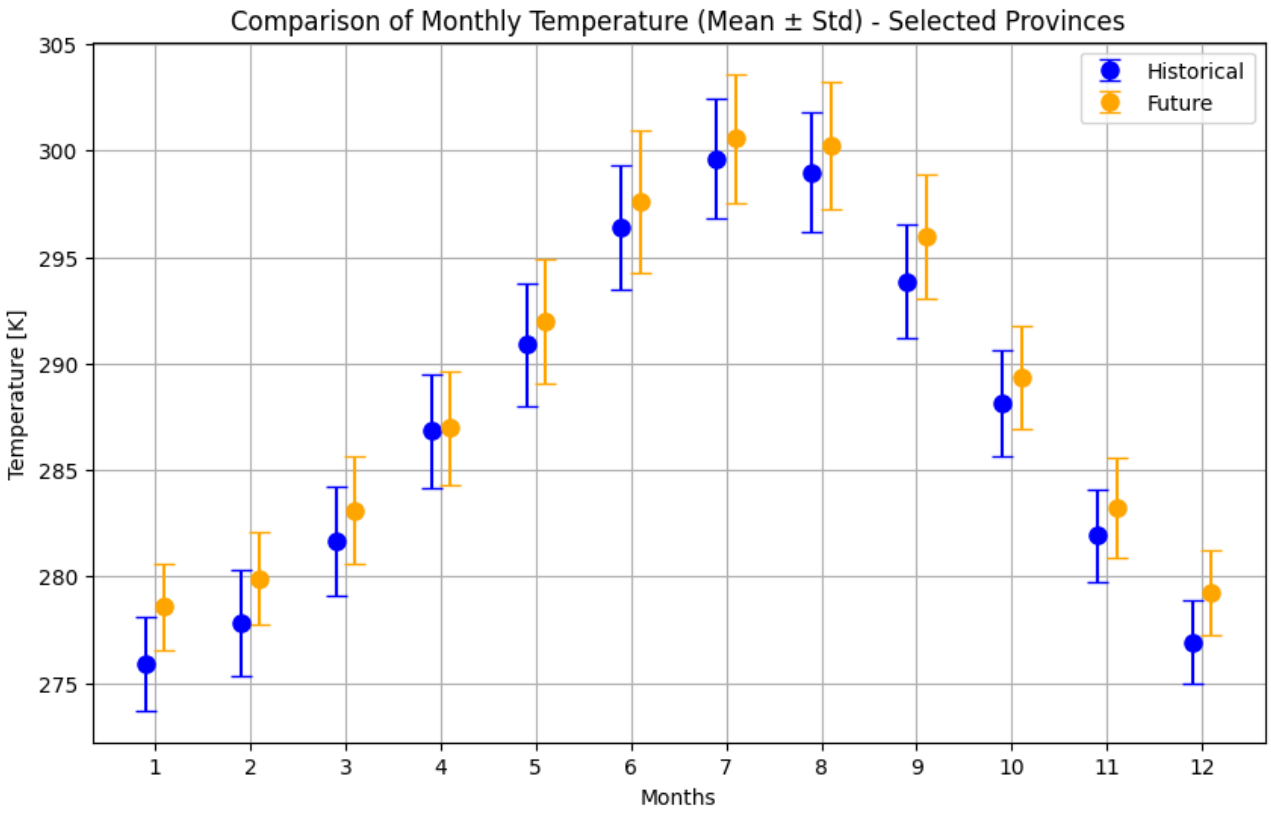


Figure 17: Monthly temperature comparison, Historical (1982-2022) vs Projection (2030-2070) mean and standard deviation

The graph in Figure 17 compares the monthly temperature data. The historical data (blue) and future projections (orange) show a clear trend of increased temperatures across most months. The warming is particularly evident in the summer months (June to August), where the mean temperature is projected to rise substantially. The standard deviations suggest that temperature variability is expected to remain relatively stable, although with a slight increase in some months, particularly in winter.

These visualizations highlight key differences between the historical and future climate scenarios for the selected provinces, qualitatively demonstrating the expected impacts of climate change in terms of both temperature and precipitation variability.

3.1.3 Variables and Target Correlation

The analysis evaluated the correlations (Pearson and Spearman) between the selected features (monthly temperature and precipitation variables spanning from March to October). Consecutive monthly temperatures showed a correlation, as expected. However, the correlation between temperature and precipitation features was less pronounced, suggesting that these variables do not exhibit strong interdependencies that would significantly impact their individual behaviour or the overall analysis.

Following this, both Pearson and Spearman correlation coefficients were calculated to explore linear and rank-order relationships between the features and the target variable (Crop Yield Anomaly). Figure 18 and Figure 19 display the results for the Pearson and Spearman correlations between the features and the target.

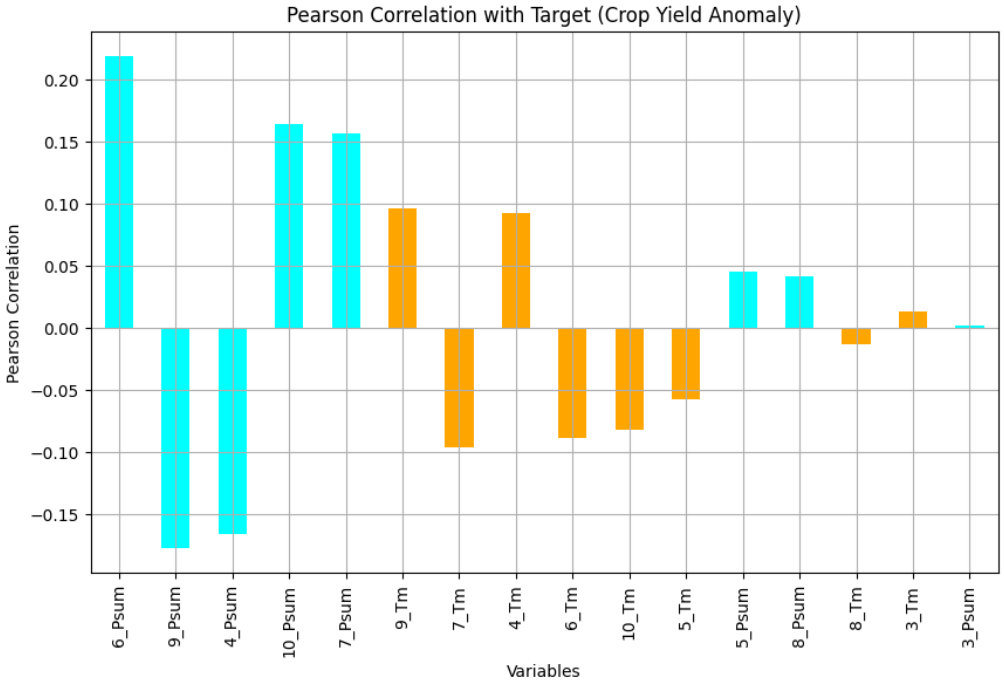


Figure 18: Variables-Target Pearson correlation

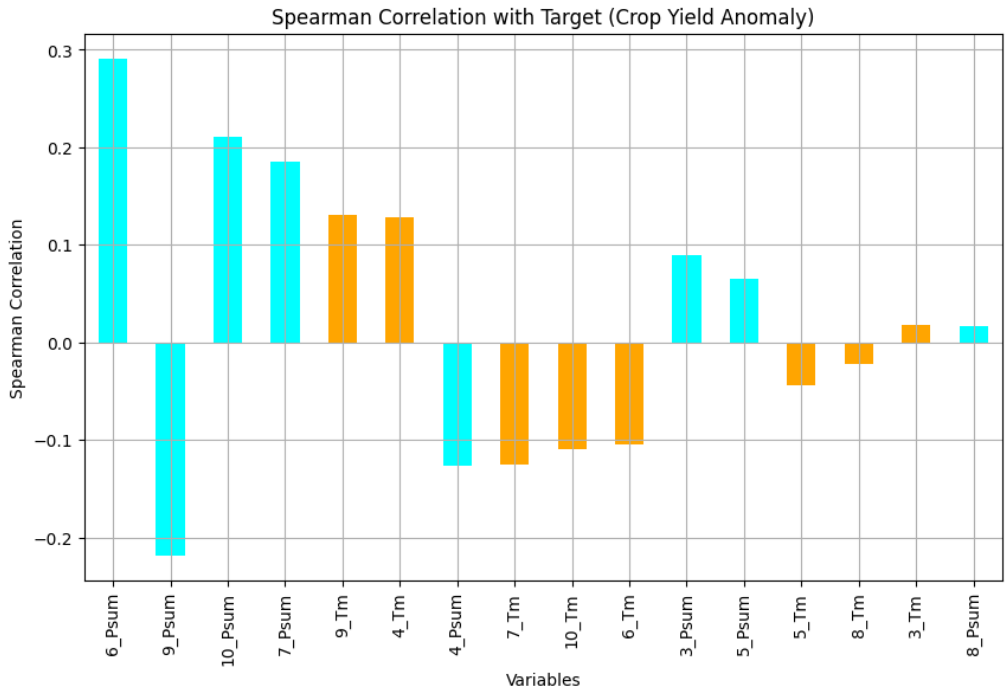


Figure 19: Variables-Target Spearman correlation

While some correlations between the features and the target are evident, none are particularly strong or highly significant. For instance, the highest correlations—such as those for precipitation in June and September (6_Psum and 9_Psum)—remain moderate and consistent with expectations from the literature. Even the best correlations are not strong enough to be considered highly predictive.

