Using machine learning and process-based crop modelling for regional scale prediction

Joseph William Gallear

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Earth and Environment

July 2023

Declaration

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

Firstly, I would like to thank my supervisors Andrew Challinor, Anthony Cohn, Netta Cohen and Julia Chatterton for support and guidance throughout my PhD. I have very much enjoyed learning from and working with you over the last few years. I would also like to thank the climate impacts group at Leeds including Stewart Jennings, Chetan Deva and Ioannis Droutsas for fascinating discussions, valued feedback, and useful advice.

Thanks to Jim Watson formerly of the climate impacts group (now working for the government of New South Wales) for providing some of the GLAM model parameter values and observed data used in this thesis. This work was funded by the Natural Environment Research Council (NERC), part of UK research and Innovation (UKRI) with an additional CASE award from Unilever PLC. I would like to thank them for making this project possible.

Personal thanks to my Friends and Family for their support throughout this project, in particular my parents for their unwavering support and the valued friendships I have made during my time at the university of Leeds.

Abstract

The aims of this thesis are to assess the effectiveness of machine learning techniques in comparison to process-based crop growth models for the purpose of prediction of the impact of climate variability on crops at the regional scale. Comparisons are made between popular and most effective machine learning methods to predict crop yields and the process-based crop growth model GLAM. Firstly, it is asked how much data is required for machine learning to outperform process-based crop modelling, and under which conditions? Secondly, the prediction performance of both methods for prediction of crop failures is compared as well as the effect of potential errors in climate data. Thirdly, machine learning and crop modelling are compared to bench-mark crop model sensitivity to climatic drivers of crop yield, hence providing a data driven approach to learn how to further improve crop model simulations. Results show that machine learning and process-based crop modelling have contrasting strengths and weaknesses. However, machine learning can be leveraged to improve process-based crop modelling through increased sensitivity to climatic drivers of crop yields. Furthermore, the effects of potential errors in data upon machine learning simulations is determined. In doing so it is shown that sensitivity of machine learning to climatological errors varies depending on model, and region, with different time-scales of effects depending on if errors are in temperature or rainfall. Overall, this thesis shows that machine learning can provide great benefit to regional scale crop yield prediction. However, due to disadvantages of reduced model interpretability and difficulty in predicting effects of extreme events, machine learning is not a perfect solution for regional scale crop yield prediction. Therefore, it is argued that machine learning should be used in cooporation with process-based crop modelling to improve understanding rather than replace existing methods or knowledge.

Contents

List of Figures

List of Tables

Abbreviations

- AgMIP - Agricultural Model Inter-comparison Project
- APSIM - Agricultural Production Systems Simulator
- CART - Classification and regression tree
- CCOEF - Pearson's correlation coefficient
- CMIP - Climate Model Inter-comparison Project
- CNN - Convolutional Neural Network
- DUL - Soil moisture drained upper limit
- ECMWF - European centre for Medium Range Weather Forecasts
- FAO - Food and Agriculture organisation of the United Nations
- FFNN - Feed Forward Neural Network
- GBM - Gradient Boosting Machine
- GCM - Global Climate Model
- GDHY - Global Dataset of Historical Yields for major crops
- GLAM - General Large Area Model for annual crops
- JULES - Joint UK Land Environment Simulator
- LPJML - Lund Potsdam Jena Managed Land model
- KNN - K-Nearest neighbours

- LAI - Leaf area index
- ML - Machine learning
- MLR - Multiple Linear Regression
- NN - Neural Network
- NRMSE - Normalized Root Mean Square Error
- NUTS - Nomenclature of territorial units for statistics
- PCA - Principal Component Analysis
- RFR - Random forest regression
- RLL - Lower limit of soil moisture
- RNN - Recurrent Neural Network
- SAT - Saturated upper limit of soil moisture
- SSA - Sub-saharan Africa
- SVM - Support Vector Machine
- W2015 - Watson et al. (2015) GLAM parameter set
- YGP - Yield gap parameter

1 Introduction

1.1 Motivation and Overview

The aims of this thesis are to assess the effectiveness, of machine learning (ML) techniques in comparison to and collaboration with process based crop growth models for the purpose of prediction of the impact of climate variability. As a case study model, this thesis uses the General Large Area Model for annual crops (GLAM) (Challinor et al. 2004) for this purpose. Although crop models are applied at a range of spatial scales, this thesis focuses primarily on the regional scale use of models. This thesis can be broken down into three key questions which can be summarised as follows:

1. Under which environmental conditions can machine learning models out-perform the GLAM process based crop model?

2. How does climate model uncertainty affect the ability of machine learning models to predict crop yield and failures?

3. Can machine learning be used to determine how & where to improve model calibration / parameterization?

The initial work in this thesis focuses on the direct comparison of a mechanistic (or process based) approach and machine learning approaches. This provides a useful premise to further investigate ways in which the powerful pattern recognition capabilities of machine learning can be appropriately leveraged to improve mechanistic crop modelling. Before this, machine learning approaches are further interrogated by investigating the relative effects of uncertainty in temperature and rainfall time series on machine learning model output predictions of crop yield. The aims, rationale and methods of each chapter are briefly summarised in the following paragraphs.

The primary research question of chapter 3 of this thesis is: can machine learning models outperform process based crop modelling and if so, how many years of data are required and which machine learning frameworks are most appropriate. To answer this, machine learning methods are compared to the existing calibration of the well established GLAM crop model from an earlier study taken from the scientific literature (Watson et al. 2015). The existing model setup allows for a range of conditions present for the testing and evaluation of both the ML models and GLAM. Included in this test dataset are the effects of the 2003 European heat wave on French maize yields. The inclusion of this extreme event in the test set leads to questions of how and why different models and approaches produce such different results when tested in extreme conditions. Contrasting model performance from different conditions leads to the construction of a meta-model tool to determine the probability each method will be the best model given particular climatic conditions.

Chapter 4 aims to compare the results of different machine learning methods used in the previous chapter against rainfall and temperature uncertainty of different magnitudes for the prediction of crop yields and crop failures. In this context, crop failure is taken as a relative measure of yield anomaly according to deviation from the historical mean yield. The purpose of this experiment is to determine the sensitivity of machine learning models to uncertainty broadly representative of climate model uncertainty, therefore showing how different ML frameworks differ in sensitivity and ability to extrapolate. This experiment is undertaken by applying a series of perturbations to rainfall and temperature time series to simulate climate model uncertainty at different timescales such as monthly offsets from the yearly mean and daily offsets from the monthly mean. This lead to a set of simulations each for perturbed magnitudes of each dimension of the perturbations and a baseline set of predictions. Each perturbed simulation was compared to the baseline simulations to identify patterns in which model performance may vary dependent on results of the climatological perturbation.

Chapter 5 builds upon the approach and findings of chapter 3 to discuss methodologies in which machine learning methods can improve the results of mechanistic crop modelling. The main approach discussed is the use of machine learning methods to benchmark process based crop models to improve calibration through identification of agro-climatic relationships and regimes which lead to poor crop model performance in comparison to the benchmark machine learning frameworks. This chapter focuses primarily on the hybrid approach, of integration of machine learning and process understanding. Hybrid methodologies of mechanistic modelling and machine learning are presented in three approaches, machine learning to assist crop model calibration, machine learning to improve prediction of crop model sub-processes and which process knowledge is of most value for machine learning methods. The answer to this third question provides additional key insight as to whether crop models and machine learning models achieve the answers they arrive at for the right reasons.

In summary, the specific research questions for each chapter are as follows:

1. Chapter 3

    (a) How many years of data are required for machine learning models to outperform process based simulations from the GLAM crop model?

    (b) Which machine learning frameworks are most skillful for crop yield estimation at the regional scale?

2. Chapter 4

    (a) What is machine learning performance for crop yield failure prediction across contrasting environments?

    (b) How would climate input data uncertainty affect machine learning model per-

formance and correct failure prediction rate?

3. Chapter 5

   (a) How do crop yield responses to climate conditions differ across models and observed data?

   (b) What value do modelled process outputs have for improving machine learning predictions?

   (c) Can insights from machine learning identify targets for crop model improvement?

1.2 Climate variability and crop yields

Variations in climate can have significant impacts on a variety of sectors such as forest fire hazards (Siegert et al. 2001), human health (Wang et al. 2016), economic and social impacts (Pielke Jr & Landsea 1999, Carleton & Hsiang 2016), as well as agriculture (Iizumi et al. 2013, 2014a), which can have knock on effects for global and regional food security, prices, well-being and nutrition (Porter et al. 2014).

Plant growth and subsequent crop yields are affected by variations in rainfall, temperature, radiation and nutrients. The relative influence of each of these factors will depend on the growth and development stage of the the plant. Growth is defined as an irreversible increase in the dry biomass of the plant resulting from the maintenance of a disequilibrium between the accumulation and loss of these 4 environmental resources (Atkinson & Porter 1996). Each of the four resources act through different mechanisms to alter growth and development. Temperature acts to alter the rates of change in the plant system (Atkinson & Porter 1996). Optimum temperature is a specific threshold value which corresponds to the maximum rate of growth. Also important are minimum and maximum temperatures,

which act as thresholds below and above which damage to a plant will occur (Prasad et al. 2017). These three temperature values are collectively known as cardinal temperatures (Porter & Gawith 1999).

Radiation has an accumulation effect on plant growth, a fraction of absorbed radiation is stored in the chemical bonds of carbohydrates, the proportion of which is determined by radiation use efficiency (RUE) (Monteith & Moss 1977). Radiation is usually expressed as a flux of energy per square unit area of ground (for example W/m^2). Solar radiation and crop yield generally have an inverse relationship, as overcast conditions generally associated with rainfall will reduce incoming solar radiation. Rainfall generally has a positive relationship with crop yields, although excess rainfall can cause waterlogging and so result in yield losses depending on the duration of the flooding event (Sullivan et al. 2001). As crop growth and yield are biological processes they are subject to limiting factors. Limiting factors are minimum requirements of resources which limit growth. Sacks et al. (2010) have presented how either temperature or rainfall can act as limiting factors and so drive crop yield variability and hence determine the optimal planting and harvest dates of crops for different regions.

Variations in climate can take place at a variety of spatial scales and will affect inter-annual variability in crop yields. For example, climate oscillations resulting from the El Niño Southern Oscillation can influence crop yield variability. (Iizumi et al. 2014a) have shown that El Niño years are associated with increases in global soybean yield variation of between $2.1-5.4\%$ and will change the global mean yield anomaly of wheat, rice and maize between $-4.3+0.8\%$. The inverse la Niña years can decrease crop yields by up to $\sim 4.5\%$. Furthermore, climate variability has been estimated to account for $\sim$ 32–39 % of observed yield variability globally, with this figure rising to >60% in important "bread-basket" regions responsible for large proportions of global production such as the maize belts of China and the mid western USA (Ray et al. 2015). Extreme events such as hot and cold

or dry spells can have severe impacts on crop yields, extreme events are better correlated with climate variability as opposed to mean changes in climate typically associated with climate change (Semenov & Porter 1995).

Large scale impacts of climate variability can result in regional scale crop failure events. Crop failures are often defined as relative decreases in from a regional threshold value which will have cascading effects on food prices and security (Mendelsohn 2007). Extreme event impacts on crops are characterized by the two most important mechanisms: drought and heat stress (Troy et al. 2015, Vogel et al. 2020, Li et al. 2023, Lesk et al. 2016, Ciais et al. 2005). The following section will describe these mechanisms and effects in more detail including joint effects.

1.2.1 Heat and drought stress mechanisms on crops

Heat stress effects on crops can be split into two categories based on the time period over which the effect takes place. Heat shock is categorized as short term extremely high temperatures, which will depend on the lethal temperature of the crop (Prasad et al. 2017). In contrast, chronic heat stress consists of moderately high temperatures above the optimum growing temperature of the crop for a longer period of time (Li et al. 2013, Prasad et al. 2017). The timing of heat shock can be particularly important. Crops such as wheat, rice and sorghum have a sensitive period lasting between 5 to 9 days before anthesis (flowering) (Prasad et al. 2017).

Similarly, drought stress will also have a more severe effect depending on the timing of the onset of drought stress relative to growth development stages. Drought around the flowering (anthesis) and grain filling stages (named terminal drought stress) in wheat is most impeding to crop yield (Farooq et al. 2014). Drought causes damage to the plant through several mechanisms including accelerated leaf senescence, oxidative damage to photo-assimilatory machinery, reduced rates of carbon fixation and assimilate allocation,

pollen sterility, reduced grain set and development and reduced sink capacity (Farooq et al. 2014). Leaf senescence is the reduction of chlorophyll content in leaves, leading to reduced photosynthetic rate (Yang et al. 2003), reducing plant growth. Drought stress around the reproductive stage can cause male sterility in wheat plants (Dorion et al. 1996), further showing the importance of the timings of such stresses.

Heat and drought stress have interactive effects on crops. Particularly, the negative effect of high temperature on carbon assimilation is increased by drought, and drought accelerates decline in water use efficiency at high temperatures. Although temperature and drought effects have a great deal of importance placed upon them, joint effects have received very little attention within the scientific literature (Urban et al. 2018). The failure to consider joint effects may be a reason why mechanistic crop models often will under-predict the effects of drought and heat stress (Heinicke et al. 2022, Schewe et al. 2019).

1.3 Climate Change and its impacts

Climate change will exacerbate the frequency and severity of extreme events and climate variability (Abram et al. 2021, Porter et al. 2014, Ossó et al. 2022, Robinson et al. 2021, Caparas et al. 2021). Greater variability in climate leads to a greater probability of occurrence of extreme weather outside of the historic record (Porter & Semenov 2005). Rainfall variability is of particular importance to agricultural gross domestic product (Thornton et al. 2014), however many regions may see the variability of rainfall increase with a warming climate (Rind et al. 1989, Zwiers & Kharin 1998). Globally, rainfall variability may increase by 3-4% $^{\circ}C^{-1}$ under a high warming, high emission scenario over land and 2-4% $^{\circ}C^{-1}$ over ocean (Pendergrass et al. 2017).

Climate changes will affect both rainfall and evapotranspiration patterns leading to further drought impacts. Modelling has predicted that severe drought, which occurs at 5%

frequency today will occur about 50% of the time by the 2050s if green house gas emissions continue to increase (Rind et al. 1990). Similarly, heat wave events will continue to rise in frequency, intensity and duration with continued warming (Thiery et al. 2021). Mean temperature increases will also have impacts such as sea level change and crop yield reductions (Lyu et al. 2014, Asseng et al. 2015). Food production and agriculture is an impact sector of high importance which has seen much research and discussion.

1.3.1 Impacts of climate change on crops

Impacts of climate change on crops vary according to a large degree of factors. Factors include crop type and cultivar, climatic region (temperate or tropical), level of adaptation, co_2 concentrations, nutrient and water resource limitations, and pests and diseases (Challinor et al. 2014). In general climate change is expected to reduce yields, especially in developing regions (Challinor et al. 2014, Asseng et al. 2015, Jägermeyr et al. 2021, Zhao et al. 2017, Müller et al. 2011, Lobell et al. 2008a).

Long term temperature increase is significant for crop yields due to effects on metabolism. Increases in temperature decrease the grain filling period of growth, meaning the crop will reach maturity faster with reduced yield. Model simulations have predicted that global wheat production is estimated to fall by 6% for each °C further temperature increase and become more variable over space and time (Asseng et al. 2015). Effects of temperature increases will have different effects depending on the optimum temperature of the crop. Wheat is generally accepted to have an optimum growing temperature range of 17-23°C over the course of the growing season, with growth stopping above 37°C (Porter & Gawith 1999). Rice and maize have higher optimum temperatures than wheat (both 30 °C) (Sánchez et al. 2014, Prasad et al. 2017), for this reason, changing crops, or developing more resilient varieties has often been included as an adaptation measure to increasing temperatures.

1.3.2 Impacts of climate change on food security

Food security is met when "all people at all times have physical and economic access to sufficient, safe, and nutritious food to meet their dietary needs and food preferences for an active and healthy life" (Porter et al. 2014). This means that food security is affected by many factors. Climate change can interact with the already existing pressures on food security creating more volatile markets and prices (Porter et al. 2014). As a country specific example, Figure 1.1 shows how the effects of climate change will interact with other pressures on UK food security, eventually leading to decreased food affordability for vulnerable groups and subsequent health impacts (Challinor et al. 2018).

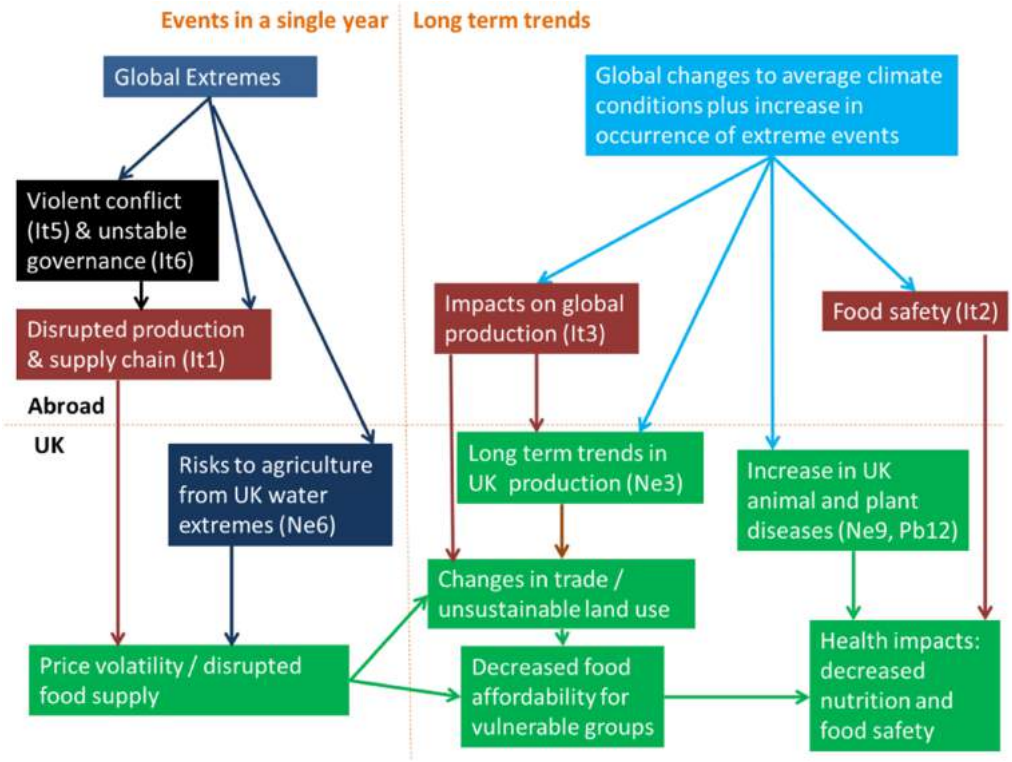

Figure 1.1: Risks to UK food security including both international ("It") and domestic sources, with risks from both natural ("Ne") and the built environment ("Pb"). Blue boxes denote direct effects of climate change, green indicates the UK food system and subsequent societal effects, brown shows international food system risks transmitted to the UK, and black indicates other factors such as wars which will have compound effects on international food system risks. Figure is taken from (Challinor et al. 2018)

The socio-economic impact of climate variability on crops comes from fluctuating food prices and famines. For this reason, even a moderate decline in yields can be devastating for household food security (Devereux 2009). For example, in 2000/2001 climate variability caused flooding in producer regions in Malawi significantly reduced maize production leading to a maize deficit of 273,000 tonnes. The resulting famine lead to the deaths of between 300 and 3000 people due to hunger and related diseases (estimates vary depending on organization) (Devereux 2002). Food price spikes have been attributed to the cause of this humanitarian disaster, seasonal maize prices have been estimated to have risen 354% from the lowest to highest month during the 2000- 2001 period (Ellis & Manda 2012). Furthermore, sensitivity of food prices to climate means that changes in climate volatility can have severe implications for poverty, for example, in Tanzania, poverty could increase by as many as 650,000 people due to an extreme annual yield decline (Ahmed et al. 2011). Similarly, in 2007, drought in South Africa and Lesotho lead to significant crop failures and severe food insecurity in Lesotho (Verschuur et al. 2021). Even in western developed countries such as the United Kingdom and France, food price shocks caused from heat wave and drought damage can have significant economic impacts (Wreford & Adger 2010, van der Velde et al. 2010).

With increased variability in climate, variability in crop yields will also increase (Devereux 2009, Porter & Semenov 2005). This will lead to an increase volatility in food prices (Devereux 2009). Increases in temperature due to climate change is expected to increase food prices, with more detrimental effects after 2050 (Schmidhuber & Tubiello 2007, Bandara & Cai 2014, Haile et al. 2017), however this may be mediated by increases in agricultural production up to 2050 (Baldos & Hertel 2014). Individual events have also been linked to climate change, and subsequent impacts on food security (Verschuur et al. 2021).

A key goal of crop modelling is to inform adaptation and food policy to enable decision makers to more readily and easily prepare for food shocks caused by climate variability,

extremes, and climate change. This can be achieved through in-season forecasting of drought, heat wave or other climate events likely to have impact on crop yields (Basso et al. 2013), the modelling of food systems using integrated assessment models (Ewert et al. 2015), to inform crop breeding and management decisions (Deva et al. 2020), and to build scenario based stakeholder driven modelling frameworks (Jennings et al. 2022).

1.4 Crop modelling

Process based (mechanistic) crop models are used to determine relationships between variations in climate and crop yields, and are used at a variety of spatial scales. Crop models are typically used to project impacts of future climate change (Jägermeyr et al. 2021, Asseng et al. 2015) and inform adaptation (Challinor et al. 2018, Minoli et al. 2019). Crop models are dynamic system models, in that internal parameters are allowed to vary and interact throughout time. Wallach et al. (2006) describe a general form of a dynamic crop model as

$$U_1(t + \Delta t) = U_1(t) + g_1[U(t), X(t); \theta] \tag{1}$$

$$\vdots \tag{2}$$

$$U_s(t + \Delta t) = U_s(t) + g_s[U(t), X(t); \theta] \tag{3}$$

Where t is time and Δt is some time increment (often days) and $U(t) = [U_1](t), ..., U_s(t)]^T$ is the vector of state variables at time t. State variables are those which are varied by the model over time such as biomass and leaf area index. $X(t)$ is the vector of explanatory variables. Explanatory variables are external influences to the model such as initial soil conditions, and daily temperature and rainfall observations. θ is the vector of parameters, which are static throughout time such as the thermal time required to reach a particular

growth stage, and g is the function which describes the relationship between each of the variables and parameters.

Crop model structures can be categorized into 3 separate archetypes. Field scale models such as APSIM (Keating et al. 2003), ecosystem models such as JULES crop (Osborne et al. 2015) and semi-empirical models such as GLAM (Challinor et al. 2004). Often crop models are used together in ensembles for the purposes of model comparison and uncertainty quantification, the ensemble mean of many crop models are typically more accurate than any one particular model alone (Martre et al. 2015, Asseng et al. 2013). Semi-empirical models such as GLAM have less parameters than models such as APSIM and use empirical relationships in some circumstances rather than process knowledge. For instance GLAM does not account for fertilizer effects and instead uses a yield gap parameter which accounts for many non-weather related effects on crops such as fertilizer application as well as pests and disease (Challinor et al. 2004).

Crop models are used to determine climate - crop yield relationships and quantify both current and future impacts of climate change on crops. Findings from crop models can be used to inform adaptation to climate change, particularly breeding adaptation in the form of new varieties (Challinor et al. 2016a), as well as responses to management adaptation (e.g. Minoli et al. (2019), Kelly et al. (2023)) and policy decisions (e.g. (Jennings et al. 2022, Foster & Brozović 2018)). To understand potential future effects of climate change, crop models will be coupled with climate model outputs (Jägermeyr et al. 2021, Osborne et al. 2007, Rosenzweig et al. 2014). Crop models are also used as part of integrated assessment models and frameworks (Ewert et al. 2015, Jennings et al. 2022). Integrated assessment has been defined as a process of combining knowledge from a range of scientific disciplines to better understand complex systems, to this end, crop growth models are combined with other modelling software or frameworks such as market models, economic models and farm response models (van Ittersum et al. 2008).

The data requirements of specific crop models vary, however generally crop models require rainfall, temperature, and solar radiation data, with some models also requiring wind speed depending on the method used to calculate potential evapotranspiration. Models may also require management data such as fertilizer input. The data requirements of crop models are described in detail in the recent work by Pasley et al. (2023). Crucially, data requirements are dependent on the complexity of the model. There is always a trade off to model complexity, an appropriate balance between model scope and sensitivity gives rise to appropriate model complexity. Increasing the model complexity can improve the versatility and scope of the model, and allow the model to account for a wider degree of variables and factors and so represent reality more completely. However, this also decreases the sensitivity of the model to individual inputs as well as increasing the model's overall uncertainty, which is the cumulative uncertainty of component processes of the model. Hence, the conceptual relationship between complexity and uncertainty arises from the optimum of the trade off between model bias (because a process is not adequately represented) and variance (due to a larger number of uncertain parameters) (Pasley et al. 2023, Challinor et al. 2018).

Often, data is difficult to acquire at the regional scale (Müller et al. 2017, Pasley et al. 2023). In many cases, simulated inputs are necessary because purely observed data is not available. This can be climate model reanalysis used for gridded temperature, or solar radiation data, or remote sensing to downscale crop yield statistical data (Iizumi & Sakai 2020), or simulated soils data (Romero et al. 2012). Using such data also brings a degree of uncertainty not present at the field scale (Challinor et al. 2018).

1.5 Machine Learning definition and uses

A very useful definition of machine learning was provided by Mitchell (1997) which states that "a computer program is said to learn from experience E with respect to some class

of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". This definition is highly informative as it emphasises that machine learning is an iterative process which will depend on experience, meaning experience through repeated exposure to a training dataset, but also experience with a broad range of data. Machine learning algorithms require data to be split into three independent groups, the training set which the machine learning model is repeatedly exposed to, the validation set which provides a simultaneous check against the results of the model using "unseen" data to prevent over fitting the model to the training set, and the test set, used to evaluate the model predictions using observed data. In this respect, training a machine learning model is a process of optimization. Optimization can be defined as the reduction of some error term E(X) between an initial set of values y_n and the ground truth or "target" values t_n. This can be summarised as

$$E(X) = \sum_{\vec{x_n} \in X} |y_n - t_n| \tag{4}$$

Meaning that the error in the data set X is the sum of the predictions y_n subtracted from the target (true) values t_n for each data point x_n in X. Machine learning architectures described in the following sections can be divided into groups based on the tasks they are used for. Supervised learning methods require that each training datum be labelled with an associated "target" value. For example in many cases in this thesis the target value will be the crop yield to be predicted. Supervised learning methods can be either regression tasks where the target data is continuous or classification tasks where the data is discontinuous and therefore is split into classes with distinct features. Unsupervised learning requires that the training data does not have associated labels. Unsupervised learning usually takes the form of clustering methods. Clustering is similar to classification in that the data is split into distinct groups (or clusters) however this sorting is determined purely by the

similarities between the variables used as input data rather than the relationship between the input variables and a target variable. Within the machine learning literature, input data is usually referred to as features. There are many machine learning architectures, a range of which will be described in the following sections. Commonly used methods for supervised learning tasks (regression and classification) include neural networks, tree based methods (decision trees, random forest and gradient boosting methods), support vector machines and distance based methods such as the K-nearest neighbours algorithm. Clustering methods often use distance based metrics and include the K-means clustering method. Some machine learning methods can also be described as representation learning which is the transformation of inputs into some intermediate form before being mapped to an output prediction. Deep learning is a form of representation learning which allows models with multiple processing layers to learn representations of data with multiple levels of abstraction (LeCun et al. 2015).

Machine learning has received much interest due to the ability for generalization and automated pattern recognition and so has seen increasing use in both a multitude of scientific fields as well as industrial sectors. With increasing quantities of data due in part to the use of remote sensing, there are new and useful prospects for machine learning to tackle the highly noisy and potentially high dimensional problems proposed by environmental datasets. In particular the generalization capabilities of machine learning may be well suited to capturing the effects of the dynamic climate system on crops which deterministic models struggle to replicate for a range of conditions.

Although machine learning is a powerful tool for prediction, machine learning models, particularly neural networks are often described as black box models (Castelvecchi 2016, Rudin 2019). This term refers to the concept that machine learning models often do not explain their predictions in a way that humans understand (Rudin 2019). However, for gaining scientific knowledge and understanding, it is important to understand model

behaviour and that predictions are explainable. For this reason, much of this thesis is focused on gaining knowledge through the interpretation of the results of machine learning methods.

1.6 Literature review

1.6.1 Statistical crop models and their comparison with ML

Before the recent growth in the use of machine learning methods to predict crop yield, statistical models were also used, sometimes to great effect, to act as yield prediction tools and crop model emulators. An important paper in this area is the Lobell & Burke (2010) study which compared the yield responses of statistical models to the Ceres wheat yield model when changes in temperature and rainfall are applied to the models. Among the conclusions of the study was the assertion that statistical models become more appropriate as the spatial scale at which output projections are required becomes broader, furthermore, statistical models were able to capture key aspects of the change in yield in response to changes in rainfall and temperature. Importantly, the most appropriate structure of statistical model depended on the yield response to be captured. For instance, a model purely trained using time series data was better at predicting precipitation responses whereas panel or cross section methods which also incorporated spatial trends were found to be more reliable for temperature responses. Statistical methods have also been used to estimate crop responses to climate change. Schlenker & Roberts (2009) found that crop yield responses increase with temperature up to a threshold temperature before resulting in harm to the crops. The statistical method used was a nonlinear regression model which depended on accumulation of heat and log transformed yield observations for specific counties.

Statistical models have continued to be used for the prediction of crop yields and agricultural productivity at the global scale (Ortiz-Bobea et al. 2021, Lobell et al. 2020, Proctor

et al. 2022, Proctor 2021). Statistical models are particularly useful because they can be used to take a more in depth look at specific effects on crop yields such as drought sensitivity (Lobell et al. 2020), or water supply (Proctor et al. 2022) or atmospheric opacity (Proctor 2021). Statistical models also have lower data requirements than both process based crop models and machine learning making them particularly useful for country level analysis.

Although statistical crop models have proven to be useful tools for crop yield prediction, this thesis instead focuses on the use of machine learning methods for crop yield prediction. This is because machine learning methods offer some significant advantages over traditional statistical models. Most importantly, machine learning methods such as random forest do not assume any particular shape of response function. This allows complex nonlinear relationships and interactions to be more easily handled (Leng & Hall 2020). Likely for this reason, machine learning methods can outperform traditional statistical approaches (Leng & Hall 2020).

1.6.2 Machine learning approaches to crop yield estimation

There are many recent studies which have used supervised learning methods to predict crop yield using a number of variables and approaches (Jiménez et al. 2009, Van Klompenburg et al. 2020*a*, Khaki & Wang 2019, Shahhosseini et al. 2019, Schwalbert et al. 2020). A survey by van Klompenburg et al. (2020*b*) has discussed both the most common machine learning frameworks used for crop yield prediction as well as the most used variables (or input features) in published studies. Most commonly used input features were temperature, rainfall and soil type. The most commonly used machine learning frameworks were neural networks, linear regression methods, random forest, support vector machines and gradient boosting trees. For model evaluation, the most common metrics found were RMSE (root mean square error) R^2 score, and mean absolute error. The survey included

studies from different spatial scales and so allowed for methods used which take advantage of high quality and quantity of data such as deep learning methods like convolutional neural networks (CNNs). Although linear regression was the second most common method found in the survey, it is the method often used to benchmark other machine learning methods, particularly to discover whether the model is required to account for nonlinear effects. As such, its place as a commonly used method does not necessarily mean it is the best method to use for crop yield prediction. Furthermore, the conclusions of the study argue that although the most common machine learning frameworks are found, this does not mean that they are the best methods to use, and in fact there is no consensus on this question. Similarly, it is also unclear how many features and which combination of features are best and which become redundant due to auto-correlation.

Studies which specifically focus on the regional scale are somewhat fewer than the overall number of studies. Many such articles have shown the great potential of machine learning even at the scale in which data may sometimes be of poor quality and relatively few in number (Leng & Hall 2020, Hoffman et al. 2018, Müller et al. 2017). A highlight of such research includes the study by Leng & Hall (2020) who found performance of a random forest ML model outperformed an AgMIP (see Rosenzweig et al. (2013)) ensemble of crop models and statistical model. Crucially however, the random forest model underpredicted the magnitude of yield variability in many locations. This is significant as it is yield variability which is most important for food security and the understanding of crop yield weather relationships. Furthermore, Shahhosseini et al. (2019) have compared several machine learning methods for estimation of crop yield and Nitrate losses. As part of this study they ask the question of how data quantity affects the model prediction skill. However the study reduces the number of data points from 2.5 to 0.5 million using simulated data. This is still a large number of data points, and so may not fully answer the question of how data quantity affects model performance.

Interpret-able machine learning methods can be used to help understand the factors and climate variables which most influence model performance. If model performance is able to explain a large degree of the variability in yield, then there is greater confidence in the interpretation of such methods. Lischeid et al. (2022) have used machine learning to attempt to better understand the relationships between driver and response variables. As part of this study it was found that air temperature and rainfall were prevalent meteorological predictors and soil moisture characteristics were comparatively less beneficial for model performance. However, due to inter-correlations between features, similar model responses were obtained using alternative predictors. For this reason, the authors of this study argued that expert knowledge is still and will always be indispensable for model interpretation regardless of the powerful predictive performance of machine learning methods. Further studies have also found some insights into modelled drivers of yield variability, for instance Bowden et al. (2023) found that variations in solar radiation was the most important climatic variable driving yield anomalies due to monsoon variability and Zhu et al. (2021b) have found that climate shocks to yield variability in Europe from 1980-2018 were mostly attributed to water limitations. Many feature importance methods and techniques such as partial dependency analysis and permutation importance analysis make the assumption that the features analysed are not correlated with other features used as inputs for the model (Molnar 2022), this can be a barrier to interpretability for climatological data in which features are often largely correlated.

Studies which use machine learning algorithms can be categorized based on the spatial scale of the input data and the type of input data used. At the smallest spatial scale, some studies use machine learning to determine crop yields in greenhouses or a single field (e.g. Pantazi et al. (2016)). Whilst others look at regions such as US counties or the Australian "Wheat belt" (Feng et al. 2019, Shahhosseini et al. 2019, Leng & Hall 2020). Type of input data can also be a key point of difference between studies. Studies which use remotely

sensed imagery or NDVI data can be very different to those which use census statistic crop yield data. Of course, different input data can lead to very different outcomes and model performance.

Some efforts have also been made to integrate machine learning models with process based models. Feng et al. (2019) adopted a hybrid approach using both the APSIM crop model (Keating et al. 2003) and random forest machine learning method which focused on using the final biomass output from APSIM, dates of the growing stages along with growth stage specific extreme event indicators. The random forest model improved the coefficient of determination (R^2) and root mean square error (RMSE) error metrics over both a linear regression model and base APSIM model. Drought was found to be the most common extreme event indicator throughout the historic period however increasing heat events were found to be the major stress affecting future yield losses. Machine learning has also been used to provide down-scaled estimates from courser resolution crop model outputs (Folberth et al. 2019). Gradient boosting and random forest model predictions both compared closely to EPIC crop model predictions in the study.

This thesis builds upon the recent research within the scientific literature in several ways. Firstly, by providing key insight into how machine learning methods may compare to an example case study model. Furthermore, the sensitivity of machine learning methods to climate input data uncertainty is assessed. Thirdly, it is determined under which conditions ML may improve upon crop model simulations, and what crop modellers can learn from the application of machine learning with process based crop models.

2 Methods

To address such questions both process based crop modelling and machine learning methods are used in this thesis. The process based model GLAM (General Large Area Model for annual crops) is used as a case study example of a process based crop model. The machine learning methods used in this thesis are also discussed in the following methods sections. It is important to address the issues which commonly arise when applying and developing such methods. Therefore, calibration and evaluation of crop models, validation of machine learning models, and difficulties when applying crop models at various spatial scales is discussed. Furthermore, maize (or corn) is used as an example crop which is used as a target (response) variable for prediction in each of the chapters. Because machine learning is crop agnostic, it is assumed that conclusions drawn from maize prediction could be applied to other crops. GLAM however requires different parameterizations for different crops. Although different parameterizations are required, model structure does not significantly and fundamentally change between crop parameterizations, and so, conclusions should be similar across crops.

2.1 The GLAM crop model

GLAM is a semi-empirical process based crop model first detailed in Challinor et al. (2004). A process based crop model is named as such because physically based processes are explicitly represented by the model, for example potential evapotranspiration determined using the Priestley-Taylor method (Priestley & Taylor 1972). Semi-empirical refers to processes in the model which are determined based on fitting the model to observed data rather than theoretically deriving parameters. A prominent example is the yield gap parameter, which is calibrated against observed crop yield data and is used to account for non-weather related factors which detrimentally affect crop yields such as sub-optimal management and pests or disease. A semi-empirical model follows the concept of appropriate complexity,

which is to model at a level of complexity appropriate to the degree of uncertainty and error associated with the parameterizations. Appropriate complexity is considered to help avoid over-fitting of model parameters (Challinor et al. 2018, 2009).

GLAM has been used in a range of locations for crops such as groundnut (Challinor et al. 2004), maize (Bergamaschi et al. 2013) and wheat (Asseng et al. 2015). GLAM calculates leaf growth rate on a daily time step, relating this to biomass then yield using a harvest index approach. Crop development is determined according to the thermal time elapsed between each growing stage. For maize, development stages are: from planting to the end of the juvenile phase, from emergence to anthesis, and from the start of the grain filling period to maturity. Thermal time is measured in growing degree days and the thermal time elapsed within a given growth stage is given by

$$t_{TT} = \int_{t_i}^{T} (T_{eff} - T_b) dt \tag{5}$$

Where t is time, T_b is the base temperature, which is the minimum temperature required for phenological development, i is the development stage number. The effective temperature T_{eff} is defined using the cardinal temperatures T_b, T_o and T_m which denote base, optimum and maximum temperatures respectively. Using these cardinal temperatures, T_{eff} is defined as

$$T_{eff} = \begin{cases} \bar{T} & T_b \leq \bar{T} \leq T_o \\ T_o - (T_o - T_b)\frac{\bar{T} - T_o}{T_m - T_o} & T_o < \bar{T} < T_m \\ T_b & \bar{T} \leq T_m, \bar{T} < T_b \end{cases} \tag{6}$$

Where $\bar{T}$ is taken either from measurements or averaged using T_{min} and T_{max}. Drought

stress is simulated using a soil water stress factor variable S_{cr}. The soil water stress factor acts to reduce the change in leaf growth rate per day. Leaf growth is represented as a leaf area index (LAI) which is defined as the area of leaf surface over the total area above ground. The change in LAI for each daily time step is therefore defined as

$$\frac{\delta L}{\delta t} = \begin{cases} (\frac{\delta L}{\delta t})_{max} \quad C_{YG} min(\frac{S}{S_{cr}}, 1) & i < 3 \\ 0 & i = 3 \end{cases} \tag{7}$$

Where C_{YG} is the yield gap parameter (otherwise referred to as YGP). When YGP is set to 1 (the maximum value) crop yields are unchanged and so therefore described as potential yields (the maximum yields which could be achieved given optimal management conditions and absence of pests). S is the soil water stress factor. There is also an alternate method to calibrate GLAM in which the yield gap parameter (YGP) is used to reduce the water holding capacity of the soil. In this method, the yield gap parameter has the following effect:

$$\theta_{dul} = \theta_{dul,max} \cdot C_{YG} \tag{8}$$

Where θ_{dul} is the drained upper limit (DUL) value of the soil after the yield gap parameter is used to reduce DUL from its maximum value. From this, the saturation limit of the soil is determined by the following empirical correlation:

$$\theta_{sat} = 0.254 + 0.787 \cdot \theta_{dul} \tag{9}$$

The absolute value of the lower limit of soil moisture (rll) is not a sensitive parameter and so this is kept at 0, model sensitivity from this parameter instead results from $\theta_{dul} - \theta_{rll}$.

This method of calibration was originally developed for the aim of increasing the standard deviation of the simulated yield time series which was underestimated by the GLAM model. This alternate method was developed in (Nicklin 2013), and is used in chapter 5 of this thesis. Of course, this method of calibration affects the soil water balance routine. It is also important to note that this is not the usual method of calibration and is only used in certain specific circumstances such as the circumstance described in chapter 4.

The soil water balance is determined each day from the soil moisture characteristics mentioned above and drainage terms. Change in moisture each day is defined by:

$$\frac{\delta\theta}{\delta t} = FD(\theta_s - \theta_{dul}) \tag{10}$$

Where FD is the drainage rate. D is determined by the drained upper limit using the following method:

$$D = C_{d1}\theta_{dul}^2 + C_{d2}\theta_{dul} + C_{d3} \tag{11}$$

Where C_{d1}, C_{d2} and C_{d3} are constants. F is defined by the following terms:

$$F = 1 - \frac{ln(Q_i + 1)}{ln(k_{sat} + 1)} \tag{12}$$

Where Q_i is the incoming water flux from the above soil layer, in the upper most soil layer Q is determined by precipitation minus runoff. k_{sat} is the saturated hydraulic conductivity, which is determined by:

$$k_{sat} = K_{ks}(\frac{\theta_{sat} - \theta_{dul}}{\theta_{dul}})^2 \tag{13}$$

Where K_{ks} is an empirical constant. Although calibration was first mentioned as an approach to calculate θ_{sat}, θ_{ll} in most cases these parameters are determined using a set of statistical correlations called pedo-transfer functions which were developed empirically to fit soil moisture characteristics to soil texture. The GLAM model uses the Saxton et al. (1986) method whereas some other crop models such as Aquacrop (Raes et al. 2009) and LPJML (Lutz et al. 2019) use the Saxton & Rawls (2006) method among others depending on region and soil type.

The soil water stress factor depends on the rate of transpiration relative to potential transpiration through the following relationship

$$S = \frac{T_T}{T_{Tpot}} \tag{14}$$

Where T_T is the transpiration rate and T_{Tpot} is the potential transpiration rate. The soil water stress factor affects leaf area growth only below the critical threshold S_{cr}. Biomass development is determined by using a limiting factor approach between transpiration efficiency (TE) and radiation use efficiency (RUE), Transpiration efficiency is crop specific (but reduced under high vapour pressure deficit conditions) and is used to convert crop transpiration to biomass. Radiation use efficiency is also crop specific, and is used to convert intercepted solar radiation to new biomass. On a daily time step, the biomass accumulated is that associated with the minimum of either TE or RUE (Osborne et al. 2013). Crop yield is determined from biomass using a harvest index approach. Terminal drought stress and lethal temperatures also work to limit growth according to extreme conditions. A lethal temperature (TRKILL parameter) causes early harvest (and so stops leaf area

and subsequent yield accumulation) when daily temperature exceeds a high temperature threshold on a particular day. Therefore, higher temperatures earlier in the season will cause an earlier stoppage of leaf area accumulation and so will have a greater effect on yield losses.

Sub-optimal conditions in the model are further represented using temperature thresholds which affect the rate of transpiration efficiency and radiation use efficiency. For hot conditions which are sub-optimal, but not hot enough to out right kill the plant, a reduction in transpiration efficiency is applied which acts to reduce the rate at which the plant will transpire, thus representing sub-optimal conditions which are slightly too hot. This is determined by:

$$ATE = TE \cdot (1 - \frac{TDAY - TETR1}{TETR2 - TETR1}) \tag{15}$$

where ATE is the actual transpiration efficiency, TE is transpiration efficiency, TETR1 is the sub-optimal high temperature threshold, and TETR2 is a lethal temperature threshold which reduces growth rate to 0 on the days in which the temperature exceeds the threshold.

Evapotranspiration is also a key model process. The calculation of potential evapotranspiration in GLAM is determined by the Priestley-Taylor equation (Priestley & Taylor 1972). The Priestley Taylor equation is defined as:

$$E^T_{pot} = E^e + T^e_T = \frac{\alpha}{\lambda} \frac{\Delta(R_N - G)}{\Delta + \gamma} \tag{16}$$

The Priestley Taylor equation defines the maximum potential evapotranspiration of a crop, where $E^e + T^e_T$ are the energy limited evaporation and transpiration rates respectively. α

is the Priestley Taylor coefficient parameterized as a function of VPD (Vapour pressure deficit), λ is the latent heat of vaporization of water, R_N is the net all wave radiation, G is the soil heat flux, γ is the ratio of specific heat of air at constant pressure to the latent heat of vaporization of water, and Δ is the change in saturation vapour pressure over the change in temperature ($\delta e_{sat}/\delta T$). Fundamentally, the rate of potential evapotranspiration is the subsequent driver of the model. Figure 2.1 below shows the general structure of the GLAM model as a flow diagram.

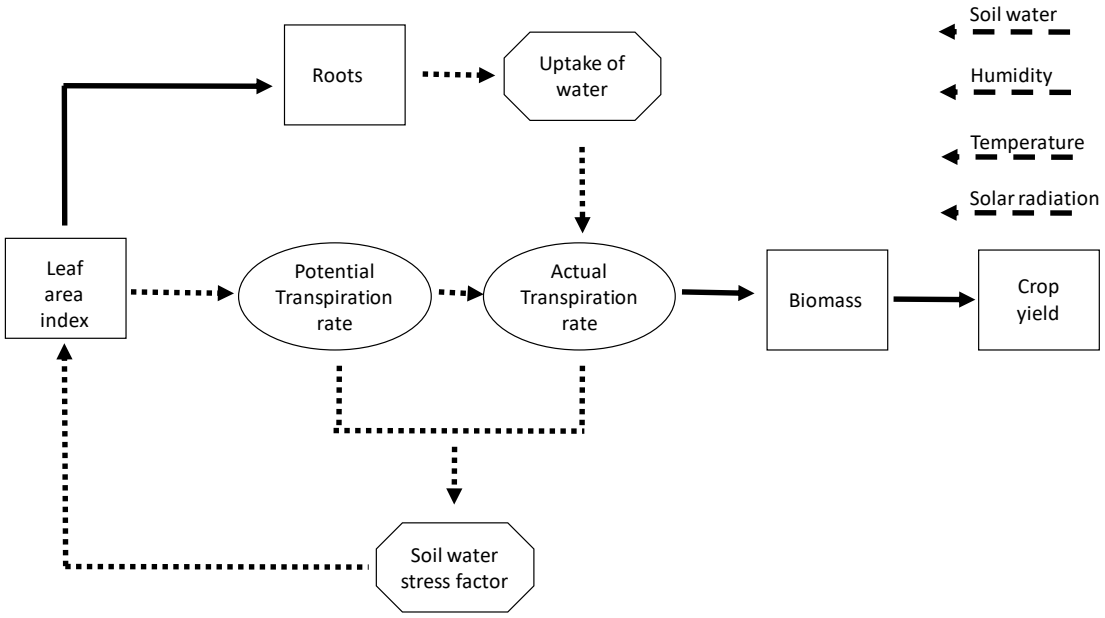

Figure 2.1: Flow diagram schematic of the GLAM crop model (taken from Droutsas et al. (2019)). GLAM state variables are represented by boxes, Rate variables by ovals and auxiliary variables represented by octagons. External variables are represented by dashed lines, mass flows by solid lines and information flows by dotted lines.

In a sequential manner, change in leaf area index feed into transpiration rates, and eventually translates to crop yield predictions via biomass using the harvest index coefficient to partition biomass into harvest-able grain yield. Leaf area index is in turn mediated by the soil water stress factor which is dependent on soil texture, hydraulic conductivity and

rainfall. Climatological inputs to the model are therefore soil water, humidity, temperature, and solar radiation. Soil texture information is also required to estimate hydraulic conductivity which determines soil moisture.

2.2 Standard GLAM calibration procedures

Calibration of GLAM parameters can vary in complexity. The standard calibration procedure used in most cases is to optimize the value of the YGP parameter which acts to reduce the simulated leaf area index of the crop to be simulated. This calibration procedure takes place at the highest spatial resolution which data allows (i.e. If crop yield data is available at the grid cell level, this is the level at which the YGP is optimized rather than some coarser aggregated resolution. The optimization process reduces the error (RMSE) between simulated yield and observed yield. Additionally, some region specific parameters which represent crop cultivars can also be calibrated. Such parameters include the thermal time required to reach each growth development stage, and transpiration efficiency value. These parameters can be changed to represent different crop varieties, some of which may be selectively bred to achieve certain characteristics such as greater yields or drought tolerance. However, these types of parameters are of course not available at a gridded resolution and so are often chosen to be broadly representative of the region of study. In chapter 5 a separate experiment is undertaken (separate from the main results) in which the YGP value is used to reduce the soil moisture to effective values rather than leaf area index. This is not a standard approach to calibrate the GLAM crop model and has not been thoroughly tested (it has only ever been used in this thesis and that of Nicklin (2013)). This method is only used for a specific experiment to determine whether improved representation of soil moisture improves the relationship between simulated yield and observed rainfall.

2.3 Choice of crop model

Many crop models are used for simulating the interacting effects of management and weather on crop yields. However, in this thesis, the GLAM crop model is chosen as a case study model. The decision was made to focus more on one particular model in order to dedicate more time to how differences in calibration and parameterization may affect the comparison between ML and the process based crop model. The GLAM crop model was chosen because it is of an appropriate level of complexity to compare to ML methods given the limited data available at the regional scale. Some other models such as APSIM or LPJML require additional data inputs such as fertilizer inputs simply not available when designing the experimental set up for this thesis (Keating et al. 2003, Lutz et al. 2019). GLAM was also explicitly designed to operate at the regional scale with outputs from climate models, which makes it ideal for regional scale prediction (Challinor et al. 2004). Although GLAM is the only process based crop model used in this thesis, the model is broadly representative of the crop modelling methodology. This is because all crop models follow the basic general formula described by Wallach et al. (2006), (equation 1) in which state variables, explanatory variables and parameters interact over some (usually daily) time step. This is in contrast to machine learning which will not follow the same formula and so will have different, behaviour, strengths, and weaknesses.

2.4 Challenges and issues when calibrating and evaluating crop models

There are several key problems for crop model calibration, which are particularly important at regional scales such as how to determine the level of complexity which is required to calibrate the crop model. This can mean how many parameters to optimize, which ones, and what are the suitable ranges to vary the parameters chosen. Common data chosen to calibrate against include phenological dates (date of flowering or maturity), crop yield,

above ground biomass, or soil moisture (Seidel et al. 2018). Seidel et al. (2018) have shown that the number of decisions that crop modellers make when calibrating crop models leads to a large number of outcomes during the calibration process. Furthermore, Wallach et al. (2021) have shown that even with the same model structure, different modelling groups make different decisions when calibrating. This is because calibration methodology is often based on expert opinion (Seidel et al. 2018) which introduces subjectivity into the model results.

Calibration can be of various degrees of detail, however it is important to consider the complexity of the problem to avoid over-fitting the model (Challinor et al. 2018). Improved detail in calibration can vastly improve results however at the cost of generalisability. Angulo et al. (2013) have shown how model performance can be vastly improved through using an empirical yield correction factor, and further improved by an extended calibration of site specific growth parameters tuned for each of a large number of climatic zones. With increasing detail, although crop model performance greatly increases, more site specific information is required. This presents a further challenge as model performance is dependent on site specific data often difficult to obtain at regional scales (Müller et al. 2017).

The GLAM crop model, like many, can be calibrated with varying degrees of detail. This is not just limited to the increments of which the yield gap parameter is varied against observed yields but also a number of parameters which can be calibrated and the method used to calibrate them. The most basic form of calibration is to vary the yield gap parameter systematically increasing the value (from 0 - 1) and record the RMSE (Root Mean Square Error) of predictions against observed crop yields. The value of YGP with the lowest RMSE is stored in a new file created which can then be used in subsequent simulations. The YGP value is determined for each location simulated and remains a constant value for future predictions. This of course means that GLAM cannot account

for increases in crop yields due to changes in management. For this reason crop yields are de-trended when evaluating against historical observed data. A simple method of de-trending is to subtract the results of a linear regression from each year simulated. GLAM calibration also can involve tuning of phenological parameters. Phenological parameters determine the amount of thermal time required to reach a particular growth stage. These parameters can be calibrated against observed yield data or in the case of (Jennings et al. 2022) these parameters can be tuned in order to achieve maximum yields and hence assume optimal crop varieties.

Crop model evaluation should be determined using historical observed data for a broad range of target variables over a range of environments, with particular focus on inter-annual variability (Challinor et al. 2018). The method, metric and importance of model evaluation depends on the purpose of the evaluation process. Crop models are often used for prediction and so evaluation is used to measure model performance. Usually crop yield is predicted however for more comprehensive model evaluation it is important to also evaluate against a wide range of observed variables to ensure the right answer is obtained, but also for the right reasons.

2.5 Challenges of spatial scale in crop modelling

Often crop models are developed at the field scale and then extrapolated to broader spatial scales, this will require additional aggregation and parameterization (Challinor et al. 2018). Increasing spatial scale will cause new factors to affect crop yields not present at smaller scales such as socio-economic factors. There are a range of methods outlined in Ewert et al. (2011) for scaling yield observations such as aggregation and de-aggregation of data, interpolation and extrapolation, as well as methods to manipulate the model itself such as using a scaling parameter, or summary model. A potential disadvantage of many scaling methods such as interpolation is the lack of consideration for spatial variability.

Sampling methods used to better represent spatial variability such as grouping data via identified homogeneous spatial zones can alleviate this problem. An example of this is the use of stratified sampling by van Bussel et al. (2016) to show how large reductions in sample size can still lead to predictions of similar model skill by organizing the sampling regime based on environmental zones with similar conditions. Variables used to group such environmental zones may include temperature, potential evapotranspiration, growing degree days and an aridity index.

2.6 Challenges and issues when comparing different crop models

Crop model ensembles show large variations in results (Asseng et al. 2015, Jägermeyr et al. 2021, Wallach et al. 2021, Palosuo et al. 2011, Bassu et al. 2014). Reasons for this are not just limited to the differences in crop model structure, but also in calibration methods (Wallach et al. 2021). For these reasons, multi-model ensemble error will depend greatly on the number and selection of which models are included in the ensemble (Martre et al. 2015). Furthermore, the variability of yield projections from crop models can be greater than variability between climate models (Jägermeyr et al. 2021).