The Spearman correlation, which accounts for non-linear relationships, shows similar trends to Pearson but highlights a stronger positive correlation for June precipitation (6_Psum) and a negative correlation for September precipitation (9_Psum). This suggests that precipitation exerts a more noticeable influence on yield anomaly compared to temperature, though the overall impact is moderate.

This outcome aligns with literature findings [3], which suggest that while these variables are important, they are not the most reliable indicators for predicting drought impact on maize yield. They were chosen for their simplicity and availability in the context of producing an initial model creation process, rather than for their strong predictive power for yield anomalies.

3.2 Model Evaluation Results

This section presents the performance of the five different models, each subjected to various preprocessing techniques, evaluated based on three metrics: Mean Absolute Scaled Error (MASE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). The models are assessed using their optimal hyperparameters, selected based on the D metric, as described in the methodology chapter. Additionally, comparison graphs, including frequency distribution histograms and cumulative frequency graphs between observed and predicted values, are provided to offer a visual assessment of the models' accuracy.

3.2.1 Models performance: Metrics

The first step in the model evaluation process required each model to achieve a MASE score below 1. Only the Linear, Extra Trees, Random Forest, and K-Nearest Neighbours (KNN) models with original features met this criterion, indicating that their predictive error was lower than that of the naive model, which generates random predictions based on the mean and standard deviation of the real distribution. For these models, additional performance metrics, such as R^2 and RMSE, are also presented in Table 3 to further assess their predictive accuracy.

The MASE metric is designed to evaluate forecast accuracy by comparing the model's prediction error to that of a naive forecast. A MASE value below 1 suggests the model's predictions are more accurate than the naive method, while a value greater than 1 indicates inferior predictive accuracy. It is worth noting that the RMSE value is very close to the standard deviation of the observed data distribution. This emphasizes the importance of knowing that the MASE is less than 1, reinforcing the hypothesis that there is indeed a meaningful correlation, as the model does not simply learn the mean and distribution of the target. The model performs better than a naive model, which relies on predicting based on the mean and standard deviation of the observed data.

For the Random Forest, Extra Trees, and K-Nearest Neighbours (KNN) models, which require hyperparameter tuning, the optimal configurations were selected based on the D

metric (2.4.1), as summarized in Table 4. This table details the key hyperparameters used for each model, alongside their corresponding D values, further supporting the model evaluation process.

Table 3: Comparison of R^2 and RMSE values across different regression models using original data preprocessing. The R^2 column includes the 90% confidence intervals, highlighting the predictive accuracy of each model.

PREPROCESSING	REGRESSION MODEL	R2	R2 90% CONFIDENCE INTERVAL	RMSE
ORIGINAL DATA	LINEAR	0.14	[0.08, 0.20]	15.1
	RANDOM FOREST	0.11	[0.05, 0.17]	14.3
	EXTRA TREES	0.19	[0.14. 0.24]	14.9
	K-NN	0.22	[0.11, 0.31]	14.3

Table 4: Hyperparameter tuning results for Random Forest, Extra Trees, and K-Nearest Neighbours (K-NN) models that achieved MASE < 1. The selected hyperparameters and corresponding

REGRESSION MODEL	HYPERPARAMETER 1	HYPERPARAMETER 2	D_{R2}
RANDOM FOREST	n_estimators = 50	Max_depth = 2	0.301
EXTRA TREES	n_estimators = 70	Max_depth = 4	0.327
K-NN	n_estimators = 9	Weights = uniform	0.372

Overall, the R2 values across all models are relatively low, indicating limited predictive accuracy in this context. However, among the evaluated models, K-Nearest Neighbours (KNN) demonstrated the best performance. Additionally, it must be noted that the RMSE values for the models were very close to the standard deviation of the observed yield anomaly distribution, suggesting that some models may simply predict the mean and standard deviation of the data. For this reason, the MASE metric, which compares the model's performance to a dummy model that predicts based on the mean and standard deviation, is particularly useful. Models with a MASE value less than 1 indicate that they

have effectively learned patterns in the data, rather than just predicting the mean and standard deviation.

The final model selection focused on KNN, which utilized monthly average temperature and cumulative monthly precipitation from March to October as the features not preprocessed. A single KNN model was applied uniformly across all provinces to maximize the amount of data available for the model. By pooling data from all provinces into one model, the analysis benefited from a larger dataset. However, this approach may lead to the loss of more localized specificity, as the model does not account for unique regional differences.

3.2.2 KNN Model: Predicted vs. Observed Data

The analysis of the predicted values generated by the K-Nearest Neighbours (KNN) model reveals several notable characteristics and discrepancies when compared to the observed data. Figure 20 and Figure 21 illustrate the frequency distribution and the cumulative frequency of the predicted values. The distribution in Figure 20 shows a concentration of values around 0%, with the majority of predictions falling within a narrow range. The standard deviation of the predicted values is 9.2%, and the range is 47.56%, indicating that the predictions are tightly clustered around the centre. This is further supported by the interquartile range of 12.6 and the 95% confidence interval of [-0.64, 1.01], which highlight the limited variability captured by the model.

Table 5: Summary of Distribution Measures for Model Predictions. The table shows the standard deviation, interquartile range, range, and 95% confidence interval of the predicted values from the KNN model

Distribution measure	**Value**
Standard Deviation	9.2 %
Interquartile Range	12.6
Range	47.56
95% Confidence Interval	[-0.64, 1.01]

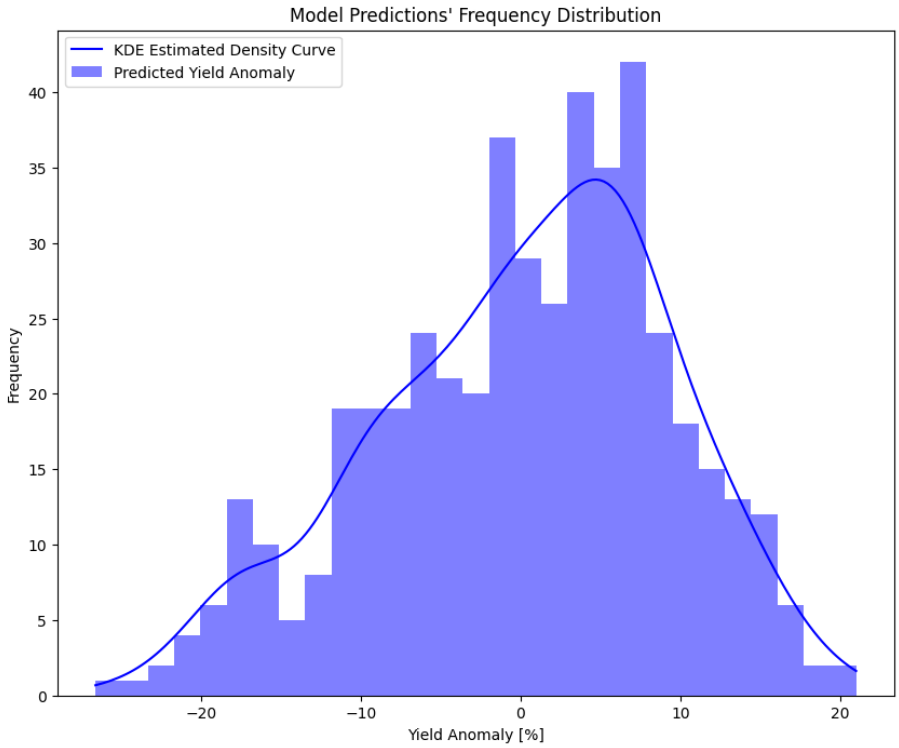


Figure 20: Frequency distribution of Model Predictions with density curve. The histogram and density curve illustrate the predicted values from the KNN model, with most values clustered near 0%

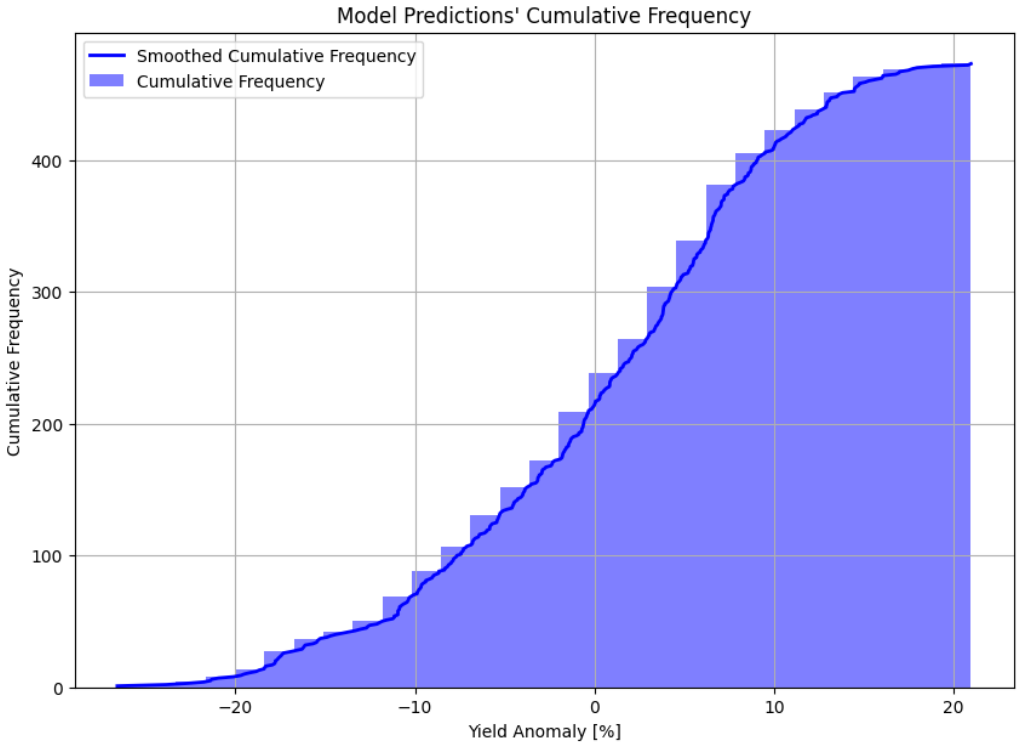


Figure 21: Cumulative Frequency of Model Predictions. The plot shows the cumulative frequency of predicted values from the KNN model, with most predictions concentrated around the centre

The cumulative frequency in Figure 21 reinforces this observation, with a steep rise in the curve indicating that most of the predicted values are concentrated around the mean, with relatively few extreme values. The predicted data exhibits a narrow spread, failing to account for a wider range of variability.

Figure 22 and Figure 23 then compare the observed values with the predicted data. The cumulative frequency comparison in Figure 23 highlights a key difference: the observed values are distributed over a much broader range, with more variability and extreme values, while the predicted data remains narrowly concentrated. Similarly, the frequency distribution comparison in Figure 22 illustrates that the predicted values exhibit a sharp peak, while the observed values are more evenly spread out.

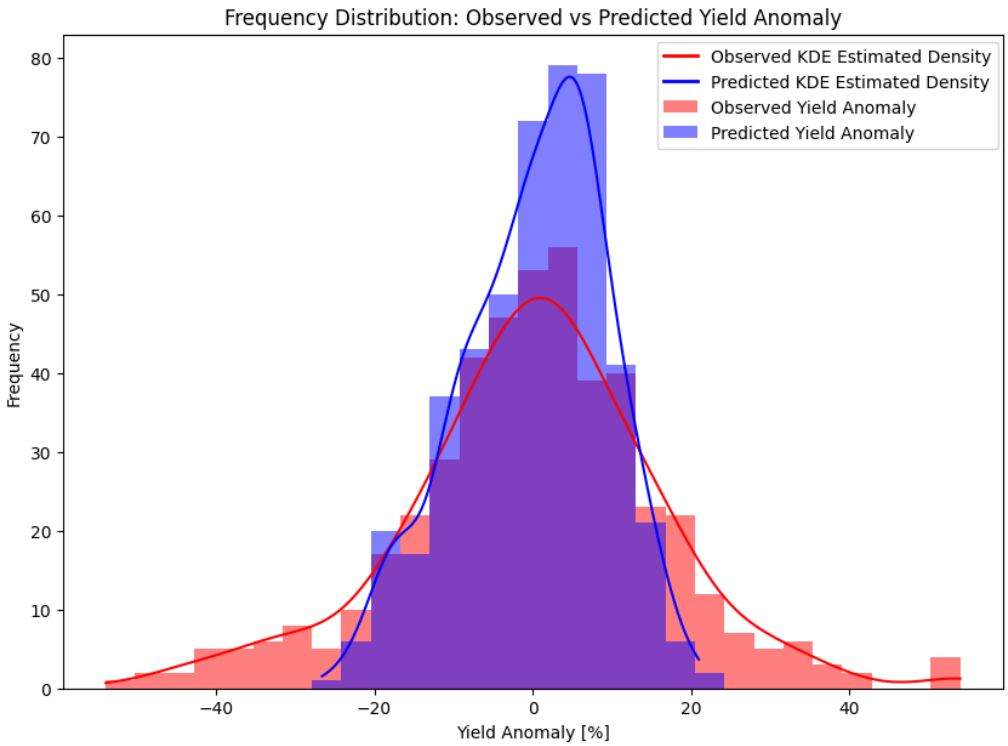


Figure 22: Frequency distribution comparison: Predictions (blue) vs Observed Values (red). The predicted values are tightly centred, while the observed values show more spread.

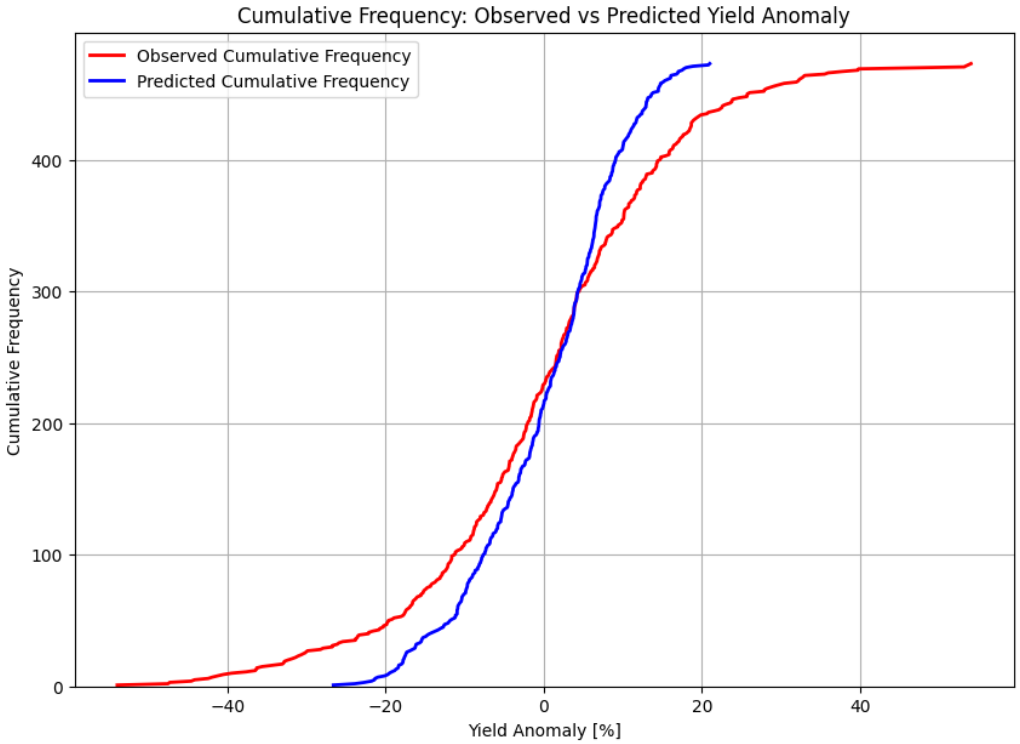


Figure 23: Cumulative Frequency Comparison: Predictions (blue) vs Observed Values (red). The predicted values are concentrated around 0%, while the observed values display greater variability

This underestimation of variability and failure to capture extreme values can be explained by two factors. First, the features used in the model may not be optimal, limiting the model's capacity to account for all factors influencing yield anomalies. Second, the relatively small size of the dataset may restrict the model's ability to generalize and capture more complex patterns in the data. Consequently, while the KNN model provides reasonably accurate central predictions, it struggles to accurately reflect the full range of observed yield anomalies.