Different crop models will use different methods to calculate processes such as yield formation, evapotranspiration, and growth. Models will also vary in the complexity of their representations. For example, some models such as LPJML (Schaphoff et al. 2018) simulate soil water dynamics for multiple soil layers; however GLAM assumes soil water characteristics such as wilting point, drained upper limit and saturation limit remain constant throughout the soil profile. Furthermore, some models will provide physical representations of processes not present in other models (or present in a less complex form) such as the effect of carbon dioxide on transpiration efficiency and effects of fertilizer application. The subject of appropriate complexity is an important issue for process based models (Challinor et al. 2018, Monteith 1996).

Differences in calibration can cause significant variation between results, even if crop model structure (the set of equations used in the model) remains the same. The cause of such variation is the large number of decisions which are required for the calibration process. Choices include defining the best parameter values, choice of parameters to estimate, and choice of optimization method. There is little consensus between modelling groups as to the appropriate answers to questions arising from model calibration (Wallach et al. 2021). The problems presented by model calibration ultimately arise due to the over parameterization and complexity of models. This, along with poor generalization between experiments and locations are key limitations which may be aided by the introduction of machine learning techniques.

2.7 Machine Learning architectures

The scope of this thesis is to evaluate and compare the results of existing machine learning methods against the GLAM process based crop model to understand how to improve crop models and discuss the uses of machine learning frameworks in the context of crop modelling for impacts of climate variability at the regional scale. Machine learning architectures range in complexity. Many studies will compare multiple model architectures to determine the appropriate level of complexity required for a specific problem. To address this, Delerce et al. (2016) has compared several of the most commonly used machine learning methods (linear models, neural networks, tree based methods, support vector machines) for regression and provided a summary of the advantages and disadvantages of each of the approaches. For example, highly important for scientific machine learning is the interpretability of the model chosen. In this respect, neural networks are often considered black box models as the individual parameters of the model cannot be easily related back to physical properties. This may be easier with some other architectures such as linear models and tree based methods however these methods are also less complex and may provide poor prediction performance by comparison if the prediction task is of substantial

complexity (as environmental datasets often are). Another key point of comparison is the ability to handle outlier data points. Support vector machines in particular are known to be sensitive to outlier values, which is not the case for tree based methods such as random forest (Delerce et al. 2016, Xu et al. 2006). This is particularly important for the prediction of climate impacts on crops due to the influence of heat and drought stress on yield. The following sections will provide a summary, uses, strengths and weaknesses of each of the ML methods.

2.7.1 Tree based methods

Tree based methods are all based on the decision tree style of architecture. Decision trees determine classification and regression using a set of choices which expand outwards from the base (root) of the tree where the first decision is made to each of the individual leaves of the tree which provide an answer to the query in the form of either a classification or regression. A common decision tree algorithm is the classification and regression tree (CART). Each decision to be made by the tree is ordered by the decision which will lead to the greatest minimization of the error for the data provided. For regression, this can be the sum of squares error or other error metric; classification tasks will use an error metric called the Gini impurity or entropy to compute information gain (Marsland 2011). Figure 2.2 shows how decision trees construct rules to build a model upon the given training data. X1 and X2 represent 2 features used as inputs to the model. Decision boundaries V and W are constructed in the feature space , this then allows a classification to be made based on the binary decision (the example is given of if $X1 > V$ and if $X2 > W$) made at each node of the tree.

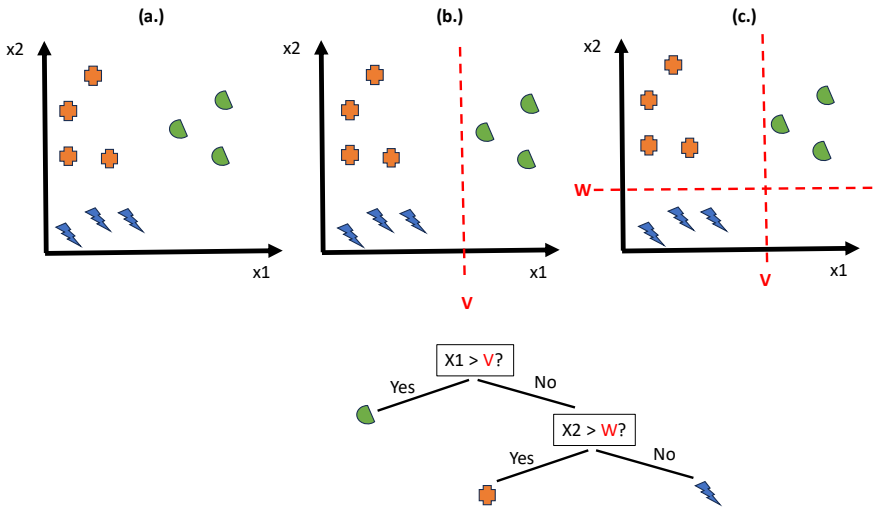

Figure 2.2: Schematic of rules constructed by a decision tree. Classification predictions are determined using splitting rules represented by either V or W. The three classes are represented by either crosses, semi-circles or bolts. The decision tree structure at the shows how the tree structure relates to the splits created in panels (b) and (c). Original figure was taken from (Marsland 2011) but re-drawn for this thesis.

Random forest (RFR) and gradient boosting machine (GBM) models are ensemble models made of a collection of a user specified number of decision trees. The main difference between these two methods is the way in which the results from each individual decision tree are aggregated. Once each individual tree is trained, an RFR approach will take either the majority vote from each tree (for classification) or a mean value (for regression). One of the key advantages of random forest is the ability to generalize through sub sampling. To increase variation in the ensemble, each tree is trained on a random subset of the data (a method called bootstrapping). Further variation is induced by only providing a random subset of the input variables at each decision tree node. By comparison, gradient boosting methods aggregate results of each individual learner consecutively using the error of the previous learner to help better train future decision trees which are added in turn. The method for gradient boosting first described by Freund et al. (1996) prescribes weights to

each data point given how poorly they were predicted by previous decision trees in the ensemble. The weights are then modified given each successive tree trained within the ensemble.

Figure 2.3 provides a schematic summary of the training process of the gradient boosting type models. Gradient boosting is a method in which decision tree 1 to k are successively trained. Each decision tree in the ensemble allocates training data samples to different leaf nodes.

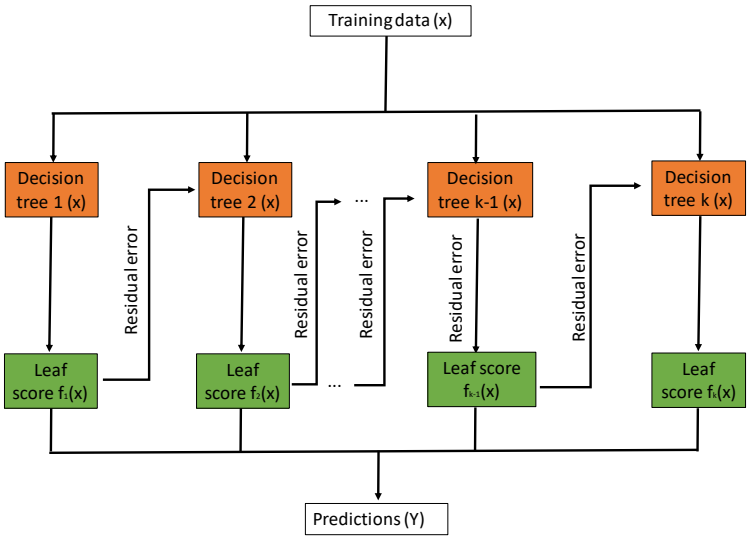

Figure 2.3: Model schematic of a gradient boosting machine model; originally figure is from Manoharan et al. (2022) but was redrawn for this thesis.

The prediction score $f(x)$ of each tree is added to create an overall prediction value for the entire dataset. The summation of information from each tree to achieve the overall prediction can be defined as:

$$y_i = \sum_{k=1}^{k} \left[\left(\alpha N + \frac{\beta |\mu|^2}{2} \right) f_k(x_i) \right] \tag{17}$$

where α and β are controlling parameters which require tuning, N is the number of leaves, and μ is the magnitude of the leaf weights.

Figure 2.4 illustrates the random forest method. In this instance, the dataset in question is split into random subsets from 1 to n. Decision trees are trained upon each data subset independently before an overall model is constructed using the predictions from each of the individual decision tree models. For regression, predictions are often averaged between each model.

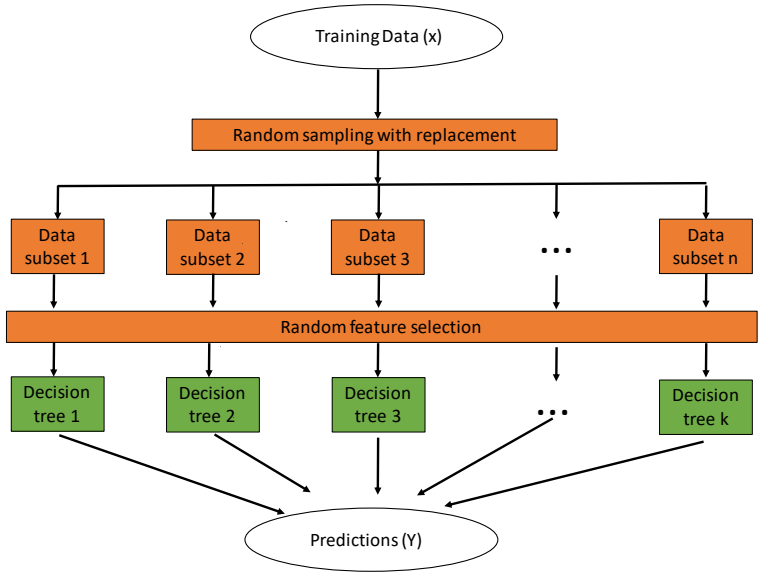

Figure 2.4: Model schematic of a Random forest model; originally figure is from Wang et al. (2018) but was redrawn for this thesis.

2.7.2 Support Vector Machines

Support vector machines (SVMs) are a popular machine learning framework. Support vector machines use an optimal hyperplane method to converge upon solutions (Cervantes et al. 2020). An optimal hyperplane is defined as the linear decision function with maximal margin between the vectors of the classes being separated. This requires only a fraction of the training data called support vectors. The support vectors can be seen as edge cases

in that they are the training examples which are most likely to be incorrectly classified. The key strength of SVM models is that the drawing of the optimal hyperplane decision boundary is only determined by the support vectors. Therefore, SVM models require far less data than comparable machine learning methods such as neural networks to produce predictions of reasonable skill (Liu et al. 2017). For many applications, a much discussed disadvantage of support vector machines is the high memory and therefore time requirement as datasets become larger (Qiu et al. 2016, Nalepa & Kawulok 2019). For this thesis however, and for the broader field of crop climate modelling, models continue to be constrained by the lack of observed crop yield data. It can therefore be argued that this much discussed disadvantage may be irrelevant for crop climate modelling.

Support vector machines were developed by Cortes & Vapnik (1995) as a method to create binary decision boundaries to separate 2 classes. The key idea of support vector machines is that the input vectors are non-linearly mapped into a higher dimension feature space. In this higher dimension feature space, a decision surface is constructed. In principle, 2 different groups of data which are inseparable in lower dimensions may become separable in higher dimensions. For inseparable problems, a soft margin was developed to draw decision boundaries with the minimal number of errors possible.

2.7.3 Distance based methods: nearest neighbours

The K-nearest neighbours algorithm is described as an instance based learning method. Instance based learning methods store each data point (or instance) before learning the relationship between the previously stored data points and any new data to generalize between them, rather than learning a general rule or function which applies to the entire data set and any new unseen data. Points are locally weighted, and the relations between them are characterized by distance in a Euclidean space. This therefore allows different approximations for each training instance. This is an advantage for complex functions which

can be also be described by a collection of less complex local approximations (Mitchell 1997).

The nearest neighbours algorithm assumes all instances correspond to points in n-dimensional space, R^n where n is the number of input features for each point. Each instance can be described by the feature vector

$$[a_1(x), a2(x), ..., a_n(x)] \tag{18}$$

where $a_r(x)$ denotes the value of the rth attribute of instance x. With this, the distance between two instances x_i and x_j can be defined as

$$d((x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2} \tag{19}$$

K-nearest neighbours can either be used for regression or classification. For continuous data new predictions are made using the mean of the k nearest training examples.

2.7.4 Artificial Neural networks

The neural network architecture was first introduced as the perceptron by Rosenblatt (1961). Rosenblatt defined the perceptron as a signal transmission network comprised of three signal generating units, sensory units, association units, and response units. The units of the perceptron are linked via a variable interaction matrix, which will depend on the past activity states of the network. The interaction matrix for the network is a matrix of coupling coefficients, with each unit being associated with with a weight or coefficient applied. Sensory units are essentially the units which receive input stimuli, (in the case of this thesis, weather input variables), association units receive the sequence of

previous signals and pass signals on to response units which transmit the signal outside the network. Each neuron in a neural network constructs a hyperplane. Across the network, the perceptron as a whole constructs a piece-wise linear separating surface.

The first incarnation of the neural network, the perceptron, did not have a method of adjusting the coefficients at each of the individual units of the network. Originally, only the coefficients of the output layer of the perceptron were therefore optimized to construct the linear decision function:

$$f(x) = \sum_i a_i z_i(x) \tag{20}$$

each coefficient (or weight) a_i is adjusted according to the output unit and some error measure to minimize error over the training data. A substantial improvement in the neural network architecture came when the back-propagation algorithm was discovered (LeCun et al. 1989, Rumelhart et al. 1986). The back-propagation algorithm is able to repeatedly adjust the coefficients of the neural network to therefore minimize a measure of error between actual and desired output vectors. In doing this, the network can develop an internal structure specific and appropriate to each task for which it is trained. In total, neural network error is defined as:

$$E = \frac{1}{2} \sum_c \sum_j (y_{j,c} - d_{j,c})^2 \tag{21}$$

where c is an index over all input/output pairs, j is an index over output units, y is the actual state of an output unit and d is the desired state of the unit (Rumelhart et al. 1986). To calculate the total error E (by using gradient descent or similar optimization algorithm) the partial derivative of E with respect to each weight in the network must

be calculated. This is done by calculating the sum of all partial derivatives from each input/output pair. In each case, the partial derivative of the error with respect to the weight are computed in 2 steps. These 2 steps are termed the forward and backward pass, with forward pass defined as when the node of network has its state determined by the input or from a previous layer. The input from a previous layer is determined by the mapping:

$$x_j = \sum_i y_i w_{ji} \tag{22}$$

where y_i is the vector of input values at each unit, and $w_j i$ is the vector of weights. This determines the weight/input combinations summed across each unit. The final step is to pass the output (x_j) through a non-linear function which has a bounded derivative such as:

$$y_j = \frac{1}{1 + e^{-x}} \tag{23}$$

In many cases the rectified linear function (relu) unit function is used which can be expressed mathematically as:

$$f(x) = max(0, x) \tag{24}$$

The rectified linear unit function simply makes the output values from the unit 0 if the gradient at this point is negative. A modified version called a leaky rectified activation has since been developed and has been found to work well for high resolution modelling tasks (Radford et al. 2015). The rectified linear unit function is often used within networks and

for regression tasks whereas a sigmoid function is often used for classification tasks. After the first stage (the forward pass) the backward pass propagates errors back from the last layer to the first to determine total error of the model. This begins by determining $\frac{\delta E}{\delta y}$ for each of the output units. For each unit, differentiating equation 21 gives:

$$\frac{\delta E}{\delta y_j} = y_j - d_j \tag{25}$$

The chain rule can then be applied to calculate the difference in error for each change in the output vector of the unit ($\delta E/\delta x_j$), this, along with substituting in equation 23 gives:

$$\delta E/\delta x_j = \delta E/\delta y_j \cdot y_j(1 - y_j) \tag{26}$$

This therefore shows how a change in the total input for the unit will affect the total error. However, since the input vector is the combination of the weights and inputs from the previous layer, to determine how the change in the weights will affect the error the following derivative is calculated:

$$\delta E/\delta w_{ji} = \delta E/\delta x_j \cdot \delta x_j/\delta w_{ji} = \delta E/\delta x_j \cdot y_i \tag{27}$$

Accounting for all connections (weights and inputs) from each unit, it can be determined that:

$$\delta E/\delta y_i = \sum_j \delta E/\delta x_j \cdot w_{ji} \tag{28}$$

For a full and detailed explanation of the back-propagation process see Rumelhart et al. (1986). Error at each unit of the neural network is reduced with respect to the target value using an optimization function, the simplest of which being gradient descent.

A key strength of neural networks is the flexibility which they afford to enable specialised networks for specific problems. Neural network architectures range widely depending on the applications of such methods. The simplest form of neural network is a feed forward fully connected neural network (FFNN). In this sense, the network is "feed forward" because the flow of information from input to output only flows in one direction (information from further layers are not added to previous layers). The network is also "fully connected" because all nodes in each layer provide input to each of the following nodes in the next layer.

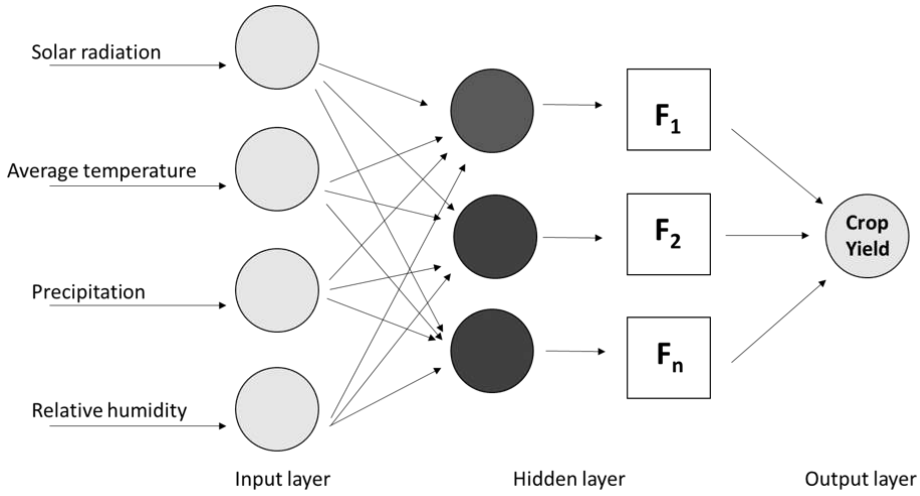

Figure 2.5: Neural networks are made of k number of hidden layers and n number of neuron nodes in each layer. The number of layers and nodes is determined using optimization. In a fully connected network, each node provides the input to every node in the following layer.

Figure 2.5 is a general schematic of a feed forward fully connected neural network. The

number of layers and neurons in each layer is decided by the user and can be optimized to improve model performance. With the addition of successive layers of neurons, the potential for universal function approximation becomes more promising. It has been shown that neural networks can approximate functions with only a single hidden layer (middle layer between input and output layer) if the number of units in the layer is sufficient (that is, that the network is wide rather than deep) (Hornik 1991). However, for complex tasks a model with many successive layers may be used. A neural network with many successive layers is often referred to as a deep learning model. With increasing size of a neural network with more layers, the number of parameters also rapidly increase. This necessitates the use of larger and larger datasets for the purpose of deep learning to constrain the many model parameters with the increasing number of neuron units.

Neural networks are flexible machine learning architectures which have more recently been further developed and adapted to excel at specific tasks. One such commonly used architecture is the convolutional neural network (CNN) which use sub-sampling to account for local variations in the dataset. CNN models are designed to process data which is in the form of multiple dimension arrays. Commonly this could be an image in the form of a 2 dimensional array or a video in the form of a 3 dimensional array. 1-dimensional convolutional neural networks could also be used for signals (LeCun et al. 2015). For example 1 dimensional CNN models have been used to detect earthquakes from waveform seismogram data (Perol et al. 2018). CNN models take advantage of concepts present within natural signals using ideas such as local connections, shared weights and pooling between nodes of the network (LeCun et al. 2015). This is particularly useful if data points which are close together within the feature space are similar or somewhat related. As such, CNN models are more easily able to identify local features such as small marks or discolouration on leaves which are an indicator of plant disease (Saleem et al. 2019). Shared weights for different neural network nodes are useful in the circumstance that local

patterns may occur in different parts of the same array. For example, when detecting plant disease from images of leaves(e.g. Saleem et al. (2019)), the same discolouration or spots which indicate disease may occur on different parts of the leaf.

Neural networks have also been adapted to process sequential inputs such as language, audio or time series such as rainfall (LeCun et al. 2015, Prasetya & Djamal 2019, Marzano et al. 2007). Once each sequential input has been processed, information from the previous instances is maintained in the form of a 'state vector' which contains information about previous instances in the sequence (LeCun et al. 2015). Although recurrent neural networks are very good at predicting the next element in a particular sequence (Sutskever et al. 2011), it was found that it was more difficult to learn longer range dependencies (Bengio et al. 1994). Therefore, the recurrent neural network was adapted by building into the network explicit memory. Resultant LSTM (Long short-term memory) networks are able to learn long term dependencies in sequential data (Hochreiter & Schmidhuber 1997).

Neural networks are used in chapters 3 and 4 of this this thesis. Much of the analysis of neural networks focuses on networks in which the nodes are fully connected and information flows from inputs to outputs (fully connected feed forward models). This choice is made often due to data limitations which is discussed in many sections throughout this thesis.

2.7.5 Multiple Linear Regression

In chapter 5 Multiple linear regression (MLR) is used as a linear comparison to the other machine learning methods. In MLR, the target value to be predicted is expected to be a linear combination of the input features. Mathematically, when there is at least 1 predictor variable x_i multiple linear regression can be defined as:

$$Y_i = x_1\beta_1 + x_2\beta_2 + ... + x_i\beta_i + e_i \tag{29}$$

Where Y_i is the target variable to be predicted, e_i is the ith error. Each β_i is a model parameter to be optimized (Olive & Olive 2017). In this thesis, least squares optimization is used to optimize the parameters of the MLR model.

2.7.6 Clustering methods

Clustering methods use a measure of similarity between points to sort data into unlabelled groups. One of the most common methods of clustering is the use of K-means clustering. K-means clustering is an iterative distance based method. The 'K' refers to the number of clusters which are specified by the user. K number of points are then chosen at random as the origin of each cluster. All instances are assigned to each cluster according to a Euclidean distance metric. The mean (centroid) of the points in each cluster is then determined. This new point becomes the new origin of the cluster, this process is repeated until the same points are assigned to each cluster following consecutive iterations (i.e cluster membership does not change) (Witten & Frank 2002).

K means is not the only clustering method. More complex methods may be used to tackle datasets with more complex relationships between inputs. Methods include hierarchical clustering and a clustering method using a neural network called a self organizing map.

2.7.7 Principal Component Analysis

This thesis uses principal component analysis (PCA) to understand similarities and structure of climatological and process model data. Principal component analysis is an analytical technique used for extracting important information from data and analyzing the

structure of the set of variables in question. This is achieved through simplification of the dataset, by projecting the data into a new coordinate system which keeps only the most important information. Axes of the new coordinate system are aligned in the direction of greatest variance, subsequent axes are orthogonal to the first and are added in descending order of variance (Abdi & Williams 2010, Witten & Frank 2002). The new set of orthogonal variables resulting from this transformation are a linear combination of the original variables defined as principal components (Abdi & Williams 2010). Figure 2.6 shows how 2 principal components may be orientated on a two dimensional data set for illustrative purposes. Principal component 1 shown in green is orientated along mode of greatest variation in the dataset, PC2 in orange has orthogonal orientation to PC1 and is orientated in the mode of the second largest degree of variation. Rainfall and solar radiation are used as illustrative theoretical examples as typically these two variables will be negatively correlated (i.e. rainfall corresponds with increased cloud cover which reduces incoming solar radiation).

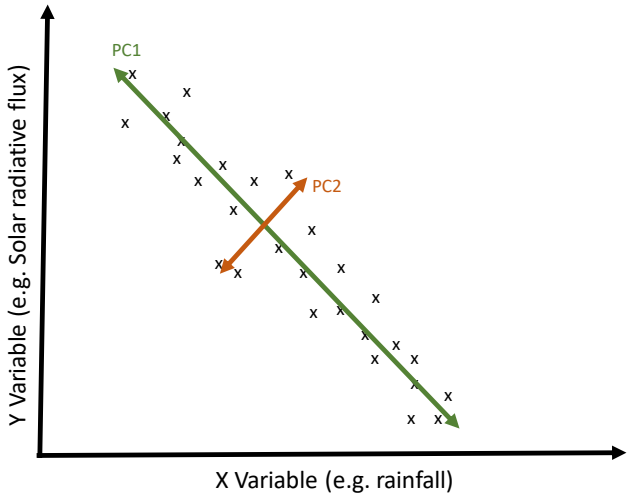

Figure 2.6: An example diagram showing a conceptual relationship between 2 correlated variables (Rainfall and incoming solar radiation) to illustrate how 2 Principal components may be orientated on a 2 dimensional dataset.

Principal component analysis requires centering of each variable before application so that the mean of each variable is 0. Furthermore, if the dataset contains variables which have different units, each element of the dataset is standardized by dividing each variable by its norm which can be achieved by the square root of the sum of all squared elements of the variable(Abdi & Williams 2010).

2.7.8 Strengths and weaknesses of ML architectures

Table 2.1 summarises the preceding sections by providing some of the key advantages and disadvantages of each of the ML methods used for regression in this thesis.

As Table 2.1 describes, there are many comparative strengths and weaknesses of the various methods used for crop yield prediction at the regional scale. These strengths and weaknesses will be leveraged to answer the research questions of this thesis in the subsequent chapters.

Table 2.1:
Comparative strengths and weaknesses of ML architectures

Method	Advantages	Disadvantages
K-nearest neighbours (KNN)	Very fast to train (Bhatia & Vandana 2010, Kumari & Soni 2017)	High computation cost when predicting large datasets (Taunk et al. 2019)
	Robust to noisy training data (Bhatia & Vandana 2010, Kumari & Soni 2017)	Easily fooled by irrelevant input features (Bhatia & Vandana 2010, Kumari & Soni 2017)
Tree methods (GBM, RFR)	Insensitive to scaling (Bhatia & Vandana 2010)	Need more computation capacity as parallel processing used (GBM) (Manoharan et al. 2022)
	Can handle continuous and discrete data equally well (Nazarenko et al. 2019)	Individual trees are prone to over-fitting (Prasad et al. 2006)
	Able to assess the significance of individual features within the model (Nazarenko et al. 2019)	Larger Random forest structures can be difficult to interpret (Prasad et al. 2006)
	Can use surrogate splits to overcome missing data (Delerce et al. 2016)	
Neural Networks	Universal approximation for nonlinear relationships (Delerce et al. 2016)	Requires high processing time for large networks (Kumari & Soni 2017)
Support vector machines (SVM)	Low risk of over-fitting and trapped at local minimum (Manoharan et al. 2022)	Black box model (poor interpretability) (Delerce et al. 2016)
Multiple linear regression (MLR)	(white box model) greater Interpretability (Delerce et al. 2016)	Non-linear relationships require transformation (Delerce et al. 2016)

3 A dual approach using a mechanistic crop model and machine learning enhances predictions across a range of conditions

3.1 Introduction

Many recent studies suggest that machine learning (ML) and process based models should be used together (Huntingford et al. 2019, Reichstein et al. 2019, Knusel et al. 2019, Kennel et al. 2016) however it is a relatively new idea to combine the approaches. In practice, ML and mechanistic crop models have only been compared in Leng & Hall (2020), although this work only compared the two approaches using one machine learning model (random forest). The study showed the potential of ML for vastly improved predictions over conventional statistical and mechanistic methods. Studies which have compared a variety of ML methods (Cai et al. 2019, Delerce et al. 2016, Newlands et al. 2019, Shahhosseini et al. 2019) to each other reveal the value in analysing multiple ML models. This is because different ML models have different strengths and weaknesses (Delerce et al. 2016, Manoharan et al. 2022, Bhavsar & Ganatra 2012). For example some methods such as support vector machines (SVMs) are more sensitive to outliers than other methods such as decision tree based methods (Delerce et al. 2016), however, SVM models are less prone to over-fitting than decision trees (Bhavsar & Ganatra 2012). Comparing multiple ML models allows for greater emphasis on which ML method is the most appropriate comparison to a mechanistic model.

Here, a single set of crop model simulations is used to benchmark the performance of several ML methods against historical observed yield data for sub-national geographic areas of France. The study period is from 1980 - 2007, the last five years of which are used as an evaluation period. This comparison is used to determine the pros and cons of each

method, including the conditions under which each method performs well. The analysis is then used to identify promising avenues for ML to be used in conjunction with mechanistic crop models in order to improve model performance. Comparisons are made against the General Large Area Model for annual crops (GLAM) (Challinor et al. 2004).

3.1.1 Research gaps addressed and aims

A comparison between process based modelling and machine learning is useful because it shows the strengths and weaknesses of both approaches. Apart from the comparison by Leng & Hall (2020) there is no comparison in the literature between process based and machine learning models for crop yield prediction as of the time of writing. Conversely, no studies compare multiple machine learning models with a process based crop model. Machine learning models range in complexity, therefore any comparison must also determine which machine learning model is of appropriate complexity for the task at hand. Also important is the quantity of data required for machine learning performance. This is important to provide a comparison under conditions which may limit ML model performance but also can be realistic for some regions if data coverage and collection is poor.

This chapter addresses the following research questions:

1. How many years of data are required for machine learning models to outperform process based simulations from the GLAM crop model?

2. Which machine learning frameworks are most skillful for crop yield estimation at the regional scale?

The first research question is important as it takes into account the data driven nature of machine learning as part of the comparison with the crop model by reducing the number of years of training data. Shahhosseini et al. (2019) have also looked into the effects of reducing the number of years of data on the performance of machine learning methods.

However, the study in question made use of synthetic data and so the number of data points ranged between 0.5 and 2.5 million. In contrast, in this study, as well as using the crop model as a benchmark, observed data is used for training rather than synthetic data, and therefore the amount of data used is considerably less. Research question 2 is what sets this study apart from that of Leng & Hall (2020) in that different machine learning methods are compared alongside the process based crop model. This is important as it accounts for the strengths and weaknesses of different machine learning methods, as described by many comparative studies such as Delerce et al. (2016), Manoharan et al. (2022), Bhavsar & Ganatra (2012).

In answering these research questions this chapter will provide an overview of the strengths and difficulties when applying machine learning to crop yield estimation. Indeed, as the application of crop models as part of wider inter-comparisons and inter-sectoral studies at regional scales continues to grow (Rosenzweig et al. 2014, Elliott et al. 2015, Schewe et al. 2019, Franke et al. 2020) and continue to be constrained by inadequate data for calibration and evaluation (Challinor et al. 2018), this approach is designed to develop a research agenda for the use of ML with crop models.

3.2 Methods

The methodological steps taken in this chapter are summarised in Figure 3.1. The experimental setup and mechanistic crop model were chosen because the data set and model have already been used in published work (Watson et al. 2015). The ML models were chosen to reflect off the shelf methods available and used in previous studies which compare the use of ML models for crop yield prediction (Cai et al. 2019, Delerce et al. 2016, Newlands et al. 2019, Shahhosseini et al. 2019, van Klompenburg et al. 2020*b*).

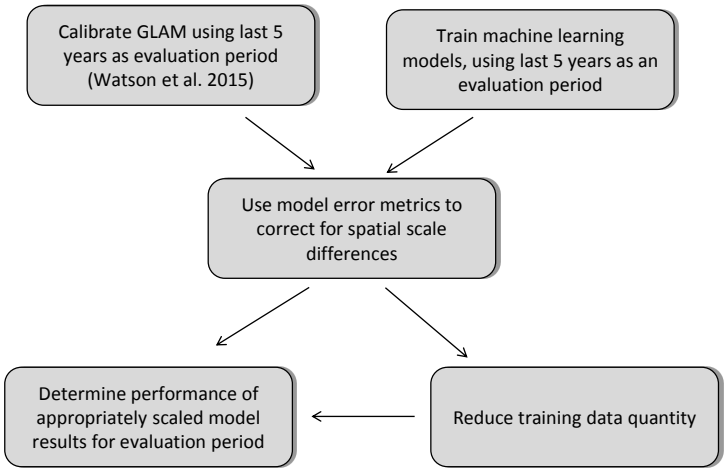

Figure 3.1: Flow chart depiction of the methodology for the study. GLAM simulations were carried out in Watson et al. (2015). Those simulations are compared to ML in this work. Model correlation coefficient was used to correct for spatial differences between weather and yield variables. models were run again, reducing the number of years of calibration data to determine effect on model performance.

Model performance for several ML methods and the GLAM crop model is defined using the Root Mean Square Error (RMSE) and correlation coefficient performance metrics. Some Figures (Figure 3.6) presented also present mean error which is simply the difference between observations and predictions across the 6 models. This additional performance metric was chosen to show the magnitude of under and over predictions in a simple manner.

3.2.1 Dataset: French maize across NUTS3 departments

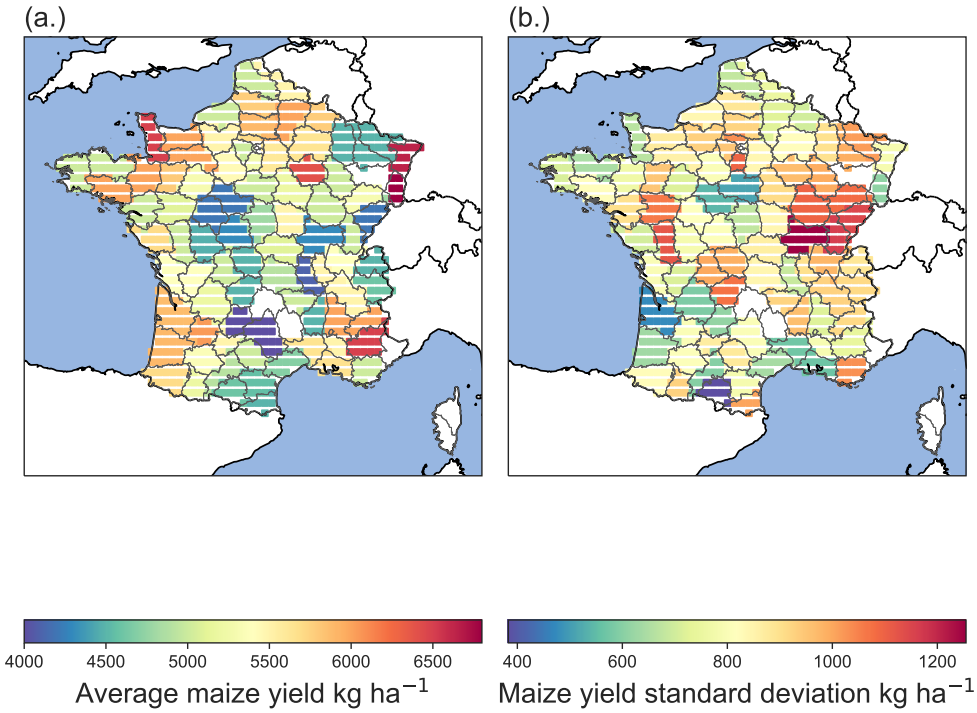

Figure 3.2: Mean (a) and standard deviation (b) of Maize yields for the across the study period (1980 - 2007).

All observed data are taken from (Watson et al. 2015), and described there. Originally, maize crop yield data was obtained from AGERESTE Statistique Agricole Anuelle for French NUTS3 departments (Watson et al. 2015). The yield data was obtained from census statistics aggregated to the French NUTS3 departments. Being census statistics aggregated from multiple sources, this can mean that weather and crop yield relationships

are more noisy than field scale obtained measurements with a nearby weather station. In total there are 96 NUTS3 departments, departments were removed with incomplete data (in which some years were missing) this lead to 86 complete departments to use for the study.

Figure 3.2 displays mean crop yield and standard deviation for each of the departments throughout the study period (1980 - 2007) The full time series is shown in Figure 3.3. Historical daily temperature and precipitation observations were obtained from the E-OBS gridded observational data set (version 7.0) described in Haylock et al. (2008). The E-OBS gridded data set is on a 0.25 x 0.25 degree latitude/longitude grid scale. This is a commonly used dataset for gridded climate impact models. Solar radiation data was taken from the ECMWF's ERA-Interim reanalysis and was re-gridded to the E-obs resolution using an area weighted average (Watson et al. 2015). Soil hydrological values consisted of saturated volume, lower limit volume, and drained upper limit, and were taken from the WISE soil database for crop simulation models version 1.1 (Romero et al. 2012).

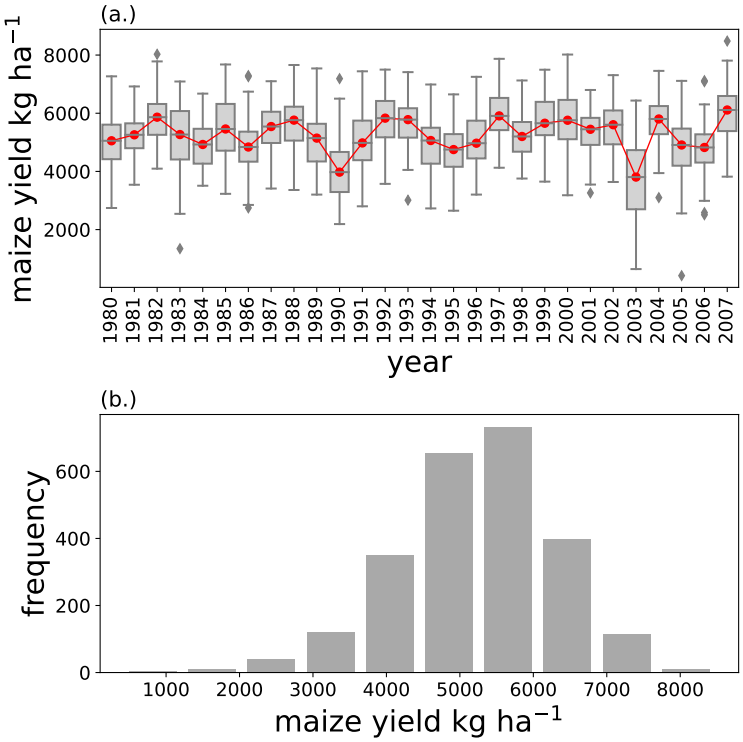

Figure 3.3: Maize crop yield distribution over time across France for the study period of 1980 - 2007. Yields were linearly de-trended using the methods described in Watson et al. (2015).

The maize yield time series was linearly de-trended in line with the start year of the data set (1980) (Figure 3.3) to be compatible for the required calibration format for the GLAM model. Of particular meteorological importance are outlier years 2003 and 2007 due to anomalous extreme heat in 2003 and exceptionally high yields in 2007. 2003 saw the lowest median maize yield (3805.5 kg ha^{-1}) and minimum of 642.8 kg ha^{-1}. This is likely due to high temperature anomalies recorded for that year throughout western Europe, leading to heat stress and low yields (Black et al. 2004). 2007 had the highest mean yields of the test period and also highest rainfall.

3.2.2 GLAM crop model calibration

This chapter follows the same GLAM calibration methodology as Watson et al. (2015), and as such, is subject to the same limitations. Principal among such limitations is the short 5 year evaluation period. GLAM was calibrated by varying the yield gap parameter (between 0-1) against observed yields to provide a correction factor for the years 1980-2002. 2003-2007 is kept as an independent test period. Although the highest resolution which crop yield data is available is for NUTS3 departments (See Figure 3.2) GLAM requires data to be presented at the grid cell scale. Therefore the data is re-gridded meaning that grid cells which fall within a departmental boundary are given the crop yield value recorded for that department. GLAM model parameter values in this chapter are taken from Watson et al. (2015) and are referred to as W2015.

3.2.3 Machine Learning methods

The machine learning methods used in this chapter were chosen to reflect a range of effective and commonly used machine learning methods. Methods were chosen to reflect a range of complexity. In this case, complexity refers to the number of parameters. Typically neural networks will have the greatest number of parameters, a nearest neighbours approach will have less parameters and so is a less complex model. All machine learning methods used in this chapter are described in section 2.7. Machine learning methods used in this chapter are: K-nearest neighbours regression (KNN) (see section 2.7.3, random forest (RFR) and gradient boosting regression (GBM) (see section 2.7.1, support vector regression (SVM) (see section 2.7.2, and neural networks (FFNN denotes fully connected, feed forward neural network) (see 2.7.4).

3.2.4 Developing a fair comparison

A fair comparison between modelling approaches aims to use consistent input data which presents several key issues, firstly calibration. The GLAM-maize simulations of W2015 were not optimised for the study region; rather an automatic calibration routine was run in order to calibrate one parameter (YGP). However, the GLAM-maize internal parameters do contain information from previous calibrations of the model in other countries. Thus GLAM-maize has the benefit of having an *a-priori* simulacrum of maize, and the ML methods do not. This disparity is redressed somewhat by not specifically tuning the GLAM representation of maize to French conditions. Hence, the use of the W2015 simulations maintains generalisability in the results - it places a relatively low bar for comparison to ML and ensures that the results are not contingent on one particularly detailed crop modelling study. Furthermore, many large and regional scale studies often omit detailed calibrations. (Rosenzweig et al. 2014, Elliott et al. 2015, Franke et al. 2020, Jägermeyr et al. 2021). Therefore, although GLAM results may benefit from more detailed regional calibration, the calibration protocol used for this work was performed to a common degree of detail.

The second issue related to input environmental data. The GLAM model makes use of daily weather input information, which is not easily interpreted in most ML methods which require 2-dimensional data in tabular format. GLAM also uses both weather and spatially varying soil moisture and empirical correction factor (YGP) parameters, meaning it incorporates 3-dimensional spatio-temporal information and crop specific information not explicitly available using the weather - crop yield relationship alone. This is important as at high resolutions spatial dependencies are more greatly influenced by local management practices, as discussed in the W2015 study. Furthermore, the magnitude of daily weather variability may affect results in a manner not explicitly represented by monthly aggregated inputs (Van Bussel et al. 2011).

Timing and statistics of weather within a growing season are also likely important for impacts upon yield (Haylock et al. 2008, Challinor et al. 2005). The input features were selected in a way to try to reflect these factors. Input data variables are described in section 3.2.7. An extra model validation step is used for ML models which is not used for GLAM. This is a method of choosing hyper-parameters which is standard for ML methods. This is used to determine optimum hyper-parameter values for the dataset.

3.2.4.1 Coordinates as input features

To enable a fair comparison between GLAM and the machine learning methods, latitude and longitude were included as input features used for training the machine learning models (along with other weather input features). Although including coordinates as features can reduce the generalization of the model for locations outside of the specific area of training, since the training and testing split is only made across time (i.e. No grid cell locations in the testing dataset are excluded from training) this should not detrimentally affect model generalization. The decision was made to do this to enable an analogous input feature to the yield gap parameter in GLAM used to calibrate the crop model (hence to enable a fair comparison). Coordinates are a continuous variable in intervals of 0.25 used to represent grid cell location, similar to how the yield gap parameter is also a spatially varying continuous variable which will vary on intervals of 0.1 or more.

3.2.5 Methods of choosing representative grid cells

Each department contains a small number of grid cells (average 11) of size 0.25 by 0.25 arc degrees containing the weather data. Thus, one yield measurement corresponds to multiple weather grid cells. The following procedure was used to select a single representative grid cell for each NUTS3 department: each model was trained (or calibrated) on all grid cells in the data set. For each model (GLAM and each ML method), the grid cells were then sorted into their respective departments. Within each department, the grid cell with the highest

correlation between the simulated and observed yield was selected as the representative simulation. Figure 3.4 pictorially displays this process.

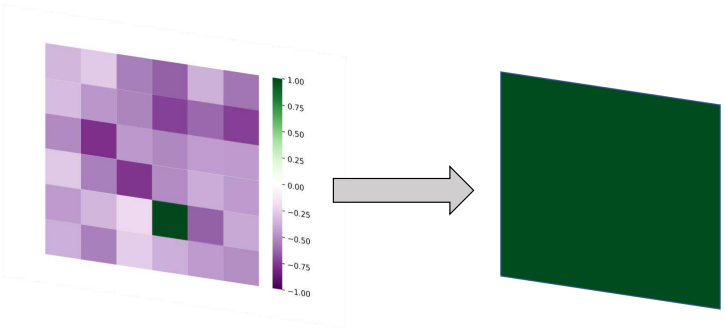

Figure 3.4: Representative grid cell scaling methodology using models. Both GLAM and ML models were calibrated / trained on all grid cells. The representative grid cell for each department was chosen according to which grid cell within the department produced the highest correlation coefficient for that department. In this example, the green grid cell (grid on left) has the highest correlation coefficient (-1 to 1) for the department and so therefore this grid cell is chosen as representative for the department (square on right).

3.2.6 Representative grid cells versus data aggregation

Although representative grid cells are used to manage differences in spatial scale, other methods are also available which have advantages and disadvantages. The main alternative method would be to aggregate up weather data instead of taking representative grid cells. This could be done by averaging temperature and rainfall across each department of which there is yield data, then training the model based on the spatially aggregated meteorology instead. An advantage to this approach is that it can account for the spread of variability across the department in the averaging. however, averaging across grid cells within the department will also result in aggregated meteorology inclusive of the errors associated with each grid cell location (Hansen & Jones 2000). Representative grid cells are a way of reducing aggregation error by ensuring that the locations chosen for the aggregated results are most representative of the real relationship between weather and crop yield (Challinor et al. 2016b). The key potential disadvantage of using representative grid cells is that

results may differ depending on which grid cell is chosen as representative. This could lead to 'cherry picked' results depending on which method is used to choose the representative locations. To address this issue, the model performance resulting from representative grid cells are compared to the grid cell chosen as least representative (that which results in the worst possible model performance) is compared in Figure 3.10. Representative grid cells have been promoted as an appropriate methodology previously in studies such as (Challinor et al. 2016b) and result in different kinds of errors than data aggregation.

3.2.7 Pre-processing and data transformations

As discussed in section 3.1.1 there are a wide number of decisions which are required when developing a machine learning pipeline. Although choice of algorithm is important, even once a model has been chosen, other aspects to the pre-processing pipeline present themselves. Training data transformation (sometimes called feature scaling) is required (for some models) if inputs are not of the same order of magnitude to avoid larger inputs biasing the model. However, there are numerous options for the method used to scale the data. Target yield distribution is also something which may be changed to alter distribution of both target values and predictions. This is common for statistical models (Lobell & Burke 2010).

Training and model validation also present further options. When training, different error metrics may penalize model error in different ways. For example, a root mean square error (RMSE) metric will penalize error at the extremes of the data to a greater degree than mean absolute error (MAE), this may be important for the influence of extreme events and outliers. For some models, such as neural networks, the duration of training is also a key factor. Training should stop when training set error has sufficiently decreased, however before error on the validation set begins to increase, indicative of over fitting the model to the training data. The number and choice of input variables will also greatly

change the model output. Testing variables for correlations and determining permutation importance of inputs is a method to show the relative weighting each of the variables is given by the model. This can provide insight into which variables to include, and how useful they are to the final prediction. If all variables are highly correlated, the option to use dimensionality reduction methods presents itself. Dimensionality reduction can improve model performance by reducing random noise and eliminating redundant input variables, as well as speed up training time. accounting for spatial scale differences is discussed in section 3.2.5.

The choice of variables to include is a key issue for machine learning simulation of crop yields (van Klompenburg et al. 2020*b*). The following variables were chosen as inputs to the machine learning models:

1. Max, min and mean temperature (°C)
2. precipitation (mm)
3. incoming long-wave radiation (Wm^{-2})
4. number of dry days (days)
5. number of days above 32 degree lethal temperature threshold
6. saturated moisture capacity of soil
7. field capacity of soil
8. Field wilting point
9. grid cell latitude
10. grid cell longitude

Although some inputs are self explanatory (e.g. precipitation, temperature, incoming longwave radiation), some rely on extra pre-processing steps. A high temperature threshold was chosen to represent the effects of heat stress. 32 °C was chosen as the temperature threshold as this was the value determined optimum for statistical model fitting by Hawkins et al. (2013). Field capacity, wilting point, and saturated moisture capacity of the soil are soil moisture characteristics. Although it could be argued that including the grid cell location as an input to the machine learning models may result in poor generalization between spatial regions, it is included as an input to provide a fair comparison against the GLAM model calibration routine (this is discussed in section 3.2.4.1).

As mentioned previously, feature scaling is required to ensure that ML models are not biased towards inputs of greater magnitude, quantile scaling is the method chosen to do this. Quantile scaling of the inputs was chosen to preserve observed distributions of rainfall and temperature. This is actually only necessary for machine learning methods which use distance based error metrics when training such as the support vector machine (Cortes & Vapnik 1995). The tree based methods do not require this step however this was carried out across all models purely for methodological consistency. Quantile scaling of inputs is carried out using the following formula:

$$\frac{x_i - u}{Q^{75th} - Q^{25th}} \tag{30}$$

where Q^{75} and Q^{25} are the 25^{th} and 75^{th} quantiles of each feature vector x and u is the median value of the feature vector.

As a final step, sampling methods may also greatly affect the results of a model. Sampling methods have long been discussed as a way to improve the prediction performance against unbalanced data (Chawla 2009). However unbalanced data problems are less well studied

for regression than they are classification. Predicting changes in crop yield is inherently a problem of unbalanced data for regression, as the most important values of interest are usually the smaller number of low yields where the crop has failed. Sampling techniques for continuous data may focus on under sampling to produce a more balanced dataset or oversampling the values within a threshold of interest (Torgo et al. 2015). When dealing with extreme events, such as the effects of the 2003 heat wave on maize yields, the effects of sampling techniques may become more important. As such, a method of under sampling is devised to determine if a balanced data distribution can improve predictions of extreme events. Yields are sorted according to a cumulative distribution shown in Figure 3.5, an equal number of samples are then taken from each bin of the cumulative distribution to create a uniform distribution to be used as training data samples. This step is undertaken for one select model to compare against the baseline model results to determine if a specialised sampling strategy for extreme events may have improved the results.

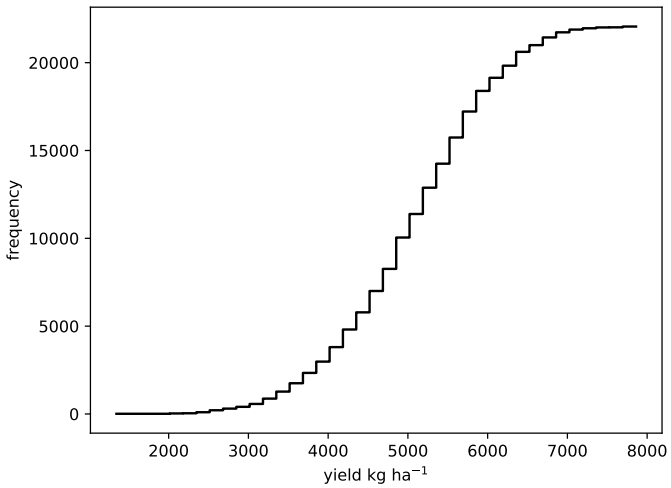

Figure 3.5: Cumulative distribution of crop yields from 1980 - 2007.

The decision to under-sample the majority of the data, rather than create new synthetic examples of crop failures was chosen as a methodological decision to not include the use

of synthetic data. In making this decision, the performance of machine learning methods can be evaluated more simply using purely observed data. This test resulted in poor performance and so the decision was made to focus model simulations without using this method.

3.2.8 Choice of temporal scale

Although some preliminary work aimed at understanding which temporal scale may be most appropriate to achieve best results, the decision was made to aggregate weather data from daily to seasonal aggregates of weather for this chapter (and subsequent chapters throughout this thesis). This is the temporal scale chosen for many when training machine learning or statistical models (Leng & Hall 2020, Shahhosseini et al. 2019, 2020, Feng et al. 2019, Lobell & Burke 2010). It was found that best performance for this dataset was achieved using seasonal aggregates of weather, although future work should certainly explore if it is possible to achieve better results under different temporal timescales.

3.2.9 Hyper-parameter optimization

Model hyper-parameters are parameters which affect overall model structure, which can be fine tuned to provide improved fit to model data. For example, the number of decision tree estimators used for the random forest model is hyper-parameter which may be optimized against a validation dataset. Hyper-parameters were chosen by optimizing values within the given ranges within each table found within this section. 2-fold cross validation was combined with some initial manual hyper-parameter tuning to select optimum hyper-parameter values. For each training set, the time series was split in half with each half becoming a single fold for cross validation. The data was split by year to ensure yield observations for the same department and year did not fall into different folds. This prevented data leakage (where information from training is passed to validation or testing) by removing spatial auto-correlation as a factor. Inputs were de-trended therefore it was

deemed unnecessary to make the years in each fold random. Model hyper-parameters showed significant noise and therefore optimum values fluctuated greatly once a local minimum value was achieved. Therefore, although exact parameter values are given for the results presented in this chapter, similar performance can be achieved using a similar set of hyper-parameter values. For each model, optimum hyper-parameter values were chosen for the full training data set and the reduced training data sets (15, 10 and 5 years). This decision was made to keep consistency with the GLAM set up which calibrates the YGP parameter for each reduced training data set.

Inputs were represented as a 3-dimensional array (of features, years, locations) which split according to year into a training and independent test data set. Training data was then flattened to create a 2-dimensional tabular format for training. Yield output is represented as a 1-dimensional array of 959 grid cells which is then sorted using a postprocessing script.

The post-processing script sorts the grid cell outputs from the models into departments then selects the grid cell within each department with the highest correlation coefficient (Pearson correlation) across the test period as the representative grid cell for each department (see section 3.2.5. In the process, the machine learning output is compared to the GLAM output. Grid cells in which GLAM failed to produce a yield value were removed for all models to ensure predictions are compared for the same locations.

When validating the support vector machine model, the only hyper-parameter which was adjusted was the choice of Kernel function. Kernel functions are dot products within a usually high dimensional feature space (Hofmann et al. 2008). The two Kernel functions chosen were linear and radial basis function (RBF). The linear kernel is the simplest Kernel function which follows the formula K(xi,xj) = xiT xj + C which is the dot product of the training vectors plus a constant. The SVM model was the fastest to run and had the least

number of hyper-parameters to tune. Table 3.1 presents the values of the kernel function used for each time the model was trained.

Table 3.1:
Support vector machine grid search optimization hyper-parameters chosen. Left to right values correspond to the full training dataset, then 15, 10 and 5 years of training data.