As shown in the scatter plot (Figure 24), there is a weak but noticeable correlation between the predicted and observed yield anomalies.

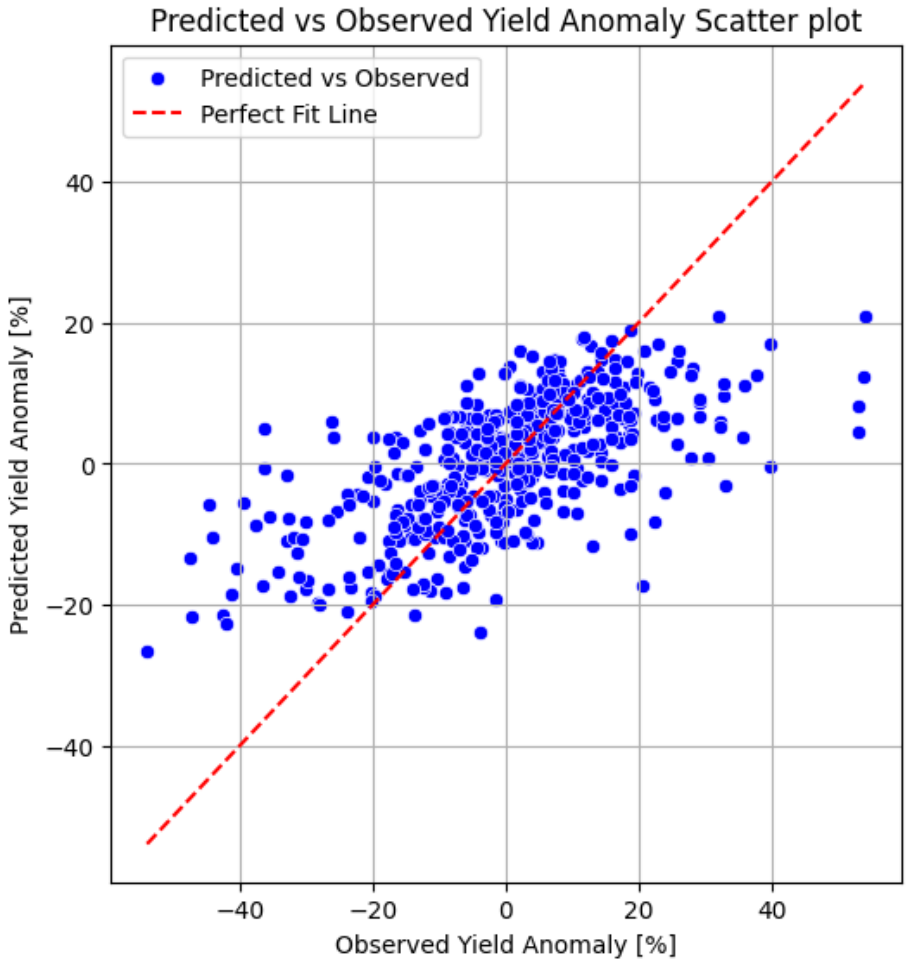


Figure 24: Predicted vs. Observed Yield Anomalies: the scatter plot compares predicted yield anomalies with observed values, with the red dashed line representing a perfect fit. While a weak correlation is evident, the model underestimates extreme values and struggles to capture the full variability of the observed data.

To further evaluate the model's performance, a binary classification of positive (gain) and negative (loss) yield anomalies was conducted. The confusion matrix in Figure 25 provides an overview of the model's ability to discern between positive and negative values. Despite struggling with extreme values, the model demonstrates reasonable performance, correctly classifying nearly 76% of the yield anomalies. Specifically, 41.0% of positive and 34.7% of negative yield anomalies were accurately predicted.

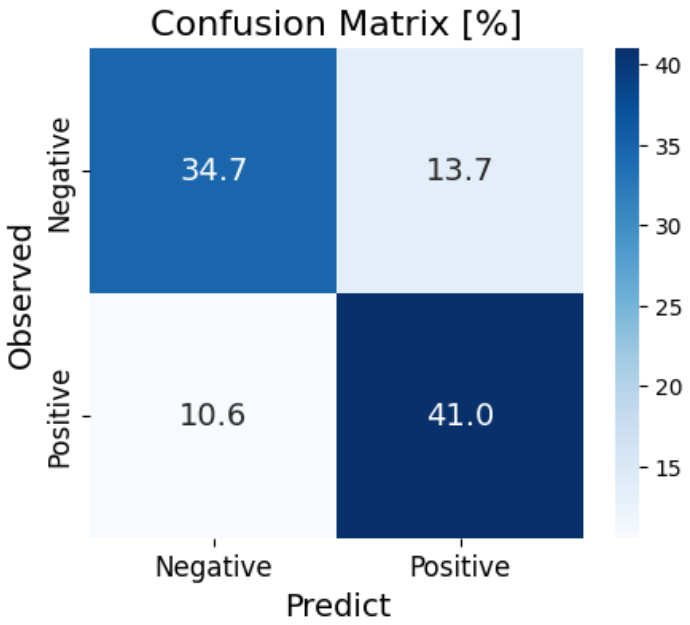


Figure 25: Confusion Matrix [%] of Predicted vs Observed Yield Anomalies. The matrix shows the model's accuracy in predicting positive (gain) and negative (loss) anomalies, with more errors in false positives

However, there is a significant error in the false positives (cases where the observed value is negative, but the model predicted positive). This is particularly concerning because accurately predicting losses (negative yield anomalies) is crucial for this study, which focuses on identifying and assessing agricultural damage. The 13.7% of cases where negative yield anomalies were incorrectly predicted as positive is an area that requires further improvement, as it directly impacts the model's utility in capturing important loss events.

While the binary classification performance is relatively strong compared to the overall predictive accuracy of the model, care was taken to mitigate overfitting throughout the analysis. Overfitting was controlled primarily through hyperparameter tuning. In the case of K-Nearest Neighbours (KNN), the key hyperparameter used to prevent overfitting was k (the number of nearest neighbours considered), as this determines the extent to which the model generalizes rather than memorizing training data. Additionally, overfitting was further minimized by employing 5-fold cross-validation during model training, which

ensured that the model's performance was evaluated across multiple subsets of the data, enhancing its ability to generalize to unseen records.

3.2.3 Feature Importance in KNN Model

This section presents the feature importance results from the K-Nearest Neighbours (KNN) model. These results provide insight into the relative influence of each climatic and agricultural feature used in the model to predict yield anomalies. By comparing the feature importance to the correlation analysis, it is possible to evaluate whether the same features that show strong correlations with yield anomalies are also prioritized by the model. This comparison helps in understanding how the model interprets the data and identifies the most influential factors affecting yield predictions.

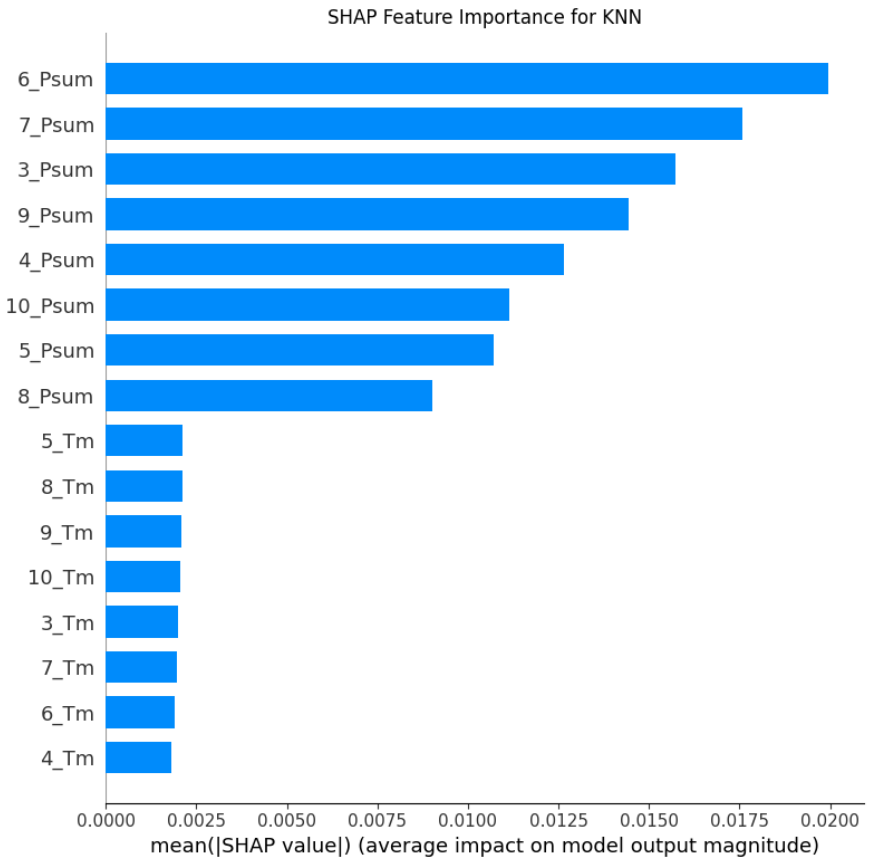


Figure 26: KNN Feature Importance (SHAP values)