Hyperparamter	Value				Description
Kernel function	Linear	Linear	Linear	Poly	Kernel type: choice from linear, polynomial, radial basis function or sigmoid

Since the initial paper describing K-nearest neighbors techniques by Cover & Hart (1967) the method has expanded and subdivided into multiple types of nearest neighbour algorithm. The algorithm hyper-parameter of Table 3.2 describes the method to search the feature space employed by the nearest neighbours model. For instance, the ball tree KNN method (value of algorithm hyper-parameter) uses a binary tree, similar to random forest decision trees, however each node in the tree is associated with a hyper-sphere in n-dimensional Euclidean space.

Table 3.2:
Hyper-parameters tuned for the KNN model with optimum values chosen. left most value is for the full training data set. Following from this are values for 15, 10 and 5 years of training data.

Hyper-paramter	Value				Description
N neighbours	46	112	112	200	Number of neighbours
Weights	Distance	Distance	Distance	Uniform	Weight function used
Leaf size	90	60	100	40	Number of samples required
Algorithm	Ball tree	Kd tree	Brute	Ball tree	Algorithm used to compute nearest neighbours

The random forest and gradient boosting models use a number of decision tree estimators. Predictions are decided from the aggregate of the ensemble members. Number of estimators was a sensitive parameter up to 200 estimators where error saturated. Max features decides the criteria to use to determine the best method to split the dataset when building

the decision trees. If Max features = sqrt then max features is determined by the square root of the number of features. Max depth default value is for there to be no maximum depth to each tree. Providing a fixed value for max depth reduces the computation time of the model. Random forest hyper-parameter values are shown in Table 3.3, gradient boosting hyper-parameter values are shown in Table 3.4.

Table 3.3:
Hyper-parameters tuned for the KNN model with optimum values chosen. left most value is for the full training data set. Following from this are values for 15, 10 and 5 years of training data.

Hyper-paramter	Value				Description
N estimators	1400	1200	1600	600	Number of decision trees
max features	sqrt	sqrt	sqrt	sqrt	number of features for splitting
max depth	80	10	10	110	Max depth of each tree
min samples per leaf	2	2	4	1	minimum number of samples required for each leaf node
Bootstrap	True	False	True	True	Bootstrap samples may be used by the tree

Table 3.4:
Hyper-parameters tuned for the KNN model with optimum values chosen. left most value is for the full training data set. Following from this are values for 15, 10 and 5 years of training data.

Hyper-paramter	Value				Description
N estimators	800	1000	1200	1600	Number of decision trees
max features	sqrt	sqrt	auto	sqrt	number of features for splitting
max depth	10	20	10	70	Max depth of each tree
min samples per split	5	2	2	2	minimum number of samples required to split a tree node
min samples per leaf	2	2	4	1	minimum number of samples required at each node
Loss function	Lad	Quantile	Lad	Quantile	Function used to minimize error

There is no common universal solution for choosing the number of layers for a neural network, likewise, several other hyper-parameters must be tuned to determine the best model. Model hyper-parameters were systematically varied however many such as the

number of layers made little difference to model performance. Number of layers was kept the same for reduced training data as changing this value had little effect on the model. The number of neurons determines the number of nodes in the neural network to which a set of weights and transformed values are found at each subsequent layer of the network. Increasing the number of nodes leads to a wider network, whereas increasing the number of layers leads to a deeper network. Neural networks learn through optimization, this is achieved by using an optimizer algorithm to minimize the error of the weights at each layer according to a loss function between the weights and values which have been transformed by the previous layer. Optimization functions were chosen from the common methods available through the Keras Python API, (ADAM, RMSPROP, NADAM). Activation functions are nonlinear transformations specified for each layer of the network. The most common activation function is the rectified linear unit (Relu), This function was used for all hidden layers, and the final layer used a linear activation function. Init mode is the hyper-parameter used to specify the initial random weights for each layer of the network such as random values from a normal distribution. Batch size was set at 10 for all models. Although the above validation process was also tried with this model, best results were found through manual tuning of the hyper-parameters.

Table 3.5:
Hyper-parameters tuned for the KNN model with optimum values chosen. left most value is for the full training data set. Following from this are values for 15, 10 and 5 years of training data.

Hyper-paramter	Value				Description
Layers	6	6	6	6	Number of sequential layers
Neurons	500	500	500	500	number of neuron nodes
Optimizer	RMSprop	RMSprop	RMSprop	RMSprop	algorithm used to reduce error
initialization	he normal	he normal	he normal	he normal	Initial distribution of neuron nodes

3.3 Results

3.3.1 Model framework comparison: GLAM versus machine learning

Seasonal error for each model is compared in Figure 3.6. In 2003, GLAM achieved the lowest Mean error. On average GLAM under-predicted yields by 97.57 kg ha^{-1} with a median under-prediction of 417.85 kg ha^{-1}. GLAM produced the lowest inter-quartile range of predictions whilst also achieving a more accurate mean error for that year. All ML models over-predicted the mean yield in 2003. Differences between mean and median over-predictions were less for the ML models than the difference between the median and mean under-prediction by GLAM, indicating a greater influence of outliers on the GLAM results. There were no significant differences between model error in 2004 and 2005. The ML models were more accurate in 2006 and 2007. All models under-predicted yields on average in 2007 however the ML models under-predicted less than GLAM.

Modest improvements are made as the number of years of training data is increased (Figure 3.7, & 3.8). Correlation with observed yields increases for KNN, RFR, with increased training data. However, there is considerable variability in predictions. GLAM performance did see improvements despite the YGP parameter being the only calibrated parameter. Generally, the largest improvements in performance from 5 training years are found in the models which achieve the lowest RMSE overall (RFR, GBM & FFNN).

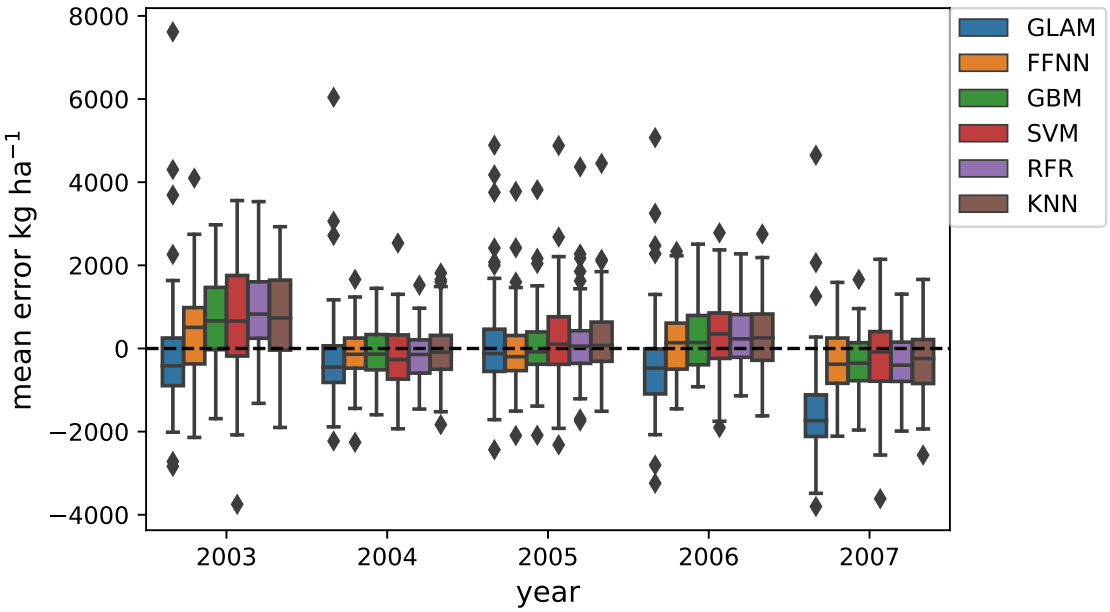

Figure 3.6: Plots (a) and (b) show the correlation coefficient when either the maximum correlating grid cell per department is chosen or the minimum. This leads to large variations in prediction performance as expected.

Changing the amount of training data results in a similar pattern pattern (Figure 3.9 with RMSE increasing as yield decreases. This effect is most pronounced in the KNN model panel. GLAM results are quite the opposite of the ML results. For GLAM, model RMSE is more inconsistent at higher yields.

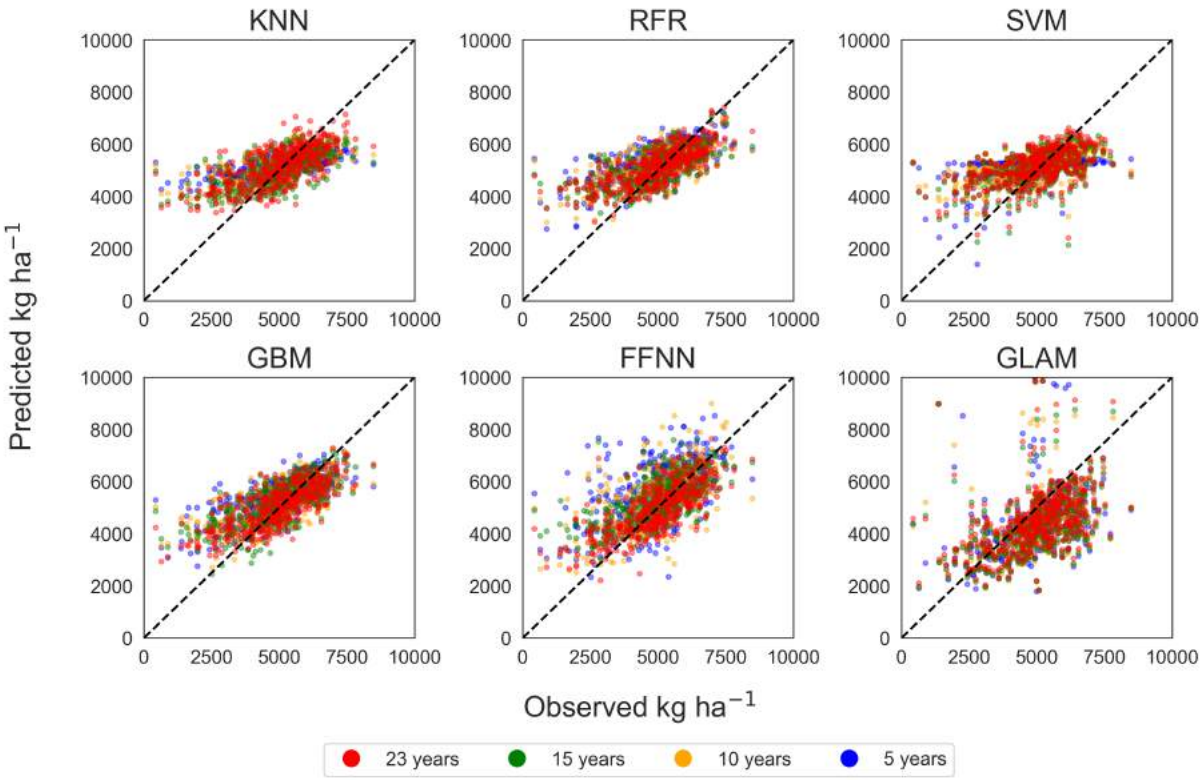

Figure 3.7: Simulated versus observed for each model with reduced data shown by colour. RMSE (top left of each panel) for each model when training data set is reduced from 23 years of data to 5 by systematically removing years in order. RMSE was normalized by the inter-quartile range of the observed data. Points signify 1 growing season for 1 department. Panels from left down designate KNN: K-Nearest Neighbours regression, RFR: Random forest, SVM: Support vector machine, GBM: Gradient boosting machine, FFNN: Feed forward neural network, GLAM: General large area model for annual crops.

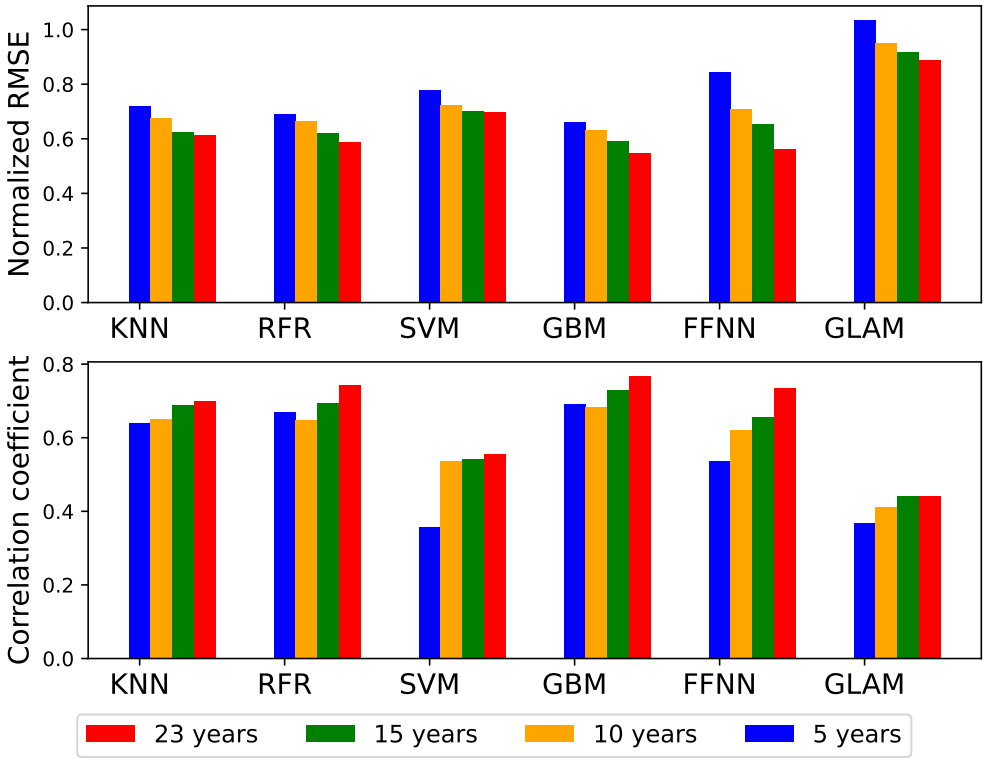

Figure 3.8: RMSE normalized by the inter-quartile range of the training data set and correlation coefficient for each model for each interval of training years.

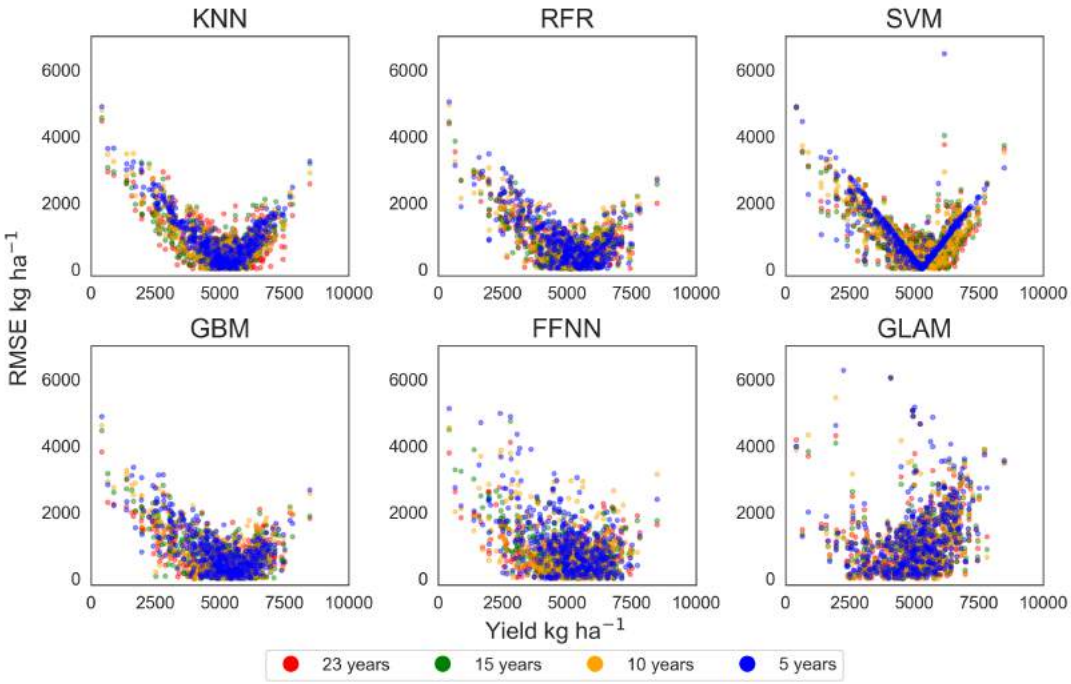

Figure 3.9: RMSE (kg ha^{-1}) plotted against simulated crop yield (kg ha^{-1}) across the 2003 - 2007 test period for each of the models. Panels from left across show KNN: K-Nearest Neighbours regression, RFR: Random forest, SVM: Support vector machine, GBM: Gradient boosting machine, FFNN: Feed forward neural network, GLAM: General large area model for annual crops.

In Figure 3.10 highest and lowest correlating grid cells for each department are compared. This shows the largest potential difference in correlation coefficient when determining a method for selecting representative grid cells. The difference between mean Pearson correlation for GLAM when taking the highest and lowest correlating grid cell was 0.26. The ML models all saw larger disparities between the lowest and highest correlating grid cells. The inter-quartile range of the correlation coefficients achieved by the GLAM model increased by 0.184 between Figure 3.10a and 3.10b. This was much smaller than models such as KNN, with the inter-quartile range increasing by 0.641 However not all ML models were so inconsistent.

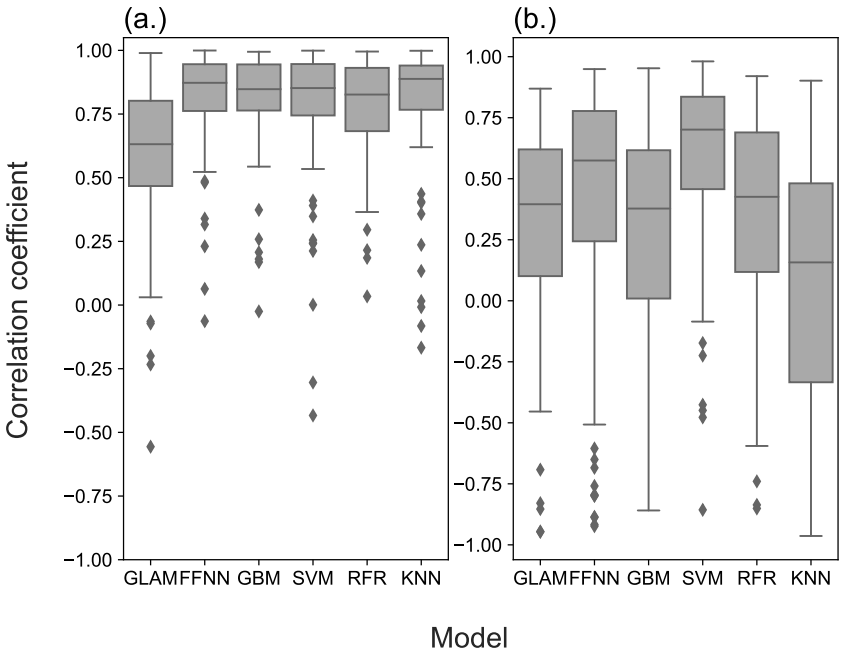

Figure 3.10: Plots (a) and (b) show the correlation coefficient when either the maximum correlating grid cell per department is chosen or the minimum. This leads to large variations in prediction performance as expected.

To further analyse how environmental conditions affect the performance of GLAM or ML, Figure 3.11 shows how the probability of GLAM being the best model changes according to environmental conditions. This was determined by training a random forest classifier on 'best model' labels (determined by RMSE). Training and testing was split randomly with 33% of the best model labels used for testing. The model achieved an accuracy score of 0.725. The model was then used to predict the probability of which model is best against solar radiation, average rainfall, and maximum temperature (a partial dependency plot or PDP). Results show that GLAM is slightly more likely to be the best model under higher solar radiation and lower rainfall.

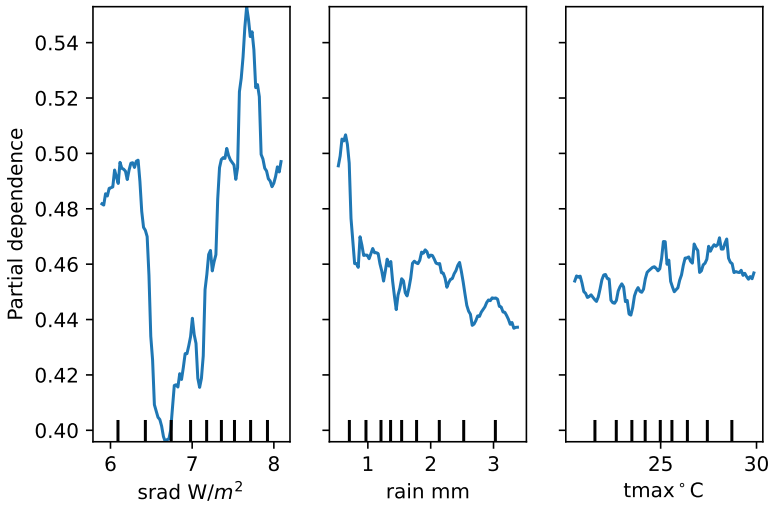

Figure 3.11: PDP of probability of best model

3.4 Discussion

3.4.1 Embedded process knowledge: How does prior parameterization improve performance for out of sample events?

Overall, although predictive performance varies between the 5 years (especially years 2003 and 2007) ML more often produced predictions closer to the observed values and achieved the lowest RMSE (Figures 3.8 & 3.6). GLAM provided the lowest mean absolute error in 2003. Differences in performance between the two methods can be related to the advantages and disadvantages of either approach. For instance embedded process knowledge improves the performance of crop models where data coverage is limited, although ML was able to outperform GLAM where data allowed.

The key advantage of mechanistic crop models over ML is the embedded process knowledge from model structure and prior parameterization. Figure 3.7 demonstrates that ML models tend to achieve poorer results below 3000 kg ha^{-1}, especially with reduced training data.

This pattern is not seen in GLAM likely due to embedded process knowledge from previous parameterization. Supplementary Figure 8.1 illustrates that GLAM is able to capture substantial negative yield anomalies regardless of calibration, therefore providing more weight to the mechanistic understanding present in the model being the source of superior model skill for low yield anomalies. Embedded process knowledge in this case refers to heat stress parameterizations which are crop specific and so cause the simulation to predict reduced yields above certain temperatures.

The neural network (FFNN) model performed the best of the ML models in 2003 (a year of high temperatures and low rainfall) and so in contrast to the other ML models does not show as significant a reduction in performance as crop yield decreases. The performance and flexible architecture of this model suggests significant potential for improvement over process based models.

However, embedded process knowledge also provides an advantage where scale differences provide a looser relationship between weather and crop yield. Figure 3.10 demonstrates this through the small change in correlation coefficient for the GLAM model between panels (a) and (b) in comparison to most ML models. Similarly, although Figures 3.7, 3.8 & 3.9 demonstrate the greater predictive performance of ML over GLAM, Figure 3.10 shows how, for ML, accurate prediction of the variability of the last five years can depend a lot more on the technique used to equate spatial scale. This statement is limited somewhat by the short 5-year period over which correlations were computed.

3.4.2 Why do ML models over-predict low yields?

Machine learning models trained in this chapter tend to over-predict yields above 4000 kg/ha. This was partly why GLAM performed better for predictions in the 2003 evaluation year. One of the key reasons for this is lack of data coverage. Machine learning models can struggle with interpolation, and so yields below that which are present in the training

data may be difficult to predict. From Figure 3.11, it shows that GLAM is more likely to be the better model for high solar radiation. This could mean that ML model struggle to represent the effects of heat stress when very high temperatures reduce crop yields. Therefore, future work could involve the creation of new features to better represent heat stress effects.

3.4.3 How much data is required for machine learning to achieve accurate predictions

The results presented indicate that even with relatively limited data, certain ML methods can provide a valuable alternative approach to mechanistic crop models. ML models achieved lower RMSE and higher correlation coefficient overall with only 5 years of training data across all tested locations within France. Results presented in Figure 3.8 indicate that overall improvements in model performance saturate toward 23 years of training data, indicating that a greater variety of explanatory variables could be more important than using a longer time series. Other studies have shown the significant potential, especially with higher order models, when using a greater diversity of inputs (Folberth et al. 2019, Shahhosseini et al. 2019).

3.4.4 Advantages and disadvantages of process based and machine learning approaches to crop yield estimation from climate variability

Advantages and disadvantages of machine learning are complementary to those of crop modelling. Table 3.6 below shows the strengths and weaknesses of either approach using both the results from this chapter as well as information from the wider scientific literature.

This chapter shows that although machine learning has the advantage over crop modelling of having greater general performance, ML models can struggle to predict the effects of

Table 3.6:
Comparitive advantages and disadvantages of machine learning and crop modelling

Method	Advantages	Disadvantages
Crop modelling	Interpretable	Better calibration requires more site specific information (Angulo et al. 2013)
	Process knowledge improves extrapolation	Calibration can be subjective (Wallach et al. 2021, Seidel et al. 2018)
Machine learning	improved general performance	Can be more difficult to predict extreme events
	Training does not require prior assumptions of the data	Interpretability not inherent in design (Razavi et al. 2022, Rudin 2019)
	Model flexibility (Razavi et al. 2022)	Difficult to extrapolate using ML (Nguyen et al. 2023)

extreme events. Furthermore, strengths and weaknesses of either approach are complementary. For instance, although improvements in crop model calibration require more specific information, improved machine learning model performance comes from a breadth of information, (i.e. greater data coverage). Furthermore, although process knowledge enables crop models to extrapolate and predict the effects of extremes, lack of process knowledge in machine learning enables models to better represent patterns purely from the data. This mitigates potential misspecification if process knowledge is not useful (Wallach 2011). Furthermore, the use of knowledge in calibration of crop models leads to subjectivity. Often models with the same structure may be calibrated in different ways (Wallach et al. 2021), this is not the case to the same degree with machine learning as model parameter values are not explicitly tied to real world model processes. However, machine learning methods also provide greater flexibility. ML models can make use of new or different types of data which may not be as easily or quickly integrated into crop modelling (Razavi et al. 2022). For these reasons, it is argued that machine learning should be used in collaboration with process based crop modelling to leverage the strengths and weaknesses of both approaches.

3.4.5 Combined approaches can leverage the strengths of both modelling methods

This work suggests several key avenues for the continuation of the use of ML with crop modelling. Most importantly, a combined approach would address the difficulty of predicting outlier years for both conditions poorly predicted by the two model types. The key avenues in which this could be achieved are:

1. Model selection based on climatic conditions - a dual prediction approach using a best case model
2. The use of meta-models to train ML using crop model outputs
3. Calibration of a crop model with ML methods
4. A combined pipeline which generates data for calibration of either modelling approach
5. A bench-marking approach in which ML models are used to determine the climatic conditions which require better process model parameterization

Of the above approaches, the simplest implementation is to base model selection on which method achieves best results per climate regime. Thresholds can be applied dynamically accounting for how each model responds to the yield climate relationships present. Post processing of crop model results may also use ML as a meta-model. This approach could make use of embedded process knowledge such as transpiration efficiency or biomass which could improve predictions of extreme events (Feng et al. 2019). A more direct integration of the two approaches would be to use ML for calibration. Some ML is used for crop model calibration currently as a way of optimizing selected parameters. However, calibration is often carried out by hand using a trial and error search (Seidel et al. 2018). Machine

learning methods are well suited to optimization tasks such as model calibration. Another simultaneous integration of the two approaches could be joint calibration using a two way pipeline. This approach would designate a 'perfect model' for where either model is most appropriate, using results to generate training data to calibrate other models. Lastly, machine learning could also be used to benchmark crop model processes. Due to the increased performance of machine learning over crop modelling, comparing crop modelling and machine learning across different climate regimes may lead to insights into how to improve crop modelling. For example, in this chapter, machine learning methods performed better in 2007 when high rainfall lead to increased yields. This could mean that the parameterization of the effect of rainfall on crop yield may require improvement in GLAM. This method of targeting improvements for crop modelling using bench-marking is explored in more detail in chapter 5. Combining approaches should work to further process understanding whilst improving pattern recognition performance.

3.4.6 How data quality and methods to change spatial scale affect results

Figure 3.10 shows that pre-processing methods to change the spatial scale of input data will affect the model performance of ML models more greatly than that of GLAM. Ultimately, changing spatial scales will affect the relationship between weather and yield (the strength of this relationship being an aspect of data quality). This therefore shows that ML models are more sensitive to the strength of the relationship between weather and yield than the GLAM crop model. Other crop models may also be similarly insensitive if parameterizations are not as accurate as that which can be obtained using machine learning. Hence, more consideration should be made to spatial scale when using machine learning methods than crop modelling methods.

3.5 Conclusions

ML and process based crop modelling offer complementary approaches for assessing climate impacts upon crops. Some ML methods such as KNN, SVM, and RFR performed worse where training data is limited. This is particularly important in the context of climate change, as a shifting distribution of climate will cause less common climatic events which are poorly represented in the immediate historic record to become more common (Porter & Gawith 1999, Porter & Semenov 2005). However this was less of an issue with FFNN and GBM models, especially when the full training data set was used. This therefore demonstrates that these two models have the most potential for improving predictions of climate impacts. Improved performance combined with the flexibility of ML make for a valuable tool. GLAM results indicate the usefulness of process knowledge for data limited conditions. The differences in the structure of the predictions may be best leveraged by using the two approaches together as part of a dual approach. Dual approaches (which are explored further in this thesis in chapter 5) are beneficial for combining process knowledge with improved pattern recognition.

3.5.1 Novel contributions of this chapter

This chapter builds on the comparison by Leng & Hall (2020) between ML and process based modelling by focusing on a specific case study of a process based crop model (GLAM) at the regional scale and comparing multiple ML methods using a small amount of observed crop yield data. Insight from this comparison should encourage crop modellers to use ML in combination with process based crop models. It is shown that even with 5 years of data, ML models can offer value as a comparison to process based crop models. This is a lot less data than other studies such as Shahhosseini et al. (2019) who also analyse the data requirements of ML models for crop yield prediction.

4 Sensitivity of machine learning algorithms to temperature and rainfall extrapolations for crop yield and failure prediction

4.1 Introduction

Machine learning models have recently been established as successful methods for the prediction of crop yield from climatological and management factors. Furthermore, such efficacy has been demonstrated in a variety of environments and conditions (Shahhosseini et al. 2019, Leng & Hall 2020, Feng et al. 2019, Delerce et al. 2016, Jiménez et al. 2009). However, as demonstrated by Leng & Hall (2020), commonly used machine learning models typically underestimate observed crop yield variability. In fact underestimation of extremes has been shown in many fields and applications such as spatial down-scaling of precipitation (He et al. 2016), prediction of climatological oscillations (Silini et al. 2021), long term rainfall prediction (Diez-Sierra & Del Jesus 2020) and air quality prediction (Castelli et al. 2020). Although this is somewhat expected, due to the problems of data imbalance in machine learning addressed by studies such as (Chawla 2009), this problem is of particular importance in crop climate modelling and climate impacts modelling as crop yield data is often a limiting factor (Lischeid et al. 2022) and the significant impact of climatic extremes (Schewe et al. 2019).

The effects of climate extremes are often very difficult to predict in a variety of sectors, most notably, agriculture, and health and ecosystem productivity. Underestimation of the effects of extremes are often related to representation of natural processes and human management in the models. For instance, resource constraints may not be accounted for such as constraints on irrigation water in crop model simulations (Schewe et al. 2019). Underestimating the effect of extremes is not just a problem for single events such as the

2003 heat wave, but on a global scale, across many events, the effects of droughts and heat waves are underestimated by crop growth models (Heinicke et al. 2022). The importance of this cannot be overstated as underestimation of the effects of extremes can lead to the underestimation of future yield decline from climate change.

Although more challenging to predict, larger crop yield anomalies are of greater practical concern due to the potential impacts on food security. On a local and country level, climatological fluctuations leading to crop failures can have a significant impact on food accessibility (Maxwell & Fitzpatrick 2012). Rapid food price spikes which could become more frequent with increased climatic warming (Caparas et al. 2021), can result in greatly increased food insecurity (Verschuur et al. 2021). At the global scale, simultaneous crop failures in various 'breadbasket' regions can aggravate impacts on food security through increases in food price volatility (Gaupp et al. 2020).

Therefore, it is necessary to further evaluate the performance and robustness of commonly used algorithms not just for the general distribution of yields, but also for yields significantly below average which are of great practical concern. In the context of this study, model robustness is evaluated by considering the effects of model input perturbations on the evaluation dataset. This is done in order to evaluate the extrapolation potential of ML methods, for both extreme events and future climate change.

Extrapolation is a key limitation for machine learning approaches. It has been argued that extrapolation is essential for simulating physical systems, and is possible because physical laws are characterized by symmetries which are consistent regardless of circumstances (Webb et al. 2023, Feynman 1966). However, extrapolation may be exceedingly difficult to achieve for machine learning methods. Haley & Soloway (1992) have argued that with only limited information provided to a model about an arbitrary function, it is impossible to know how the function may behave outside of the training domain. Hence, if the data

provided does not provide enough information to describe the function entirely, predictions which are too extreme may not be predicted well enough.

Machine Learning methods have been used to extrapolate for future changes in yields by authors including Leng & Hall (2020), Feng et al. (2019). This introduces additional uncertainty as future yields may be affected by warming induced novel climates (in which mean and/or weather variability exceed that of the present climate variability for a particular location) (Porter & Semenov 2005, Williams & Jackson 2007). Therefore, extreme weather events may be more frequent and more extreme than the training data sourced from the present day (Mitchell et al. 2006). This will increase uncertainty of the model simulations depending on the ability of the machine learning method to extrapolate results beyond the training data for increased temperatures in potential future novel climates. Machine learning models differ in structure and complexity, and so it can be generally assumed that ability to extrapolate will also vary between machine learning models. Therefore, it is useful to evaluate how different models may respond to test data in which the input features have been perturbed to increase variability.

4.1.1 Research aims

Here, we investigate the effects of systematic changes to temperature and rainfall across a range of temporal scales to determine the relative effect on model performance, thereby determining model behaviour for various magnitudes of input extrapolation. This is undertaken to show how machine learning models behave for temperature and rainfall extrapolation, and the robustness to input data uncertainty. Multiplicative scaling factors are applied across daily, monthly, and yearly timescales before aggregating inputs to create seasonal features and compare the effects on model performance both across extremes and the entire distribution of observed crop yields. In doing so, the following research questions are addressed:

1. What is machine learning performance for crop yield failure prediction across contrasting environments?

2. How would climate input data uncertainty affect machine learning model performance and correct failure prediction rate?

Here, effects of uncertainty are assessed by using perturbations to compare model performance when input variability is extrapolated against baseline predictions. This is to show potential variability in model results depending on future uncertainty in climate model projections. The term uncertainty is often quantified by variation within an ensemble of climate models and is associated with future changes in climate. The use of perturbations is described in section 4.2.3 and the appropriateness of the magnitude of the perturbations is discussed in section 4.2.4. The perturbation scheme is applied to temperature and rainfall data. Crop failures are methodologically defined in section 4.2.2. Machine Learning methods used are briefly described in section 4.2.5. To compare relative changes in model performance, change in model performance is defined relative to a baseline set of simulations.

4.2 Methods

Figure 4.1 describes the methodological steps taken in this chapter. Observed weather data is used to predict crop failures using machine learning methods (and GLAM, described in 2.1, is shown as a benchmark). Next, climatological data is perturbed using a scheme taken from (Watson et al. 2015). The objective of the perturbations is to simulate uncertainty in rainfall and temperature inputs. Finally, perturbed predictions from the test data are compared against the baseline predictions both for crop failures and crop yields. To test model extrapolation, uncertainty through the perturbations is simulated only in test data, and training data is kept unperturbed. This is to simulate the potential effects which may arise from differences in unseen data relative to the training data, such as future climate

change scenarios or forecasting unseen climate extremes.

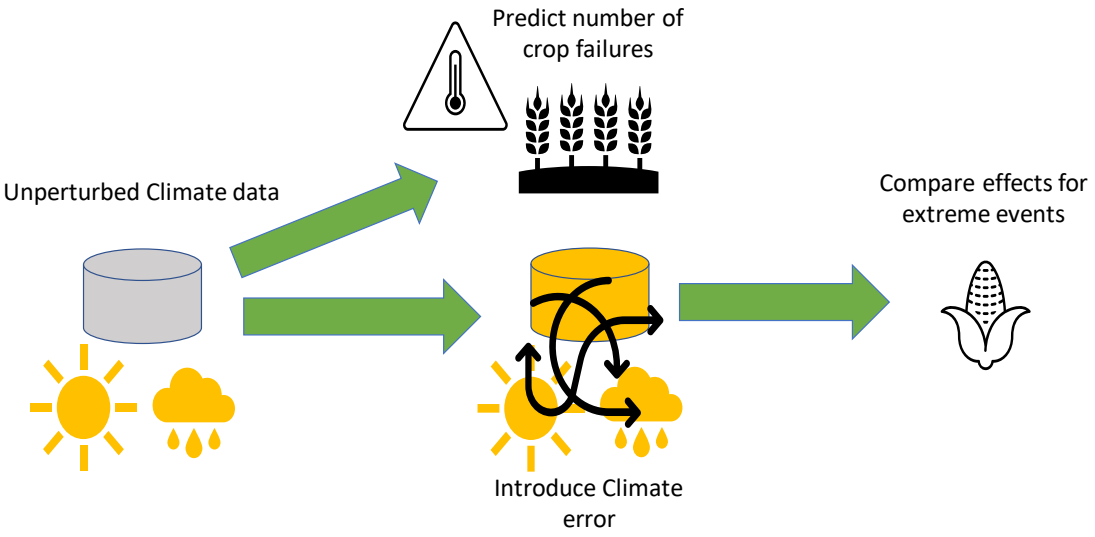

Figure 4.1: Flow diagram describing the methodological workflow of chapter 4

4.2.1 Data

Model results across 2 contrasting datasets are compared. The 2 countries used differ significantly in terms of climatology, and management. Firstly, the French maize (FMA) data set was obtained from AGERESTE Statistique Agricole Anuelle for French NUTS3 departments (Watson et al. 2015). Historical daily temperature and precipitation observations were obtained from the E-OBS gridded observational data set (version 7.0) described in Haylock et al. (2008). The E-OBS gridded data set is on a 0.25 x 0.25 degree latitude/longitude grid scale. The 2 datasets are on different spatial resolutions (0.25 and 0.5 degree resolutions respectively), and across different years (with some overlap).

The SAM dataset developed by Iizumi & Sakai (2020) is different to the FMA dataset in that the source of yield estimates is derived from remotely sensed estimates of NPP (Net

Primary Productivity). This dataset (hereafter referred to as GDHY which denotes global dataset of historical yields) is a hybrid dataset of agricultural census statistics and remote sensing. The authors describe a 4 stage process which was used to build the dataset, which can be summarised as follows:

1. Annual crop yield statistics are obtained from the United Nations Food and Agriculture organization statistical database (FAOstat)

2. Grid cell level (0.5 arc degrees) net primary productivity (NPP) was determined using remotely sensed Leaf area index (LAI), fraction of photo-synthetically active radiation (FPAR), reanalysis solar radiation and reported crop specific radiation use efficiency.

3. Harvest area maps, and crop calendars, were used to determine where and when a crop of interest was grown

4. If a crop has more than one season in a given year, the share of production by different cropping seasons was used to find a production weighted mean from the given seasons.

The method in which the datasets were constructed is hugely significant for this analysis as the representativeness of crop yields within a given grid cell will affect the relationship between observed yields and variations in weather. Climate data used is the EWEMBI dataset developed by Lange (2016, 2018). This dataset includes minimum and maximum temperature, rainfall, and solar radiation data. EWEMBI is a reference climate dataset which merges the E2OBS, WFDEI, and ERAI data sets and then bias corrects the merged data.

The technological yield trend component across time was removed from the data by fitting a LOWESS regression to the observed yields before subtracting the residuals from the

observed data. To ensure the mean of the de-trended yields was consistent with the mean of the observed yields, the mean of the observed yields plus the residuals was subtracted from the mean of the residuals. The method of de-trending crop yields was adapted from Cleveland (1979). The decision to de-trend crop yield data was made to make it easier to assess the machine learning models ability to predict effects of errors in inter-annual variability without the complication of a trend component to the data which is likely determined by factors (such as management) lacking from the observed data. Baseline model performance is also compared against the process based crop model GLAM (General Large area model for annual Crops) (see section 2.1) which requires yield data to be de-trended.

Figure 4.2 presents some of the substantial differences between the France and South Africa datasets used. On average, irrigated area per grid cell is 22.21 % in France and 10.9% in South Africa (maximum irrigated areas of 79.15 % and 72.34 % respectively). There is also larger variation between the proportion of irrigated area per grid cell in France, the standard deviation and inter-quartile range of the percentage of irrigated areas per grid cell being 16.07, 26.18 and 14.29, 16.57 (for France and South Africa). Although the standard deviation of the irrigated areas is only slightly higher in France, the larger inter-quartile range (and larger minimum) in France shows that the most irrigated cells, are much more intensely irrigated in comparison to the rest of the country than in South Africa. This is important for the comparison of rainfall input errors between the two countries. With this in mind, it would be expected that rainfall errors would have a decreased influence when irrigation is more intense. Panels a.ii., a.iii. and b.ii., b.iii. of Figure 4.2 show how the distribution of maize crop yields is significantly different between the two countries. Across all years and locations, the maize crop yields in France follow a normal distribution (Skew of -28.99 and kurtosis of 15.90, an omnibus test of normality defined by $skew^2 + kurtosis^2$ shows a normality p value of 4.19e-238). The South Africa maize yield distribution has a

much higher skew and kurtosis and does not follow a normal distribution. The difference in distributions as well as the strengths of the weather/crop yield relationships (panels a.iv. and b.iv.) mean that there is little generalization between the 2 datasets.

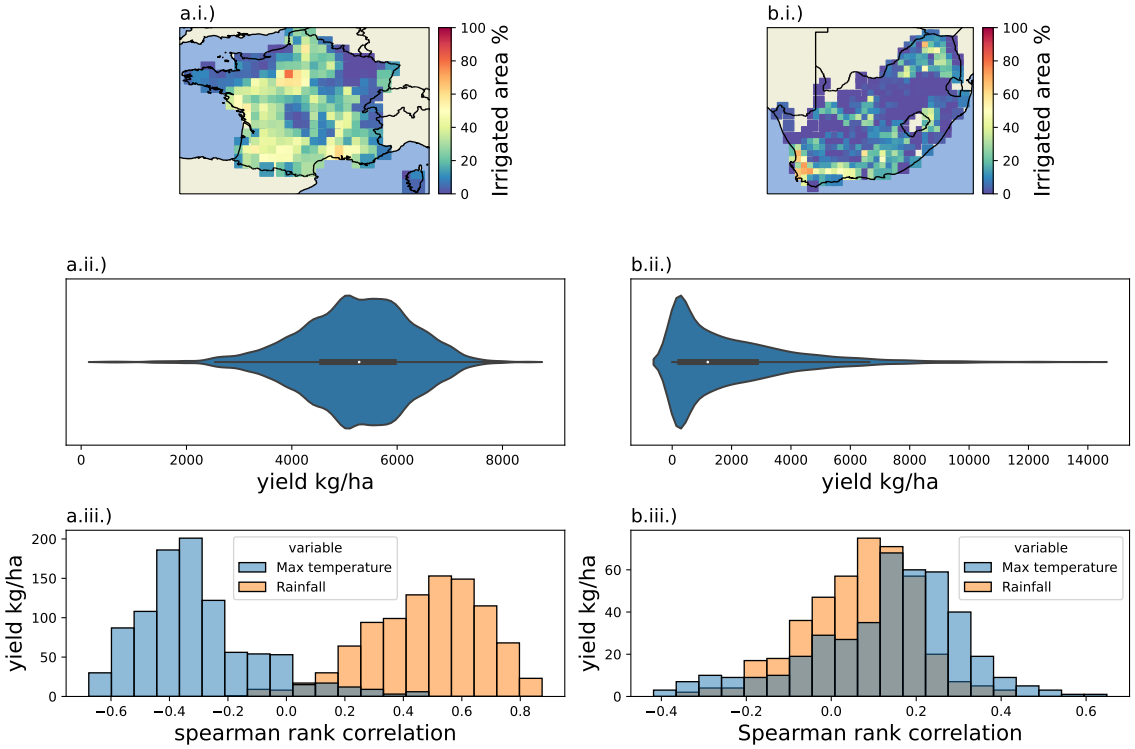

Figure 4.2: A summary of some of the characteristics of the two datasets used in this chapter. Panels a.i. and b.i. show the spatial distribution of irrigated cropland area as a percentage of the total area in each grid cell in France and South Africa respectively. Panels a.ii. and b.ii. show violin plots of the distribution of observed maize crop yields. France has a normal distribution of observed crop yield data whereas South Africa does not. Panels a.iii. and b.iii. are histograms of the Spearman rank correlation between observed crop yields and maximum temperature and rainfall respectively for each grid cell across years.

This is in agreement with Figure 4.3 which shows very little overlap between principle components 1 and 2 of the 2 countries. The PCA was applied using maximum temperature, rainfall, and crop yield data to coincide with the temperature and rainfall perturbations discussed in this chapter.

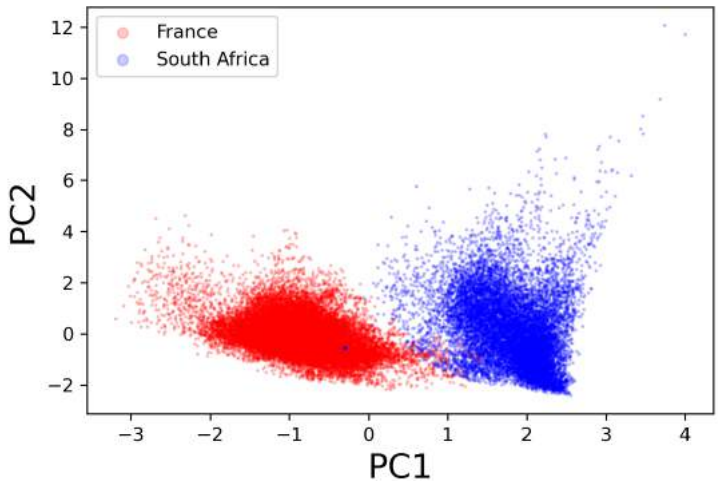

Figure 4.3: Principal component analysis of the maximum temperature, rainfall and crop yield data from both the France and South Africa datasets. Degree of overlap between the 2 datasets indicates level of generalization between them.

Further to the analysis from Figure 4.2, Figure 4.4 shows the average spatial correlations between temperature and yield across France and South Africa. Clearly, there is a much stronger relationship across space between climate and crop yields in South Africa than France. Therefore, The French dataset has stronger temporal correlations between climate and crop yield, but the South Africa dataset has stronger spatial correlations between climate and yield.

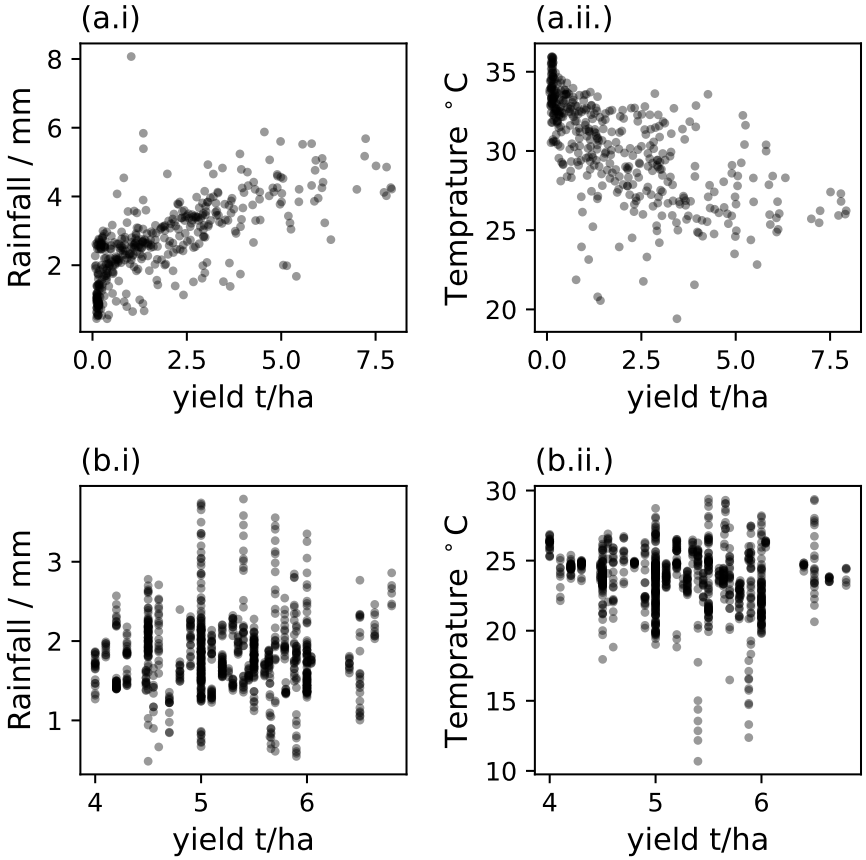

Figure 4.4: Spatially averaged maximum temperature and rainfall plotted against spatially averaged crop yield. In South Africa rainfall most strongly correlates with crop yield across space (Pearson's correlation of 0.58), Panels a.i and a.ii. show the relationship between rainfall and yield and maximum temperature and yield respectively in South Africa. Panels b.i. and b.ii. show the same relationships but for France.

4.2.2 Defining crop failures

At the field scale, a crop failure has been defined as the complete loss of crops on a farm (Mendelsohn 2007). However at the regional scale, when crop yield data is often informed by census statistics, the definition of failure becomes more complex and subjective. ultimately a crop failure should be defined as the threshold at which reductions in yield have a wider societal or socio-economic impact, whilst also taking into consideration the

historical distribution of crop yields for the location to determine what 'failure' really means in the context of the historical mean and standard deviation. This gives rise to 2 different methods of either using an absolute or relative definition of failure. An absolute crop failure threshold is reflective of the yield value below which a farmer would need to break even on costs (Jennings et al. 2022). This could also be indicative of how a change in yield will affect the price of food, particularly within the country of interest however food price volatility is subject to several influential factors such as importer/exporter market share, import deficit, diets and consumption as well as equilibrium effects in international markets and so it can be difficult to attribute a single climatic event which results in yield instability to price changes (d'Amour et al. 2016).

A relative definition of crop failure takes into account the historical precedent of the available data. As such, the number of failures will always be proportional to the distribution of the data overall. This is of particular importance when considering how the number of failures will affect a models ability to predict such failures. There are multiple choices of both definition and magnitude of threshold to use when defining relative yield failure thresholds. Firstly, failure can be defined as below a threshold relative to mean yield for a particular location. Challinor et al. (2010) defined crop failures as yields falling below 2 standard deviations from the mean of a particular location. This study also presented a crop failure index, which is the de-trended yield for a particular year (representing an expected harvest based on the long term trend) divided by the actual harvest for the year. In contrast, Gaupp et al. (2020) set a failure threshold at the lower 25^{th} yield deviation percentile.

Here, crop failures are also defined as below a percentile threshold of the observed maize yield distribution across time for each location. The main reason for choosing the percentile as the method to categorize thresholds is the varying skewness of the 2 yield distributions. Although a threshold based on subtracting some magnitude of the standard deviation from

the mean has been used in many cases (Challinor et al. 2010, 2016b, Goulart et al. 2021, Yang et al. 2020) and could be appropriate for the France dataset, the South Africa yield dataset has considerable skewness which leads to negative thresholds in some cases when using this method. Figure 4.2 panel (b.iii) and (a.iii.) show the differences in crop yield distribution between the two countries, which lead to the decision to use the percentile definition of crop yield failure. To keep consistency and compare crop yield failure rates, regardless of observed yield distribution a percentile threshold was used for the definition of crop failures for both datasets.

The definition of crop failures aims to create a grid cell definition of failure analogous to notable failure events in the historic record. Here the 25th percentile threshold is chosen as the definition of crop failure, meaning, when the observed crop yield falls below the 25th percentile of the observed yield distribution across time for a particular grid cell, a crop failure is said to have occurred. 2 extra thresholds which are more extreme are also tested, the 5th and 10th percentiles. 2 extra thresholds were chosen to ensure conclusions are robust across a range of crop yield failure definitions. Figure 4.5 shows the number of crop failures defined for each year at the 25th percentile threshold. Several years stand out as years in which a significant number of grid cells are failing. Most prominent are 2003 and 1990 in France, and 1991, 1994, and 2006 in South Africa. Some of these years represent well studied significant historical failure events which lead to large socio-economic impacts. For instance the 2003 heat wave in France lead to a 55% reduction of maize production at the European union level (van der Velde et al. 2012). This lead to financial losses of 265 million Euros which resulted in significant intervention from the French government and substantial financial aid (COGECA 2003). Furthermore, in 2007 (crops planted in 2006 in South Africa were harvested in 2007 due to December - February growing season). a 31% decrease in crop yield production in South Africa lead to significant food insecurity for 400,000 people in Lesotho, which heavily depends on South Africa for maize imports

(Verschuur et al. 2021). Analysis of crop yield failure prediction rate takes place at the grid scale. In comparison to baseline simulations, crop yield failure prediction rates are compared with increasing temperature and rainfall errors in the results and discussion sections of this chapter.

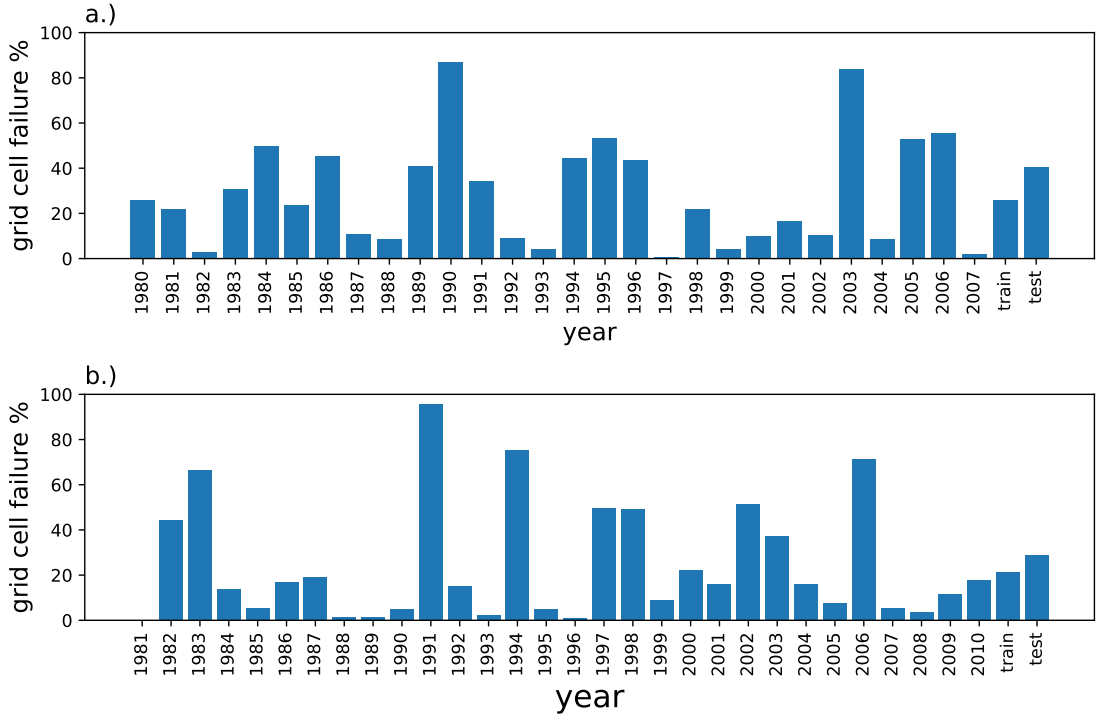

Figure 4.5: Bar plot showing number of grid cells categorized as crop failures each year as a percentage of the total number of grid cells. Panel a.) France data time series and panel b.)South Africa. Last 2 bars show percentage of failures in training and testing datasets. Definition of crop failure set to below the 25th percentile of the observed historical crop yield for each grid cell.

Model performance at predicting crop failures can be measured using several complementary metrics. These metrics are defined according to various combinations of the categories found in Figure 4.6. The confusion matrix is a way to categorize all four possible outcomes of a binary classification. The four components are True positives, (TP) (observed crop failure, which is correctly predicted as a crop failure), False positives (FP) (An incorrectly

predicted crop failure), False negatives (FN) (An observed crop failure which is not correctly predicted) and True negatives (TN) (a non-crop failure which is predicted as such. False positives and False negatives can also be called type I and type II errors.

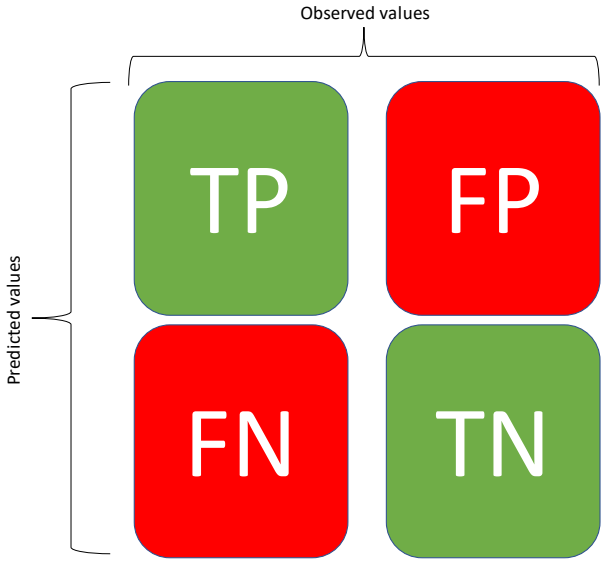

Figure 4.6: Confusion matrix used as a basis to define crop failure performance metrics.

Combining these metrics more useful explanations of model performance can be derived, namely, the True positive rate, false positive rate, recall and precision. The True positive rate is simply the number of true positives as a ratio of the total number of observed positives, similarly, the false positive rate is the number of false positives as a ratio of the observed number of positives. Recall is the number of true positives as a ratio of the number of true positives plus the number of false negatives. Precision is the number of true positives as a ratio of the number of true positives plus the number of false positives. Formally, recall is defined as:

$$recall = \frac{TP}{TP + FN}$$

Similarly, precision is defined as

$$precision = \frac{TP}{TP + FP}$$

In this manner, recall can be thought of as the performance of the model in proportion to the bias towards predicting the negative class (non failures), and precision can be thought of as the performance of the model in proportion to the bias at predicting the positive class (failures). As well as recall and precision, the correct crop failure percentage, and false alarm rate are also used. These two metrics are simply the True positive rate and false positive rate as a percentage. For reference, the correct crop failure percentage and false alarm rate are defined as:

$$correctcropfailure\% = \frac{TP}{TP + FN} X100$$

And

$$Falsealarmrate\% = \frac{FP}{TP + FP} X100$$

Of course by these definitions, recall and true positive rate are the same. However, the distinction is made between them to directly compare True positive rate in a clear way with false positive rate, whilst also comparing recall and precision in a similar manner.

4.2.3 Perturbation scheme

The perturbation scheme was chosen in order to simulate the effects of increases in variability at different temporal scales. For this purpose, the temperature and rainfall perturbation schemes used were taken from Watson et al. (2015). In contrast to the Watson et al. (2015) method, the effects of rainfall and temperature perturbations on model behaviour are evaluated separately rather than together. As the perturbation method is the same as the Watson et al. (2015) study the results here can be compared with the process based model simulations in the previous study. The method was replicated and adapted for use with the South African Maize data along with the original French maize data. Each temperature time series is deconstructed into a mean, and 4 components of variability in the following manner

$$z(y, m, d) = \mu + \alpha_y + \beta_m + \gamma_{ym} + \delta_{ymd}$$

Each of the terms can be defined as

$$\begin{aligned}
\mu &= z(\cdot, \cdot, \cdot), \\
\alpha_y &= z(y, \cdot, \cdot) - \mu, \\
\beta_m &= z(\cdot, m, \cdot) - \mu, \\
\gamma_{ym} &= z(y, m, \cdot) - (\mu + \alpha_y + \beta_m), \\
\delta_{ymd} &= z(y, m, d) - (\mu + \alpha_y + \beta_m + \gamma_{ym}
\end{aligned}$$

Where $\cdot$ denotes the mean across the missing index in $z(y, m, d)$. Once terms have been

defined, a vector of coefficients Θ is used as a multiplicative adjustment to perturb each of the variance components and mean of the input time series to get the new time series $z*$:

$$z^* = (y, m, d, \theta) = \mu_\theta \mu + \alpha_\theta \alpha_y + \beta_\theta \beta_m + \gamma_\theta \gamma_{ym} + \delta_\theta \delta_{ymd} \tag{31}$$

Where:

$$\theta = [\mu_\theta, \alpha_\theta, \beta_\theta, \gamma_\theta, \delta_\theta] \tag{32}$$

Setting $\theta = [1, 1, 1, 1, 1]$ will reconstruct the original time series from the deconstruction. The deconstruction components each therefore can be used to adjust the time series across different time scales. Adjusting μ changes the overall mean of the dataset across all years, α_y adjusts the average deviation from μ for year y, β_m adjusts the average deviation for month m averaged across all years, thus, the sequence of β_m values represents the average seasonal cycle. γ_{ym} is the year dependent deviation from the mean seasonal cycle, and δ_{ymd} denotes the daily deviation from the monthly mean. The components of the deconstruction are illustrated in Figure 4.7 for the example of minimum temperature for one grid cell.

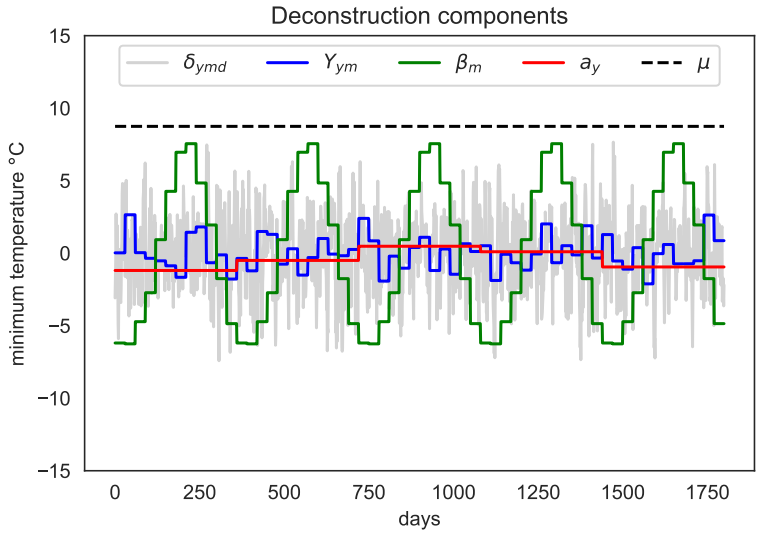

Figure 4.7: Temperature time series deconstructed into mean and components of variability used in the perturbation scheme for minimum, mean, and maximum temperature.

For each grid cell, components of temperature variability were adjusted by up to 3 times their original value. Parameters were adjusted in a uniform manner up to this value. μ was adjusted by 45% of the observed value. The maximum perturbed values for each component are shown in Figure 4.8. Each component was perturbed independently to gain perturbed estimates of yield across 5 parameters and 10 increments of each parameter.

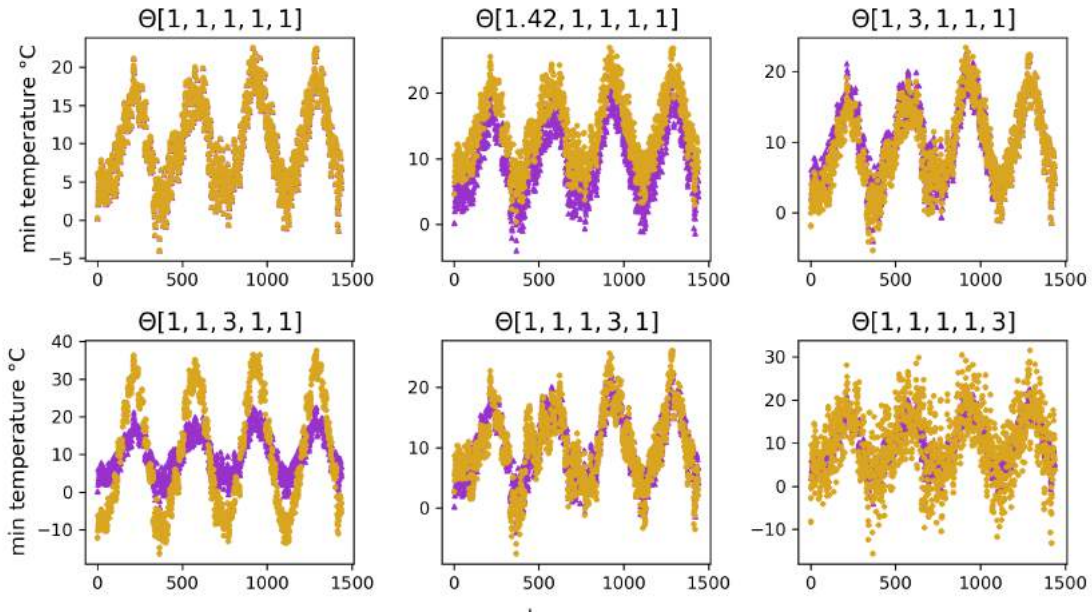

Figure 4.8: Maximum value for each temperature perturbation, with golden points representing the perturbed data, and purple points showing the original data. The top left panel shows the 2 superimposed onto each other as all components of θ are 1.

For rainfall perturbations, there are several different factors to consider. Firstly, using the same method would lead to negative values of rainfall which would be unrealistic. Secondly, rainfall on a daily scale follows a logarithmic distribution with many days without rainfall. With this in mind, Firstly, daily variability was averaged and the log of the resulting time series was taken such that:

$$z(y, m) = log[P(y, m, \cdot)] \tag{33}$$

The resulting time series is therefore deconstructed into the mean and following components of variability:

$$z(y,m) = \mu + \alpha_y + \beta_m + \gamma_{ym} \tag{34}$$

In a similar manner to the temperature perturbations, the components of rainfall variability are determined by

$$\mu = z(\cdot,\cdot),$$
$$\alpha_y = z(y,\cdot) - \mu$$
$$\beta_m = z(\cdot,m) - \mu$$
$$\gamma_{ym} = z(y,m) - (\mu + \alpha_y + \beta_m)$$

Once each of the rainfall components is determined, the transformed rainfall time series is created by:

$$P^*(y,m,d) = \zeta_{ymd} exp[z^*(y,m)], \tag{35}$$

Where:

$$\zeta_{ymd} = \frac{P(y,m,d)}{P(y,m,\cdot)} \tag{36}$$

Both the rainfall and temperature perturbation method are taken from Watson et al. (2015). The rainfall transformation results in the time series shown in Figure 4.9. Similar to Figure 4.8, Figure 4.9 shows results of an unchanged time series against a time series with each of the perturbed parameters set to the maximum value used for the analysis.

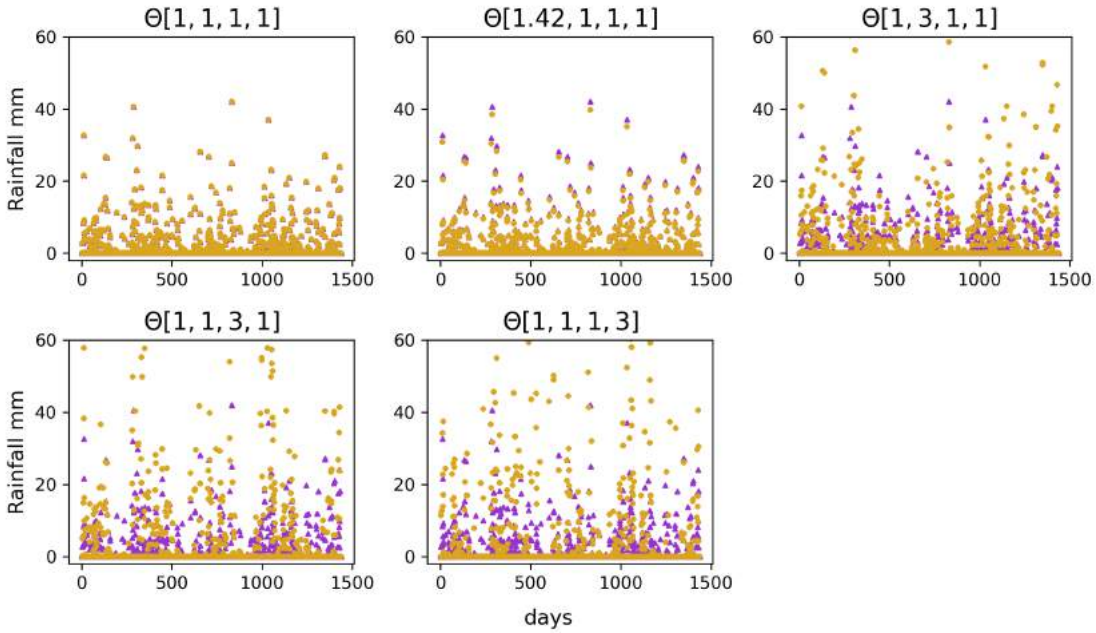

Figure 4.9: The maximum value for each rainfall perturbation, with golden points representing the perturbed data, and purple points showing the original data. The top left panel shows the 2 superimposed onto each other as all components of θ are 1.

4.2.4 Timescale and rationale of perturbations

Perturbations in this chapter are broadly associated with uncertainty across climate model projections. Woldemeskel et al. (2016) have compared uncertainty across an ensemble of climate model projections using CMIP5 models. The authors found that for South Africa and NEU (the region studied which included France) model uncertainty in temperature, described using a square root error variance metric, increased to about 3°K in NEU and between 1-2 °K for South Africa for the period of 2070-2090. Typically, uncertainty was found greater in colder regions and less in warmer regions (providing a contrast between France and South Africa). The magnitude of perturbations used in this chapter, encompass the magnitude of uncertainty for both regions for the maximum timescale (2070-2090) in the study.

Very recently the new climate model ensemble projections (CMIP6) have been published. CMIP6 saw significant improvements in spatial resolution, physical parameterizations, and the inclusion of new components such as nutrient limitations on the carbon cycle (Fan et al. 2020). However, perhaps due to the incorporation of additional components and processes, there are greater differences between GCM projections in CMIP6 than CMIP5, leading to total uncertainty in CMIP6 being 1.20 - 1.93 times higher than in CMIP5, with inter-model variance, 1.38 - 2.07 times larger in CMIP6 (Zhang & Chen 2021a).

Therefore, To encompass a wide range of variability in climate model projections, perturbations were designed to increase variability up to 3 times the observed values (for $\alpha_\theta, \beta_\theta, \gamma_\theta$ and δ_θ). Mean values (μ_θ) were increased by up to 45%. Importantly, each dimension ($\mu_\theta, \alpha_\theta, \beta_\theta, \gamma_\theta, \delta_\theta$) was perturbed individually to result in 10 x 5 sets of model simulations for each model, this allows for ease of comparison between the dimensions, between models, environments and temperature and rainfall.

4.2.5 Machine Learning methods

Machine learning methods were chosen to reflect those which have been used for crop yield prediction tasks within the literature. The decision was made to choose such commonly used algorithms so that the results would be more relevant to the many users who are familiar with such algorithms and may decide to use them for forecasting, crop yield prediction, or climate change impact projection tasks. The algorithms used are the k-nearest neighbours algorithm (KNN), support vector machine (SVM), tree based algorithms random forest (RFR) and gradient boosting machine (GBM), and a neural network. The neural network configuration used is of simple fully connected feed forward architecture. Machine learning models are the same as those used in the previous chapter.

4.2.5.1 feature selection

To keep consistency between datasets, only data which was available for both was used. Furthermore, to ensure focus on the temperature and rainfall perturbations, (and considering substantial correlations between both and incoming solar radiation), only maximum and minimum temperature, rainfall, the number of days over 32 °C and the number of days without rainfall were used for both the baseline predictions and perturbed data. Figure 4.10 shows how as a function of both temperature and rainfall, solar radiation can vary greatly. Pearson's correlation coefficient between solar radiation and maximum temperature is 0.38 in France and 0.40 in South Africa, and correlation between rainfall and solar radiation is -0.52 and -0.74 in France and South Africa respectively. Considering the perturbations are applied to rainfall and temperature in this chapter, solar radiation is removed as a variable as without independent perturbation of solar radiation as well, the presence of this variable would have a varying effect on the differences in model performance resulting from rainfall and temperature perturbations. The number of days above 32 degrees and days without rainfall were included as variables because they have been shown to have a substantial effect on model performance as shown in chapter 3. The number of days above 32 degrees also changes with each perturbation, this leads to significant changes in the representation of extreme events in the input to the models, potentially affecting the models ability to predict crop failures due to heat stress.

The key difference between the experimental setup in this chapter and the previous is that the ML models used in this chapter were not trained using using the latitude and longitude coordinates or solar radiation whereas these three variables were included as inputs in the previous chapter. The decision was made to remove the latitude and longitude as inputs for this chapter as they were shown to be important for model performance for the dataset used in the previous chapter, however they are unrelated to weather variability. Since latitude and longitude can have a strong effect on model performance these two

variables were removed as this would then ensure that model performance would be more dependent on temperature and rainfall. Also removed as inputs for this chapter are soil moisture characteristics DUL, SAT, and RLL. This data was removed as it had very little effect on the model performance in the previous chapter and the data was not readily available for both datasets.

Table 4.1 summarises the differences between the ML model feature inputs between this chapter and the previous. As described in the previous chapter, tmax, tmin and tmean are the maximum, minimum and mean daily temperature averaged over the growing season, srad is the total incoming longwave solar radiation across the growing season, rain is the accumulated rainfall across the growing season, D>32 is the number of days across the growing season which record a maximum temperature above 32 degrees, ddays is the number of days without rainfall across the growing season, latitude and longitude are the grid cell coordinate locations, and DUL, RLL, and SAT are the drained upper limit, wilting point and saturated moisture of the soil.

Table 4.1:
Machine learning models used in both this chapter and the last with Xs denoting where an input feature has been used.

input feature	Chapter 3	Chapter 4
tmax	X	X
tmean	X	-
tmin	X	X
srad	X	-
rain	X	X
D>32	X	X
ddays	X	X
latitude	X	-
longitude	X	-
DUL	X	-
RLL	X	-
SAT	X	-

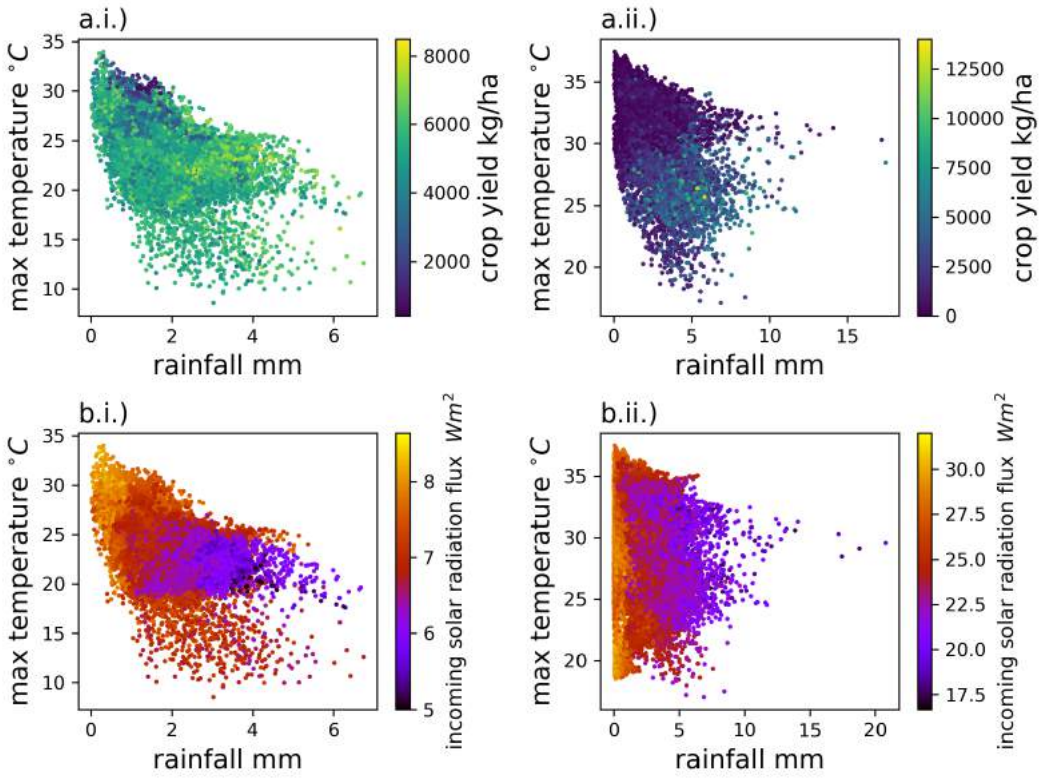

Figure 4.10: Correlations between input rainfall and temperature as a function of crop yields (a) and incoming solar radiation flux (b). (i) shows France data, (ii) shows South Africa data.

4.2.6 Process based crop model

To evaluate crop failure prediction performance in comparison to baseline simulations, a process based crop model is used as a benchmark. The GLAM crop model (General large area model for annual crops) (Challinor et al. 2004) has been used to predict crop yields across a range of spatio-temporal conditions for maize crops (Watson et al. 2015, Challinor et al. 2016a, Bergamaschi et al. 2013, Falconnier et al. 2020). A description of the GLAM model is found in section 2.1. GLAM was chosen as it is able to produce predictions of crop yield at the regional scale and has been previously tested in the locations chosen for this chapter (Watson et al. 2015, Jennings et al. 2022). For maize, France has been delineated

as exhibiting entirely temperature limited conditions, whereas South Africa has a mixture of temperature and rainfall limited conditions (Sacks et al. 2010). GLAM represents non-optimal temperature conditions by using critical thresholds, below (or above) which a linear reduction in transpiration efficiency (TE) is applied. For maize, the cold threshold is 18 °C and hot threshold being 32 °C. Transpiration efficiency determines change in biomass for each daily time step, reductions in transpiration efficiency will therefore reduce yield through this mechanism. High temperature stress is also represented by the acceleration of leaf senescence and a lethal temperature threshold. The effects of changes in rainfall are represented by using a soil water stress factor parameter which affects growth of leaf area index below a critical threshold value. A terminal drought stress parameter is also used to affect growth in extreme conditions. There is currently no parameterization in GLAM for representation of excess rainfall leading to water logged crops.

4.2.7 Process-based crop model calibration

Across the two datasets, different calibrations are required. For the French maize dataset, simulations from Watson et al. (2015) are used as the benchmark. For this calibration, the YGP parameter was calibrated using increments of 0.01 from 0 to 1, all other parameters were set to their default values in GLAM aside from transpiration efficiency (TE) and the maximum value of normalized transpiration efficiency (TEN MAX). These two parameters have been found to have significant effect on simulated yield. TE was set to 5.45 pa and TEN MAX was set to 6 (kept constant for each location in France). These values are taken from Tallec et al. (2013) and are more realistic for temperate regions such as France. The planting date (start of crop yield simulation) was set to April after Birch et al. (2003).

For South Africa, although a previous calibration exists using crop yield data aggregated for the entire country (Jennings et al. 2022), 3 calibrations were tested across 2 spatial

scales, and variations of the YGP parameter. The best calibration was subsequently chosen for subsequent analysis after initial comparisons. Figure 4.11 shows a comparison of 3 GLAM simulations. column (a) shows the results of the calibration using the country level (FAO) data, gridded to the 0.5 degree grid cell scale. This results in the same yield value, and hence the same YGP value, being used for each grid cell location across the country. The value determined by the calibration was 0.08. The YGP parameter was varied between 0.02 and 1 using intervals of 0.02 to determine this value. By contrast, Column (b) shows the results of a calibration against the GDHY dataset (Iizumi & Sakai 2020) in which yield spatially varies at the 0.5 degree scale. Section 4.2 goes through the differences between the GDHY dataset and the FAOstat data at the country scale. In the case of column (b), the YGP parameter was calibrated from 0.01 to 1 using intervals of 0.01. Column (c) displays the results of an uncalibrated GLAM simulation, meaning that the YGP parameter was not calibrated and so the YGP value is 1 meaning that the parameter has no effect on the results of the model. This parameter setting is often called "potential yields" meaning that projected crop yields are determined for the hypothetical scenario that no spatial limiting factors such as non-optimal management or pests and diseases reduce crop yields.

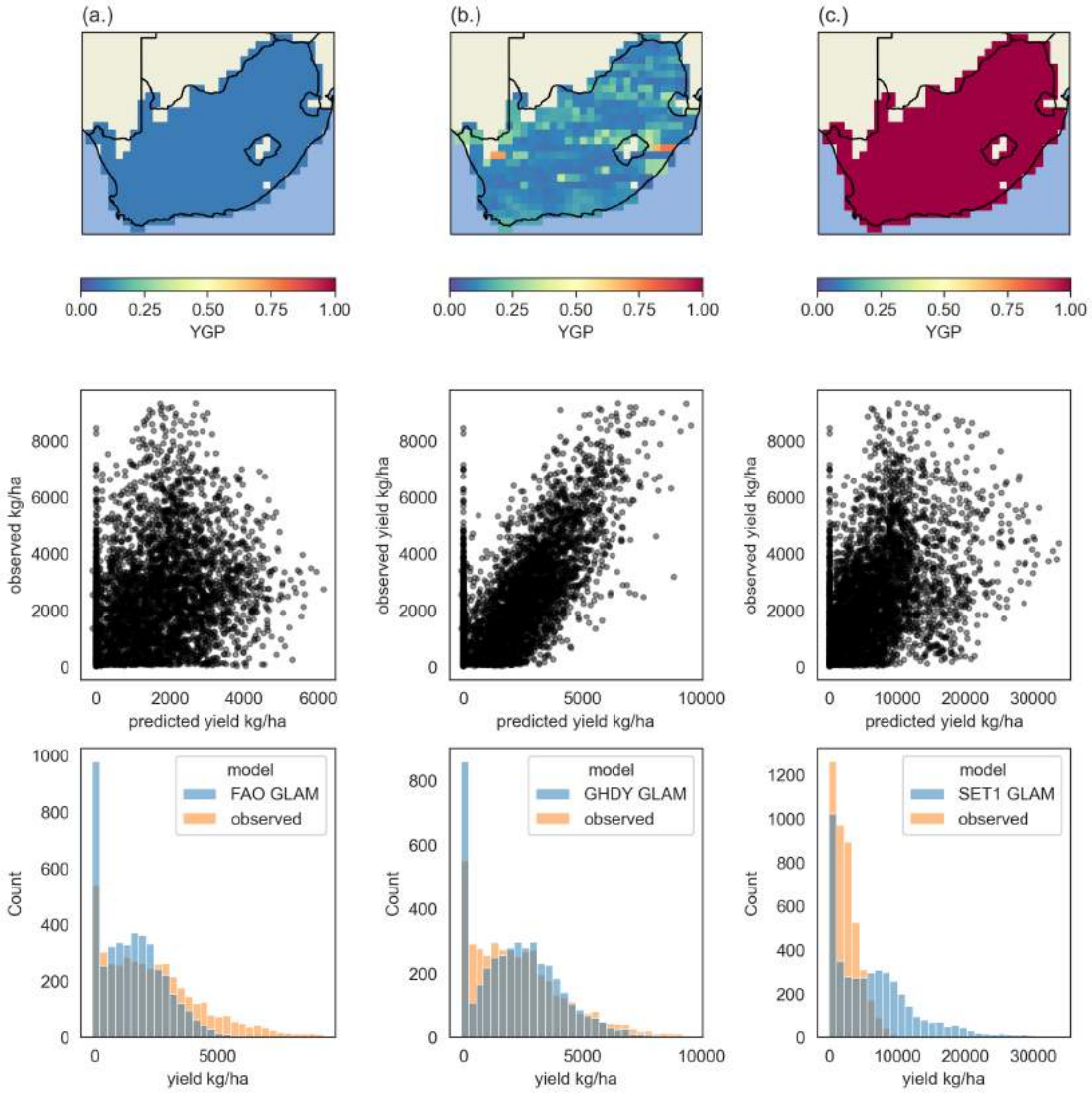

Figure 4.11: GLAM calibrations across 3 spatial scales. Column (a) shows 1 YGP value for the entire country (FAO GLAM), (b) shows a YGP value per grid cell (GDHY GLAM), (c) shows results where YGP is uncalibrated (SET1 GLAM). Each column shows a map of YGP values, scatter of simulated versus observed yield, and histograms of predicted and observed yield.

GLAM model performance appears to be much more skilful when the YGP is calibrated per grid cell. This shows that more specific crop yield data improves GLAM model simulations when spatially distributed (similar to the conclusions of (Angulo et al. 2013)).

For South Africa, all other GLAM model parameters were taken from (Jennings et al. 2022). Most crop parameters apart from thermal time requirements were taken from Asfaw et al. (2018) and thermal time requirements for the region were taken from Durand et al. (2018). Parameters were checked to ensure that they produced potential (no YGP calibration) yields that were realistic (according to the yield gap atlas see van Bussel et al. (2015a)). Testing and evaluation periods were chosen based on the period which had the best correlation between observed crop yield (from FAO) and weather data. This decision was made because some parts of the time series had very poor correlations between weather and yield which would make comparisons difficult. Planting dates were chosen based on dates which achieved greatest yield assuming optimal timing of planting date, this resulted in a planting date of December (Jennings et al. 2022).

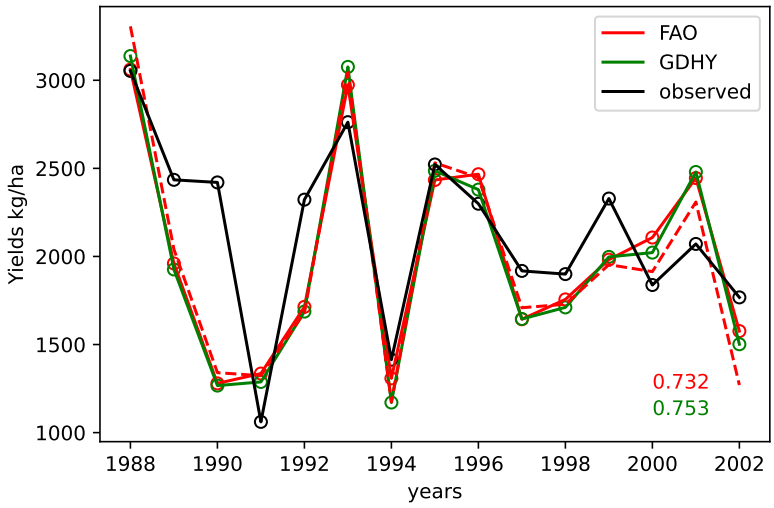

Figure 4.12: GLAM model performance when calibrated using country level crop yield data from FAO (Food and agriculture organisation of the United Nations), and GDHY (Global dataset of historical yields) for major crops (Iizumi & Sakai 2020), model performance is measured as the pearson correlation coefficient in green for GDHY and red for FAO calibration.

To compare skill at the country scale, from grid cell level predictions, several transfor-

mations are applied. Firstly total production per grid cell is determined by multiplying predicted yield by the harvested area of the grid cell. Harvested area data was acquired from Portmann et al. (2010). Data available is for the year 2000, hence this may be a data limitation of the method depending on how much harvested areas have changed since this year. Production per grid cell is then added before being divided by the total growing area of the country. The most valuable aspect to this method of plotting the data is that crop yield predictions and observations are area weighted. This can be important given the often substantial correlation between growing area and yields. This is often due to the idea that areas of land in which crops are grown tend to be more productive, a phenomenon known as the agricultural niche effect (Challinor et al. 2015). When considering GLAM performance it is interesting to note the vast difference in skill when compared against grid cell level and country level data. Since Figure 4.12 is weighted by growing area, and Figure 4.11 is not, it is most likely the cause of this disparity. Furthermore, Watson et al. (2015) has also shown that GLAM performance can more consistent when predicting using grid cells of higher growing area.

4.3 Results

The following results show how ML model results can be significantly affected by input data uncertainty. Across timescales, uncertainty in the magnitude of inter-annual variability has the greatest effect on both general model performance as well as crop failure prediction. Baseline predictions show the large differences in model performance across environments, highlighting the need for testing across a range of conditions.

4.3.1 Baseline simulations

Baseline simulations demonstrate the wide range of machine learning performance across regions presented. Both datasets show bias against crop failures. Yield variability, both observed and predicted is greater in South Africa and mean yields are in general lower in

South Africa (Figure 4.13).

Figure 4.13 shows that GLAM tends to under-predict yields in both datasets, with a large standard deviation. Many machine learning models over predict yields, this is more prominent in France than in South Africa. As observed yields are normally distributed in France, however in South Africa the distribution shows far greater skew, a larger number of lower observed values mean that some machine learning models (NN, SVM, KNN) do not significantly over predict yields in South Africa. A notable difference between the two datasets is the large difference in both observed and modelled standard deviation. Machine learning models in South Africa show greater skill in reproducing the observed standard deviation across the dataset. This is also shown by the higher correlation coefficient scores in 4.2. In fact, in general, machine learning models show greater performance using the South Africa dataset than the France dataset.

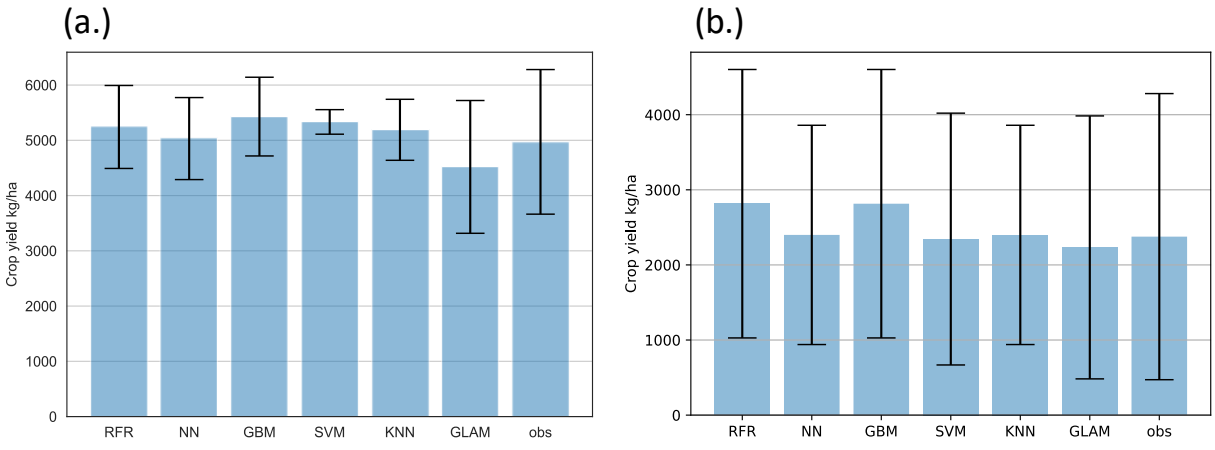

Figure 4.13: Mean yields as predicted by all models against observed and predicted standard deviations shown as error bars. Mean and standard deviation are taken across years and locations simulated for each respective test period. Panel (a) shows the model predictions for the French maize dataset. Panel (b) shows the model predictions for the South African maize dataset.

When comparing between South Africa and France in Figure 4.14 model performance appears to vary for similar reasons. Although observed values continue to decrease below 4000 kg/ha many ML models do not predict much lower than this value. In contrast, in South Africa many observed values of around 0 are predicted much higher by all models. This was also the conclusion drawn in chapter 3. Since this conclusion is common across both datasets, even though the distribution is more skewed in South Africa, it is evident that machine learning models struggle to predict the very extremes of variability regardless of the distribution of the data leading to more extremes.

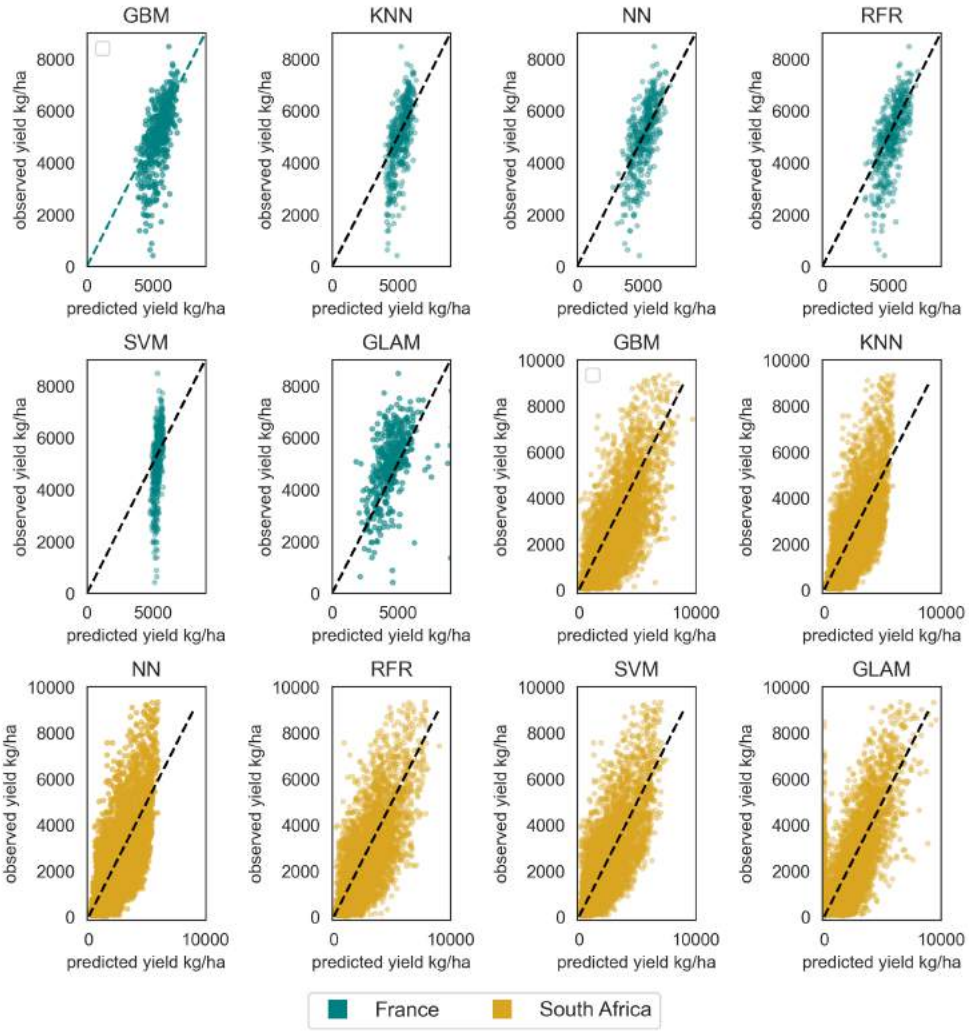

Figure 4.14: Baseline scatter plot for both France and South Africa with all models. ML models are purely driven by weather variables, with no solar radiation input as described in section 4.2.5.1.

A key aspect of model performance is the ability to capture the inter-annual variability of crop yields. Figures 4.15 & 4.16 show how each of the models compare to the inter-annual variability per grid cell. In France, the GBM and neural network models best capture the general trend of higher variability in the north east and south west of the country. In South Africa, interestingly, in many cases, modelled inter-annual variability is actually

greater than observed. This is unusual, particularly for machine learning models as model performance will depend on data coverage, with less coverage at the extremes (Leng & Hall 2020). However, GLAM, the GBM model and the RFR model all predict larger coefficient of variance than which is observed in the west of the country. It is important to note that in the west of the country, water availability is a factor limiting crop growth (Sacks et al. 2010), although irrigation is not as intense in South Africa (see section 4.2.1) some irrigation in the west could suppress the variability which is not accounted for in the models. The machine learning models in particular have no parameterization for irrigation and are purely weather driven and so would not distinguish between irrigated and non-irrigated grid cells. The gradient boosting and random forest models may be used to confirm the need for irrigation, presenting the difference between irrigated and non-irrigated scenarios when compared against observed.

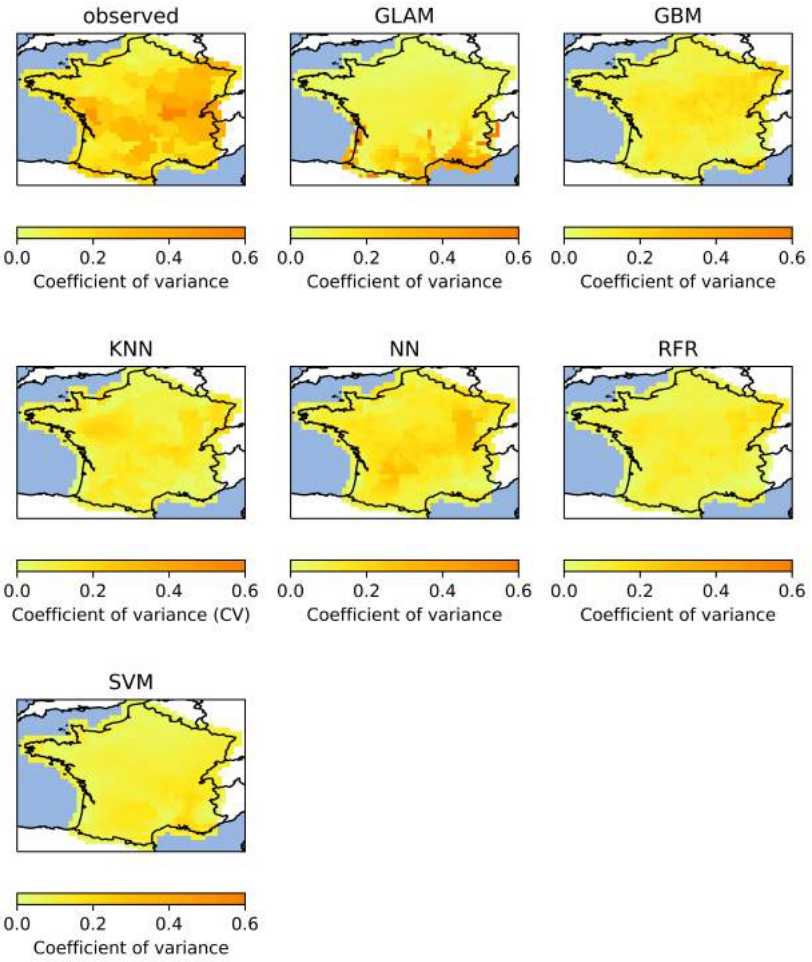

Figure 4.15: Spatial distribution of maize variability both observed and predicted by each of the models used within this chapter in France. Yield variability is measured by coefficient of variance (CV) which is the standard deviation of yields divided by the mean across time per grid cell.

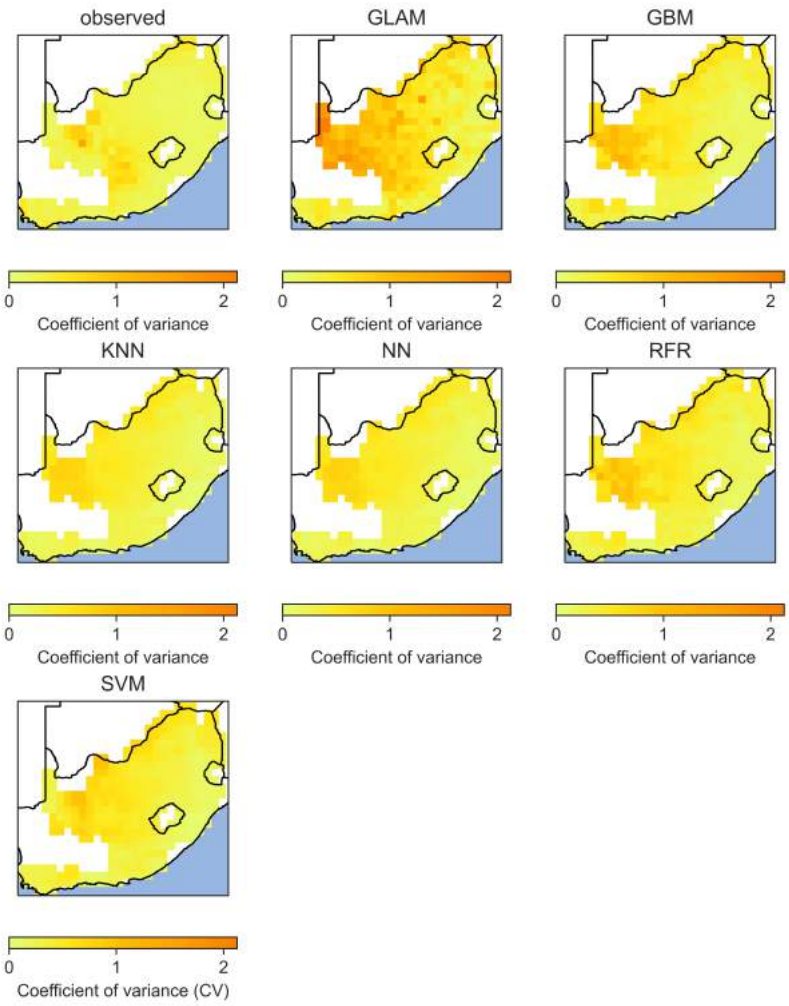

Figure 4.16: Spatial distribution of maize variability both observed and predicted by each of the models used within this chapter in South Africa. Yield variability is measured by coefficient of variance (CV) which is the standard deviation of yields divided by the mean across time per grid cell.

Table 4.2:
Model performance metrics between baseline simulations of all models and observed data in both France and South Africa. Metrics used are RMSE: root mean square error (normalized by the inter-quartile range of the observations), CCOEF: pearsons correlation coefficient.

model	France		South Africa	
	NRMSE	CCOEF	NRMSE	CCOEF
GBM	0.709	0.643	0.554	0.734
KNN	0.662	0.657	0.479	0.767
NN	0.614	0.687	0.479	0.745
RFR	0.632	0.648	0.496	0.751
SVM	0.794	0.530	0.462	0.767
GLAM	0.890	0.443	0.534	0.712

When comparing model performance for predicting crop failures (Figure 4.18), GLAM is able to predict crop failures consistently both at larger and smaller thresholds, whereas for all ML models correct crop failure percentage (true positive rate) decreases substantially when the failure threshold is reduced from the 25^{th} percentile to the 10^{th} percentile then the 5^{th} percentile. In South Africa the number of False alarms also decreases, meaning simply less crop yields are predicted as crop failures, correct or not. In France False alarm rate increases however true positive rate decreases between the 25^{th}, 10^{th} and 5^{th} percentiles, meaning the models become more biases towards false alarms as the crop failure threshold decreases. In France the GLAM correct crop failure percentage decreases as the false alarm rate increases, however in South Africa, correctly predicted crop failures increase with the decreasing failure threshold and the false alarm rate stays roughly the same, meaning the model is biased towards predicting crop failures. As well as Figure 4.18, Figure 4.17 also shows how even though ML model performance degrades with each decrease in the crop failure threshold, GLAM performance is slightly more consistent. The largest difference is that GLAM recall decreases in South Africa between the 25^{th} and 5^{th} percentile thresholds more than precision. This means that more False negatives are predicted at smaller thresholds but false positives either stay the same or decrease. Of the

machine learning models, the neural network model is the most precise in France, even at smaller thresholds, the model with the highest recall changes between random forest and gradient boosting depending on the threshold. In South Africa, the support vector machine is often the more precise model, and more so at the lower thresholds.

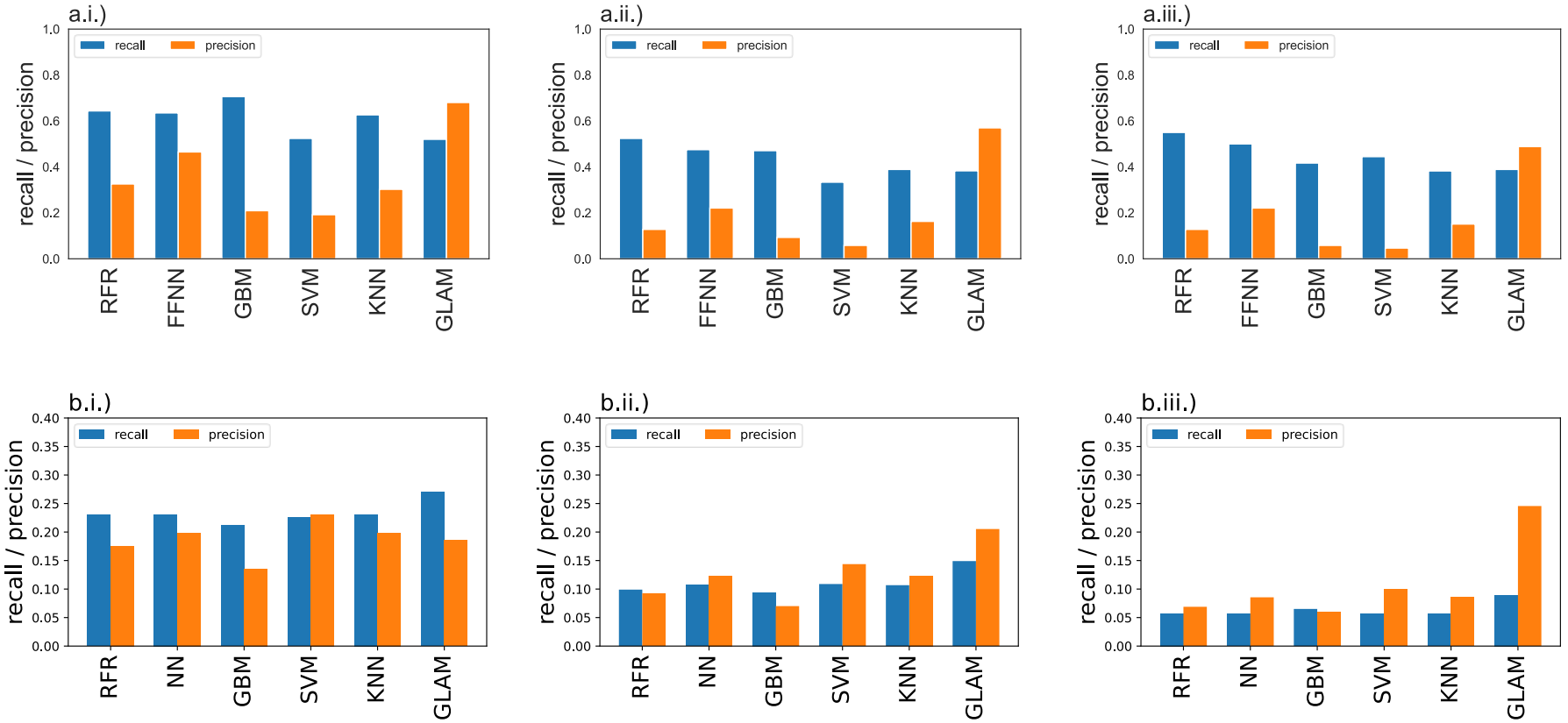

Figure 4.17: recall and precision for models evaluated using the France (a) and South Africa (b) datasets, with (i) showing results for the below 25th percentile definition of crop failure, (ii) shows the 10th percentile definition of crop failure and (iii) shows results for very extreme crop failures only, below the 5th percentile of the observed yield.