The analysis of feature importance for the KNN model, as shown in the SHAP (SHapley Additive exPlanations) plots (Figure 26, Figure 27), reveals that precipitation-related variables have the greatest impact on predicting crop yield anomalies. Specifically, the precipitation values from months June (6_Psum), July (7_Psum), March (3_Psum), and September (9_Psum) are among the most influential features. In contrast, temperature variables play a smaller role in the model's predictions.

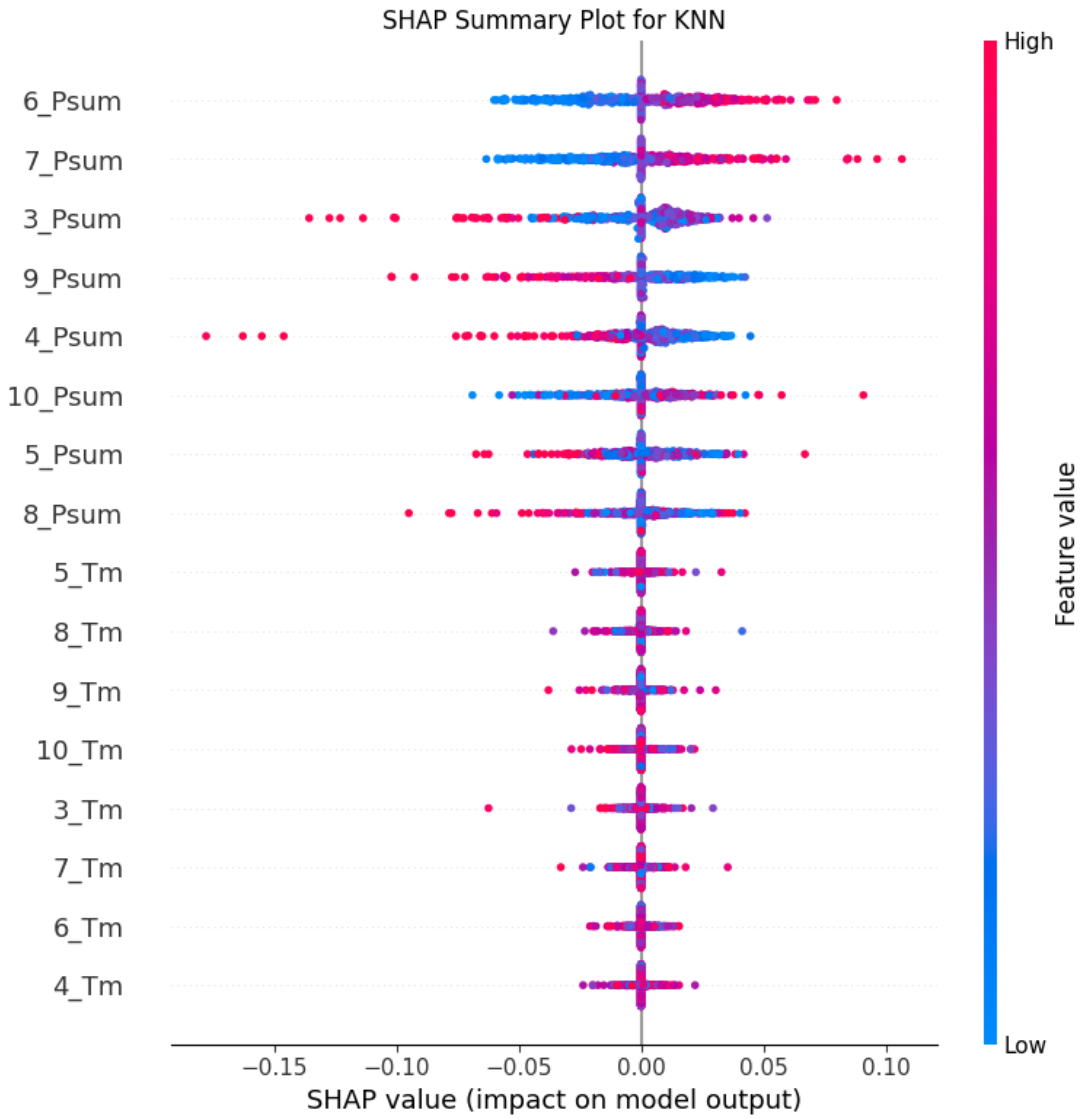


Figure 27: SHAP Summary Plot for KNN, showing the impact of each feature on yield anomaly predictions. Higher precipitation values (magenta) in June and July have the greatest influence.

The SHAP Summary Plot (Figure 27) provides further insights into how individual feature values influence the model's predictions. Features like 6_Psum and 7_Psum show a clear pattern where higher values (represented by the magenta points) tend to positively impact the model's output (i.e., higher yield anomaly predictions), while lower values (blue points) have a negative impact. This is consistent with the idea that higher precipitation during these months supports better crop performance [28]. On the other hand, temperature variables, such as May and September temperature, exhibit a smaller range of influence, further confirming that temperature has a limited effect on the model's predictions compared to precipitation.

This is fairly different from the results of the Pearson and Spearman correlation analyses (Figure 18, Figure 19), where temperature variables, especially September, July and April mean temperatures, show relatively higher correlations with the target variable (crop yield anomaly).

While both the feature importance and correlations highlight the relevance of precipitation, the KNN model places more emphasis on precipitation during key months like June, July, and September, reflecting the model's sensitivity to seasonal rainfall patterns. This aligns with the existing literature [3], [28], which emphasizes the dominant role of precipitation in determining crop yields, particularly for maize. On the other hand, the higher correlation of temperature variables in the correlation analysis suggests that temperature has a stronger direct linear relationship with yield anomalies, even though its influence is less prioritized by the KNN model.

3.3 Inference Results on RCP 8.5 Projection Using KNN

This section presents the results of the inference on the RCP 8.5 projection, using the K-Nearest Neighbours (KNN) model to estimate the average yearly yield anomaly. Although the model has several limitations at this stage, it is still applied to provide insights into potential future outcomes. The analysis focuses on understanding how future climate conditions under the RCP 8.5 scenario may impact agricultural production, with particular

attention to trends and deviations in yield anomalies. These results, despite the model's constraints, allow for an initial evaluation of possible future scenarios.

Table 6: Summary of key statistical measures for RCP8.5 scenario; 2030-2050 and 2050-2070 periods comparison

Distribution measure	Value (2030-2050)	Value (2050-2070)
Standard Deviation	7.3%	7.7%
Interquartile Range	9.85	9.41
Range	46.82	47.71
95% Confidence Interval	[-1.43, -0.24]	[-2.35, -1.10]