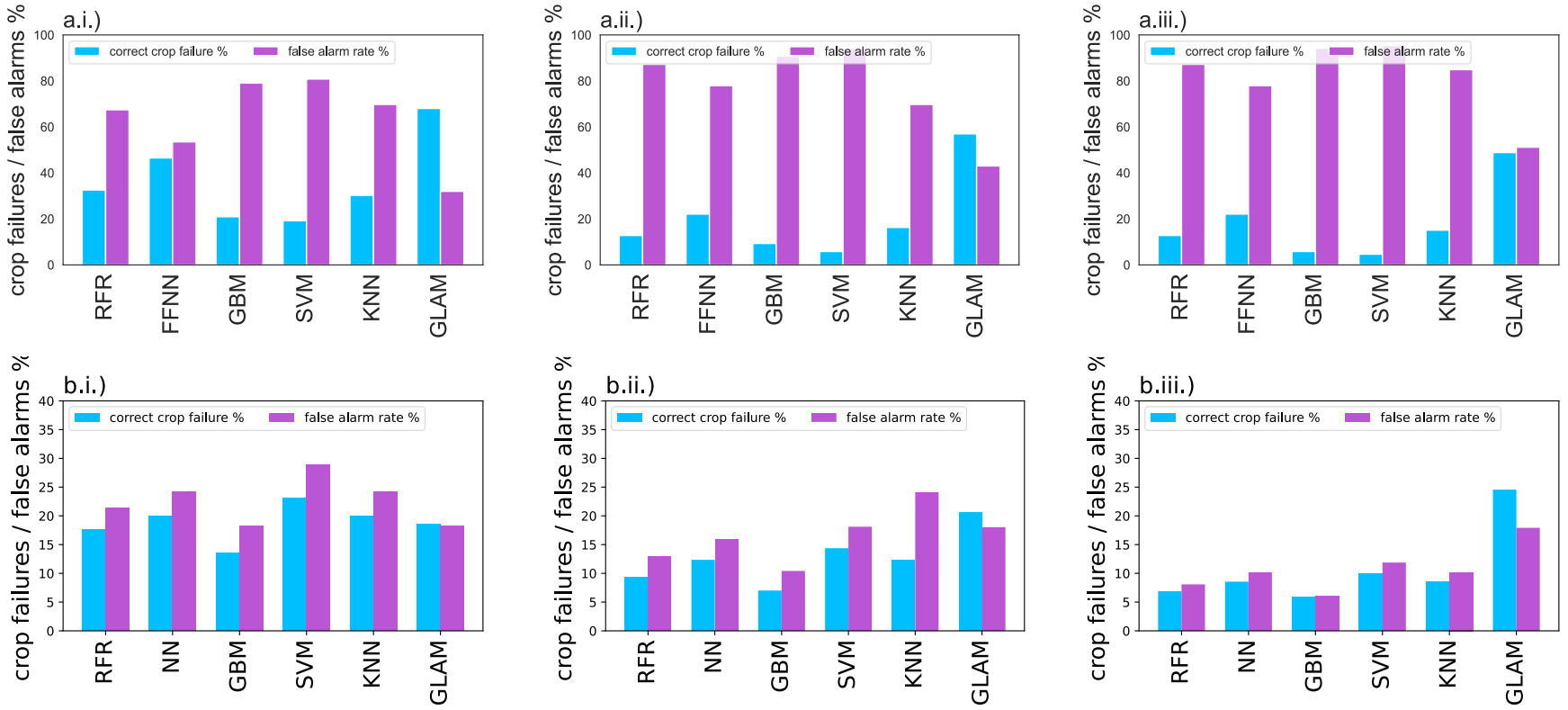

Figure 4.18: correctly predicted crop failure and false alarm percentage for models evaluated using the France (a) and South Africa (b) datasets, with (i) showing results for below 25th percentile definition of crop failure, (ii) shows the 10th percentile definition of crop failure and (iii) shows results for very extreme crop failures only, below the 5th percentile of the observed yield.

4.3.2 Effects of input data uncertainty on model performance

Results from the perturbation experiment are shown here. RMSE, pearson's correlation coefficient metrics, as well as the percentage of correctly predicted crop failures with associated false alarm rate (number of incorrectly predicted crop failures) are shown to vary with magnitude of perturbation. The effects of perturbations on the prediction of crop failures are presented in Figure 4.21. Crop failures are shown as the number of correctly predicted crop failures as a percentage of the total number of crop failures in the observed data.

Figure 4.19 shows how each of the models respond to perturbations of increasing magnitude across the timescales described in section 4.2.3. There are several interesting implications which can be drawn from this Figure. Firstly, the models tend to agree that uncertainty in mean temperature is more important than uncertainty in mean rainfall for model performance. Errors in mean temperature are associated with μ in row b of both 1 and 2. In most cases, increases in the magnitude of mean temperature error also increases model error, however this is more so for South Africa than for France. This is in contrast to errors in inter-annual variability, which affect models both through temperature and rainfall. The models therefore suggest that they are more sensitive to changes in rainfall extremes on inter-annual timescales than mean changes. Overall, models are more sensitive to uncertainty in temperature than rainfall. Mean increase in RMSE across all models with maximum input temperature perturbation is 0.0368 with increases in mean error and error in inter-annual variability increasing model error by 0.359 and 0.0735 respectively. Errors in inter-annual variability of rainfall by comparison increase model output error by 0.0140 in France.

For South Africa temperature uncertainty is also more important however RMSE is more similar between rainfall (0.0896 RMSE) and temperature (0.1394 RMSE). Models are also

more sensitive to the changes in γ (which is the year dependent deviation from the mean seasonal cycle) in South Africa than in France. Uncertainty tends to have a more significant effect on model performance in South Africa than in France. For instance, correlation coefficient reductions in South Africa from the most sensitive model (SVM) are greater than the most sensitive model in France (Neural network). Largest differences between correlation coefficients are the change between 0.687 to 0.402 when the uncertainty in the inter-annual variability of temperature is increased for the neural network model in France, and the decrease from 0.767 to -0.154 when the uncertainty in the inter-annual variability of temperature is increased in South Africa for the SVM model. In both cases, the model which performs best in the baseline also has the greatest reductions in performance when uncertainty increases. This could be due to differences in management intensity such as greater irrigation intensity in France. It is important to note that differences between the countries may be due to the differences in how the yield data was collected rather than country specific differences.

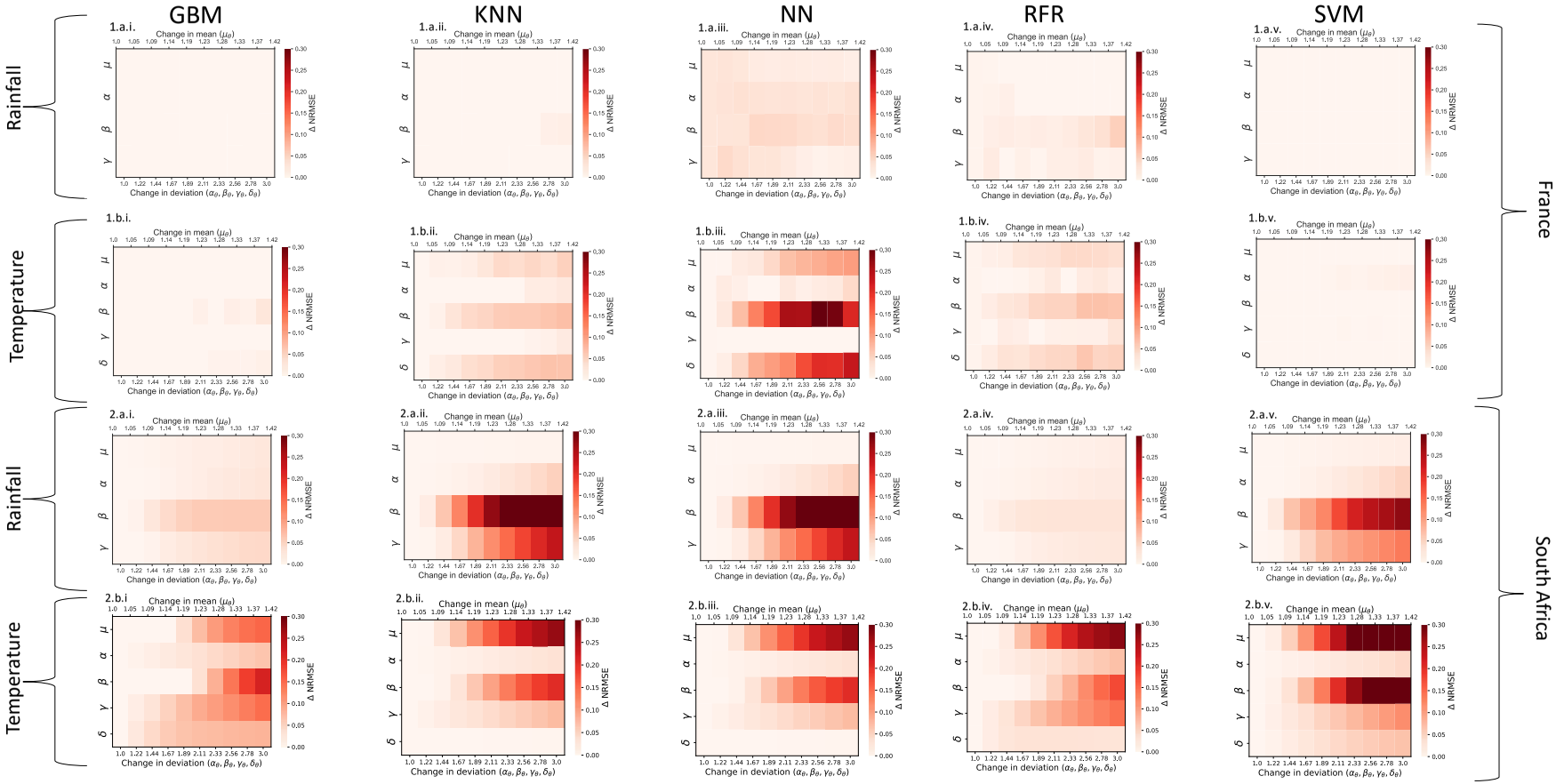

Figure 4.19: Changes in model RMSE with increasing perturbations across models, the 2 environments (France and South Africa) with changes to temperature and rainfall respectively. Model acronyms are GBM: gradient boosting machine, KNN: K-nearest neighbours, NN: Neural network, RFR: Random forest, SVM: Support vector machine. For each panel, number 1 or 2 denotes results for France or South Africa respectively, a or b denotes rainfall or temperature perturbations respectively, and numerals i to v represent each of the machine learning models tested ranging from GBM to SVM, left to right in alphabetical order.

Figure 4.20 gives some indications as to the causes of deviations in model RMSE and correlation coefficient. For instance, changes in mean and inter-annual variability in South Africa result in stronger reductions in the standard deviation of model predictions in South Africa than in France. This coincides with changes in model RMSE and correlation coefficient also being greater overall in South Africa. In many cases, strong changes in RMSE coincide with strong decreases in predicted standard deviation (e.g. temperature in South Africa and panel 2.a.v). However this is not always the case, evident from panel 1.b.iii for example which sees increases in standard deviation across all timescales although RMSE shows a clearer trend of increasing with increases in the perturbed mean (μ), uncertainty in the inter-annual variability (β), and daily variability from the monthly mean (δ). The reasons behind why the standard deviations of the outputs may change in response to the perturbations are explained in section 4.4.2.

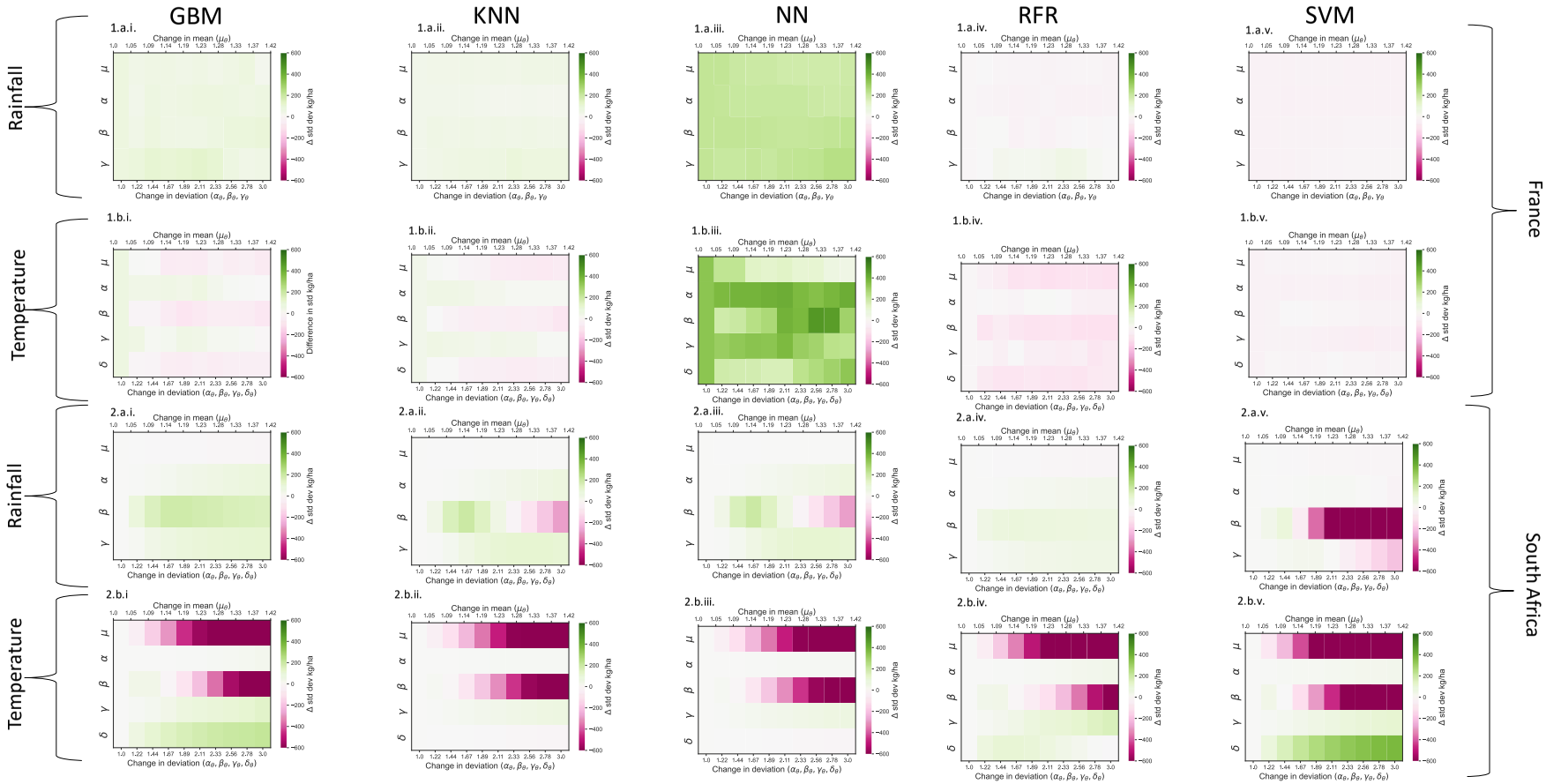

Figure 4.20: Changes in standard deviation across models, the 2 environments (France and South Africa) with changes to temperature and rainfall respectively. Model acronyms are GBM: gradient boosting machine, KNN: K-nearest neighbours, NN: Neural network, RFR: Random forest, SVM: Support vector machine. For each panel, number 1 or 2 denotes results for France or South Africa respectively, a or b denotes rainfall or temperature perturbations respectively, and numerals i to v represent each of the machine learning models tested ranging from GBM to SVM, left to right in alphabetical order.

Figure 4.21 and 4.22 show the correctly predicted crop failure rate (True positives) and falsely predicted crop failure rate (False positives) respectively. These two Figures are complementary as True positives and false positives are related via the confusion matrix presented in Figure 4.6. Combined, the Figures show that generally, with increased magnitude of perturbed variability, both correctly and incorrectly predicted failures both increase. Hence the number of predicted failures increase regardless of whether they are correct or not.

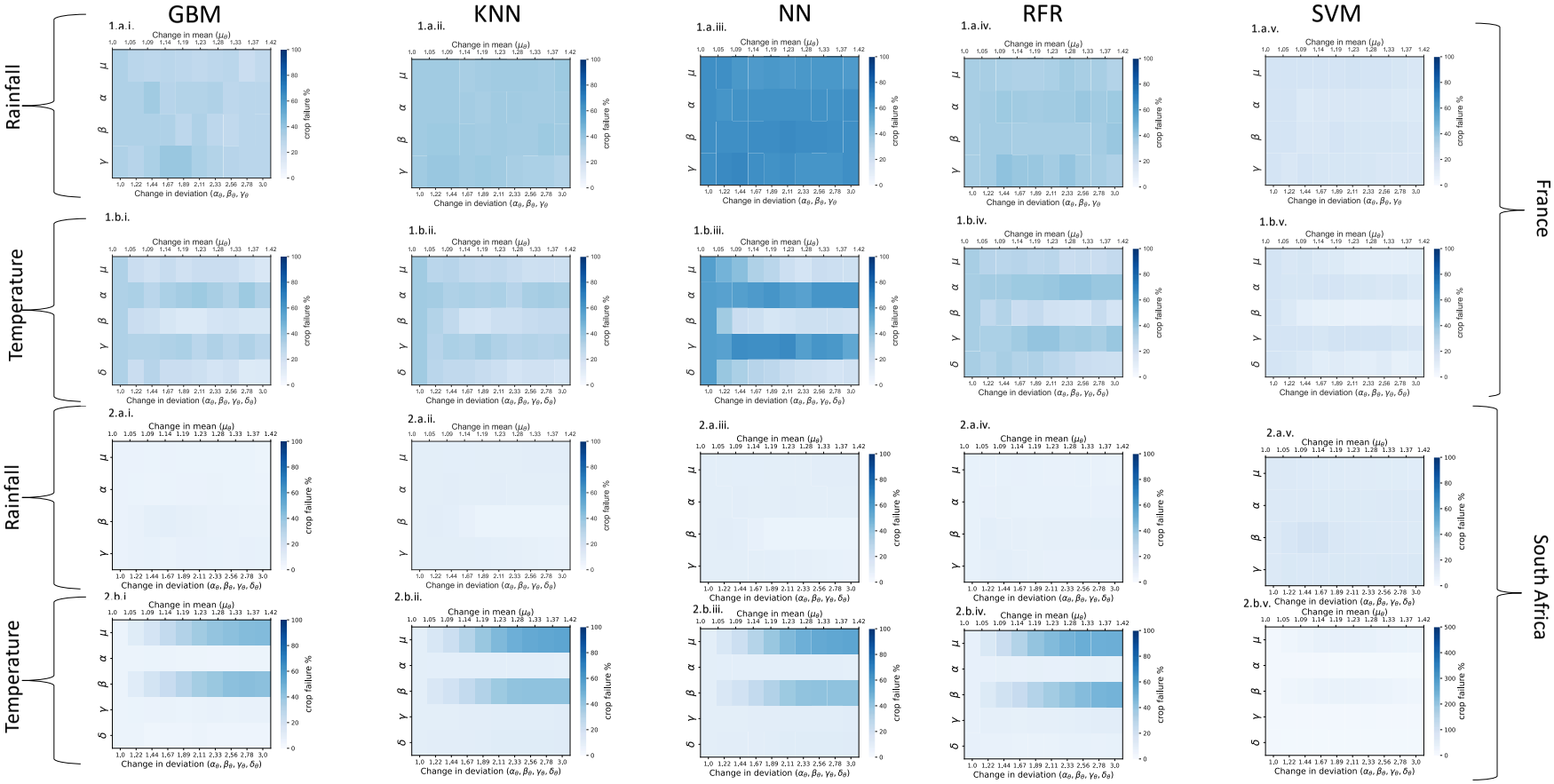

Figure 4.21: Correctly predicted crop failures as a percentage of the total number of observed crop failures across models, the 2 environments (France and South Africa) with changes to temperature and rainfall respectively. Model acronyms are GBM: gradient boosting machine, KNN: K-nearest neighbours, NN: Neural network, RFR: Random forest, SVM: Support vector machine. For each panel, number 1 or 2 denotes results for France or South Africa respectively, a or b denotes rainfall or temperature perturbations respectively, and numerals i to v represent each of the machine learning models tested ranging from GBM to SVM, left to right in alphabetical order.

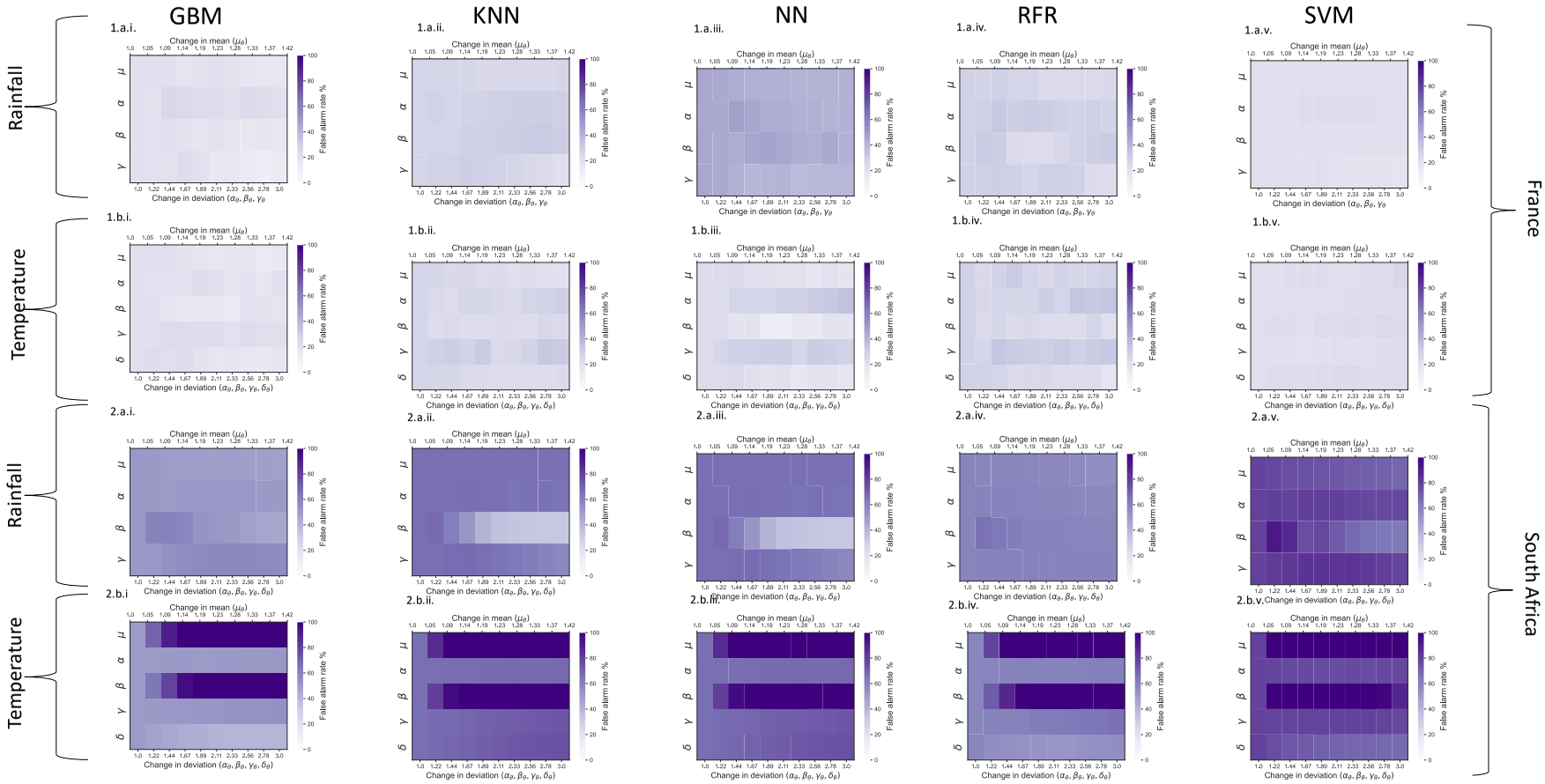

Figure 4.22: Falsely predicted crop failures as a percentage of the total number of observed crop failures across models (False alarm rate), the 2 environments (France and South Africa) with changes to temperature and rainfall respectively. Model acronyms are GBM: gradient boosting machine, KNN: K-nearest neighbours, NN: Neural network, RFR: Random forest, SVM: Support vector machine. For each panel, number 1 or 2 denotes results for France or South Africa respectively, a or b denotes rainfall or temperature perturbations respectively, and numerals i to v represent each of the machine learning models tested ranging from GBM to SVM, left to right in alphabetical order.

4.4 Discussion

Results are split into baseline results, crop failure prediction and the effects of the simulated errors on model predictions. The discussion of each of these topics follows the same structure.

4.4.1 Reasons for Machine learning performance against contrasting environments

Overall, it is clear that model performance varies across datasets. In particular, although neural networks were more effective at prediction of crop failures in France, model performance in South Africa is weaker in comparison to other models, particularly the support vector machine. Fundamentally, crop climate relationships differ across the 2 contrasting dataset used, and so it is not surprising that a model which performs best in one dataset performs worst in the other by comparison. Analysis of the inter-annual relationships between temperature, rainfall and crop yield (Figure 4.2 shows that the rainfall relationship with crop yield is more positive and the temperature relationship with crop yield is more negative in France. This means that there is a stronger signal between climatology and crop yield anomalies across inter-annual timescales. Spatially however, there is a stronger relationship between rainfall and crop yield in South Africa. This indicates that all ML models are better at capturing spatial relationships than temporal relationships, this is most the case for the SVM model. Neural networks trained exclusively on either spatial or temporal crop yield data have been shown to favour training on spatial relationships over temporal relationships, in fact training purely on temporal data causes model performance to be more inconsistent and so dependent on the dataset used (Guo et al. 2014). This is most likely the reason why ML models perform better overall in South Africa. However, the same study, also showed that although spatial training resulted in a more consistent model, adequate model performance only applied for normal or average years

(Guo et al. 2014). Therefore, This goes some way to explaining why although overall model performance is greater in South Africa, this does not translate to better prediction of crop failures.

Due to the greater spatial relationships in the South Africa dataset, this could mean that the SVM model is more easily able to select support vectors which are representative of the overall dataset, whereas in France, weaker correlations across space produce more inconsistent training data. Therefore, since the SVM model trains on a small subsection of the data, this means it is more likely to produce results unrepresentative of the larger overall dataset (Deka et al. 2014). In France, Neural networks correctly predicted more crop failures than other models. Predicting crop failures is essentially a test of extrapolation, (due to the effects of heat waves etc) so the best models in this respect are the best at extrapolation. Typically, many machine learning Models such as SVM (Deka et al. 2014) Random Forest (Mendez & Lohr 2011) neural networks (Nguyen et al. 2023) are known to be particularly poor at extrapolation and the prediction of extremes. Although crop failures in France were better predicted by the neural network, this was not the case in South Africa, indicating that the performance of the model is more inconsistent across datasets than the tree based models, (Random forest and gradient boosting models). Overall, model performance indicates that the random forest, KNN and gradient boosting models may be the best models for crop yield prediction as they are the most consistent across datasets. The SVM model is quite different in structure to these models and so in future studies it may be useful to include this model as well to provide a contrasting model type which is relatively easy to run and can produce results of sufficient model performance.

4.4.2 How does input data uncertainty affect machine learning model performance

Uncertainty resulting from the perturbations was greater in South Africa even though the pattern of the effects was the same. Therefore, the models are more sensitive to climate for the SAM dataset. Consistently across both countries, it was found that increases in mean temperature affects model RMSE more so than mean changes in rainfall. For rainfall, perturbations increasing the inter-annual variability had a greater effect. This will affect both the number of high rainfall days, as well as the number of days with very little rainfall. Therefore, because this is the dominant mode of increase in RMSE, this shows that the extremes of rainfall have a disproportionate effect on model predictions. This is more so the case for rainfall than for temperature. The implications of this are that extremes in rainfall are predicted to have a greater effect on modelled yield than extremes in temperature. In reality this will depend on the limiting conditions to growth which occur with geographic spread. For example, in France extreme rainfall effects due to drought have declined in importance in recent years due to the increased use of irrigation (Hawkins et al. 2013). In South Africa, the west of the country is typically drier and crop growth is therefore limited by water availability, this is less the case in the east of the country (Sacks et al. 2010). Therefore, the models generally agree with expected conditions in reality as extreme rainfall event conditions have a greater effect in South Africa.

Figure 4.20 shows that models respond differently to the increases in rainfall and temperature perturbations. In some cases, increasing temperature and rainfall perturbations increased the standard deviation of the outputs, whereas in other cases model output standard deviation decreased. Fundamentally, the effect of increasing the standard deviation in the input data will depend on the parameter interactions within the models. If there is limited parameter interaction, it may be the case that an increase in standard deviation

of the inputs results in an increase in standard deviation of the outputs. However certain parameters may affect the model's ability to predict an increased standard deviation from an increased standard deviation in the inputs. For example, if a parameter exists which decreases yields above a certain temperature threshold such as maximum cardinal temperatures in the GLAM model (Challinor et al. 2004), depending on the strength of the effect of such a parameter, this will decrease the standard deviation of the outputs. In the perturbation scheme, increases in the β parameter increase the number of days with both very high and very low temperatures, this affects yields through the threshold input feature which is the number of days above 32 degrees, this can dampen yield variability through the reduction in yield.

Crop failures are poorly predicted in both datasets by machine learning models. However, Figure 4.21 indicates poor model performance between the countries may be for different reasons. In France, if the magnitude of input perturbations in temperature are increased in both the mean, daily and inter-annual variability the number of correctly predicted crop failures further decreases. Decreases are more substantial for models which are better at predicting crop failures in the baseline simulations. This is in contrast to the effect in the South African maize dataset. In general the percentage of correctly predicted crop failures actually increases with increases in mean temperature and temperature variability. The reasons for this difference between the datasets is ultimately due to what the true and modelled mechanisms are for crop failure, and whether these are in fact the same. Because as extreme temperatures increase, so do crop failures in South Africa (both correctly and incorrectly predicted), the modelled mechanism for crop failure is extreme temperatures. However, since crop failures decrease in France, high temperatures are not the modelled mechanism of crop failure. In fact, increases in mean and temperature variability result in yield increases, meaning that all machine learning models have determined that the modelled mechanism for crop failure is extreme low temperatures. However, in reality,

although low temperatures can affect crop yield to a significant extent, particularly due to growth limitation effects or frost damage (Challinor et al. 2004, Barlow et al. 2015, Monfreda et al. 2008), considering models perform worst in the year 2003 (Figure 3.6 in chapter 3 demonstrates this) which was a year of significant heat extremes, causing reductions in primary productivity across Europe (Ciais et al. 2005) this is unlikely to be the significant driving mechanism of crop failures in France. This goes some way to explain poor model performance in the country. By contrast, because increases in temperature in South Africa also increase the predicted instances of crop failures it is more likely to be the true mechanism contributing to crop failure in the country. This goes some way to explaining why models achieve better RMSE and correlation coefficient scores in South Africa. Therefore, it may be argued that to improve machine learning model performance in France, a further parameterization of heat stress effects is required. This could be either a new feature which extracts such information such as a parameterization of thermal time or a measure of error for increases in temperature as an input feature.

Furthermore, Across both datasets, uncertainty in rainfall has far less effect on crop failure prediction than temperature. For this reason, the models agree that crop failures are more driven by temperature effects than rainfall effects. For France, this coincides with conclusions drawn by Hawkins et al. (2013) in which the relative effect of high temperatures are shown to overtake the effect of droughts around the year 2000. Since the test data used for this analysis is from 2003 - 2007 this agrees that temperature has the dominant effect. Another possible reason for these disparities in importance is that both countries are majority limited by temperature rather than rainfall. According to Monfreda et al. (2008) maize crops in both France and South Africa are temperature limited overall (although in South Africa there is a significant proportion of the country to the west which is rainfall limited). If rainfall is not a limiting factor, increasing mean rainfall will not have a significant effect on crop growth. This is shown in the results, as for both countries,

changes in rainfall do not significantly affect crop yield failure prediction rate.

When comparing Figures 4.19 and 4.21 it is made evident that although increases in rainfall uncertainty can affect overall model performance (Figure 4.19) this does not necessarily mean changes in model performance at the extremes. This contrast is particularly noteworthy in South Africa, in which large increases in RMSE with rainfall perturbation magnitude do not coincide with increases or decreases in the number of correctly predicted crop failures. This could be due to the limiting conditions presented by Monfreda et al. (2008), which may cause extremes to be less affected by changes in rainfall.

4.4.3 ML models and extrapolation

Results suggest that neural networks may be more likely to extrapolate future yields than the other models tested. This may be the case because this model architecture shows greater increases in standard deviation with perturbations in France (Figure 4.20). However, this is not also seen in South Africa. The prediction of crop failures is also most frequently predicted than other models with increased temperature perturbations. Across models, the SVM and GBM models show the least change (either positive or negative) with increases in temperature and rainfall. This shows that although they may be robust models for present day climates, future changes in temperature and rainfall as a result of novel climates may be mis-represented the most if crop yield change is projected with these models. Some models lack a strong response to particularly large increases in rainfall and temperature (e.g. SVM and GBM models in France) The reason for this could be that these models struggle to predict extreme values outside of the initial training range. This is more likely to be the case with the SVM model as training occurs on smaller subsets of the data (the support vectors).

4.4.4 Differences between FMA and SAM datasets

Fundamentally the French maize (FMA) and South African Maize (SAM) datasets are very different. This is because the French maize dataset has been compiled using census statistics of crop yields per administrative region (or department as they are known in France) and the South African Maize dataset has been compiled using remotely sensed data to downscale country level statistics with growing area data. Although there are uncertainties associated with both datasets, the FMA dataset could be considered closer to truly observed data, whereas the SAM data has undergone a greater degree of processing to arrive at modelled estimates of crop yields. The most important limitation of the SAM dataset is the use of static growing area data which does not change over time. This could mean that the spatial variability of yield across south Africa across time may not be as realistic as would be found with observed data. Although this is a limitation of the data, crop specific growing area is very difficult to source and so will be a common limitation of remotely sensed data sources.

It is shown that model responses are typically stronger for the SAM data set than for the FMA dataset. Of possible reasons for this, one of the most likely is that relationships between weather and yield are stronger and so the models are able to explain more of the variation in yield due to the weather (which is subsequently perturbed leading to a stronger response to the perturbation). In South Africa, there is a stronger variation between wet and dry climates than in France, as well as less widespread irrigation. Typically, the west (and more strongly north west) is very dry with a strong rainfall gradient between this region and the south east which is more suitable for growing maize crops. Hence due to this spatial relationship which is more easily captured by the models, models are more sensitive to perturbations in South Africa. However, it is important to note that because greater model skill in South Africa is achieved due to the more pronounced spatial relationship between rainfall and yields, because the South Africa data set is modelled data, this could

be the cause of why model performance is improved over the FMA dataset. As discussed in section 4.2.1 spatial down-scaling is achieved using estimates of NPP and leaf area index derived from remote sensing. The relationship between NPP and rainfall is very likely different (and as shown here, likely more sensitive) to that of rainfall and crop yield. Therefore, models trained on the FMA dataset are more likely to be a reflection of general ML model performance for crop yield prediction. A useful future comparison would be a field scale comparison against the other two data sets. Field scale crop yield data with a nearby weather station would have the least uncertainty in comparison to other data types. However, this too would bring comparison difficulties as different factors often govern field scale and regional scale crop-climate relationships (Ewert et al. 2011).

4.4.5 The robustness of machine learning algorithms to input data uncertainty and the value of comparison and combination of models

Robustness in the context of this chapter is defined as the ability of machine learning models to effectively predict crop yields and crop yield failures regardless of input uncertainty. Robustness should also mean the ability of a model to predict out of sample. Many, such as Hendrycks et al. (2019) define robustness as learning in the presence of corrupted labels. But here, it is shown that even with a pre-trained model which is not subject to the same data corruption, model evaluation can lead to different conclusions depending on the model used, dataset, and prominence of temperature and rainfall uncertainty in the test data. In this respect, GBM and RFR (gradient boosting and random forest) models showed the least change across model input uncertainty across evaluation against the 2 datasets. However, considering in most cases, machine learning models exhibited the same behaviour but to different magnitudes this suggests that a combination of models is best used such as a weighted model ensemble to account for the range of sensitivities across model perturbations.

As shown by the results in this chapter, magnitude of sensitivity to rainfall and temperature perturbations varies not just between models but also between datasets. As such, the most sensitive model in one dataset may not be the most sensitive in another. Therefore, as well as improved performance over individual models (Shahhosseini et al. 2020) combining models into weighted model ensembles may also be a valuable method to address the varying sensitivities of models to different temperature and rainfall errors which will have a varying effect depending on the specifics of the dataset. This statement broadly agrees with the results of Pham & Olafsson (2019), who showed that weighted ensembles are able to outperform other methods in the presence of spurious data. The results in this chapter give rise to the question of whether this statement holds true for different timescales of input uncertainty. Given the differences between models from different datasets, (i.e. the model which is most sensitive and achieves the best performance metrics is different) this is likely to be the case. Ensemble learning methods may provide further advantage by improving the ability to predict extreme events, some studies have shown that training each model in an ensemble on a different balanced subset of the data can reduce the bias against minority classes in unbalanced datasets (Sagi & Rokach 2018). Therefore, further work should assess the effectiveness of the models evaluated in this chapter against the robustness of weighted machine learning ensembles for prediction of the effects of extreme events.

4.4.6 How changes in training procedure may affect results of ML models

Here ML models were trained to predict yield as a continuous variable and then post-processing sorted simulated and observed yields into either crop failure or no crop failure. Although some preliminary work did explore training models to classify crop failures instead of predict yield value, the decision to train models for regression was made for three key reasons. Firstly, it is important to compare the models to the existing benchmark tool

for predicting crop failure (GLAM) in order to provide context for the results. Therefore, models were trained for regression for a comparison. Secondly, in order to have confidence in model results general model performance was assessed or the same models. This required the comparison of yield prediction between GLAM and ML models trained for regression. Thirdly, a key aim of the analysis was to show the effects of extrapolation of weather data data outside of the range of training data. The effect on yield predictions was easier to compare with conventional process based models when using models for regression.

Certainly however, training the ML models for classification should be an avenue for potential future work along with a comparison of sampling strategies. Some changes in the training procedure may improve the ML models ability to predict crop failures. This may include sampling for extreme events and training models for classification whilst using a loss function designed to penalise incorrect classifications of the minority class more strongly. However, there is a balance to be struck between training for a specific task and the bias introduced by doing so against general yield prediction. This balance raises the question of whether a fair comparison between ML and process based crop modelling is ever even possible because although ML may be able to perform specific tasks related to crop modelling better (e.g. predict yield variability, or predict mean yield or predict the spatial variability in yield), to perform some tasks better than crop models may require a ML model trained in a completely different way. For example, although the ML models in this thesis tend to predict yields better than GLAM overall, it may require models trained using different sampling methods in order to predict crop failures better, however this may also introduce bias and so make the current models worse at yield prediction.

4.4.7 Broader implications of the effects of uncertainty on ML model predictions

As discussed in section 4.4.2, rainfall and temperature uncertainty have contrasting effects both at different timescales and across environments. The broader implications for this means that uncertainty from climate models will have differing effects on both spatial and temporal scales. The method of the perturbations used here is to introduce uncertainty by overemphasising climate extremes. It has been shown that in the present day climate, CMIP6 models exaggerate the magnitude of daily temperature anomalies and under or over estimate precipitation depending on region (Di Luca et al. 2020, Dong & Dong 2021). Models also tend to overestimate evapotranspiration in both France and South Africa (Wang et al. 2021) which will have cause exaggeration of drought. These deficiencies are likely to increase for future scenarios with increasing uncertainty in the future. If so, this too will have knock on effects for ML models used to determine effects of future changes on yields.

Uncertainty from climate models will likely exacerbate uncertainty in impacts projections when using ML models, with greater uncertainty given that magnitude of effects varies depending on ML framework chosen. The larger the magnitude of uncertainty the more the uncertainty in the output will be compounded by the models. Similarly, uncertainty in the frequency of heatwaves on the inter-annual time scale will also affect models to a considerable degree. Given this, and given the differences between models, if machine learning methods are to be used to forecast future impacts with climate model inputs, this should only be attempted with the use of an ensemble of ML models of varying complexity. Using an ensemble is the best way of accounting for the differing sensitivity of model responses to perturbation induced uncertainty. If only one ML model is used, uncertainty may be very different depending on which model is chosen.

The results of the input data perturbations are designed to be broadly comparable with that of (Watson et al. 2015) who used a similar method to determine relative effects of temperature and rainfall errors on a statistical regression model and the process based crop growth model GLAM, also used in this study. From this study, both GLAM and the statistical model are affected most strongly by perturbations of the mean and inter-annual variability of temperature and are less affected by errors in precipitation. Interestingly, although GLAM uses daily input data in a more explicit way, pereturbations in daily variability are not as significant as mean and inter-annual perturbations. However, errors in daily variability have been shown to have a greater effect on model performance if the process parameterization is more complex. For example, there is a greater difference between model results for models using Farquhar method of photosynthesis calculation and/or detailed parameterization of LAI than a simpler aggregated LAI parameterization when using aggregated instead of daily temperature data (Van Bussel et al. 2011). Therefore, machine learning models are broadly affected by temperature and rainfall uncertainty in similar ways to the GLAM process based crop model. It would be interesting to determine whether these effects are consistent across other process based crop models which may not be semi-empirical and could utilise more complex methods such as the Farquhar photosynthesis approach.

4.5 Conclusions

Across environments, uncertainty in temperature affects model performance more substantially than uncertainty in rainfall. This is especially so for crop failures, with the proportion of correct failure predictions more substantially affected by perturbations in the mean and inter-annual variability of temperature than rainfall. At smaller, daily time scales, the effects of temperature uncertainty may vary more by environment (or dataset). A limitation of this exercise may be that both South Africa and France are overall temperature limited environments, and so this may be the cause of why errors in rainfall

have comparatively little effect. However, South Africa does contain some regions which are rainfall limited to the west, this may contribute to why RMSE increases are more pronounced with increases in rainfall in South Africa.

Uncertainty in input data can have significant effects which vary by the model framework chosen, timescale, as well if uncertainty is in temperature or rainfall. These results have implications for model extrapolation, both for future projections and to predict the effects of extremes when using models for forecasting. Given projected future increases in extremes (Porter & Semenov 2005) ML models require further development in order to be used reliably to predict future crop failures. For long term projections, given uncertainty in climate models, the use of a single machine learning approach cannot be recommended to project future changes in yields. This is because uncertainty affects different models which achieve similar baseline performance to different magnitudes, meaning that there is greater uncertainty added purely based on which ML model has been chosen. Therefore, a method to remediate this is to use an ensemble of machine learning methods which can account for the uncertainty from ML model choice. Machine learning methods also do not account for CO_2 fertilization effects, which has the potential to mediate yield decline (Lobell & Field 2008b). Therefore, future work should seek to integrate machine learning with process knowledge as well as incorporate the use of a range of ML models to mitigate the uncertainty associated with ML model choice.

4.5.1 Novel contributions from this chapter

From this chapter, the following contributions are made to the scientific literature:

- The robustness of different ML models to uncertainty in test data is presented, uncertainty effects are shown to vary across models, timescales and environments
- Machine learning model performance is shown to be very different for crop failure

prediction than for crop yield simulations

The analysis in this chapter is useful for many applications in which machine learning is used to determine impacts of future climate change. In particular, if extremes of heat and rainfall are underestimated (represented by the β perturbation) in future climate model simulations, this will affect different ML models to different degrees. This has important implications for model selection in many applications such as future crop yield estimates (Leng & Hall 2020, Feng et al. 2019) among other fields and applications (Park & Lee 2021, Jung et al. 2021, Nwokolo et al. 2023).

5 Machine learning and crop model bench-marking to improve yield predictions

5.1 Introduction

Crop models have been used extensively to provide climate impact projections (Jägermeyr et al. 2021, Rosenzweig et al. 2014) and inform adaptation to climate change at regional, national, as well as field scales (Asseng et al. 2019, Challinor et al. 2009, 2018, 2016a, 2015, Tariq et al. 2018, Reidsma et al. 2010). However, to effectively inform adaptation to climate change, crop model predictions must be robust in the current climate to ensure confidence in future projections and recommendations. As shown in chapter 3 and chapter 4 of this thesis as well as Leng & Hall (2020), machine learning models are able to outperform process based crop models in a range of environments by achieving greater overall performance through improved generalization. This strength of machine learning is actually a weakness of crop modelling, as more location specific information is required to improve calibration which in turn reduces generalization (Angulo et al. 2013). Machine learning can be used to improve crop models through either assistance in calibration, improvement of parameterizations or the use of machine learning to benchmark and evaluate existing crop model parameterizations and calibrations to identify which areas to focus on for future model improvement and development.

Calibration is the process of estimating model parameter values to reduce the error between model results and measured data (Wallach et al. 2021). Parameters are never precisely known and so calibration only ever approximates the true values. This is especially the case for parameters which are difficult to estimate, such as root zone penetration (Wallach 2011) or parameters which lack observed reference data at large spatial scales (Müller et al. 2017). Because crop models consist of several smaller coupled process models (e.g. model for predicting evapotranspiration, model for predicting change in biomass etc) since

the parameter values for the process models are not known, parameters from each process model are used in error to some degree. Since there is substantial uncertainty in the value of the parameters from the individual process models, model error aggregates, leading to model misspecification (Wallach 2011). Calibration therefore is a complex task which often does not have a single agreed upon correct answer (Wallach et al. 2021).

Due to the discussed problems arising from parameter uncertainty, crop model calibration does not have a consistent methodology. Wallach et al. (2021) showed how a variety of outcomes result from the large number of decisions which are required to calibrate crop models. The results of this study showed how even with the same model structure, modelling groups came to different decisions when calibrating and therefore model results were also different. Decisions which are required when calibrating a crop model include: which variables to include in the calibration process, whether to estimate all parameters together, or group parameters based on effect, measures of error (e.g. crop yield, development stage etc), choice of objective function (e.g. sum of squared errors, weighted sum of squares etc) or whether to use a frequentist or Bayesian approach to optimization. The vast majority of modelling groups used expert knowledge to choose which parameters to estimate. These decisions lead to different results even with the same model structure.

As with the other chapters in this thesis, this chapter focuses on the regional scale approach. This is as opposed to the field scale approach in which calibration data is all located at a field site. When scaling up predictions from fields to farms to regions which could potentially contain many farms with a diversity of farming practices, due to the integration of complexity when increasing spatial and temporal scale, calibration brings with it a different set of difficulties at larger spatial scales (Ewert et al. 2011, Challinor et al. 2018). This is partly because fundamentally, models designed to model crop yields at the field scale are used at the larger gridded scale (Challinor et al. 2018). This can lead to further model bias depending on the spatial variability of weather and soil (Hoffmann

et al. 2016). Therefore, although the same question of how to calibrate a crop model presents itself at large scales, there are additional challenges including lack of observed reference data (Müller et al. 2017), uncertainty from choice of data aggregation method (Ewert et al. 2011) and degree of spatial heterogeneity (Hoffmann et al. 2016, Angulo et al. 2013). Furthermore, spatial aggregation of weather data can affect its temporal properties depending on climate, orography and the variable in question (Rajulapati et al. 2021). Most importantly, extreme values decrease in frequency with increasing data aggregation (Rajulapati et al. 2021). This is particularly important considering crop models have difficulty capturing the effects of extreme events (Schewe et al. 2019). Unsurprisingly, the greatest differences between aggregated and fine scale resolution data is found for rainfall data (Hoffmann et al. 2017). Such problems add up to cause the evaluation of crop model skill to be more difficult at larger spatial scales (Challinor et al. 2018).

Improving parameterizations is a problem intrinsically related to improvement of calibration methodologies. Machine learning may be especially useful for the improvement of empirical parameterizations. For example, as mentioned previously, the GLAM crop model is a semi-empirical process based model. This means that some sub-processes within the model were developed according to measured relationships between observed data rather than based on theory or mechanistic understanding. For example, the effect of changing CO_2 on the growth rate of crops is accounted for by adjusting the transpiration efficiency and maximum value of potential transpiration parameter so that yield, biomass and transpiration response ratios are within the range of values determined by Free air CO_2 enrichment (FACE) experiments detailed by Kimball (2016). A second example is that the effects of sub-optimal management are considered by adjusting the yield gap parameter, against the spatial variability of observed crop yields. Furthermore, the determination of soil moisture characteristics (the drained upper limit, lower limit and saturation limit of the soil) are determined by a set of statistical correlations between

soil texture, and structure with water potential and hydraulic conductivity (Saxton & Rawls 2006). In fact, although process based, other crop models also rely on empirical relationships such as that between soil moisture and texture characteristics (e.g. LPJML, Aquacrop) (Lutz et al. 2019, Raes et al. 2009).

Although a relatively new concept, machine learning parameterizations of crop model sub-processes or state variables has been shown to provide great value where used appropriately. For example, Droutsas et al. (2022) have integrated a random forest and extreme gradient boosting machine learning methods into the GLAM-parti crop model (Droutsas et al. 2019). The method used daily weather input variables: minimum and maximum temperature, solar radiation, vapour pressure deficit, thermal time, and the ratio of photosynthetic organ mass to total above ground biomass to predict response variables of growth stage, radiation use efficiency, and change in harvest index per daily time step. The integrated model was able to significantly improve model performance against biomass, yield as well as anthesis and maturity date response variables.

Bench-marking using machine learning methods is a method to evaluate crop models by determining the reasons why model performance may vary across spatial variations in climate and management. In doing so, the efficacy of crop model calibrations and parameterizations can be tested. If machine learning methods out perform process based crop models for a given dataset, it is important to ask why this is the case and if improved calibrations or parameterizations can be targeted to improve crop model performance based on what machine learning can reveal about the given dataset and the relationships between both observed and modelled yields and the climatological data used for predictions. For example, Across spatial domains, bench-marking can be used to determine the spatial specificity of calibration, in particular, what is the spatial scale at which to adjust potential yields to the given observed values by accounting for non-optimal management, and what is the most appropriate spatial scale to determine the extent of simulated varieties

(i.e. how do you represent which cultivars are likely to be grown where?).

Furthermore, bench-marking is also used to explore the ways in which information from GLAM may be used to improve machine learning predictions. One way to do this is to determine which particular variables from GLAM may be important for model predictions. Information from GLAM which may be useful for machine learning could include planting dates and estimated crop duration, and information on heat stress parameterizations such as temperature at which crops are killed by lethal temperatures.

5.2 Research questions and aims

Through the informing of calibration and parameterization, bench-marking can lead to improvements in understanding of how processes are represented and the way in which they could be represented better in crop models. The following questions are posed by the work in this chapter:

1. How do crop models and ML represent climate-yield relationships differently?

2. How can process based models be used to improve machine learning predictions?

3. Can insights from machine learning be used to inform crop model improvement?

To enable integration and comparison between machine learning methods and process based crop models, it is important to assess if both methods use climate information in the same way. This is for 2 reasons, firstly if machine learning and crop models predict the same outcomes, are the predictions given due to the same reasons? If so this suggests that relationships identified are robust across methods. Conversely, if machine learning methods predict different responses to the process based model, and the model performance of the machine learning model is better, this suggests that there are relationships learned by the machine learning methods which are not fully utilized by the process based model,

which enable improved performance. This method of bench-marking acts as a comparison between not just machine learning and mechanistic models, but also different architectures of machine learning models, as climate - yield responses may not be the same across different model architectures (Lischeid et al. 2022). This is why modelled yield responses are compared against different machine learning architectures, namely, random forest, support vector regression and a multiple linear regression (for baseline linear comparison).

Furthermore, another reason for bench-marking is to improve the predictions from machine learning methods through the use of GLAM. Incorporation of information from GLAM into machine learning is not just useful for improving machine learning model performance but can also be used as an independent model which determines the effectiveness of GLAM predictions based on further agro-climatic information. For instance, if biomass estimates from GLAM lead to better ML predictions in one country than another, then it could be inferred that biomass GLAM predictions are also more skillful where this is the case. Furthermore, if the soil water stress factor is more important in one country than another, this could indicate that observed crop yield response is more water limited in that region. Machine learning is useful for this purpose as it provides an independently constructed model to cross refer the importance of such variables. Machine learning methods with process based information incorporated into the inputs are referred to in the following sections as hybrid ML. This is because outputs from a mechanistic model are used to construct the features.

Once the differences between yield responses to climate conditions across models are determined, this can then be used to determine the causes of differences between model performance. For instance, if the crop yield response in one country is more closely captured by a machine learning model than by GLAM, this could be due to calibration or parameterizations which are sub-optimal and so require additional development.

The results from this chapter lead into the analysis undertaken in appendix 9. This shows a method in which ML can be used to predict intermediate variables used as part of crop modelling. In this example soil moisture characteristics are predicted using soil texture information with a random forest model. The purpose of this test is to demonstrate potential improvements in the correlation between simulated yield and rainfall, and so model skill in rainfall limited environments.

5.3 Methods

In summary, the methodology of this chapter is to use machine learning methods to benchmark crop model performance under different climate conditions. From this an assessment can be made as to which conditions are associated with the strongest crop model errors and how could model performance be improved. Figure 5.1 shows the general methodological steps to this chapter. Step A is the step of training and evaluating machine learning methods. As part of this step, machine learning methods are trained and evaluated using different mechanistic inputs to answer the question of How can process based models be used to improve machine learning predictions (research question 2)? This is achieved using a combination of feature importance analysis, and principal component analysis (PCA) to distinguish highly correlated features. For more on this method see section 5.3.5. Once step A has been completed, Step B is the step of Comparison between machine learning and GLAM performance across different climate conditions. To do this 2 approaches are used, firstly agro-climatic variables are defined, then correlated against model performance across the best performing machine learning model and GLAM. Secondly, correlations across both space and time are used to determine if the drivers of modelled yield variability in response to spatial variability of climate are the same across both machine learning and mechanistic crop modelling methods. This will answer research questions 2 and 3 of this chapter. Step C is the step of identification of how to improve model performance for the conditions found in step B. This applies to both how to improve machine learning and

GLAM using information from the bench-marking undertaken in step B.

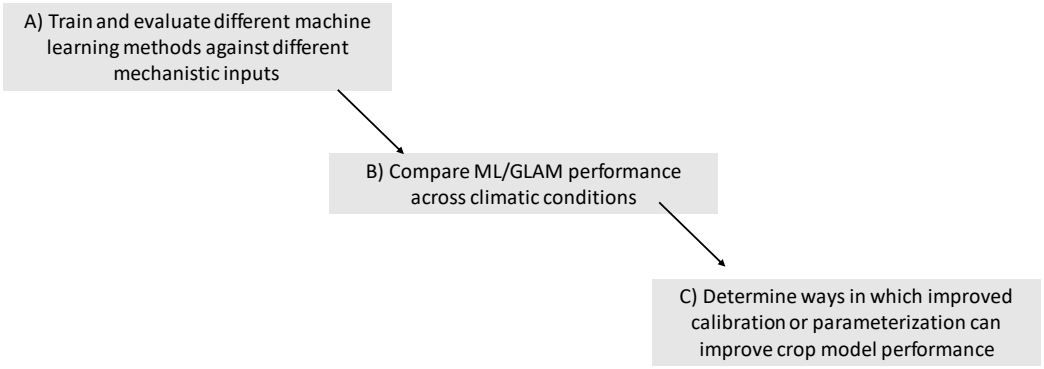

Figure 5.1: A summary of the methodological steps in this third chapter.