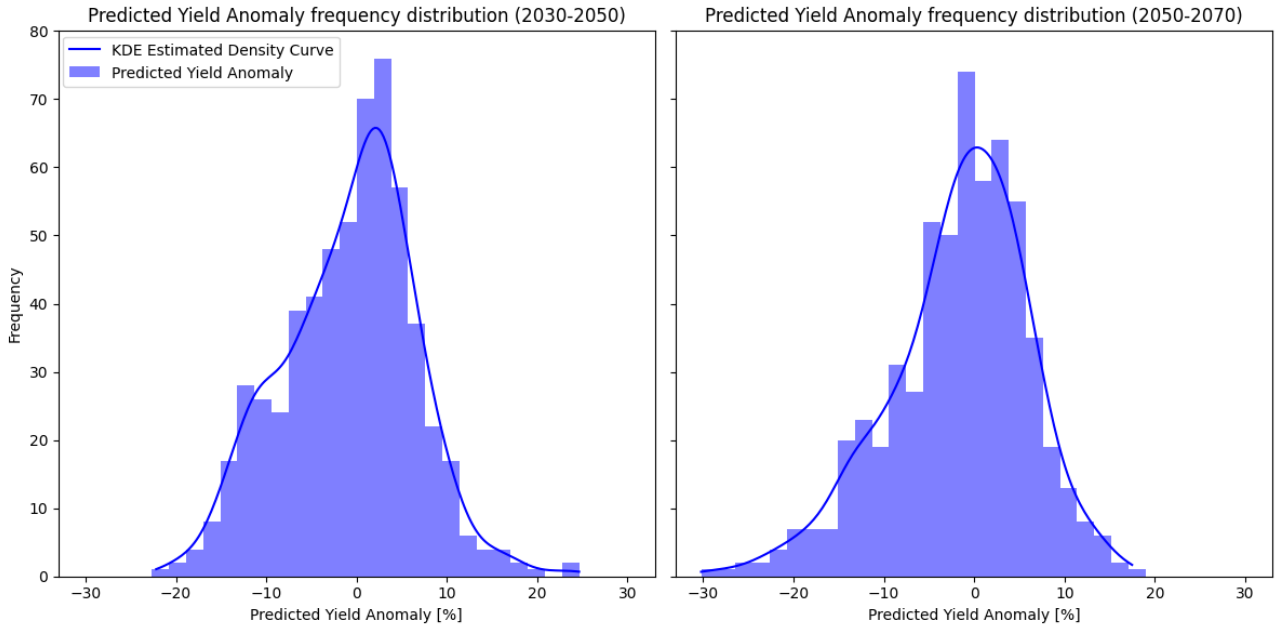


Figure 28: Frequency distribution of predicted yield anomalies for the periods 2030-2050 and 2050-2070 under the RCP 8.5 scenario

From both Table 6 and the distribution plot (Figure 28), it is evident that, despite the narrower spread compared to the present-day historical data, the projected distributions for the future (2030-2050 and 2050-2070) show a shift toward negative values compared to frequency distribution of baseline scenario (2006-2022). This shift indicates a consistent trend towards yield reductions in future periods. The narrower spread of the future projections is consistent with the model's limitations, as discussed in the model evaluation section, where it was noted that the model is less capable of capturing extreme events. In addition, the 95% confidence intervals for the future periods further

confirm this shift towards negative anomalies. Both future periods (2030-2050 and 2050-2070) show confidence intervals entirely in the negative range, indicating a greater certainty of negative yield anomalies in the future.

The reduced spread and the tighter confidence intervals for future periods compared to the present-day data suggest that the model predicts fewer extreme high or low yield years, but consistently negative yield anomalies under the RCP 8.5 scenario.

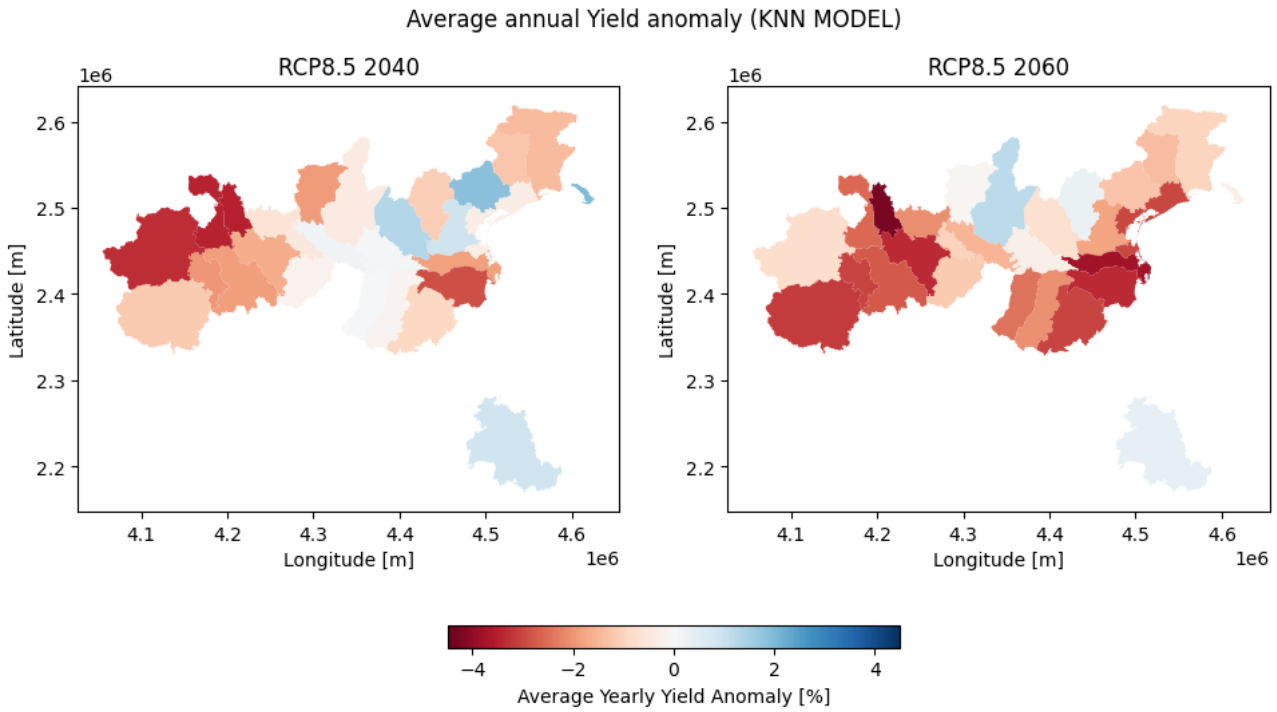


Figure 29: Average annual yield anomaly for selected provinces under the RCP 8.5 scenario, centred around 2040 and 2060

The spatial distribution of the average annual yield anomaly in the maps (Figure 29) aligns with the earlier analysis. The 2040 projection shows several provinces experiencing mild to moderate negative yield anomalies, with some provinces showing resilience with near-zero or slightly positive values. However, the 2060 projection reveals a worsening scenario, with a much broader area affected by negative anomalies. Many provinces face increasingly severe reductions in yield anomaly, with values dropping below -4% in some areas. This spatial pattern reinforces the trend observed in the yield anomaly distributions: as the timeline progresses towards 2060, the risk of yield reductions becomes more pronounced across key agricultural regions.

4 Results Discussion and Conclusions

The findings from the model evaluation and projection analysis show that, although the model achieved a low overall accuracy, it performed reasonably well in distinguishing the binary difference between loss and gain events. This distinction provides valuable insights into drought vulnerability and potential agricultural risks, even with limited predictive accuracy.

It is important to note that the model predicts a more constrained distribution, failing to adequately capture the extremes of yield anomalies. This limitation, consistent with the model's overall performance, contributes to an underestimation of the most severe impacts on crop yield. While this narrower distribution reflects the model's inability to detect extreme values, it still offers a useful perspective on more moderate trends in yield variation, aligning with broader patterns identified in the literature.

These results, however, are largely in line with expectations given the methodological limitations of the study. Several factors likely contributed to the model's constrained predictive capability. First, the absence of bias correction in the climatic data likely introduced errors into both the historical validation and future projections, limiting the accuracy of predictions. Additionally, the short time series of yield data and the limited spatial coverage constrained the model's ability to generalize, making it more difficult to capture extreme variations in yield anomalies.

The choice of features also played a critical role in shaping the model's performance. Based on the existing literature, it is known that the features selected for this analysis, while relevant, may not fully capture the complexity of factors influencing agricultural yield. For example, the absence of more refined variables related to evapotranspiration or soil moisture could have hindered the model's ability to accurately predict extreme losses or gains.

Moreover, the risk of spatial overfitting in the model's training process may have influenced the results. Provinces that are geographically and climatologically similar were

likely overrepresented in the training set, which could have limited the model's ability to generalize to other regions with differing conditions.

Furthermore, the yield anomaly may potentially be influenced by climatic factors beyond drought, such as extreme weather events like floods or hailstorms. However, the present study intentionally focused on assessing maize vulnerability specifically to drought conditions. As such, the contribution of other climatic events to yield anomalies was not considered. Including these factors would pose considerable challenges, as it would require disentangling the influence of various events and distributing their contributions to the overall agricultural anomalies.

Additionally, external factors such as irrigation practices and market dynamics could significantly impact yield outcomes. These influences were also beyond the scope of this analysis, given the specific focus on drought vulnerability. Irrigation, in particular, can play a crucial role in mitigating yield losses during periods of low rainfall, but reliable data on irrigation usage is often scarce and difficult to model. Similarly, market fluctuations, including changes in crop prices or demand, could alter planting decisions and yield results in ways that are not directly related to climatic factors.