This chapter uses crop yield and reanalysis weather data (as well as soils data) for 4 countries across sub-Saharan Africa, namely Malawi, South Africa, Tanzania and Zambia. These four countries were chosen to build upon model simulations performed to inform future food policy found in Jennings et al. (2022). Both GLAM and machine learning methods are used to simulate crop yields across all four countries.

5.3.1 Model comparisons

To answer the research questions in section 5.2 multiple comparisons are made both between machine learning models and between machine learning models and crop modelling. Figure 5.2 summarises the model comparisons made in this chapter, with each comparison listed in Table 5.1. Comparisons are made between GLAM and ML across common test periods for each country. Test periods were chosen based on the most significant correlations between crop yield data from FAOSTAT (2022) and area weighted weather variables. This method to choose the training and testing split is chosen because some sections of the observed yield time series have very poor correlations with weather vari-

ables and therefore there is little signal for the model to compare to (Jennings et al. 2022). Common test periods were taken from Jennings et al. (2022). A separate comparison is made between ML models to show performance across various train-test splits in a cross validation process. ML models were tested across each 5 year window of the time series of data, with training sets comprised of each year left out of each 5 year window. This extra step was undertaken to ensure that ML model performance was not significantly different across the common test periods than the rest of the dataset (Figure 5.15).

Between machine learning models, the comparison is made between models trained on purely climatological inputs and machine learning models trained on both climatological inputs and information from the GLAM crop model. This comparison is made to answer research question 2. Machine learning inputs are listed in Table 5.2. Research question 2 is further addressed in section 5.4.2. Between machine learning methods and GLAM, comparisons are made for calibrations across 2 spatial scales (grid cell and country). The grid cell spatial calibration results from the calibration of the yield gap parameter to capture mean yields for each grid cell. The yield gap parameter is an empirical correction factor used to tune values of the leaf area index to mean observed yields described in section 2.1. The country spatial calibration uses 1 value of the yield gap parameter per country to capture mean yields on average across the country. Generally, mean yields are greater in South Africa than the other three countries, therefore, the country level calibration was chosen (instead of no calibration) to ensure more realistic crop yields per country without tuning yields to each particular grid cell. To provide an analogous comparison to the spatially varying calibration parameter used by GLAM, machine learning methods were trained both with and without latitude and longitude coordinates as inputs.

Climate and crop yield relationships are compared for both spatial and temporal relationships separately. This distinction is made for two reasons. Firstly, spatial and temporal relationships in crop - climate relationships are shown to be marginally different in the

data set used. Secondly, model comparisons focus both on how to better calibrate to capture mean yields (which requires assessment of spatial correlations) and model skill at capturing the inter-annual variability is also compared (which requires assessment of temporal correlations). Crop climate relationships are compared to answer research question 1. This question is therefore addressed in section 5.4.3 which is split into spatial correlations (section 5.4.3.2) and temporal correlations (section 5.4.3.1) separately.

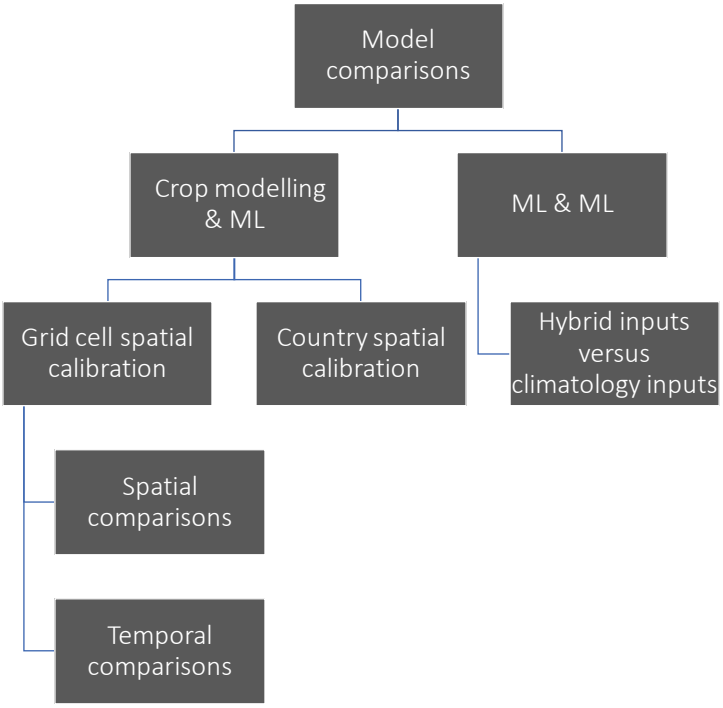

Figure 5.2: A summary of the model comparisons made in this chapter between both crop modelling (GLAM) and machine learning (ML) and between different ML models.

Comparisons between models are made with the view to answering research question 3 and so identify targets for crop model improvement. Crop model improvement is further explored in results sections 5.4.6. Section 5.4.6 was developed after the initial model comparisons made in the preceding sections. As a result of the comparisons made, questions arose as to whether GLAM is too insensitive to the effects of rainfall. Therefore, this sec-

tion shows the attempts to increase the sensitivity of GLAM to rainfall by changing model parameters and removing the effect of crop duration on yield. Firstly, it was test tested whether the effect of duration has a dampening effect on the relationship between rainfall and simulated yield. When this was discovered to not be the case, this lead to a further test to see if changing the method of calibration to one which instead of reducing LAI reduces the water holding capacity of the soil would improve this relationship. Methods of calibration are described in section 2.1.

Table 5.1:
Model comparisons made in this chapter with corresponding sections and Figures, RQ (Research question addressed) for each comparison is shown in the right most column.

Comparison	section	Key Figure	RQ
Crop modelling and ML general model comparison	5.4.1	5.16	3
ML model cross validation	5.4.1	5.15	3
Hybrid and climatology ML	5.4.2	5.16	2
Calibration comparison	5.4.4	5.32	1
Temporal weather-yield relationships	5.4.3.1	5.30	1
Spatial climate-yield relationships	5.4.3.2	5.28	1
Model skill and agro-climatic relationships	5.4.5	5.35	3
fixed duration model sensitivity	5.4.6	5.37	3
YGP calibration method	5.4.6	5.38	3

5.3.2 Data

The four countries chosen for this analysis span a range of climatological regimes, and levels of crop management which affects both the technological trend in the observed time series of crop yields, and the magnitude of yields. This affects the levels of calibration required to accurately represent yields under sub-optimal management. The analysis of climate-crop relationships requires a spatially explicit dataset and so crop yield data is taken from the dataset developed by Iizumi & Sakai (2020). The same crop yield data used for the SAM dataset in the previous chapter is also used in this chapter. The use of this dataset is expanded to 3 other countries, namely Malawi, Tanzania and Zambia. This dataset, called the GDHY dataset, is subject to the same uncertainties discussed in

section 4.4.4 of the previous chapter. For this region, in the absence of observed census statistic data, the GDHY data set was the best spatially explicit yield dataset available at time of analysis. However, some uncertainties associated with this dataset include lack of dynamic harvested area data (which is difficult to find) as well as correlations between NPP and weather. Yields are modelled yields rather than observed (either in the field or through census statistics). Hence, correlations with weather data are somewhat different to what they may be if yield data was compiled from census statistics. However, without this data also available to compare at the time of analysis, it is difficult to say what the true differences are. Potentially, modelled GDHY yields may be more sensitive to weather variability than census statistics. However, the relationship of truly observed yield and weather may not be so different to that of weather and the GDHY data given that remotely sensed indices have been shown to provide substantial explanatory power for yield estimation (Wall et al. 2008, Aranguren et al. 2020). Remotely sensed data has also been used as substitute for yield data for crop simulation models in other studies to some success (Moriondo et al. 2007, Kim & Kaluarachchi 2015).

In sub-Saharan Africa (SSA) food production and economic prosperity is dependent on weather sensitive agriculture (Rowhani et al. 2011, Jury 2002a, Stige et al. 2006, Msowoya et al. 2016). The four countries chosen for this chapter, Malawi South Africa, Tanzania, and Zambia all share similar dependence on weather variability for crop yield and productivity (Rowhani et al. 2011, Jury 2002a, Amadu et al. 2020, Msowoya et al. 2016, Mulenga et al. 2017), in particular, the lack of irrigation in the region leaves many small holder farmers vulnerable to climate shocks to crop yields due to insufficient rainfall (Rowhani et al. 2011, Jury 2002a, Amadu et al. 2020, Msowoya et al. 2016, Mulenga et al. 2017, Stige et al. 2006).

The GDHY dataset provides an estimate of the spatial extent of crop productivity across each of the four countries. South Africa has some substantial differences both in terms

of climatology and management from the other 3 countries used in the analysis. Most notable, is the substantial difference in yield levels between the most productive areas of South Africa and the other three countries, as well as the intensity, (in terms of fraction of area which is used to harvest crops) of the most intensely farmed regions. Harvested area data for the time period was taken from Sacks et al. (2010). Growing area from this data source was used to mask grid cells to determine growing regions for each country. Figure 5.3 demonstrates this through the distributions of harvested areas across each of the countries. Evidently, there is a large disparity between the most intensely harvested areas in South Africa and the least, with the distribution of harvested areas reflecting this with an exponential decline in the distribution of harvested areas from the smallest upwards. Although Tanzania, and Zambia also show similar distributions, the decline in number of locations with increasing harvested area is not as pronounced as in South Africa. Notably, this decline is the least pronounced in Malawi.

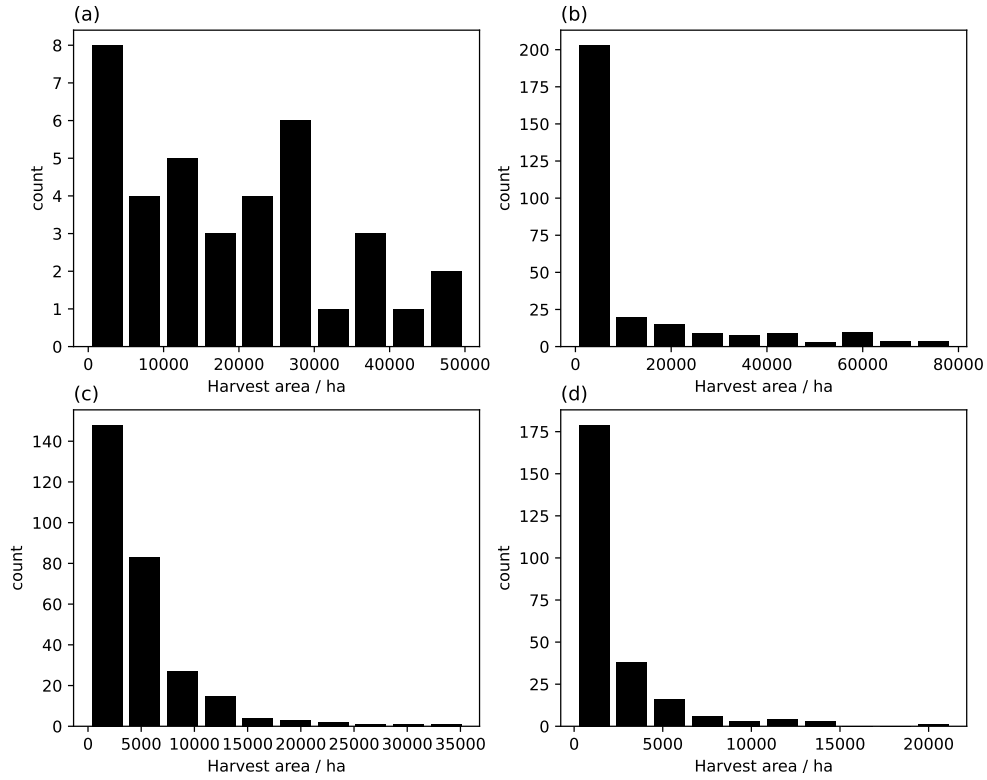

Figure 5.3: Distribution of harvest area across each of the countries for the year 2000 from Sacks et al. (2010) Countries are labelled according to panels (a.) Malawi, (b.) South Africa, (c.) Tanzania, (d.) Zambia.

The distribution of crop yields (Figure 5.4) interestingly does not show quite the same distributions across the four countries as harvested areas. Crop yields across locations and time in Zambia are normally normally distributed, however, the other 3 countries do not have normal distributions in yields. This suggests that there is little agricultural niche effect (Challinor et al. 2015). The absence of this effect would suggest that there is little relationship between average yield and harvested area. Figure 5.5 also suggests this. In South Africa in particular, the areas of highest crop yield are not the same as the the

areas of highest harvested area however there is some overlap between high harvested area locations and high yielding locations in Malawi.

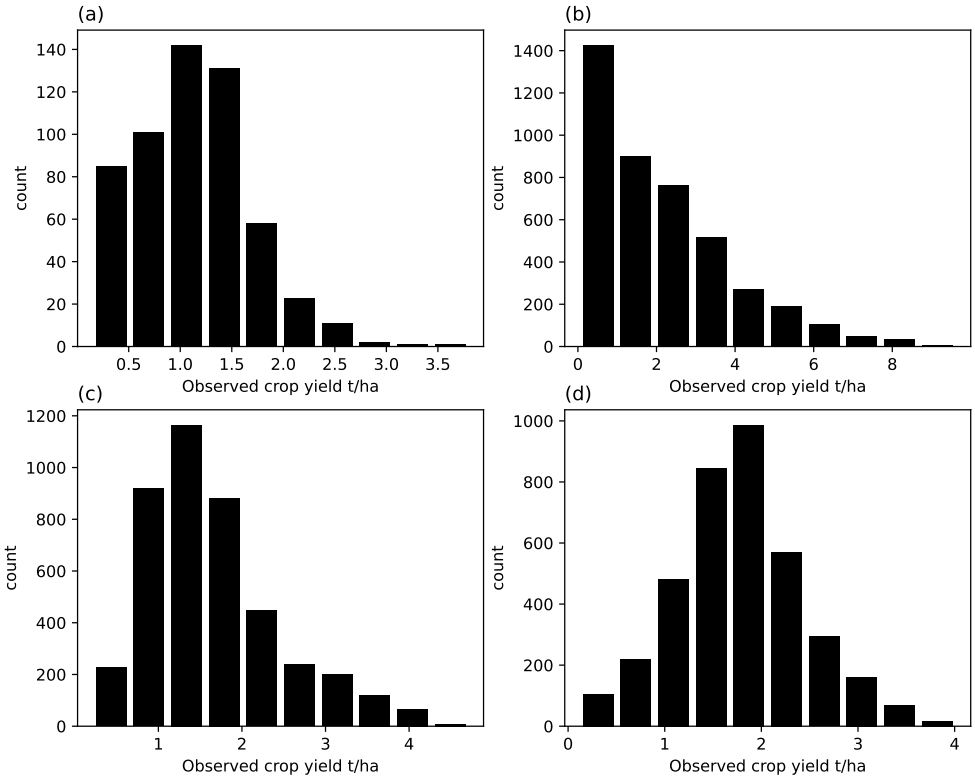

Figure 5.4: Distribution of the observed yield dataset produced by Iizumi & Sakai (2020) for each of the 4 countries. Countries are labelled according to panels (a.) Malawi, (b.) South Africa, (c.) Tanzania, (d.) Zambia.

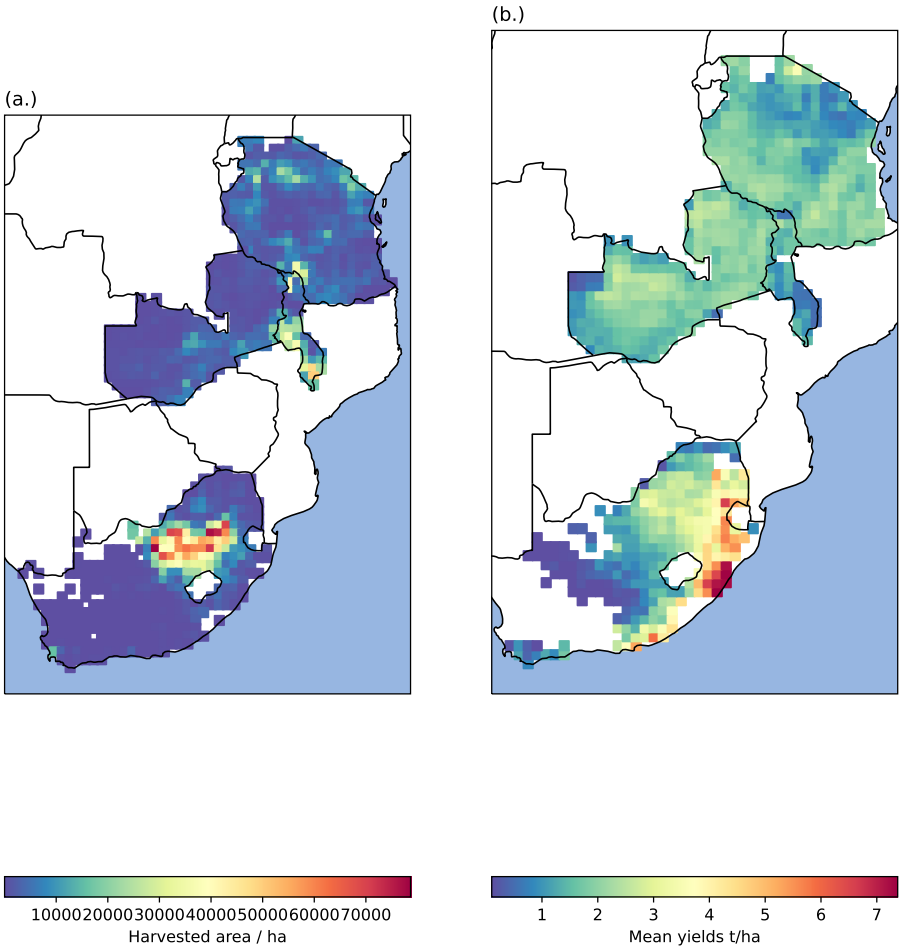

Figure 5.5: a.) Map of harvested area across each of the 4 countries for the year 2000 (b.) Average crop yield for the study test period 1990 - 2002.

In Malawi, maize, is cultivated by a majority of small holder farmers, accounting for 76% of small scale agricultural plots with an average size of 0.32 ha in 2014 and 2015 (Amadu et al. 2020). In terms of nutrition, Almost half of calorie intake in Malawi is met by the direct consumption of maize or maize products. Furthermore, over a third of the Gross domestic product (GDP) of Malawi is related to agricultural activity (Warnatzsch & Reay 2020) . Therefore Maize is a highly important crop for the country. The area of Malawi is 118,500 km^2, making it the smallest country of the 4 case study countries chosen for the dataset used in this chapter (Adhikari & Nejadhashemi 2016). Average yields are low in Malawi at 2.1 metric tons per hectare, lower than the average for African countries (3 t/ha) and much lower than globally important maize producer, the United States (11 t/ha) (Amadu et al. 2020).

The exponential distribution of crop yields in South Africa with location (Figures 5.5 and 5.4 may be due to policies which disproportionately favour larger farms over small holder farmers (Fischer & Hajdu 2015). In South Africa, average maize yields across the dataset are 2.29 t/ha making it the largest yielding country, followed by Zambia (1.83 t/ha), Tanzania (1.54 t/ha) then Malawi (1.27 t/ha). Whilst also having the highest average yields, South Africa also has the largest inter-quartile range between lowest and highest yielding areas (2.43 t/ha) and Tanzania has the lowest inter-quartile range between high and low yielding areas (0.61 t/ha), Malawi and Zambia have the second and third highest inter-quartile ranges between high and low yielding areas (0.79 t/ha and 0.66 t/ha respectively). Hence, South Africa is both the highest yielding country but also that which has the greatest spatial yield variability.

Ultimately, the GDHY dataset depends on the FAOstat agricultural statistics database. Figure 5.6 shows the yield anomaly time series for each country studied in this chapter. Yield anomaly is given as standard deviations from the historical mean of each time series. Of note, is that there appears to be a stronger correlation between the yield time series

in Malawi, South Africa and Zambia than in Tanzania. 1991/92 saw a large synchronous negative crop yield anomaly in the three aforementioned countries. The large positive yield anomaly in Tanzania between the years 2000-2005 appears unusual.

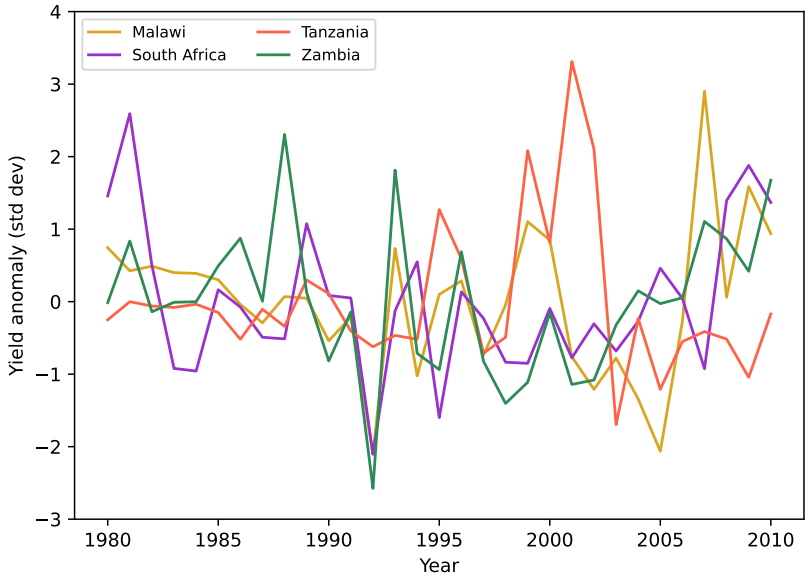

Figure 5.6: Anomaly from mean country level yield across the time series of each country taken from the FAOstat database.

Similar to Malawi, both Zambia and Tanzania rely heavily on rain-fed agriculture (Mulenga et al. 2017, Rowhani et al. 2011). However unlike the other three countries, Tanzania has 2 prominent rainfall regimes. In the north there is a bimodal precipitation regime with long rains occurring from between March and May and short rains experienced from October to December (Rowhani et al. 2011). Several studies have confirmed the influence of large scale climate events such as the El Niño Southern Oscillation (ENSO) and the North Atlantic oscillation on the east and south east African climate (Stige et al. 2006, Rowhani et al. 2011, Giannini et al. 2008). In Malawi, Jury & Mwafulirwa (2002b) have mapped correlation of the ENSO with rainfall indices and have found strong regional ex-

pression of ENSO. In Tanzania, El Niño years are associated with above average rainfall whereas la Niña is associated with below average rainfall, this is due to the changing of the onset of the rainy season, with an earlier onset during El Niño years and later onset during La Niña years (Kijazi & Reason 2005). Studies have also associated El Niño with reduced frequency of "wet spells" over Zambia, associated with flooding (Hachigonta & Reason 2006). The onset of the rainy season is normally during October or November although this is characterized by a large degree of inter-annual variability (Hachigonta et al. 2008). In South Africa, rainfall patterns may be linked to "ENSO like" decadal and multi-decadal patterns sea surface temperature and and circulation patterns (Reason & Rouault 2002). The western region of South Africa is generally drier than the east of the country. The western Cape province has a historical problem of severe droughts (Mahlalela et al. 2019).

In the EWEMBI dataset used, mean rainfall variability differ greatly across the four countries. Inter-annual variability in rainfall is on average greatest in Tanzania. Figure 5.7 shows rainfall anomalies across time, with the collection of grid cells each year making up each boxplot, the median across all locations is shown as a red dotted line. Of note, is the large rainfall anomaly in Tanzania in 1997 and substantial anomalies in 1991, 1999 and 2005 in Malawi. Figure 5.8 is in many ways the inverse of Figure 5.7, in that there are large positive anomalies of solar radiation in 1990 and 1991 in Malawi and a large negative anomaly in solar radiation in Tanzania in 1997. The analogous Figure pertaining to temperature (Figure 5.9), correlates less with rainfall variability.

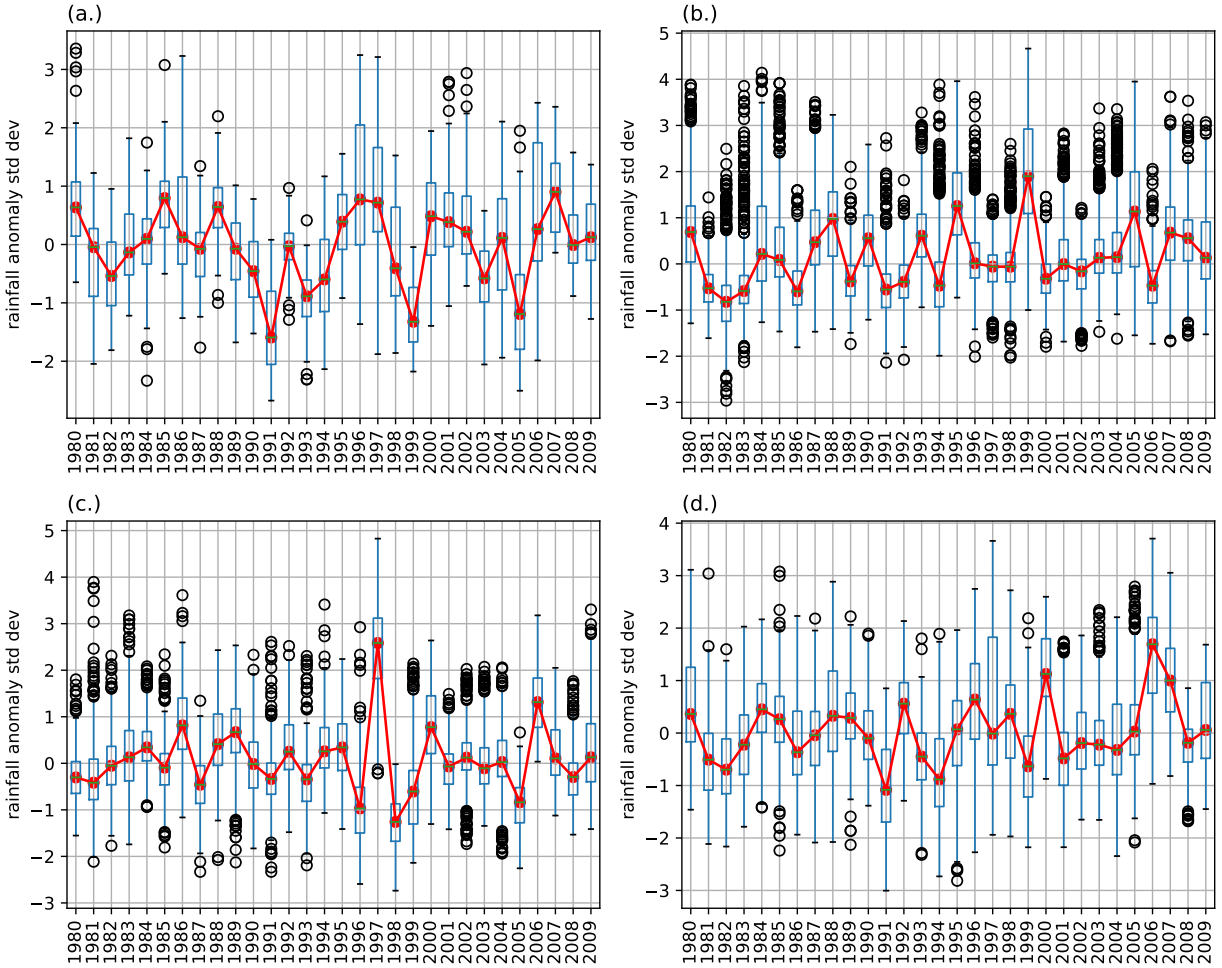

Figure 5.7: total Rainfall anomaly distribution every year during the growing season of December - April for each of the 4 countries. Anomaly is determined as standard deviations from the mean of each grid cell location. (a.) Malawi, (b.) South Africa, (c.) Tanzania, (d.) Zambia.

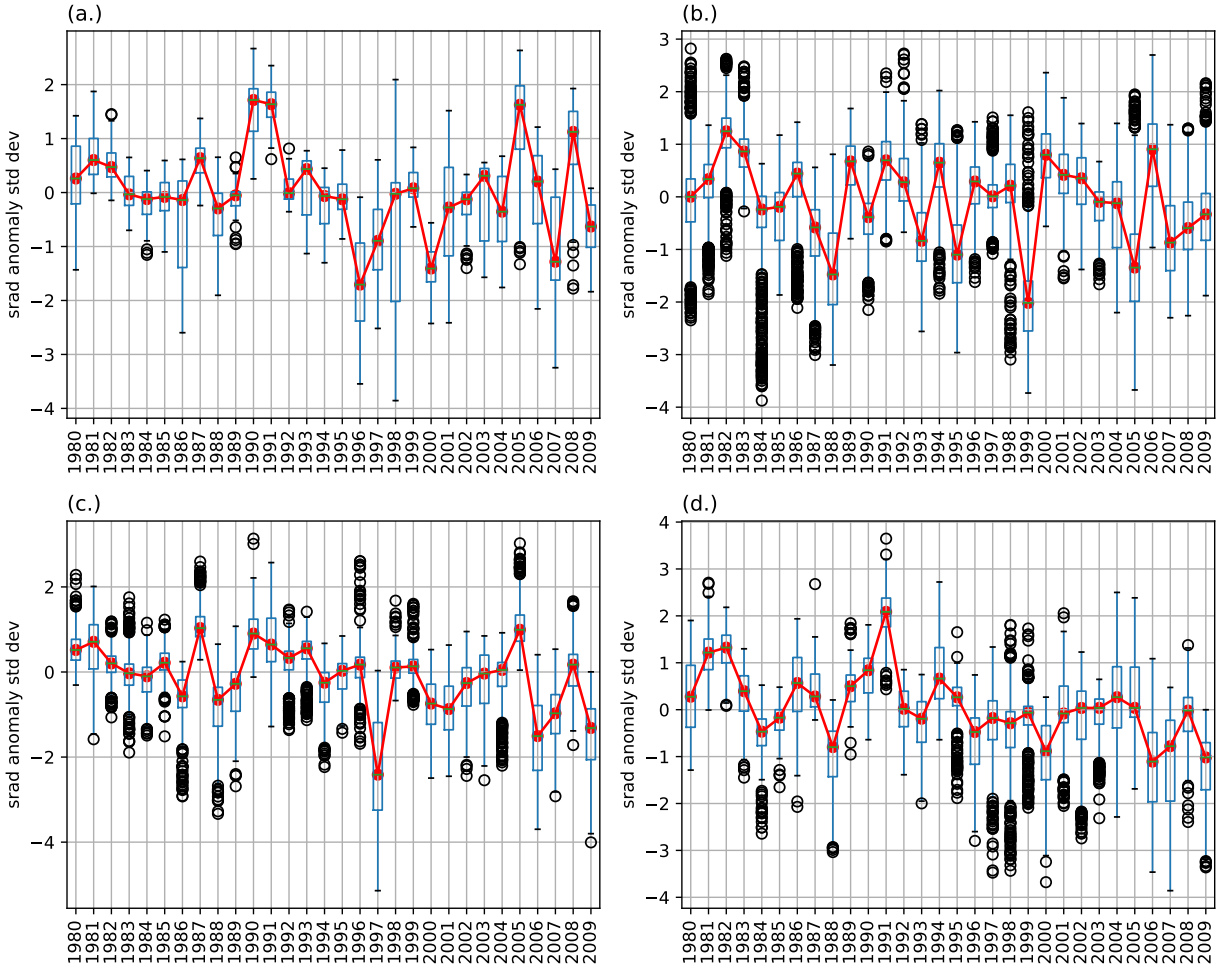

Figure 5.8: long-wave solar radiation anomaly distribution per year during the growing season of December - April for each of the 4 countries. Anomaly is determined as standard deviations from the mean of each grid cell location. (a.) Malawi, (b.) South Africa, (c.) Tanzania, (d.) Zambia.

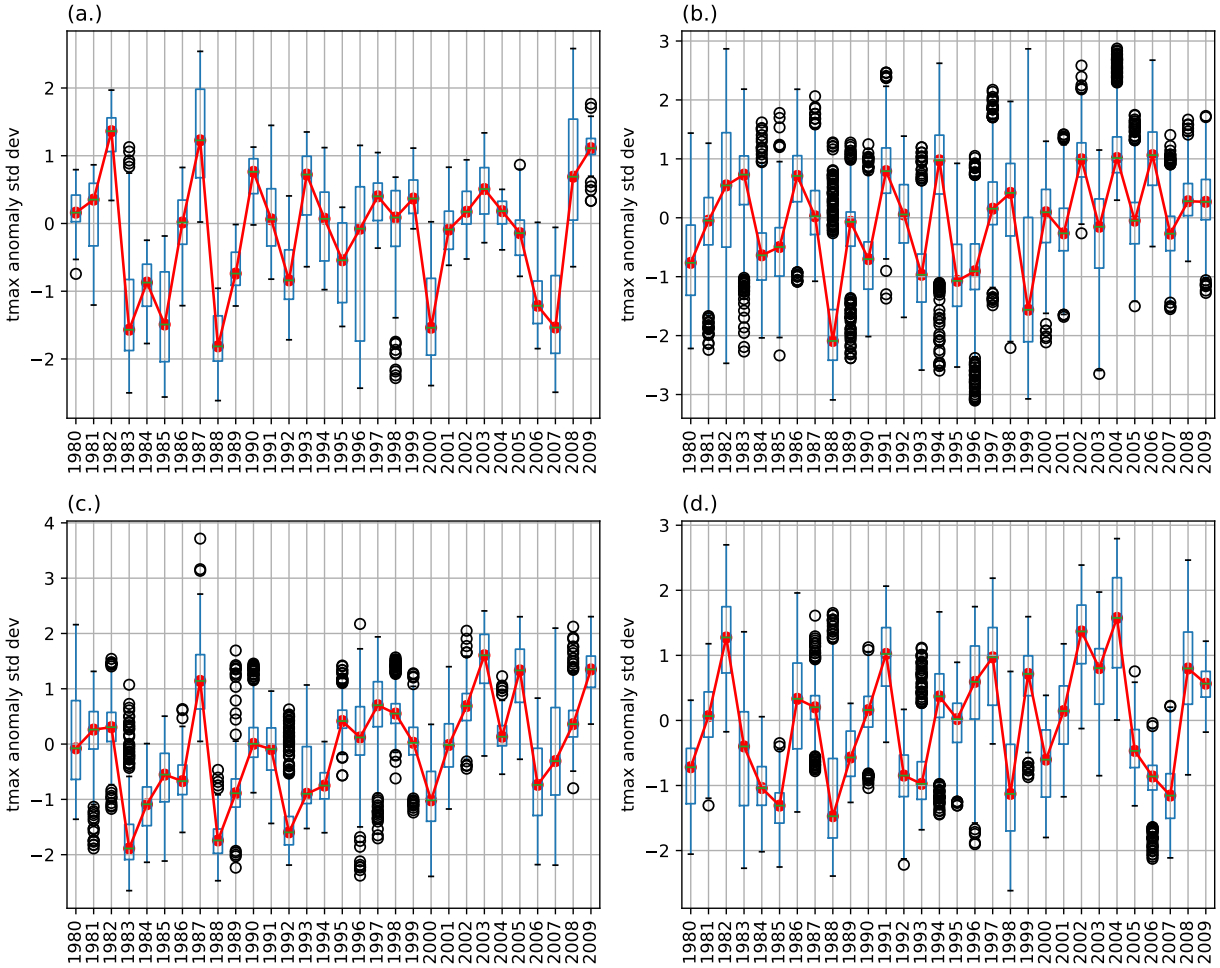

Figure 5.9: mean daily maximum temperature anomaly distribution each year during the growing season of December - April for each of the 4 countries. Anomaly is determined as standard deviations from the mean of each grid cell location. (a.) Malawi, (b.) South Africa, (c.) Tanzania, (d.) Zambia.

5.3.3 Model choice and set up

In this chapter, the GLAM crop model (General Large Area Model for annual crops) is compared to predictions from a random forest model (RFR), and support vector machine (SVM), as well as a multiple linear regression (MLR). GLAM is described in detail in section 2.1 whereas descriptions of machine learning methods can be found in section

2.7. Random forest and the support vector machine models were chosen as they are both very different machine learning models in terms of model structure and so this allows for greater generalization between methods. Multiple linear regression is used as a baseline comparison to assess the need for non-linearity for improved model performance. GLAM is used as the case study process based crop model because it has been used for maize yield simulation in the study region (Jennings et al. 2022) and is designed to be used with an appropriate level of complexity at the regional scale (Challinor et al. 2018, 2004).

Models were compared for the evaluation periods used in (Jennings et al. 2022) for each of the four countries (Malawi, South Africa, Tanzania, Zambia). The full set of years, which ran from 1981 to 2009 was therefore split in half for model calibration / testing and evaluation. For example, in South Africa, models were tested for the years 1988 to 2002, and so therefore model training / GLAM calibration was undertaken for the years 1981 - 1987 and 1989 - 2009. GLAM is usually only capable of calibration using years prior to testing and so a new method was developed for the model to enable a leave-one out calibration style approach for consistency with the machine learning models. This method consisted of masking out years in the evaluation period so that the calibration routine skips the years required for evaluation.

GLAM calibration consisted of applying a calibration routine to optimize the Yield gap parameter (YGP) which is an empirical correction factor static throughout time and is used to account for the effects of management on crop yields (Challinor et al. 2004). GLAM model calibration was undertaken for each grid cell location, however some comparison simulations were also developed using a country level calibration (that is, 1 value of the YGP parameter per country) to understand the effect of spatially varying the YGP parameter on the relationship between climate and predicted yield. As an analogous comparison, machine learning models were trained both with and without grid cell location coordinates as input features. This comparison was made to include the effects of an empirical location

specific yield correction factor similar to the yield gap parameter. Adding and removing these coordinates and making the comparison of spatially varying the YGP enables determination of the appropriate level of importance of the YGP in GLAM and understand the effects of the YGP on crop - climate relationships. All other GLAM parameter values aside from the YGP were taken from Jennings et al. (2022).

5.3.4 Model performance metrics and use of correlations to measure model skill

In this chapter, numerous performance metrics are used to assess model performance of both GLAM and ML models for the 4 Southern African countries to be studied. Firstly, in section 5.4.1 model performance is compared per country using RMSE and correlation coefficient (Pearsons correlation). Both these metrics are used to give a comparative measure of overall model bias (RMSE) and ability to capture the observed variability. In section 5.4.3 Correlations between observed weather and crop yield (both observed and modelled) are compared. This is useful for assessing the sensitivity of the relationships parameterized by GLAM and the ML models (e.g. if the correlation between rainfall and modelled yield is less than the observed correlation then sensitivity could be improved through changes to the soil water balance. Thirdly, in section 5.4.5 the correlation between model skill (correlation coefficient between modelled and observed yield) and the correlation between observed yield and weather (either maximum temperature or rainfall) is used. Although unconventional, this metric is appropriate and informative for noisy data such as the yield data used in this chapter. This is because it can be used to show how model performance is dependent on the observed strength of relationship between observed yield and weather variables. Therefore, this metric is useful because it takes into account the effects of data quality on model skill.

5.3.5 Feature importance with correlated features to assess influence of ML inputs

Various methods of feature importance analysis are useful in answering research questions 1 and 2, however meteorological variables are often highly correlated which makes the use of standard feature importance methods such as permutation importance and partial dependence analysis difficult without any data transformations. This is because high correlations between predictors prevent the attribution of model importance to be assigned to any one particular variable. Furthermore, partial dependence analysis with correlated features can result in unrealistic data points being used to construct the partial dependence plots (Molnar 2022).

In this chapter, due to the high correlation between variables, feature importance is determined through a two step process. Firstly, Principal component analysis (PCA analysis, described in section 2.7.7) is used to create new orthogonal features, this remove correlations between predictors. Then, once the model has been trained using the new features, the correlation between Principal components and features is then used to qualitatively infer feature importance. For example, if Principal component 1 is the most important dimension of the variance, and the seasonal sum of rainfall is the feature which most strongly correlates with principal component 1 , then it can be inferred that the seasonal sum of rainfall is likely the most important variable to determine yield as recognised by the model. Feature importance analysis with PCA analysis is used in this chapter to determine the relative variable contributions to machine learning model performance from a selection of climate variables, indices and outputs from the GLAM mechanistic model. Table 5.2 shows the selection of variables chosen to evaluate using this method.

Table 5.2:
Selection of variables evaluated as inputs to machine learning models. Each variable has a brief description, and origin. Variable origin is simply whether it was obtained from GLAM as a process based model output or if it was originally obtained from the EWEMBI climate dataset (Lange 2018)

variable	Origin	description
tmax	EWEMBI	Maximum daily temperature averaged over the growing season
tmin	EWEMBI	Minimum daily temperature averaged over the growing season
rain	EWEMBI	Rainfall total across the growing season
srad	EWEMBI	total incoming solar radiation across the growing season
dry days	EWEMBI	Total number of days without rainfall
hot days	EWEMBI	Total number of days in which tmax exceeds 32 degrees
pday	GLAM	Simulated planting date
durations	GLAM	Simulated crop duration from planting to harvest
swsf	GLAM	Soil water stress factor
cumpotT	GLAM	Cumulative potential transpiration
ET	GLAM	Evapotranspiration
drainage	GLAM	simulated soil water drainage
HI	GLAM	Simulated harvest index
Biom	GLAM	Simulated biomass
TTMJUV	GLAM	Accumulated thermal time to the end of the juvenile stage

Model variable importance is determined using the permutation importance algorithm. The permutation importance algorithm is defined as follows:

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j} \tag{37}$$

The algorithm follows that for each input feature j and for each repetition in K input feature j of dataset D is randomly shuffled to produce a a corrupted version of feature j. Then the score $s_{k,j}$ is then determined for each iteration and feature. Importance ij is then determined from the some of the importance's as shown above.

Although weather data was derived from the EWEMBI dataset (Lange 2018), dates are required to be selected to filter the weather data from 365 values (one value per day for each year) to be only days which are included in the maize growing season. Growing season dates were chosen to correspond to the crop calendars from the FEWS NET (Famine Early Warning Systems Network) (FEWS NET 2023) archive or was taken from the planting and harvest dates from the GLAM crop model. Machine learning models described as hybrid used the planting and harvest dates from GLAM. Machine learning models which were not hybrid models (only using climate information to predict yield) used a 3 month planting and harvest window taken from FEWS NET (FEWS NET 2023).

GLAM simulated planting date was calculated by Jennings et al. (2022) using a range of values selected from FAO crop calendars and communications with local experts. Planting date was selected in combination with the simulated variety. The simulated variety is determined by varying thermal time parameters (discussed in section 2.1) TLIMJUV, TLIMSIL, TLIMPFL, TLIMPTA and PPSEN. Both variety and planting date values were determined by finding the combination of variety and planting window that resulted in the highest simulated yield in each grid cell. In this respect, This method to simulate planting and variety assumes optimal management with regards to these decisions. Simulated

planting and harvest date per grid cell are shown in Figure 5.10.

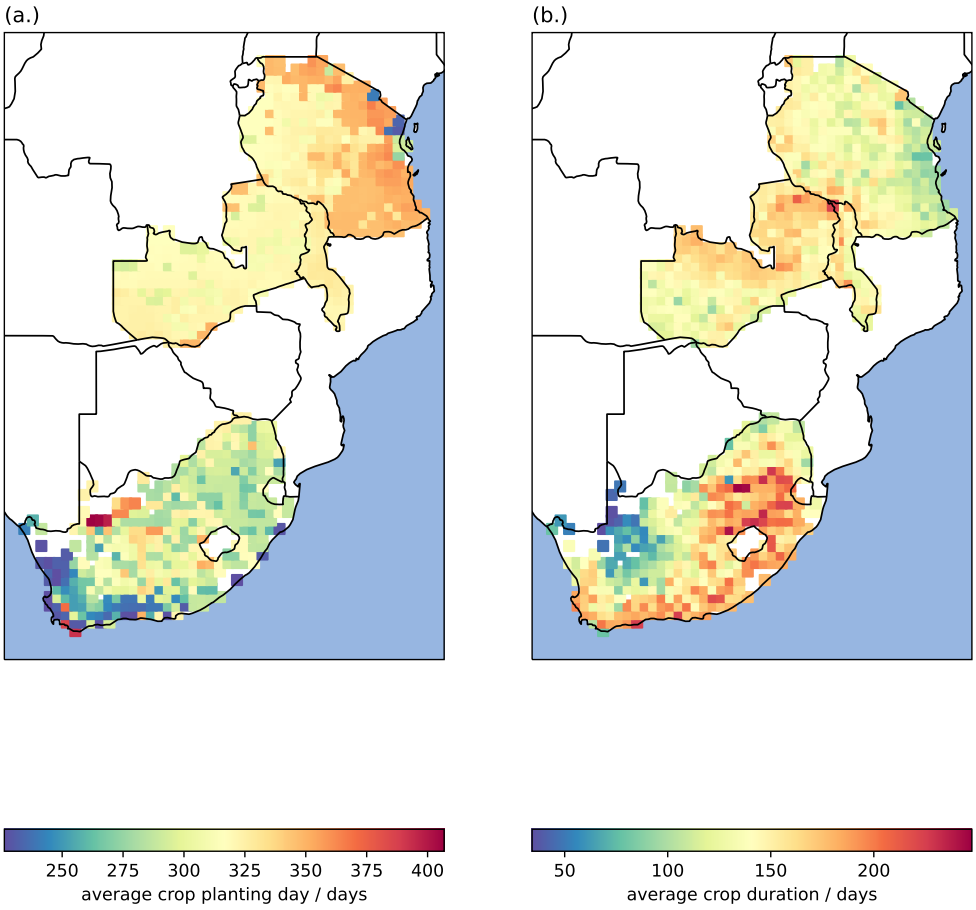

Figure 5.10: a.) Average planting day across the study period as simulated by the GLAM crop model. (b.) Average crop duration from planting to harvest as simulated by the GLAM crop model. Values are averaged between the years 1979 - 2010.

5.3.6 Methods to compare machine learning and GLAM

Machine learning and GLAM are compared in two ways. Firstly, a comparison is made between the relationships between crop yield and climate variables taken from the EWEMBI dataset (Lange 2018). Particular focus is placed on the relationship between predicted yield and rainfall. These comparisons are made to answer the question of How do different models respond to climate conditions? As outlined in the research aims of this chapter.

Secondly, agro-climatic characteristics are correlated against model performance metrics for both the GLAM crop model and the best machine learning model. Characteristics used for this approach include averages of climate variables, as well as measures of climatic variability (namely the coefficient of variation) and measures of the crop yield - climate signal (i.e. the correlation between observed yields and climate characteristics such as rainfall). This is particularly important as a correlation between this variable and model performance would suggest the degree to which model performance depends on the strength of the observed relationship between crop yield and the climate. Table 5.3 lists the variables used for this analysis.

Table 5.3:
Variables chosen to determine spatial effects of climate on model performance as well as performance metrics. Variables used in Figure 5.34 are listed here.

variable	description
t av	Average maximum temperature
lat	latitude
swsf cv	coefficient of variation in the soil water stress factor
GLAM RMSE	GLAM model Root Mean Square Error
r av	average rainfall
swsf av	Average seasonal total soil water stress factor

clay	Soil clay percentage
lon	longitude
RF ccoef	Correlation coefficient between random forest predicted yields and observed yields
ry ccoef	Correlation coefficient between rainfall and observed crop yield
GLAM Y ccoef	Correlation between GLAM predicted yields and observed yield
yield cv	coefficient of variation in observed yield
r cv	coefficient of variation in rainfall
cv dur	coefficient of variation in GLAM predicted crop duration
ty ccoef	correlation coefficient between average seasonal daily maximum temperature and observed yield
sand	Soil sand percentage
s av	Average daily incoming longwave solar radiation
sy ccoef	Correlation between solar radiation and observed yield
RF rmse	Random forest model Root Mean Square error
s cv	coefficient of variation in solar radiation
t cv	coefficient of variation in temperature
av dur	Average GLAM predicted crop duration

5.3.7 De-trending of yield data

Crop yield trends in each country will be, to varying degrees, affected by trends in technological improvements to crop management. Such trends are not simulated by the GLAM crop model and so de-trending is a method of removing technological trends. As shown from Figure 5.11 in all countries aside from Tanzania, there are some positive trends over time in mean yields.

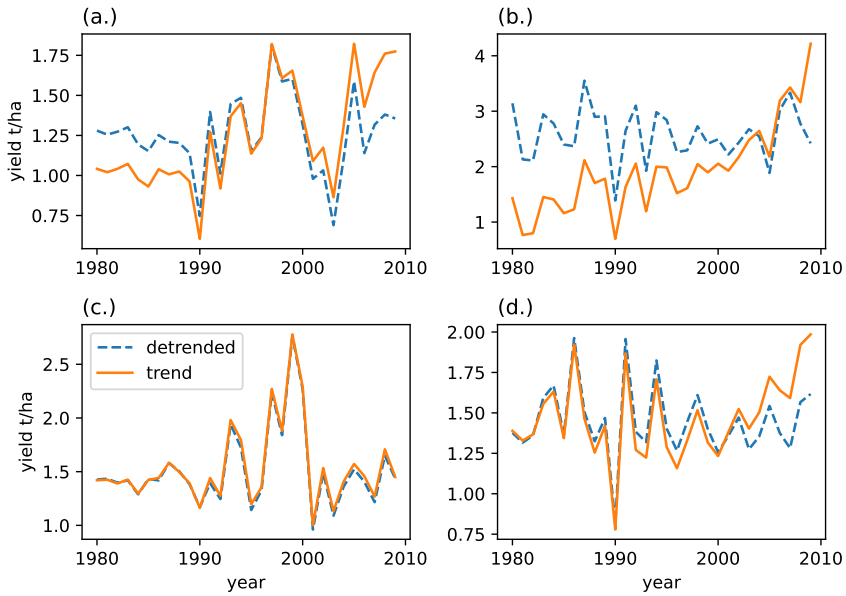

Figure 5.11: Trended and de-trended mean yield across the study period for each of the four countries, (a.) Malawi, (b.) South Africa, (c.) Tanzania, and (d.) Zambia.

Yields were de-trended by fitting a lowess function to the observed yield data (Cleveland 1979). De-trended values are produced by the difference between the observed and lowess fitted values. The mean of the residuals produced by the difference between the lowess fitted values and the observed values is then subtracted from the de-trended yield values so that the mean of the de-trended yields is the same as that of the observed yields. Trends across time and the effect of this de-trending can be compared using both Pearson and Spearman rank correlations. The most obvious trend is shown in South Africa (panel b of Figure 5.11). The Pearson and Spearman rank correlations of the raw data over time are 0.822 and 0.836 respectively. When de-trended, the correlations over time become statistically insignificant (p value of 0.9005 and 0.9006 respectively). A Trend across time is also significant in Malawi, the Pearson and Spearman rank correlations over time are 0.665 and 0.653 respectively. Again, trends become statistically insignificant after de-trending. Although there is some trend over time in the Zambia yield time series, this only

takes place from about the year 2000, and so the de-trending process removed this aspect from the data. Hence, although the trend from the year 2000 on-wards has a Pearson correlation of 0.916 and a Spearman rank correlation of 9.15, the overall trend across time is 0.298 (Pearson) and 0.284 (Spearman rank). Therefore, the de-trending process clearly has differing effects each country, as strong trends are present in the Malawi and South Africa time series, only a partial trend is present in the Zambia time series, and there is no significant trend in the Tanzania time series.

5.4 Results

The following section describes the results of the model simulations for this chapter. Firstly, a direct comparison is made between benchmark machine learning methods and GLAM. Secondly, to better understand the reasons behind the differences in model performance, the following sections will provide further detail between correlations with input and output variables, as well as an assessment of agroclimatic conditions which most greatly affect model performance.

5.4.1 Bench-marking of overall model performance

Figure 5.12 shows the spatial distribution of correlation coefficients across time for each grid cell (prediction skill of inter-annual variability in yield). The best performing model can depend on the country being assessed. Notably the random forest model predicts the inter-annual variability best in Tanzania, but the support vector machine model offers better predictions for South Africa. All models appear to predict the inter-annual variability in yield well in Malawi, but poorly in Zambia. In Zambia, all models show negative correlations with yield across time.

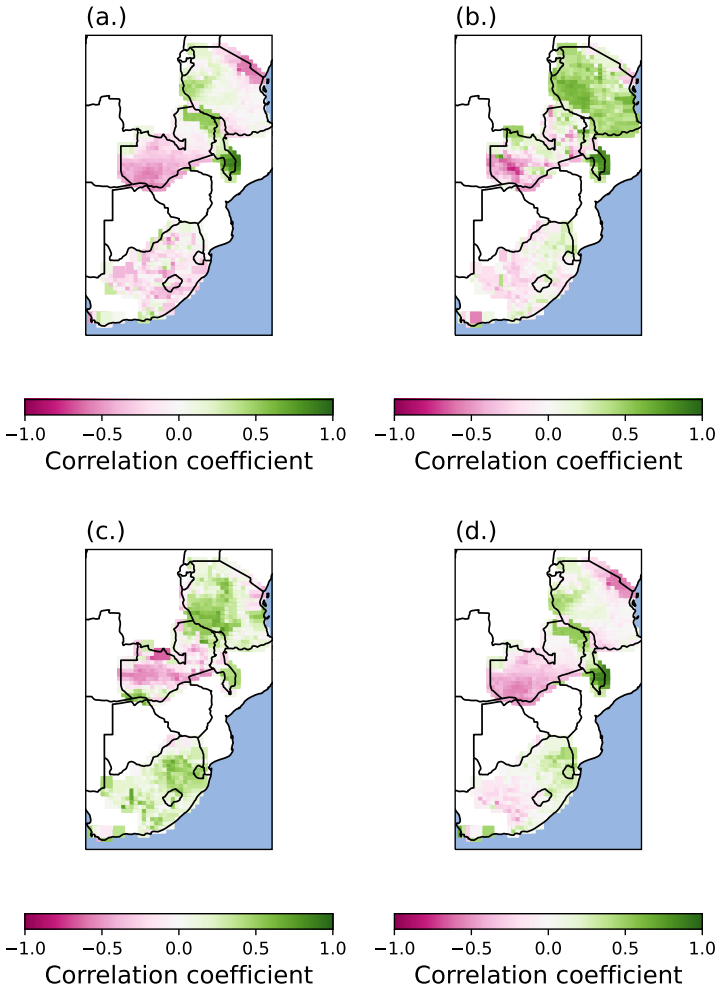

Figure 5.12: Correlation coefficients between predicted values by each of the models (a) : GLAM, (b) Random Forest, (c) Support vector machine, (d) Multiple linear regression and observed crop yield data from the GDHY dataset (Iizumi & Sakai 2020)

For RMSE (Figure 5.13, models all predict largest RMSE on the eastern coast of South Africa, (which has the highest yields in the dataset), and some models show larger RMSE

for eastern Zambia which is generally drier and has the lowest maize yields in the country.