Taken together, these limitations explain the relatively narrow distribution of predicted values and the model's underperformance in capturing extreme events. The results are, therefore, not unexpected, given the methodological constraints, and provide a clear direction for future work aimed at refining the model and improving its generalizability.

4.1 Recommendations for Future Work

Moving forward, several steps could be taken to improve the model and address its limitations. A priority would be to reduce the number of features used in the model. Given the relatively small size of the dataset, reducing the feature set could help minimize noise and improve the model's generalizability, particularly in future projections.

In addition to this, bias correction of the climatic data should be applied to both historical and projected datasets. This would enhance the model's accuracy and reliability, as bias-

corrected data would provide a more accurate foundation for future risk assessments. Furthermore, refining the features used in the model, such as incorporating more sophisticated variables like the Standardized Precipitation Evapotranspiration Index (SPEI), could allow for a more nuanced understanding of drought impacts, enhancing the overall quality of the predictions.

Additionally, using aggregated cereal data, which offers a more extensive historical record, could provide more robust insights. However, this approach comes with challenges, as it involves combining crops with different seasonal cycles, potentially complicating the interpretation of the results.

To further refine the model, one potential improvement would be to use a mask for the climatic data to focus only on cultivated areas. By limiting the data to relevant agricultural zones, the model could better reflect the actual conditions impacting maize cultivation, potentially improving accuracy and relevance in the predictions.

Another critical area for future work is refining the preprocessing strategies. The current preprocessing techniques were not as effective as anticipated, and it is likely that a more carefully designed approach could yield a significant improvement in the model's performance.

Once a more satisfactory model is achieved, it will be possible to conduct statistical inference to derive probability distributions with greater accuracy. This will be particularly useful for identifying extreme events through return periods, enabling a more accurate assessment of the risk of extreme yield anomalies

While the current regression model has been relatively successful in distinguishing between gain and loss in yield anomalies, it struggles to capture the full distribution of yield anomalies, particularly the extreme values, often predicting outcomes closer to the mean. As a potential future improvement, the adoption of multiple classification models, each trained on a specific yield anomaly threshold, could allow the model to better capture the distribution of anomalies, particularly at the extremes. By estimating the probabilities of exceeding or falling below different thresholds, it would be possible to

generate a more nuanced probabilistic distribution of predicted yield anomalies. This shift from regression to classification could improve the model's ability to handle varying degrees of risk.

4.2 Conclusions

This study has laid the groundwork for assessing maize vulnerability to drought under future climate conditions using a data-driven approach. While the results reflect certain limitations, they provide useful insights into how climate change may impact agricultural production in Italy's most productive provinces. Despite the model's relatively low overall accuracy, it demonstrated some capacity to distinguish between loss and gain events, which is a valuable step toward identifying areas at greater risk of agricultural damage due to drought.

One significant contribution of this work is the development of an initial framework for modelling yield anomalies at the provincial level. Although a single model was used across all provinces, the predictions are still specific to each province, reflecting their unique climatic conditions. This approach enables localized assessments while maintaining a consistent methodology, providing valuable insights into how regional differences may affect agricultural vulnerability to drought.

However, the model's limitations, such as its inability to capture extreme values and its relatively narrow predicted distributions, underscore the need for further refinement. The projections of average annual yield anomaly does not contradict existing literature [3], but the underestimation of extreme yield losses suggests that the true extent of future agricultural risk may be higher than predicted. The methodological constraints, including data limitations and the choice of features, contributed to these results, but the model still represents a quite meaningful step forward in understanding the impacts of climate change on maize cultivation.

Looking ahead, several key improvements have been identified. These include refining the feature set, applying bias correction to the meteorological data, and reducing the number

of features to improve generalization. Additionally, the proposal to shift from a regression-based approach to a classification model offers a promising direction for future work. By classifying yield outcomes at various thresholds, the model could generate a probabilistic distribution of predicted anomalies, providing a more detailed risk assessment. This approach could better capture the complexity of yield outcomes.

In conclusion, this study demonstrates the potential of provincial-level modelling for drought vulnerability assessments in agriculture. While the model's current limitations require further exploration and refinement, its ability to provide localized insights highlights its relevance for future agricultural strategies, particularly in the context of climate adaptation. By continuing to refine the model and expand its scope, future work could contribute to more accurate and actionable predictions, supporting efforts to mitigate the growing risks posed by climate change to the agricultural sector.

Reference List

[1] K. Calvin *et al.*, 'IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.', Intergovernmental Panel on Climate Change (IPCC), Jul. 2023. doi: 10.59327/IPCC/AR6-9789291691647.

[2] F. Giorgi and P. Lionello, 'Climate change projections for the Mediterranean region', *Glob. Planet. Change*, vol. 63, no. 2, pp. 90–104, Sep. 2008, doi: 10.1016/j.gloplacha.2007.09.005.

[3] L. Rossi *et al.*, 'European Drought Risk Atlas', JRC Publications Repository. Accessed: Sep. 22, 2024. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/handle/JRC135215

[4] S. Liu and F. Qin, 'Genetic dissection of maize drought tolerance for trait improvement', *Mol. Breed.*, vol. 41, no. 2, p. 8, Jan. 2021, doi: 10.1007/s11032-020-01194-w.

[5] D. B. Lobell *et al.*, 'Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest', *Science*, vol. 344, no. 6183, pp. 516–519, May 2014, doi: 10.1126/science.1251423.

[6] K. E. Trenberth, 'Climate change caused by human activities is happening and it already has major consequences', *J. Energy Nat. Resour. Law*, vol. 36, no. 4, pp. 463–481, Oct. 2018, doi: 10.1080/02646811.2018.1450895.

[7] V. Masson-Delmotte *et al.*, Eds., *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021. doi: 10.1017/9781009157896.

[8] P. Faggian, 'Future Precipitation Scenarios over Italy', *Water*, vol. 13, no. 10, Art. no. 10, Jan. 2021, doi: 10.3390/w13101335.

[9] 'Climate change, impacts and vulnerability in Europe 2016', European Environment Agency. Accessed: Oct. 02, 2024. [Online]. Available:

https://www.eea.europa.eu/publications/climate-change-impacts-and-vulnerability-2016

[10] ‘European Climate Risk Assessment’, European Environment Agency. Accessed: Oct. 02, 2024. [Online]. Available: https://www.eea.europa.eu/publications/european-climate-risk-assessment

[11] H. M. D. Goulart, K. van der Wiel, C. Folberth, J. Balkovic, and B. van den Hurk, ‘Storylines of weather-induced crop failure events under climate change’, *Earth Syst. Dyn.*, vol. 12, no. 4, pp. 1503–1527, Dec. 2021, doi: 10.5194/esd-12-1503-2021.

[12] K. S. Harmanny and Ž. Malek, ‘Adaptations in irrigated agriculture in the Mediterranean region: an overview and spatial analysis of implemented strategies’, *Reg. Environ. Change*, vol. 19, no. 5, pp. 1401–1416, Jun. 2019, doi: 10.1007/s10113-019-01494-8.

[13] J. Spinoni, J. Vogt, G. Naumann, P. Barbosa, and A. Dosio, ‘Will drought events become more frequent and severe in Europe?’, *Int. J. Climatol.*, vol. 38, pp. 1718–1736, Mar. 2018, doi: 10.1002/joc.5291.

[14] D. Bonaldo *et al.*, ‘The summer 2022 drought: a taste of future climate for the Po valley (Italy)?’, *Reg. Environ. Change*, vol. 23, no. 1, p. 1, Dec. 2022, doi: 10.1007/s10113-022-02004-z.

[15] Joint Research Centre (European Commission) *et al.*, *Drought in Europe: March 2023 : GDO analytical report*. Publications Office of the European Union, 2023. Accessed: Oct. 02, 2024. [Online]. Available: https://data.europa.eu/doi/10.2760/998985

[16] M. Bozzola and T. Swanson, ‘Policy implications of climate variability on agriculture: Water management in the Po river basin, Italy’, *Environ. Sci. Policy*, vol. 43, pp. 26–38, Nov. 2014, doi: 10.1016/j.envsci.2013.12.002.

[17] R. R. Heim, ‘A Review of Twentieth-Century Drought Indices Used in the United States’, Aug. 2002, doi: 10.1175/1520-0477-83.8.1149.

[18] A. F. Van Loon *et al.*, ‘Drought in a human-modified world: reframing drought definitions, understanding, and analysis approaches’, *Hydrol. Earth Syst. Sci.*, vol. 20, no. 9, pp. 3631–3650, Sep. 2016, doi: 10.5194/hess-20-3631-2016.