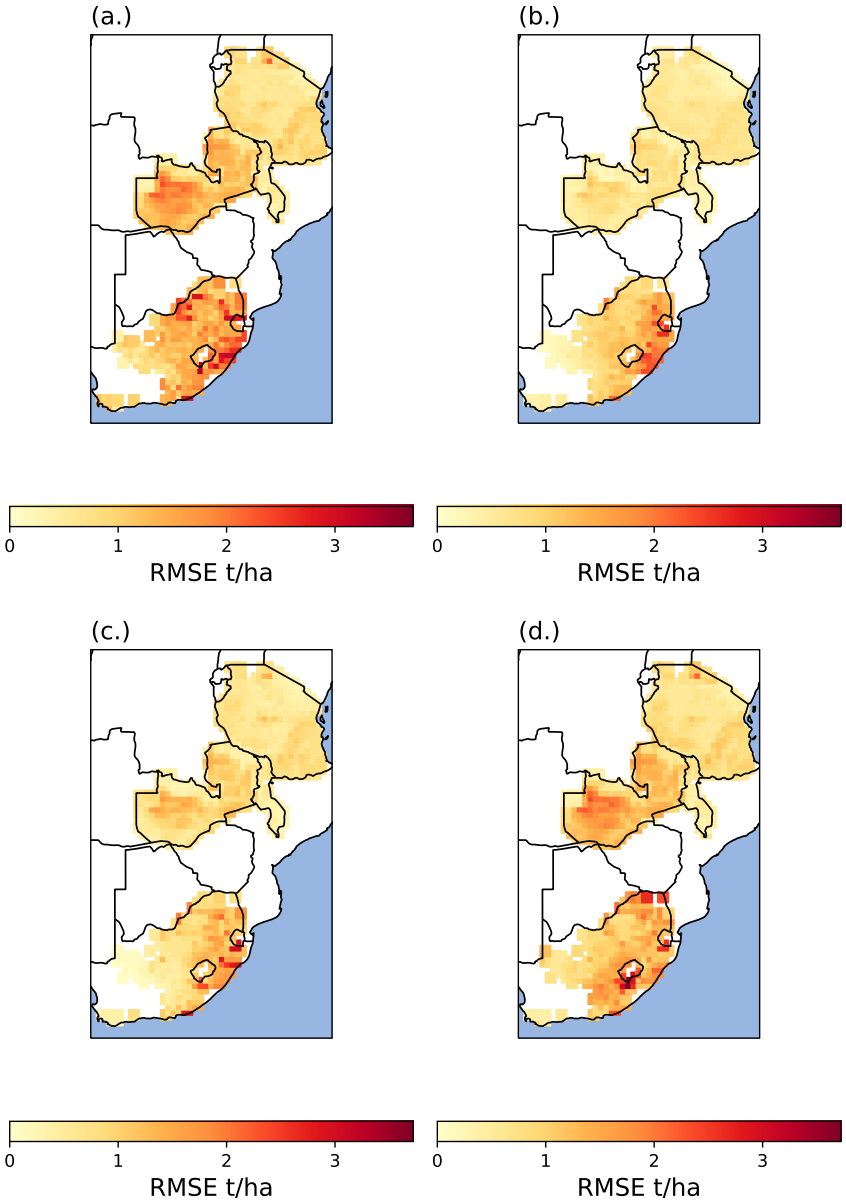

Figure 5.13: RMSE between predicted values by each of the models (a) : GLAM, (b) Random Forest, (c) Support vector machine, (d) Multiple linear regression and observed crop yield data from the GDHY dataset (Iizumi & Sakai 2020)

As well as correlation coefficient and RMSE, mean yield predictions are compared between the observed data and models in Figure 5.14. Notably, only the support vector machine and random forest models are able to capture the relatively low yields present in north east Tanzania and come closer than other models to predicting the low yields in Zambia. In South Africa however, GLAM is able to predict the stronger distinction between the high yielding zone in the east of the country and lower yields associated with the dry region in the west, meaning central South Africa is better predicted.

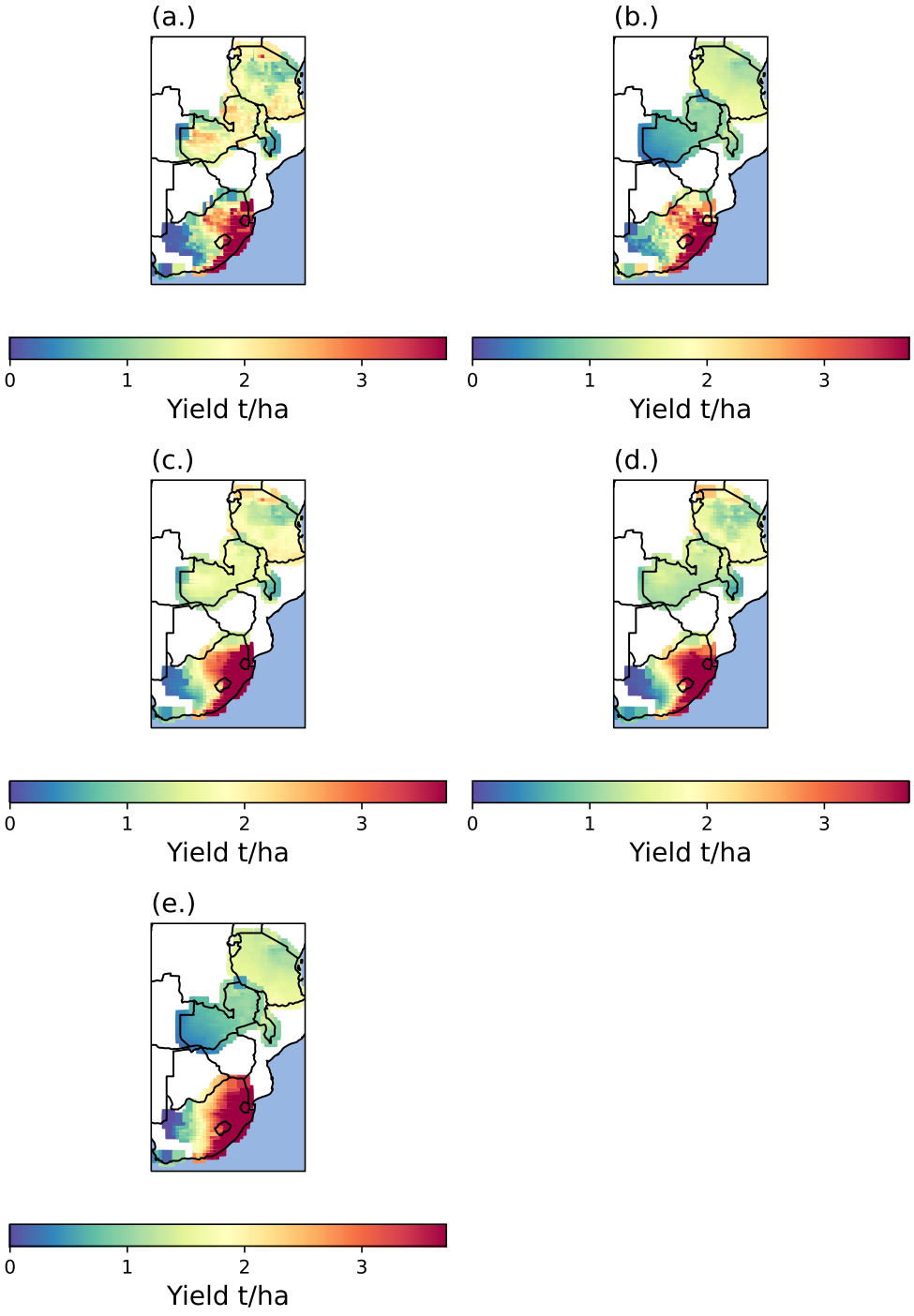

Figure 5.14: Mean predicted values by each of the models (b) : GLAM, (c) Random Forest, (d) Support vector machine, (e) Multiple linear regression and observed crop yield data from the GDHY dataset (Iizumi & Sakai 2020), Observed mean yield is shown in Panel (a).

Table 5.4 shows correlation coefficient and RMSE for each country and model. models used grid cell coordinates as inputs and were the climatology only inputs. The ML models performed better in general across all countries than the GLAM benchmark. Models are compared across hybrid and climatology models in Figure 5.16 which displays correlation coefficient and RMSE error metrics for each model. Panels (a.i.) and (b.i.) show the results of training ML models purely on climate data found in Table 5.2, whereas panels (a.ii.) and (b.ii.) show the results of training models on both climate information and GLAM variables (Variables used are also in Table 5.2). A comparison between the 2 sets of panels in this Figure shows that overall for South Africa GLAM outputs slightly improve Random Forest model and MLR predictions, but do not improve the performance of the support vector machine. Furthermore, information from the GLAM model can be detrimental to ML model performance. This is seen in Malawi, Tanzania and Zambia to a greater or lesser extent.

Across countries performance across all models is variable. Best model performance is seen in South Africa across models. The random forest model also performs better than all other models in Malawi and Tanzania. Performance of models in Zambia is poor, potential reasons for this are discussed in section 5.5.7.

Figure 5.15 shows that ML model performance in the common test periods for each country is in line with typical model performance across the dataset. Figure 5.15 shows that when splitting the time series of data into sequential 5 year test periods ML model performance is most most consistent in South Africa. Since the ML models achieve acceptable performance across countries which is consistent across years and (crucially) outperform GLAM there is no need for tuning of hyper-parameters.

Table 5.4:
Model performance metrics between simulations of all models and observed data for all four countries studied in this chapter across the common GLAM and ML test periods. Metrics used are RMSE: root mean square error (normalized by the inter-quartile range of the observations), CCOEF: pearsons correlation coefficient.

Country	Metric	GLAM	RFR	SVM	MLR
Malawi	NRMSE	1.137	0.691	0.977	1.539
	CCOEF	0.107	0.765	0.618	0.422
South Africa	NRMSE	0.581	0.454	0.400	0.564
	CCOEF	0.609	0.852	0.856	0.762
Tanzania	NRMSE	1.193	0.924	1.155	3.120
	CCOEF	0.333	0.635	0.505	0.304
Zambia	NRMSE	1.149	2.280	3.114	2.414
	CCOEF	0.161	0.559	0.140	0.084

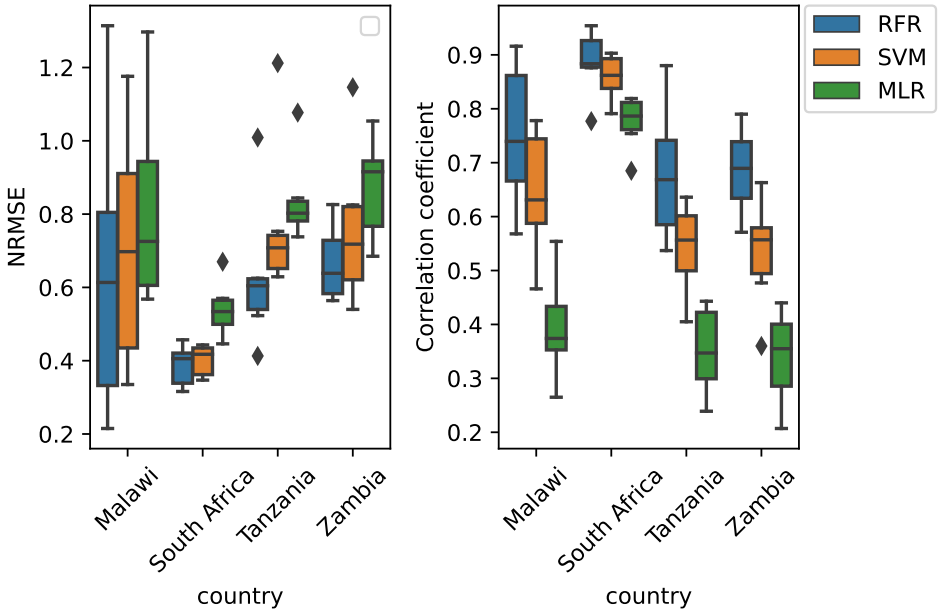

Figure 5.15: ML model cross validation across each 5 years of the dataset for each country.

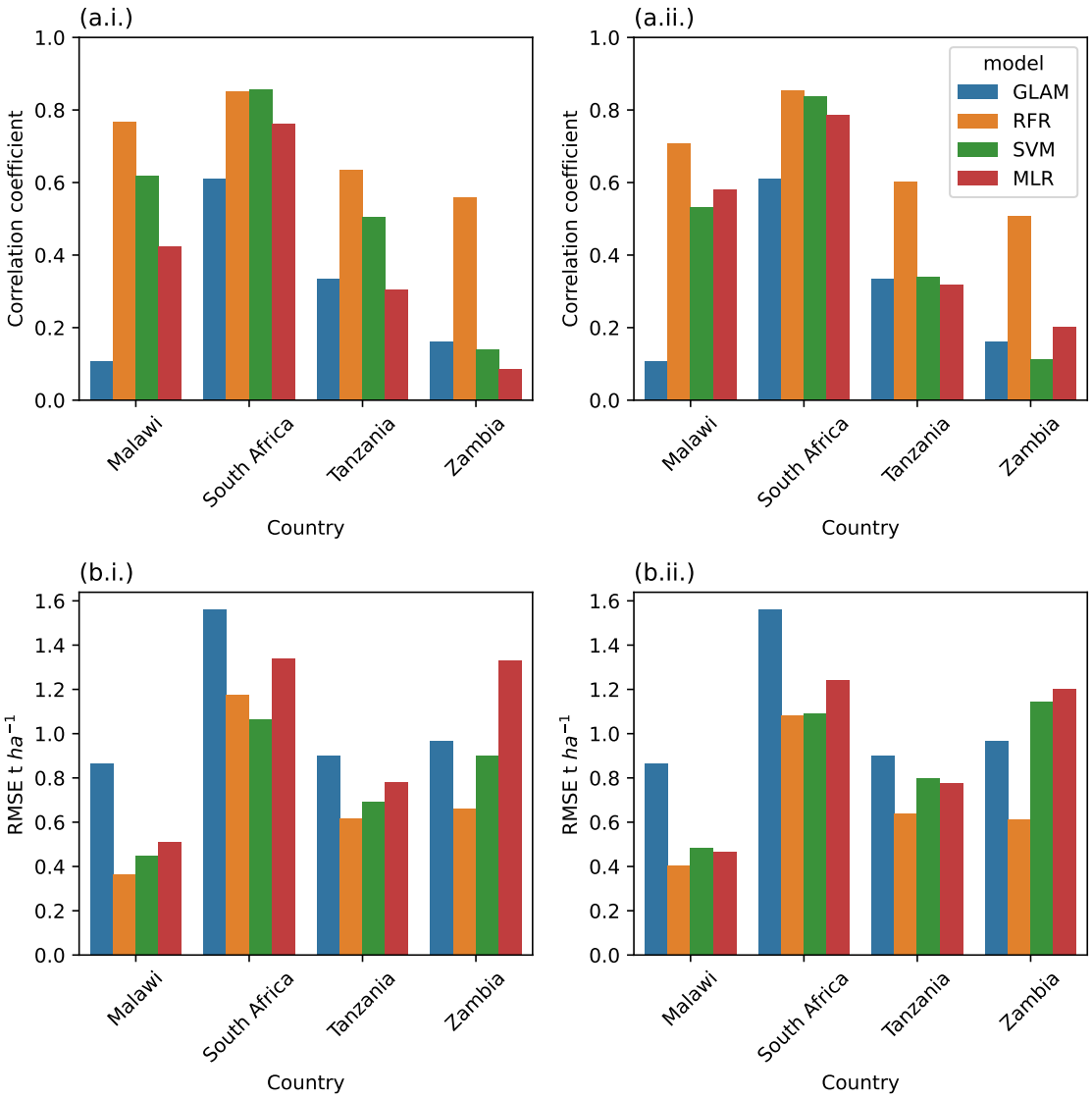

Figure 5.16: Bar plot of (a) Correlation coefficient and (b) Root mean square error (RMSE) for each model tested per country. Numeral (i) denote that machine learning models were trained purely using climate data, and numeral (ii) denote that machine learning models were trained using both climate data and GLAM variables. Table 5.2 refers to each of the variables used for both approaches.

5.4.2 What value do modelled process outputs have for improving machine learning predictions?

Figures 5.17 to 5.20 show the correlations between each of the considered input variables for training the machine learning methods in this chapter. Table 5.2 provides a key for the variables used. Crucially, the relationships between yield and climatological variables are not the same for each country. This meant that during the training process, it was found that training models on each country individually resulted in improved model performance over models trained on all countries together, even when model performance for a country was particularly poor (as was the case in Zambia). This suggests that in fact, more data does not always equal a better model because relationships between weather and yield will change between countries or regions and generalization across such a wide geographical area may not always be beneficial.

The correlation matrices show that weather relationships with yield are much stronger in South Africa than the other countries, in particular, the effect of rainfall on yield. Across all countries, thermal time outputs (TTMJUV, TTMPTA, TTMASI, TTMPFL) are highly correlated. These parameters are also strongly negatively correlated with vapour pressure deficit in Malawi and Zambia. Yield has the weakest correlations with weather input variables in Zambia, likely being the reason behind poor model performance. Yield is strongly negatively correlated with dry days and in South Africa, and moderately positively correlated with rainfall, indicating the importance of a dry day indicator for model performance.

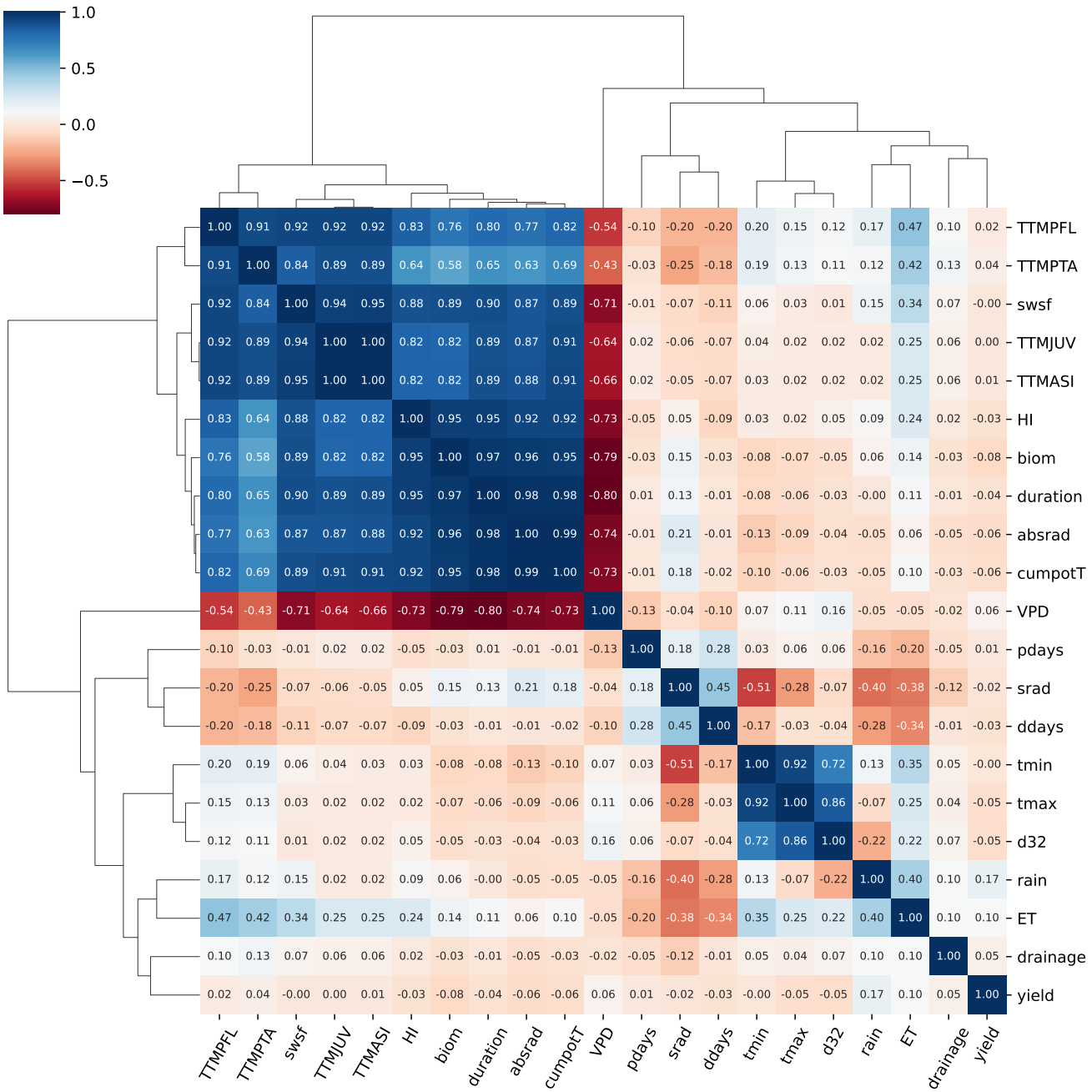

Figure 5.17: Correlations and hierarchical clustering between each each of the variables considered for the machine learning models and observed yield and coordinate location in Malawi. Observed yield was removed from the dataset before model training.

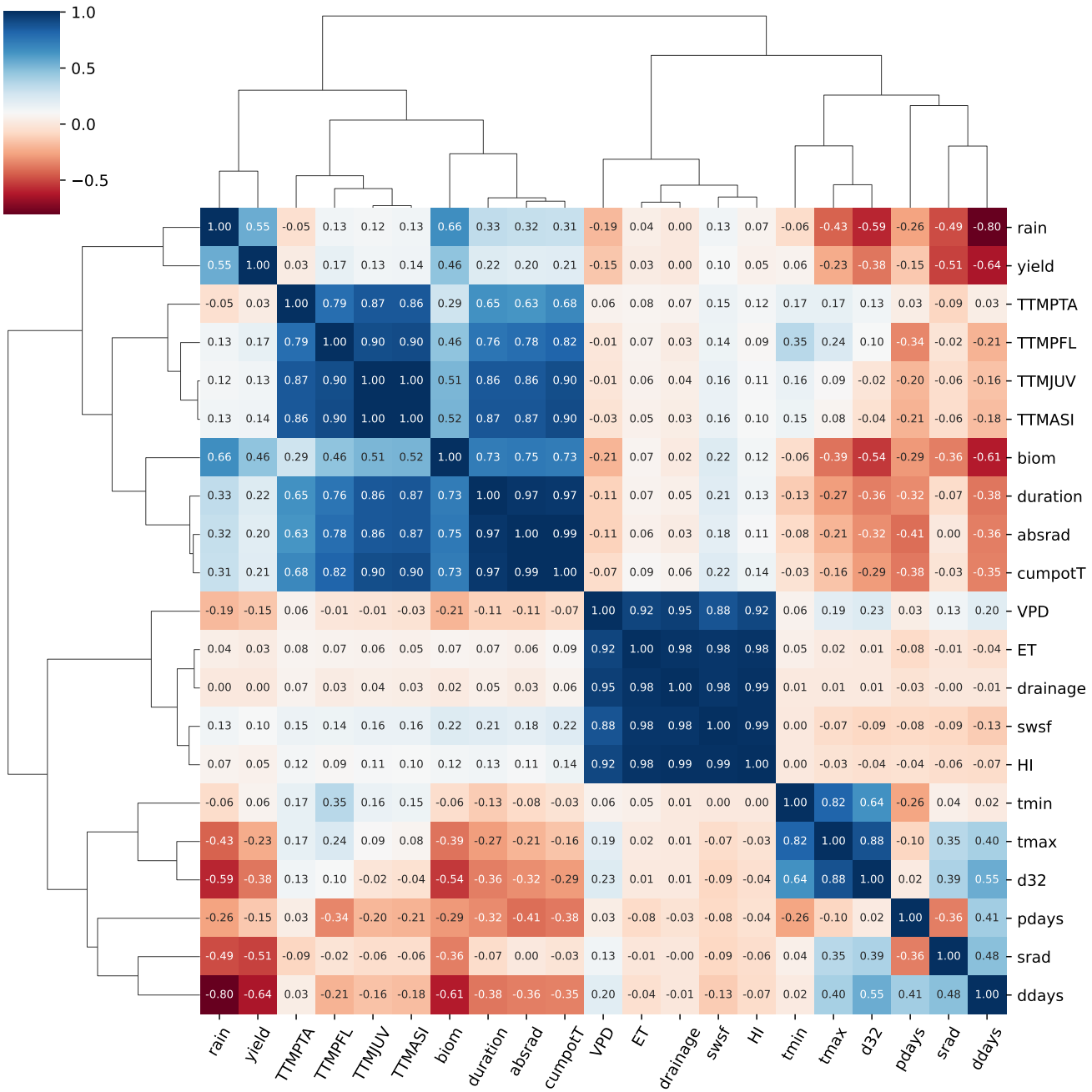

Figure 5.18: Correlations and hierarchical clustering between each each of the variables considered for the machine learning models and observed yield and coordinate location in South Africa. Observed yield was removed from the dataset before model training.

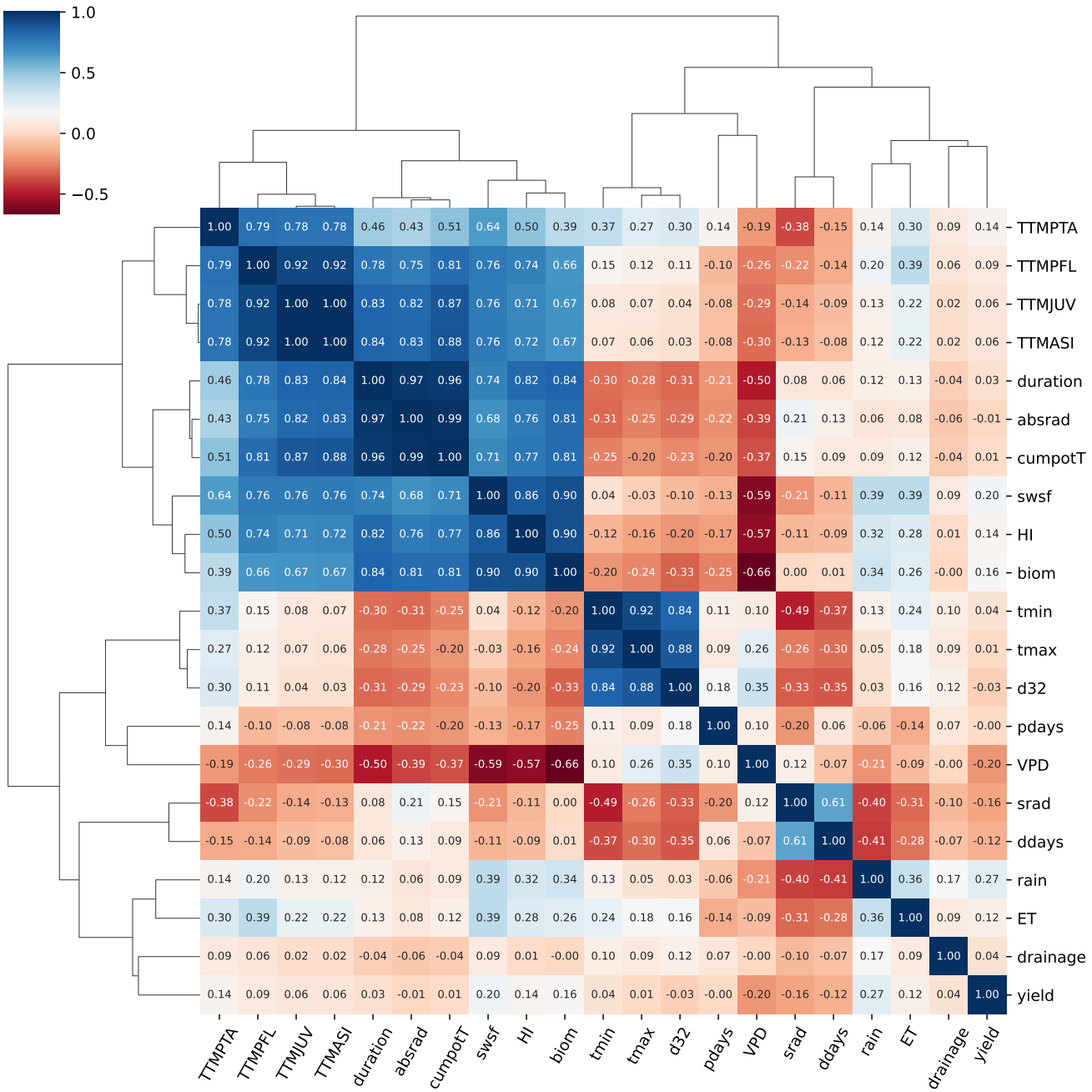

Figure 5.19: Correlations and hierarchical clustering between each each of the variables considered for the machine learning models and observed yield and coordinate location in Tanzania. Observed yield was removed from the dataset before model training.

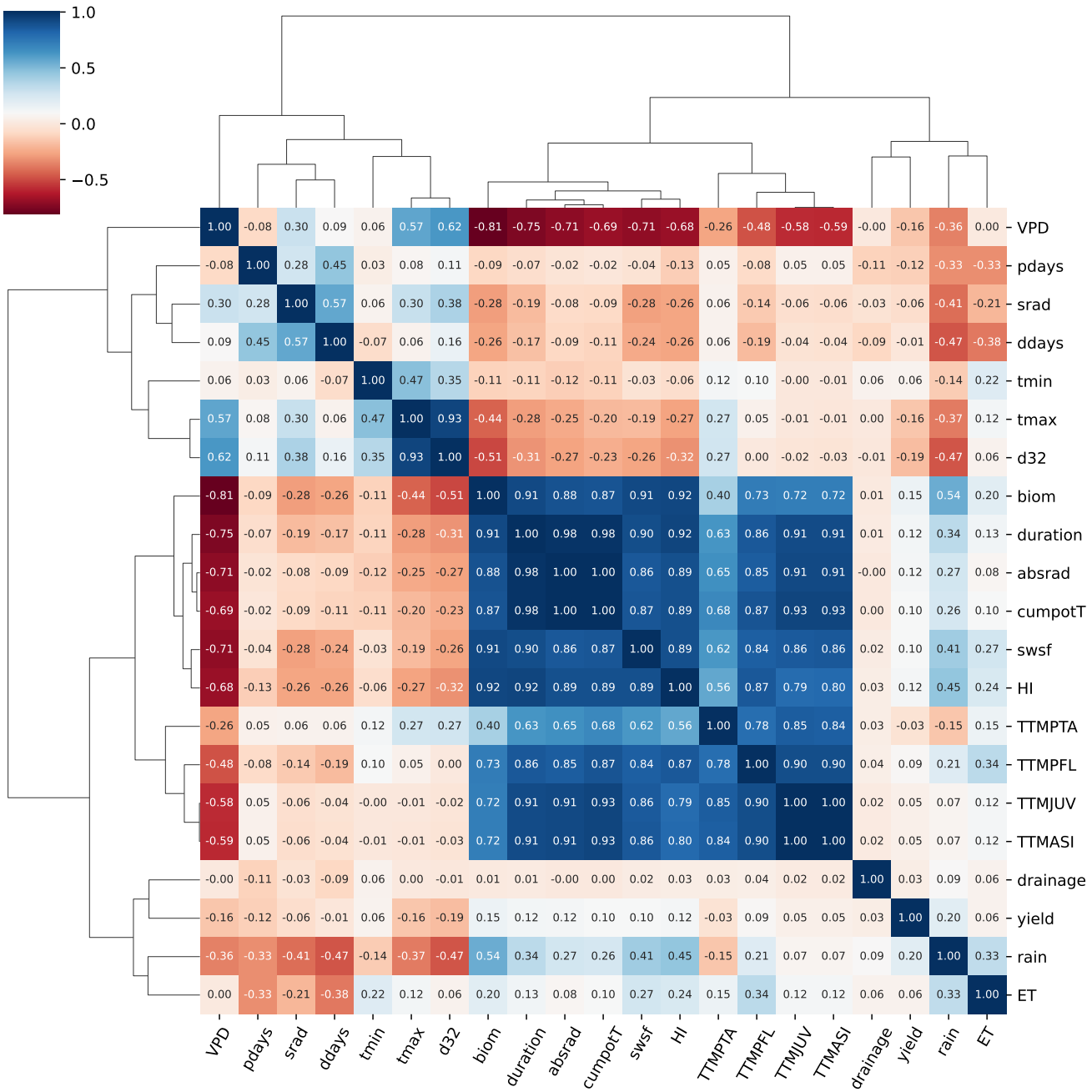

Figure 5.20: Correlations and hierarchical clustering between each each of the variables considered for the machine learning models and observed yield and coordinate location in Zambia. Observed yield was removed from the dataset before model training.

PCA analysis was applied to each country separately. Figure 5.21 shows the variance explained ratio for each principal component from the first to the 20th for each country. In South Africa (panel b) although Principal component (PC) 1 explains the least variation of all of the first PCs for each country, PC2 explains comparatively more variation and less principal components are required overall to explain the total variation of the dataset.

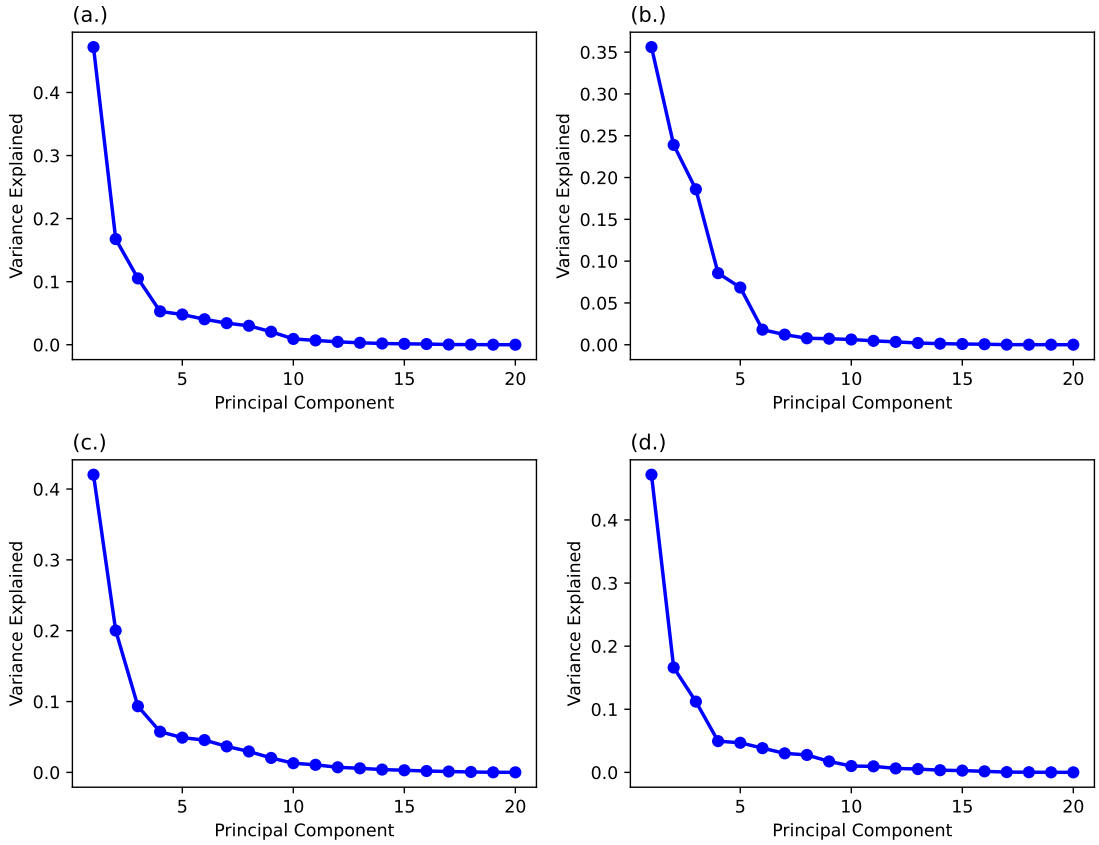

Figure 5.21: Variance explained ratio across 20 principal components using the variables consider for machine learning analysis. Figure annotations denote each of the four countries of the analysis, namely (a) Malawi, (b) South Africa, (c) Tanzania, (d) Zambia.

After PCA analysis is used to determine degree of variance explained for PCs for each country, a random forest, support vector machine and multiple linear regression are trained

using all principal components. Figures 5.22 to 5.25 show both model performance and feature importance for each Principal component when training using this method. Results presented here are used as a separate test and are not used as part of any other analysis in this chapter. Model performance is significantly reduced in comparison to baseline model simulations presented in section 5.4.1. Notably, feature importance differs between the 3 models and per country. In Malawi and south Africa, feature importance is very different for the three models. However, in Tanzania, feature importance for the random forest and support vector machine models are a lot more similar, this is also the case in Zambia. Although feature importance is more similar between RFR and SVM models in these two countries, model performance is particularly poor and so for these two countries, it cannot be said which variables may be more useful for model performance.

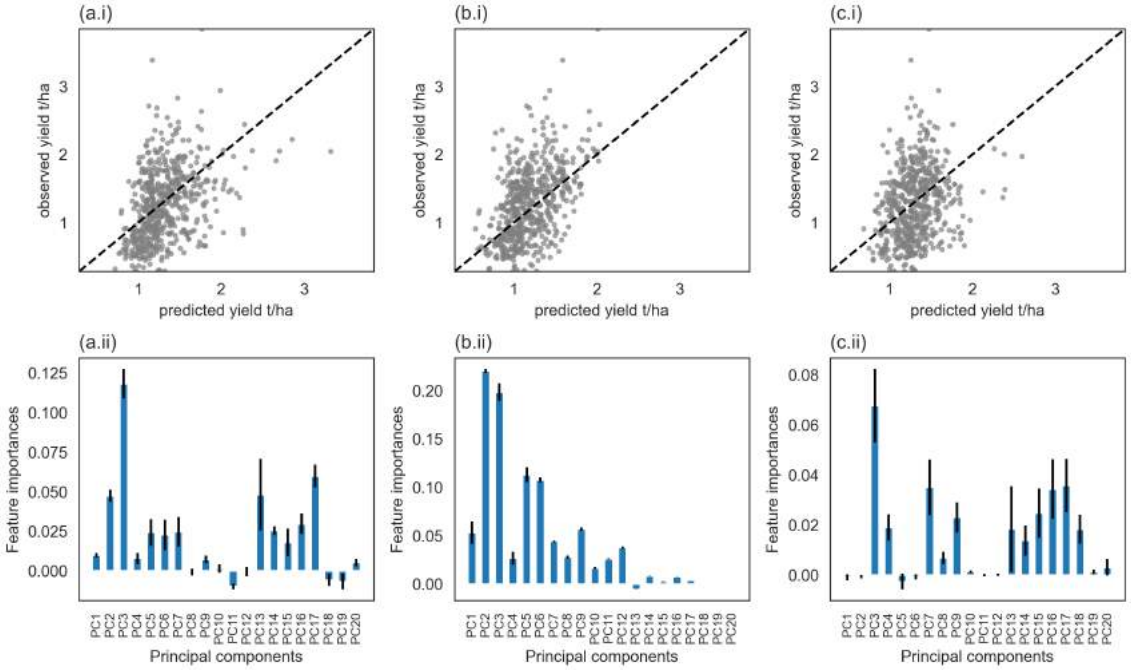

Figure 5.22: Feature importance for Principal components with scatter of predictions resulting from the model trained. Results are shown for 3 models trained and tested using data from Malawi only. Column (a) shows results of a random forest model, (b) for a support vector regression, and (c) for a multiple linear regression. row (i) shows a scatter plot of predicted values from each model against observations, row (ii) shows the permutation feature importance for each Principal component.

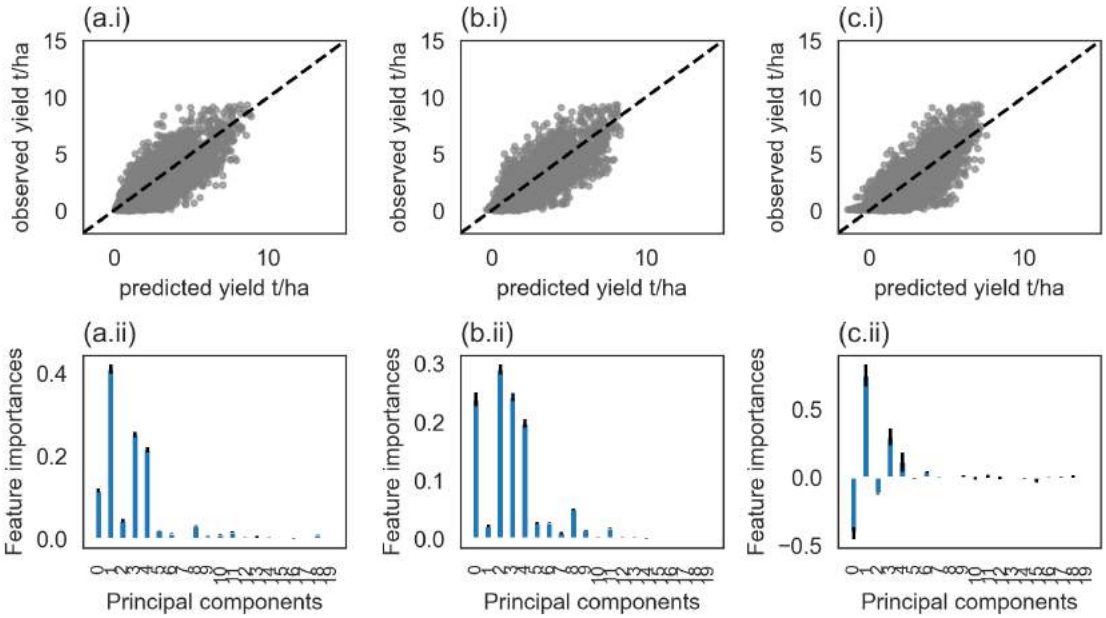

Figure 5.23: Feature importance for Principal components with scatter of predictions resulting from the model trained. Results are shown for 3 models trained and tested using data from South Africa only. Column (a) shows results of a random forest model, (b) for a support vector regression, and (c) for a multiple linear regression. row (i) shows a scatter plot of predicted values from each model against observations, row (ii) shows the permutation feature importance for each Principal component.

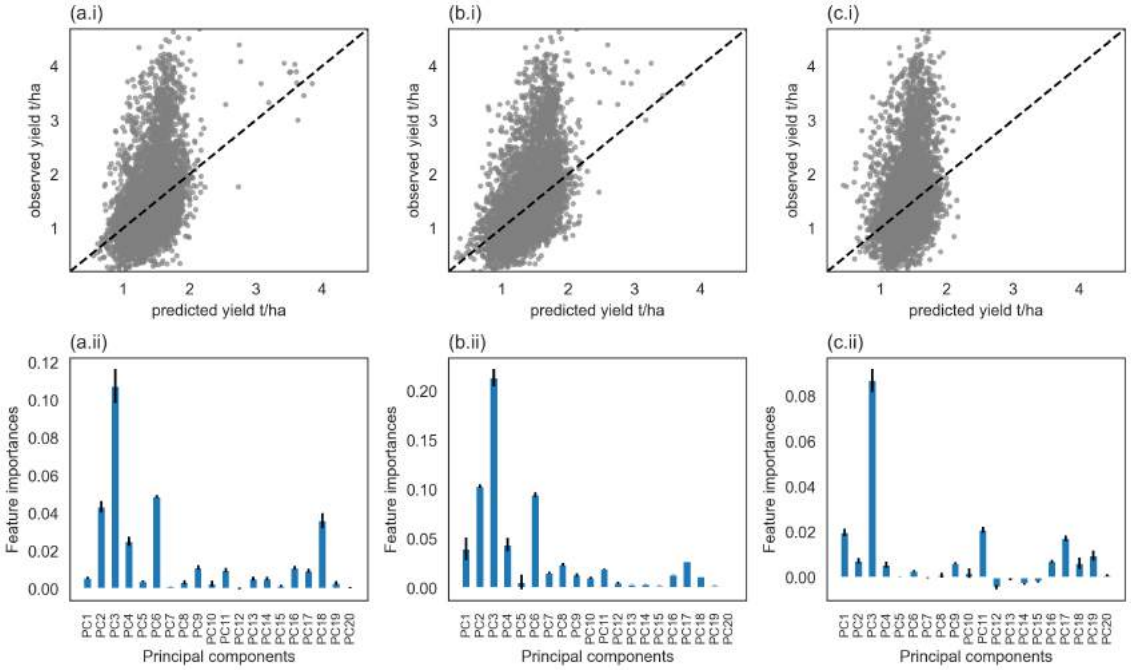

Figure 5.24: Feature importance for Principal components with scatter of predictions resulting from the model trained. Results are shown for 3 models trained and tested using data from Tanzania only. Column (a) shows results of a random forest model, (b) for a support vector regression, and (c) for a multiple linear regression. row (i) shows a scatter plot of predicted values from each model against observations, row (ii) shows the permutation feature importance for each Principal component.

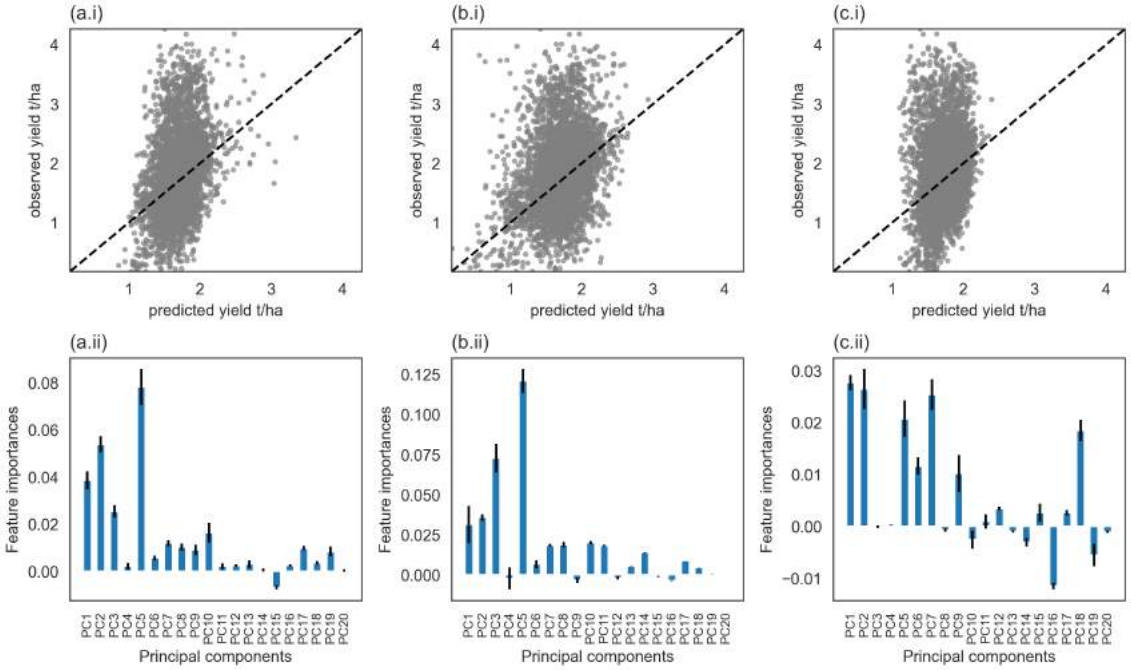

Figure 5.25: Feature importance for Principal components with scatter of predictions resulting from the model trained. Results are shown for 3 models trained and tested using data from Zambia only. Column (a) shows results of a random forest model, (b) for a support vector regression, and (c) for a multiple linear regression. row (i) shows a scatter plot of predicted values from each model against observations, row (ii) shows the permutation feature importance for each Principal component.

In Malawi the SVM model favours minimum and maximum temperatures which are correlated more strongly which PC2, and the random forest model favours rainfall more as a predictor (which has a stronger correlation with PC3). In South Africa, the random forest model more strongly favours PC2 which is most correlated with evapotranspiration, vapour pressure deficit, harvest index and soil water stress factor. By contrast, the SVM model more favours PC1 and 3 most correlated with cumulative potential transpiration, the number of days above the 32 degree temperature threshold, average maximum daily temperature and minimum daily temperature.

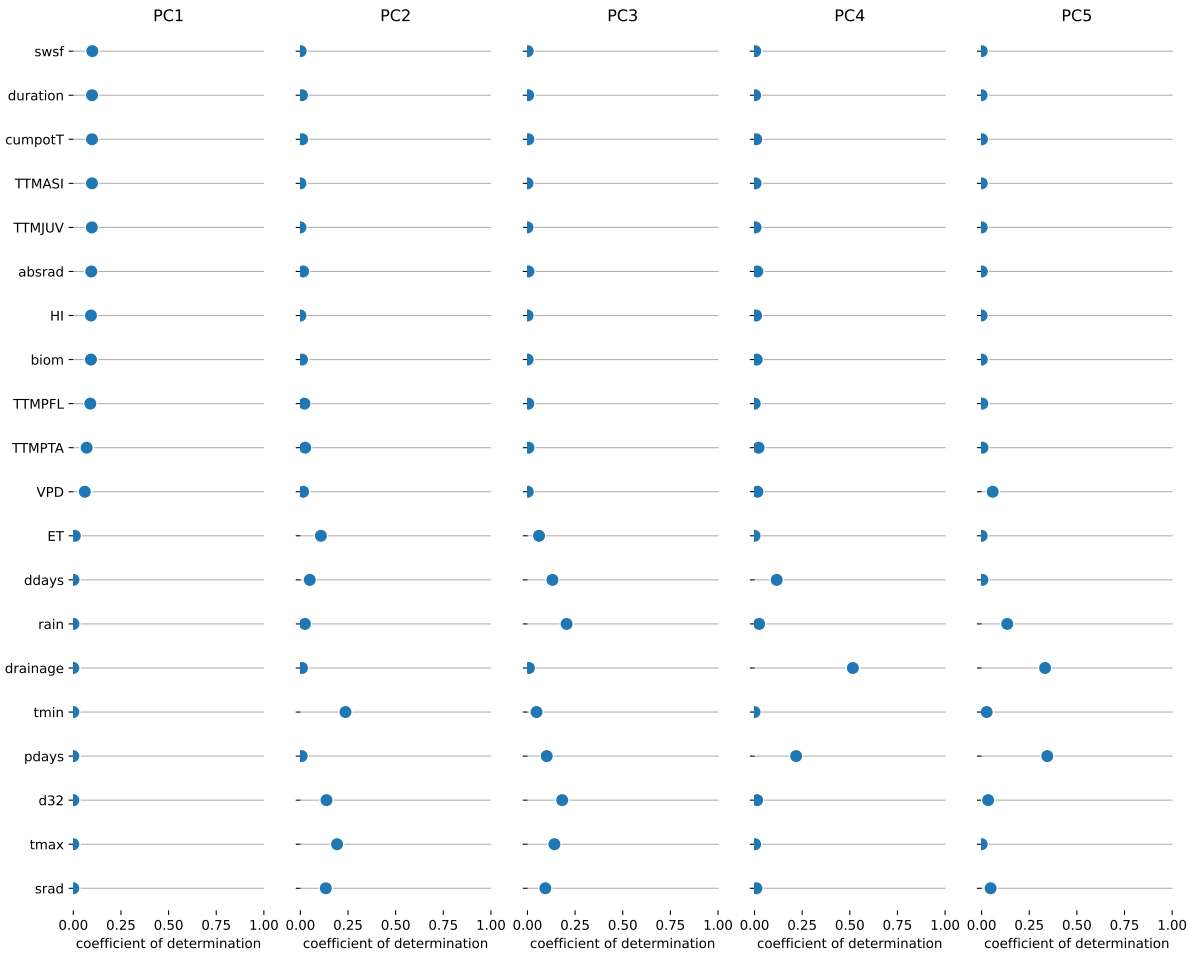

Figure 5.26: Coefficient of determination between each of the input variables and principal components 1 to 5 construct from a PCA decomposition. See Table 5.2 for a description of each of the variables used for the analysis. Data is shown for Malawi only.