[19] A. AghaKouchak *et al.*, ‘Anthropogenic Drought: Definition, Challenges, and Opportunities’, *Rev. Geophys.*, vol. 59, Apr. 2021, doi: 10.1029/2019RG000683.

[20] Y. Li, K. Guan, G. D. Schnitkey, E. DeLucia, and B. Peng, ‘Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States’, *Glob. Change Biol.*, vol. 25, no. 7, pp. 2325–2337, 2019, doi: 10.1111/gcb.14628.

[21] S. M. Vicente-Serrano, S. Beguería, and J. I. López-Moreno, ‘A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index’, Apr. 2010, doi: 10.1175/2009JCLI2909.1.

[22] Z. Zajac *et al.*, ‘Estimation of spatial distribution of irrigated crop areas in Europe for large-scale modelling applications’, *Agric. Water Manag.*, vol. 266, p. 107527, May 2022, doi: 10.1016/j.agwat.2022.107527.

[23] S. Daryanto, L. Wang, and P.-A. Jacinthe, ‘Global Synthesis of Drought Effects on Maize and Wheat Production’, *PLOS ONE*, vol. 11, p. e0156362, May 2016, doi: 10.1371/journal.pone.0156362.

[24] P. Ranum, J. P. Peña-Rosas, and M. N. Garcia-Casal, ‘Global maize production, utilization, and consumption’, *Ann. N. Y. Acad. Sci.*, vol. 1312, no. 1, pp. 105–112, 2014, doi: 10.1111/nyas.12396.

[25] O. (ESS) LavagnedOrtigue, ‘FAO. 2023. Agricultural production statistics 2000–2022. FAOSTAT Analytical Briefs, No. 79. Rome.’.

[26] A. Maresma, A. Ballesta, F. Santiveri, and J. Lloveras, ‘Sowing Date Affects Maize Development and Yield in Irrigated Mediterranean Environments’, *Agriculture*, vol. 9, no. 3, Art. no. 3, Mar. 2019, doi: 10.3390/agriculture9030067.

[27] V. Benďáková, H. Nagy, N. Turčeková, I. Adamičková, and P. Bielik, ‘Assessing the Climate Change Impacts on Maize Production in the Slovak Republic and Their Relevance to Sustainability: A Case Study’, *Sustainability*, vol. 16, no. 13, Art. no. 13, Jan. 2024, doi: 10.3390/su16135573.

[28] ‘FAO. Maize, Crop Informations, Food and Agriculture Organization of the United Nations, Land & Water’. Accessed: Sep. 26, 2024. [Online]. Available: https://www.fao.org/land-water/databases-and-software/crop-information/maize/en/

[29] ‘ISTAT - Superfici e produzione - dati in complesso’. Accessed: Sep. 15, 2024. [Online]. Available: https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,Z1000AGR,1.0/AGR_CRP/DCSP_COLTIVAZIONI/IT1,101_1015_DF_DCSP_COLTIVAZIONI_1,1.0

[30] M. Todorovic, A. Caliandro, and R. Albrizio, ‘IRRIGATED AGRICULTURE AND WATER USE EFFICIENCY IN ITALY’.

[31] M. Fader, S. Shi, W. von Bloh, A. Bondeau, and W. Cramer, ‘Mediterranean irrigation under climate change: more efficient irrigation needed to compensate for increases in irrigation water requirements’, *Hydrol. Earth Syst. Sci.*, vol. 20, no. 2, pp. 953–973, Mar. 2016, doi: 10.5194/hess-20-953-2016.

[32] ‘Assessing Climate Change Impacts on Irrigation Water Requirements under Mediterranean Conditions—A Review of the Methodological Approaches Focusing on Maize Crop’. Accessed: Oct. 01, 2024. [Online]. Available: https://www.mdpi.com/2073-4395/13/1/117

[33] D. Lobell and S. Gourdji, ‘The Influence of Climate Change on Global Crop Productivity’, *Plant Physiol.*, vol. 160, Oct. 2012, doi: 10.1104/pp.112.208298.

[34] ‘CMCC DDS ERA5 downscaling @2.2 km over Italy’. Accessed: Sep. 15, 2024. [Online]. Available: https://dds.cmcc.it/#/dataset/era5-downscaled-over-italy/hourly

[35] M. Raffa *et al.*, ‘Very High Resolution Projections over Italy under different CMIP5 IPCC scenarios’, *Sci. Data*, vol. 10, no. 1, p. 238, Apr. 2023, doi: 10.1038/s41597-023-02144-9.

[36] ‘Crop production (apro_cp) Eurostat metadata’. Accessed: Oct. 06, 2024. [Online]. Available: https://ec.europa.eu/eurostat/cache/metadata/EN/apro_cp_esqrscn2_it.htm

[37] ‘Copernicus Climate Change Service, Climate Data Store, (2021): Temperature and precipitation climate impact indicators from 1970 to 2100 derived from European climate projections. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)’, DOI: 10.24381/cds.9eed87d5. Accessed: Oct. 06, 2024. [Online]. Available: https://cds.climate.copernicus.eu/datasets/sis-hydrology-meteorology-derived-projections?tab=overview

[38] M. Raffa *et al.*, 'VHR-REA_IT Dataset: Very High Resolution Dynamical Downscaling of ERA5 Reanalysis over Italy by COSMO-CLM', *Data*, vol. 6, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/data6080088.

[39] 'Highlander project'. Accessed: Sep. 14, 2024. [Online]. Available: https://highlanderproject.eu/

[40] H. Hersbach *et al.*, 'The ERA5 global reanalysis', *Q. J. R. Meteorol. Soc.*, vol. 146, Jun. 2020, doi: 10.1002/qj.3803.

[41] M. Raffa, A. Reder, M. Adinolfi, and P. Mercogliano, 'A Comparison between One-Step and Two-Step Nesting Strategy in the Dynamical Downscaling of Regional Climate Model COSMO-CLM at 2.2 km Driven by ERA5 Reanalysis', *Atmosphere*, vol. 12, no. 2, Art. no. 2, Feb. 2021, doi: 10.3390/atmos12020260.

[42] 'CMCC DDS Climate Projections RCP4.5 and RCP 8.5 downscaled @2.2 km over Italy'. Accessed: Sep. 18, 2024. [Online]. Available: https://dds.cmcc.it/#/dataset/climate-projections-rcp85-downscaled-over-italy/historical

[43] 'Statistical regions in the European Union and partner countries — NUTS and statistical regions 2021'. Accessed: Oct. 05, 2024. [Online]. Available: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-20-092

[44] J. R. Lamichhane *et al.*, 'Relay cropping for sustainable intensification of agriculture across temperate regions: Crop management challenges and future research priorities', *Field Crops Res.*, vol. 291, p. 108795, Feb. 2023, doi: 10.1016/j.fcr.2022.108795.

[45] 'RandomForestRegressor - Scikitlearn documentation', scikit-learn. Accessed: Sep. 18, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[46] 'ExtraTreeRegressor - Scikitlearn documentation', scikit-learn. Accessed: Sep. 18, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.tree.ExtraTreeRegressor.html

[47] ‘KNeighborsRegressor - Scikitlearn documentation’, scikit-learn. Accessed: Sep. 18, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

[48] A. J. Dobson and A. G. Barnett, *An Introduction to Generalized Linear Models*, 4th ed. New York: Chapman and Hall/CRC, 2018. doi: 10.1201/9781315182780.

[49] J. Avila, ‘scikit-learn Cookbook - Second Edition’. [Online]. Available: https://www.oreilly.com/library/view/scikit-learn-cookbook/9781787286382/

[50] A. Liaw and M. Wiener, ‘Classification and Regression by RandomForest’, *Forest*, vol. 23, Nov. 2001.

[51] P. Geurts, D. Ernst, and L. Wehenkel, ‘Extremely randomized trees’, *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[52] K. Taunk, S. De, S. Verma, and A. Swetapadma, ‘A Brief Review of Nearest Neighbor Algorithm for Learning and Classification’, in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[53] ‘XGBoost Documentation — xgboost 2.1.1 documentation’. Accessed: Sep. 18, 2024. [Online]. Available: https://xgboost.readthedocs.io/en/stable/

[54] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[55] C. Molnar, *9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning*. Accessed: Oct. 05, 2024. [Online]. Available: https://christophm.github.io/interpretable-ml-book/shap.html

[56] ‘An introduction to explainable AI with Shapley values — SHAP latest documentation’. Accessed: Sep. 18, 2024. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

[57] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, ‘Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research

Frontier]', *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 59–76, Nov. 2018, doi: 10.1109/MCI.2018.2866730.