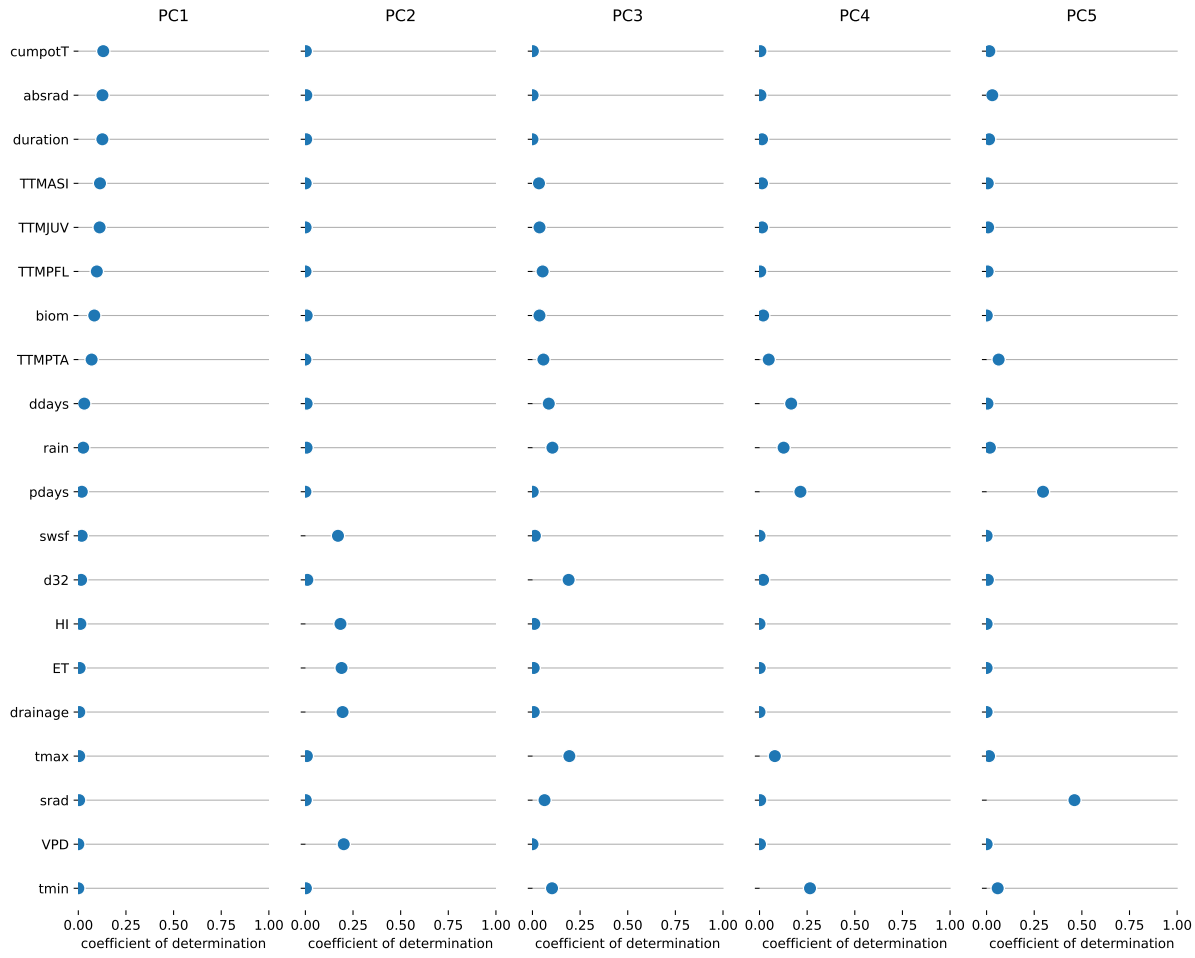

Figure 5.27: Coefficient of determination between each of the input variables and principal components 1 to 5 construct from a PCA decomposition. See Table 5.2 for a description of each of the variables used for the analysis. Data is shown for South Africa only.

Feature importance here is shown as a way of determining the possible importance placed upon different data sources for different models. Rainfall and temperature climate variables are most important due to the effects on crops due to heat and drought stress as well as the effect of temperature on crop physiological development, governed by thermal time. Hence, in the following section, the importance of rainfall and temperature is determined for both GLAM and the machine learning models by presenting the correlations between predicted yield and rainfall, as well as temperature whilst also taking into account the

scale of GLAM calibration. Feature importance is not shown for Zambia and Tanzania because the models trained using principal components for those countries resulted in very poor performance and so feature importance is not as useful to understand.

5.4.3 How do crop yield responses to climate conditions differ across models and observed data?

5.4.3.1 Part 1: Temporal correlations

In this section, modelled and observed crop yield responses to weather and climate variability are compared both across time and space separately. Figures 5.28 and 5.29 show the observed and predicted correlations between rainfall, and solar radiation. This Figure was constructed by applying a Pearson's correlation between crop yield and rainfall across time per grid cell. Generally, the modelled relationships between rainfall and yield are more positive for machine learning model predictions than GLAM, and the relationships between modelled yield and solar radiation are more negative for machine learning models than GLAM. Unusually, the observed yield response to rainfall is generally slightly negative and solar radiation generally slightly positive. The potential reason for this is addressed in section 5.5.6. Interestingly, GLAM predicts a positive relationship between solar radiation and observed yield in northern Tanzania which has a bi-model rainfall distribution.

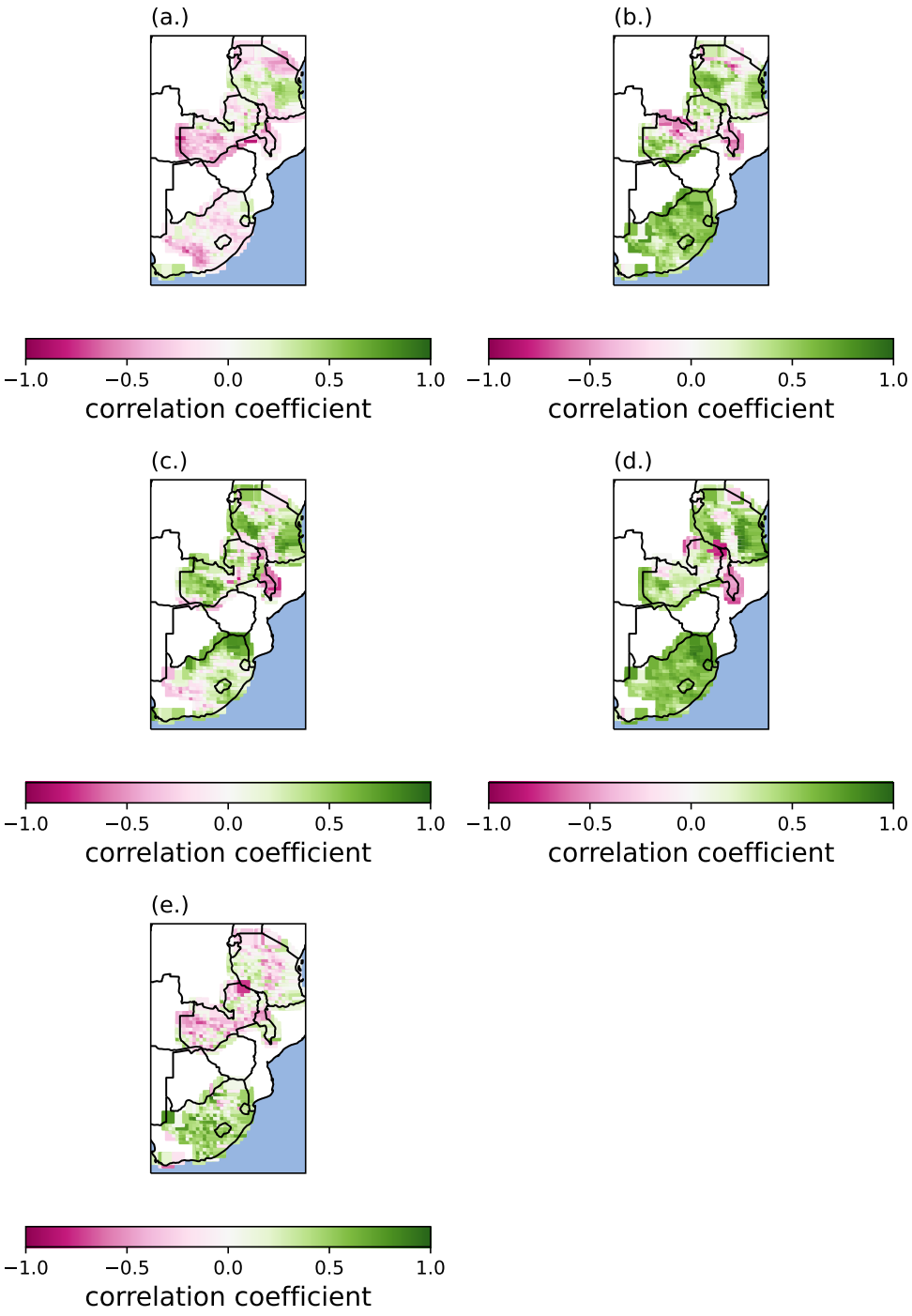

Figure 5.28: Pearson's correlation coefficient between the inter-annual variability in rainfall and maize yield for each grid cell location in the GDHY dataset. (a) Observed yields, (b) Random forest model, (c) Support vector machine, (d) Multiple Linear regression (d) GLAM

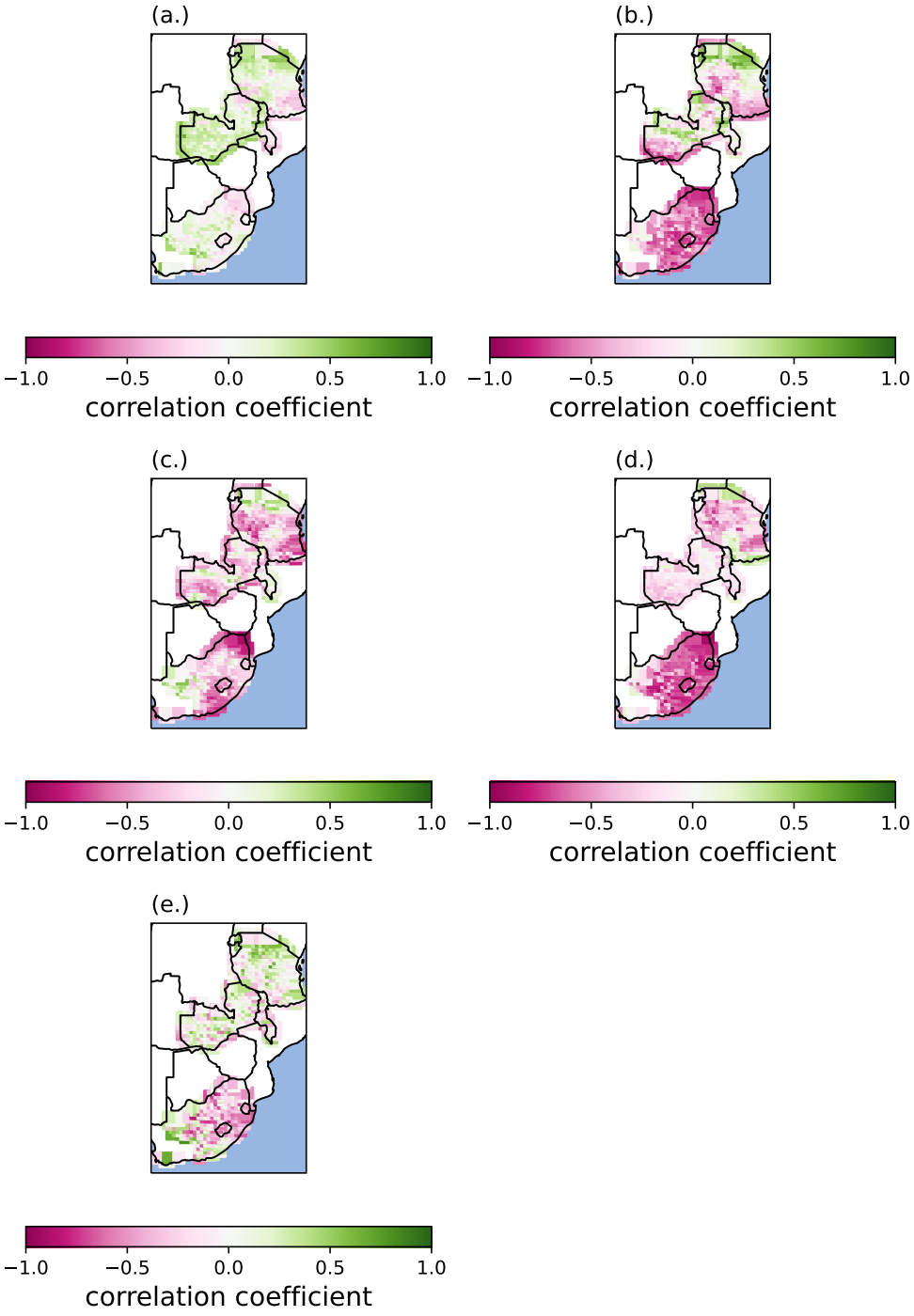

Figure 5.29: Pearson's correlation coefficient between the inter-annual variability in incoming long-wave solar radiation and maize yield for each grid cell location in the GDHY dataset. (a) Observed yields, (b) Random forest model, (c) Support vector machine, (d) Multiple Linear regression (d) GLAM

5.4.3.2 Part 2: Spatial correlations

For this chapter however, model performance is compared across average conditions to understand the effects of crop model calibration and parameterization on model performance. Therefore, it is important to focus on the relationship between climate and crop yields across space for this purpose. Figure 5.30 shows the relationship between average seasonal rainfall and maize yield. Although all models predict the strong positive correlation between maize yield and rainfall in South Africa, only the random forest and SVM ML models also capture the positive correlation in Malawi, furthermore, ML models also predict correlations with rainfall closer to observed in Tanzania. In Zambia there is little observed correlation between average rainfall and crop yield. Figure 5.31 shows very weak observed correlation between average daily maximum temperature and yield. The ML models predict a strong negative correlation between maximum temperature and yield in Zambia, although this is not present in the observed data.

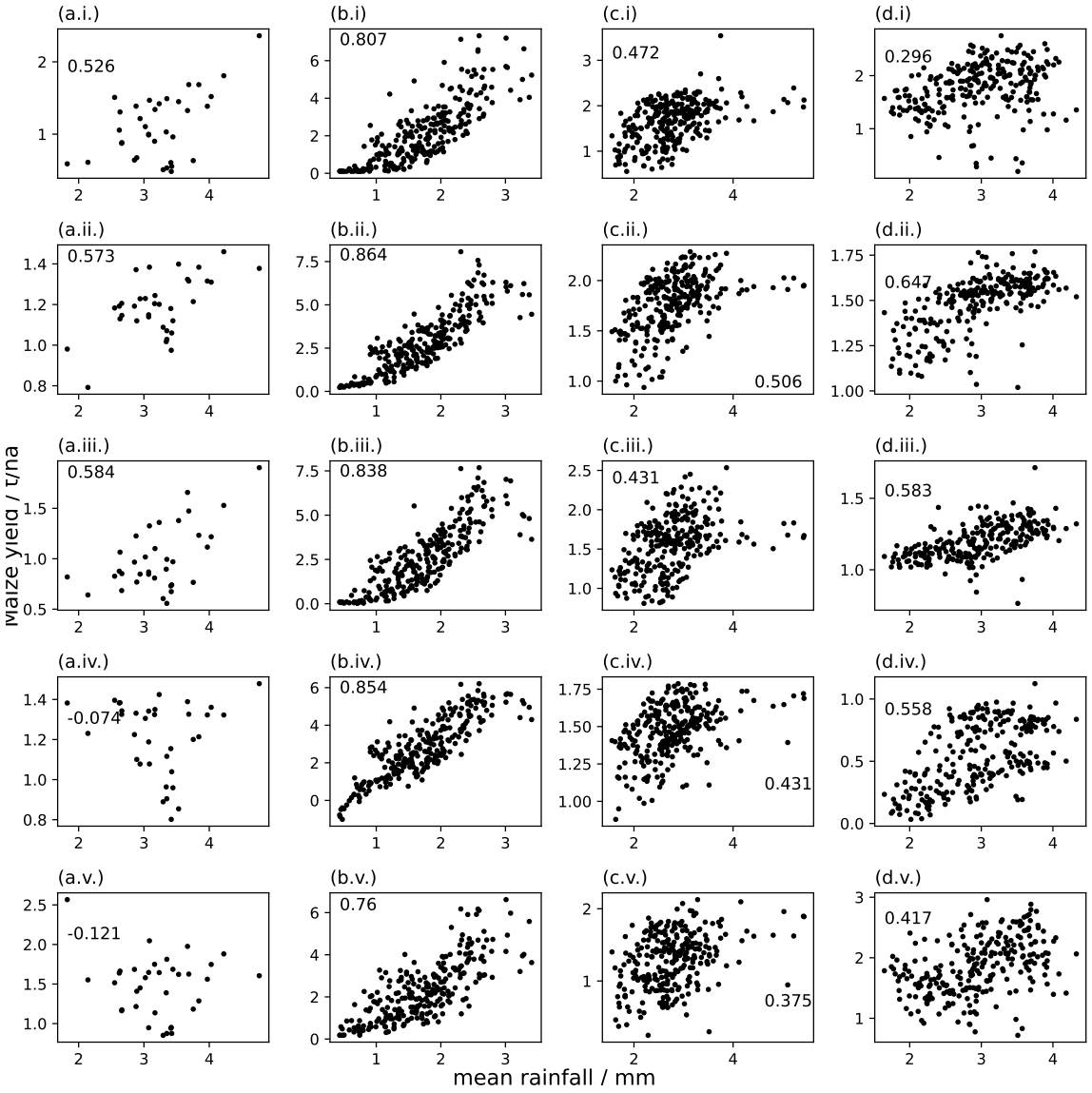

Figure 5.30: Relationship between rainfall and yield across countries and models, Each column represents a different country studied, namely: (a) Malawi, (b) South Africa, (c) Tanzania, (d) Zambia. The first row (row (i)) shows the relationship between observed mean rainfall and mean yield for each grid cell. Subsequent rows denote each model, namely: (ii) Random forest model, (iii) Support vector regression, (iv) multiple linear regression, (v) GLAM crop model.

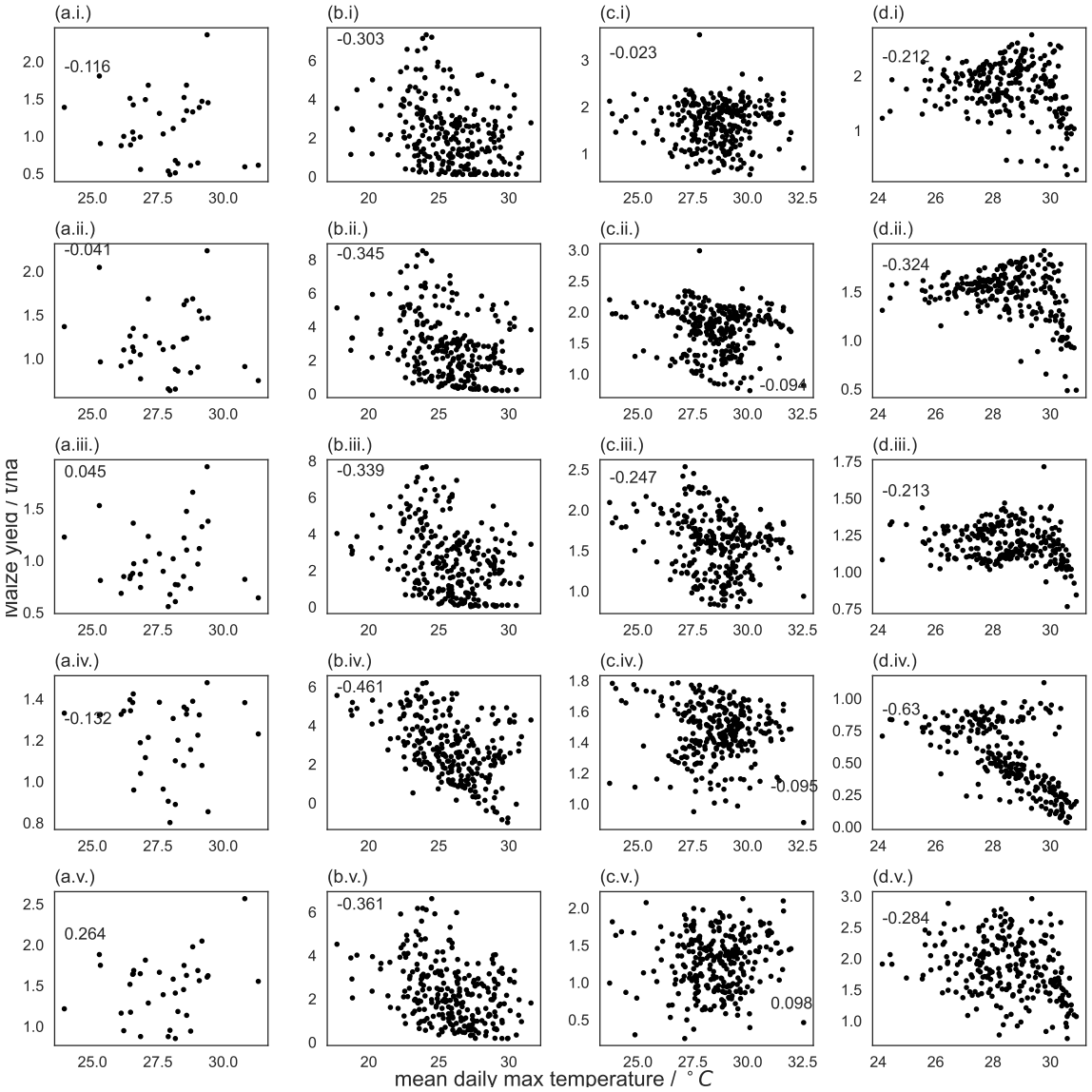

Figure 5.31: Relationship between average daily maximum temperature and yield across countries and models, Each column represents a different country studied, namely: (a) Malawi, (b) South Africa, (c) Tanzania, (d) Zambia. The first row (row (i)) shows the relationship between observed mean rainfall and mean yield for each grid cell. Subsequent rows denote each model, namely: (ii) Random forest model, (iii) Support vector regression, (iv) multiple linear regression, (v) GLAM crop model.

5.4.4 Climatological relationships are affected by model calibration

Relationships between observed and predicted yield will be affected by the scale of calibration between observed and modelled data. Hence, as a further test to the relationships shown in the above Figures, the correlation between predicted yield and observed rainfall was compared between a calibration which allowed the GLAM yield gap parameter to vary per grid cell, and a simulation in which one value of the YGP was used for the entire country. For an analogous comparison, machine learning models were compared with and without latitude, longitude coordinates as input features. Figure 5.32 shows the results of this comparison. Crucially, the correlation with rainfall substantially increases in all countries when the YGP is allowed to vary per grid cell. Conversely, the random forest machine learning model does not see as significant reductions in correlations with rainfall when the coordinates are removed as input features. Most notably, the correlation between GLAM crop yield and rainfall decreases from 0.76 to 0.53 when the yield gap parameter is no longer allowed to vary per grid cell in South Africa. These results show the significant effect the yield gap parameter has upon the correlation between predicted yield and rainfall.

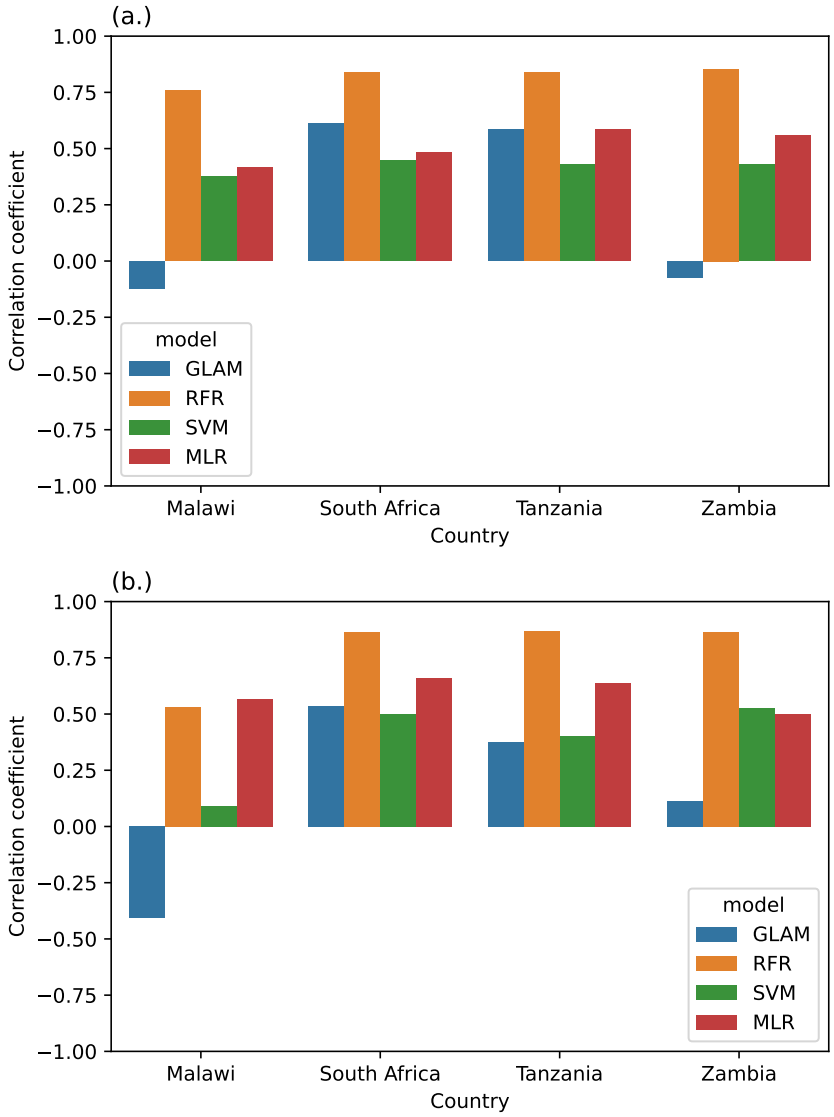

Figure 5.32: Correlation coefficient between predicted yield and rainfall for (a) simulations in which the yield gap parameter was allowed to vary per grid cell, and latitude longitude coordinates were included as input features to the ML models for an analogous comparison, (b) 1 value of the YGP per country and with coordinates removed as input features to the ML models.

Furthermore, it is not only the spatial correlation which is affected by the variation of the YGP, but also inter-annual correlation between rainfall and yield. Figure 5.33 shows the

comparison between model predictions when the YGP parameter is allowed to vary per grid cell (a) and kept constant per country (b). Notably, Results for Zambia produce more negative correlations between predicted yield and rainfall when calibration is per grid cell, and the correlation for some grid cells in South Africa is less positive.

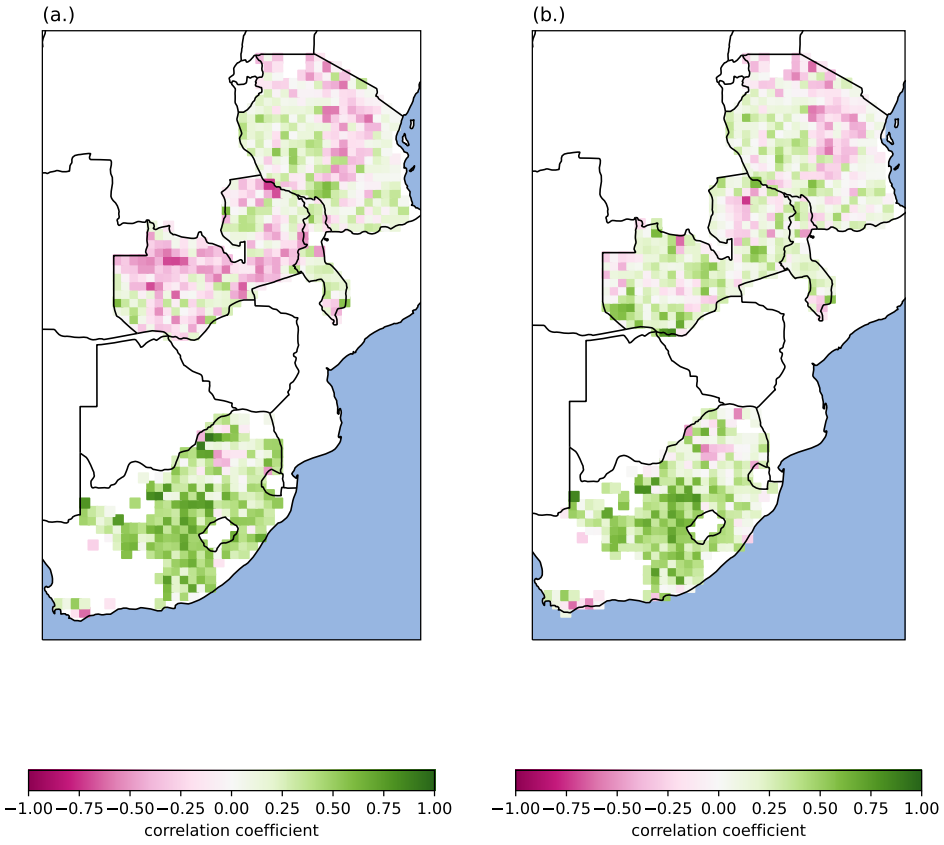

Figure 5.33: Correlation coefficient between GLAM predicted yield and observed rainfall for each of the grid cell locations across time. Panel (a) shows the correlation coefficient when the YGP parameter is allowed to vary for each grid cell, Panel (b) shows the results of the same correlation if the YGP is kept constant with 1 value per country.

The effects of input weather on predictions will affect model performance through the strength of the relationships on predicted yield. Where strong correlations exist between observed yield and weather, it is more likely to create a better model. However, if this is not the case, and stronger correlations in the observed data do not translate to improved model performance. This shows an opportunity for model improvement through a stronger representation of input/output relationships where they are strong drivers of observed yield.

5.4.5 Agro-climatic conditions affecting model performance

Figure 5.34 shows the correlations between each variable considered to assess the relationships between agro-climatic conditions and model performance. There are several relationships and patterns which are shown through this Figure. Firstly, the correlation between observed yield and rainfall has a stronger correlation with random forest model skill (measured by correlation coefficient) than GLAM model skill. This suggests that the strength of the relationship between rainfall and yield influences machine learning model skill more than GLAM model skill. This result suggests that GLAM performance may be improved by increasing the sensitivity of the model to rainfall variability. Conversely, both GLAM and the random forest model achieve poorer performance if the correlation between yield and temperature is more positive. This shows that at more negative correlations between temperature and yield, model correlation coefficients are greater. Complementing this analysis, Figure 5.35 shows the correlations between the relationships of rainfall and temperature with observed yield and model performance. Panel (a) of Figure 5.35 clearly shows the lack of relationship between GLAM crop model performance and rainfall-yield signal, in stark contrast to panel (b) of the same Figure which shows the strong positive correlation between the rainfall-yield signal and model skill. Furthermore, This Figure also shows that GLAM is more sensitive to temperature effects on yield than rainfall effects. Correlation between GLAM model skill and the temperature-yield signal is -0.47,

much stronger than the correlation with the yield-rainfall signal. In more detail, Figure 5.35 panel (a) shows that the GLAM correlation coefficient ranges from approximately 0.6 to -0.6 as the correlation between temperature and rainfall ranges from -0.6 to 0.6.

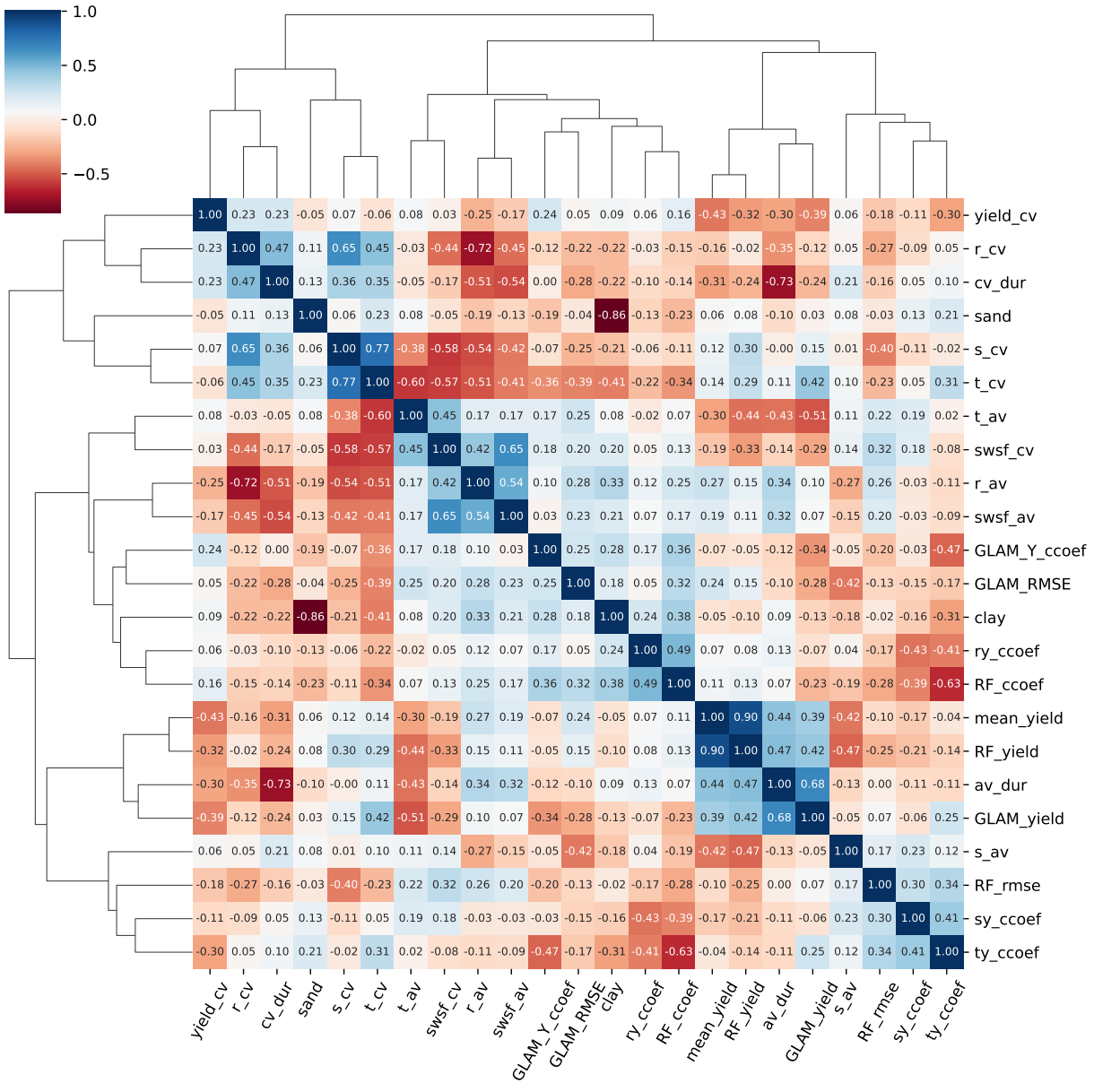

Figure 5.34: Correlations between each of the variables considered for the assessment of agro-climatic relationships on model performance and target variables RMSE and correlation coefficient. Total number of points was 857, a description of each of the variables is found in Table 5.3

Figure 5.34 also shows that GLAM RMSE is positively correlated with average rainfall,

temperature and the soil water stress factor, however negatively correlated with average solar radiation. The strongest correlations with GLAM RMSE are with solar radiation and rainfall, all indicating that GLAM RMSE is greater in wet conditions with less solar radiation. By comparison, the results of the random forest model do not show as a strong relationship between model RMSE and rainfall. Model RMSE is however positively correlated with a positive relationship between temperature and yield as well as the coefficient of variation in the soil water stress factor.

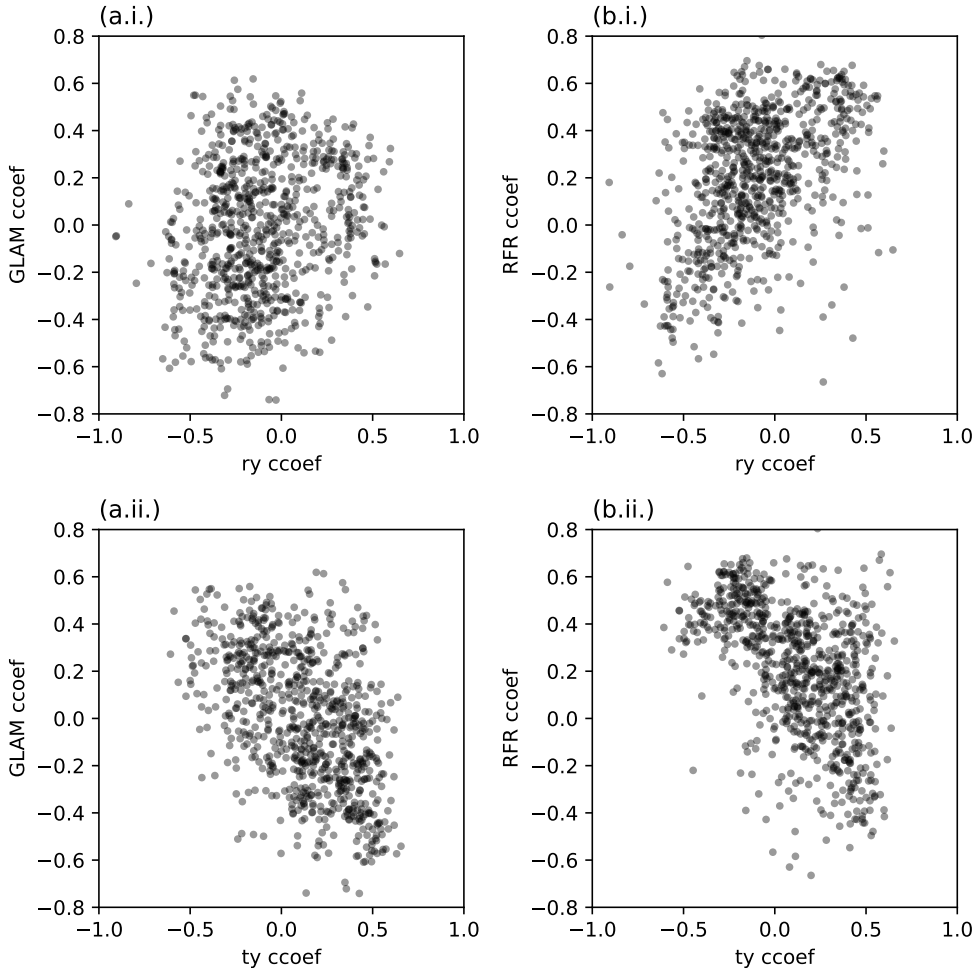

Figure 5.35: A comparison between model skill and the strength of effect of temperature and rainfall on observed crop yield. Panel (a.i.) and (b.i.) show the correlation between rainfall and observed yield plotted against the correlation between either GLAM predicted yield and observed yield (a) or Random forest predicted yield and observed yield (b). Panels (a.ii.) and (b.ii.) also show the model skill of GLAM and Random forest on the Y axes, but instead plotted against the correlation between temperature and observed yield.

5.4.6 GLAM model changes to improve the yield, rainfall correlation

Evidence from the previous sections show that GLAM is not as sensitive to the effects of rainfall variability as machine learning models. Furthermore, the strength of the relationship between observed crop yield and rainfall affects machine learning models much

more strongly than the GLAM crop model. However, GLAM model performance was affected by the strength of the relationship between temperature and yield. In GLAM, daily temperature variability drives model predictions through the effects of crop duration (number of days to crop maturity), Figure 5.36 shows the relationship between predicted yield and predicted crop duration. Duration has a significant effect on predicted yield. However, it is not clear that duration has such as strong effect on yield in reality. This Figure also shows the comparison between observed yield and predicted crop duration as a comparison. Unusually, there is no relationship between observed yield and predicted crop duration. Although it is expected that GLAM is overestimating the effect of duration (because the effect of rainfall is underestimated and there is a trade off between the two effects), it is also expected that some relationship should exist between observed yield and modelled crop duration (Asseng et al. 2015).

It is worth noting that simulated durations shown in Figure 5.36 can be quite high in places. South Africa having the highest durations. Although durations over 200 days are not that common, this can occur and be due to sub-optimal temperatures which persist throughout the growing season. Although grid cells simulated were masked by growing area, this still included some grid cells which are not part of the key maize growing region in the south east throughout this period. Therefore, it is not unexpected that some locations may provide longer durations than usually expected.

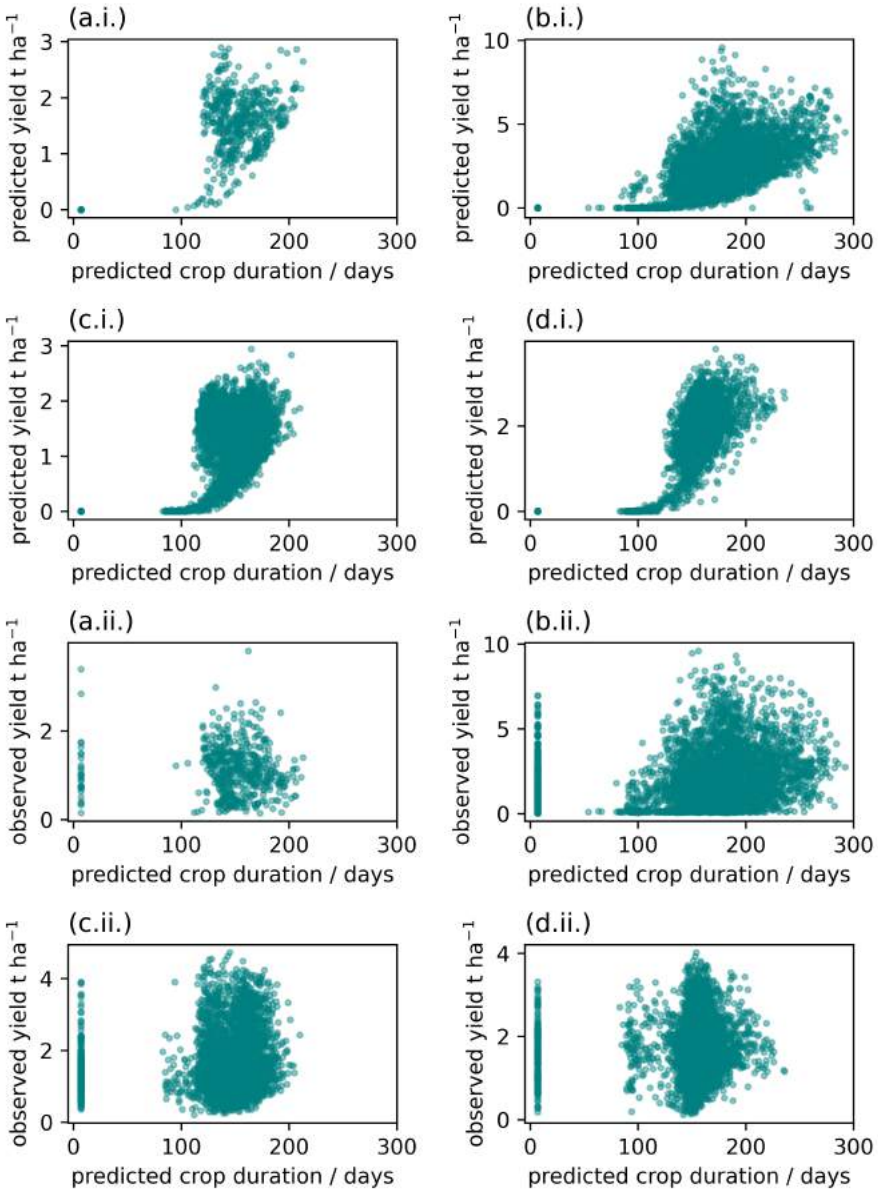

Figure 5.36: Relationship between the crop yield and predicted crop duration in number of days from planting to harvest. Results are shown for each country individually (a) Malawi, (b) South Africa, (c) Tanzania, (d) Zambia. Panels with numeral (i) denote that predicted yield is on the Y axis, whereas panels with numeral (ii) denote that observed yield is on the Y axis.

As a sensitivity analysis experiment, the effect of crop duration on yield was removed from

the GLAM model. This was achieved by setting the crop growth stage to change manually at specified dates throughout the growing season regardless of how much temperature is accumulated. This resulted in a growing season which lasts 110 days, with each growth stage lasting either 30 or 20 days leading up to the fixed harvested date. The choice of the growing season length is arbitrary as the purpose of the test is to determine if removing the variation in crop duration will improve the correlation between predicted yield and rainfall. It was also necessary to remove several model processes which interact with the timings of different crop growth stages or cause early harvest such as the TRKILL parameter which causes early harvest if temperatures exceed the lethal limit, and model processes which account for heat and drought stress around flowering (HTS, WS paramters). Crucially, the effect of the yield gap parameter was also removed by setting the value to 1 in all grid cells. This ensured that no reductions in predicted yield were made based on observed values of yield. The results of this test are shown in Figure 5.37. The test showed that in fact, fixing the simulated crop duration in GLAM did not have a significant effect on the correlation between GLAM model skill and rainfall correlation. More prominent are the effects of the yield gap parameter, when uncalibrated, GLAM model skill is more dependent on the correlation between rainfall and yield.

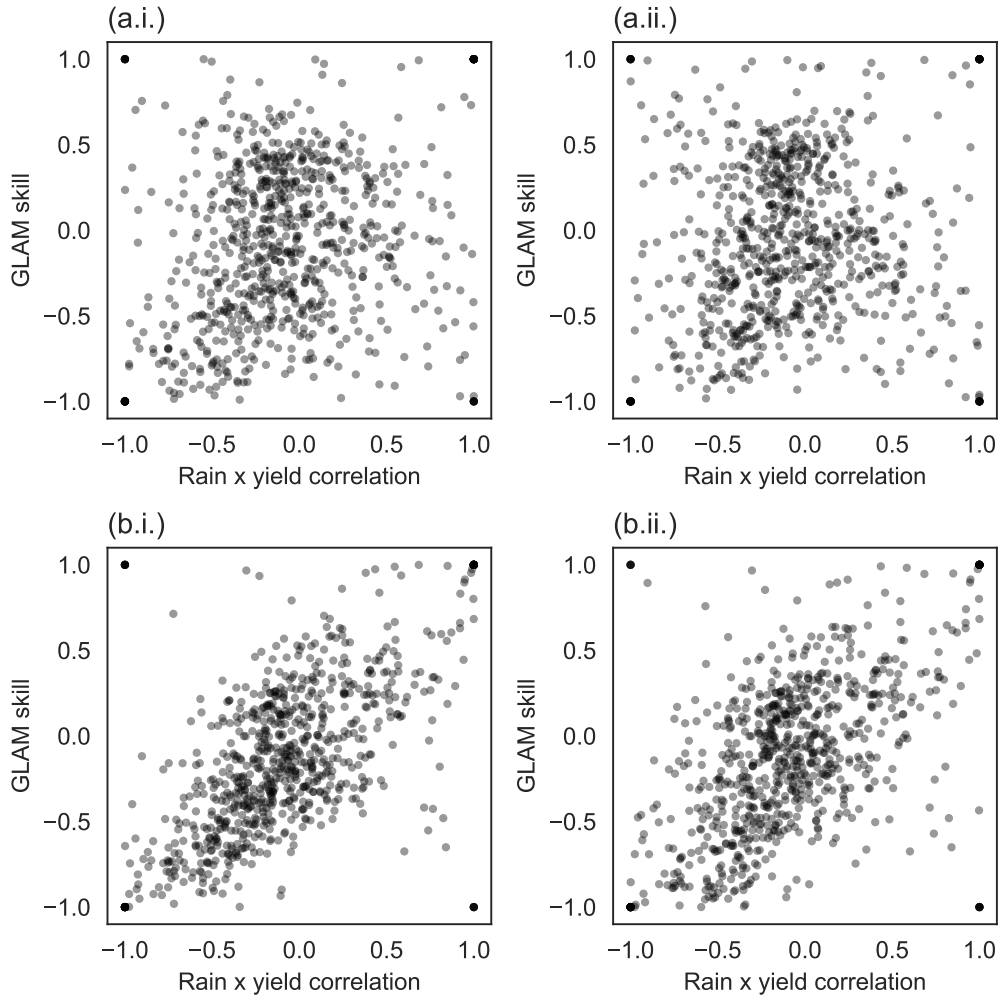

Figure 5.37: GLAM model skill against the correlation between rainfall and observed crop yield. Panels (a.i.) and (a.ii.) denote GLAM model simulations with calibrated YGP per grid cell, (a.i.) is the control simulation in which the duration is allowed to vary as normal (a.ii.) is the fixed duration simulation. Panels (b.i.) and (b.ii.) show the same results but with the YGP parameter left uncalibrated.

In summary, the results of this test show that although the effects of rainfall on simulated yield in GLAM may still require better representation, the answer is unlikely related to the effects of duration being too dominant on simulated yield. However other parameter interactions may likely affect the effect of rainfall x yield correlation on model skill. This is

shown by the large effect the yield gap parameter has on the correlation. Further research should seek to understand the effects of parameter interaction on the correlation between model skill and correlation with rainfall signal.

The initial fixed duration test lead to the idea that improvements in the accuracy of the GLAM soil water balance through more accurate prediction of soil moisture characteristics may be the factor which instead improves the correlation between predicted yield and rainfall. Therefore, an alternate method of calibrating the model was used which adjusts the YGP parameter in order to reduce the available soil water rather than leaf area index (see methods section 2.1). This method to calibrate the GLAM crop model is not commonly used, and does not result in improved model performance. Figure 5.38 shows that correlations between rainfall and simulated crop yield increase when adjusting the soil moisture parameters.

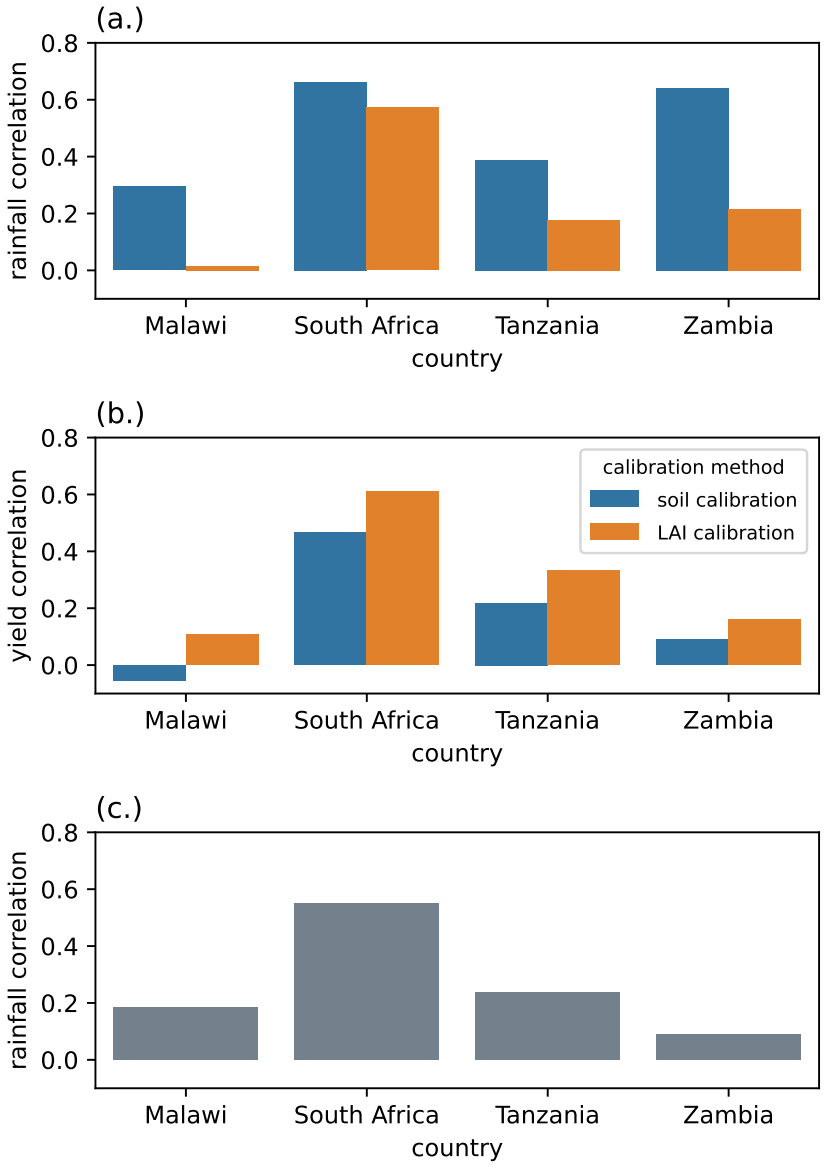

Figure 5.38: Panel (a) Correlation coefficients between rainfall and simulated crop yield between the default LAI method of calibration, and the soil moisture method of calibration. Panel (b) Correlation coefficients between observed and simulated yield using both of the calibration methods. Panel (c) Correlations between observed yield and rainfall. All correlations are across both time and space.

The result of this calibration also lead to an increase in the correlation between GLAM

model skill and the relationship between rainfall and observed crop yield (the rainfall-yield signal). Figure 5.39 illustrates that the soil moisture calibration improves the correlation between rainfall and simulated yield from 0.208 to 0.467.

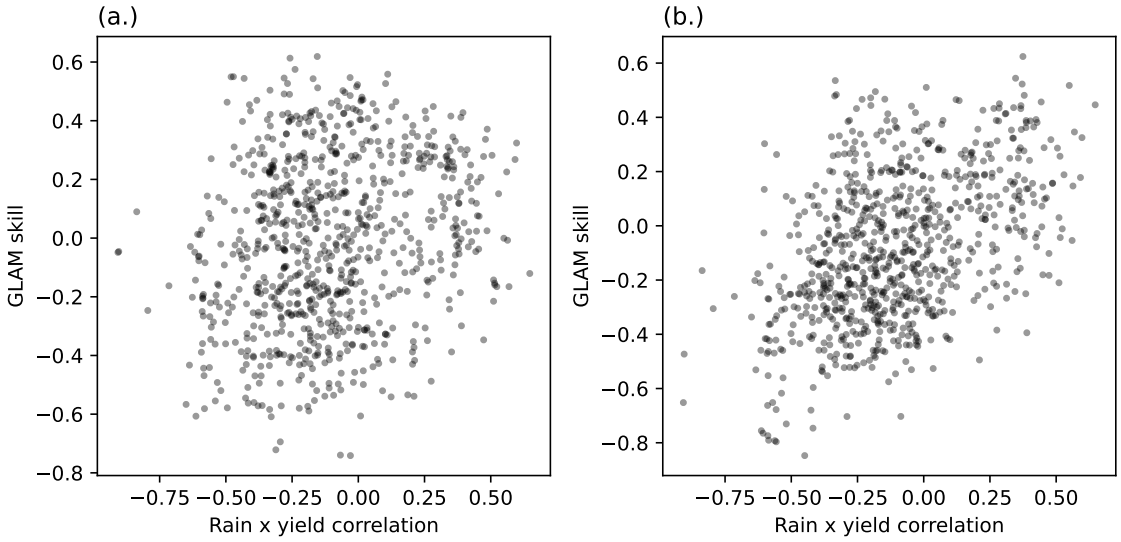

Figure 5.39: GLAM model skill against the correlation between rainfall and observed crop yield. Panel (a) is the control simulation in which the YGP is calibrated against the effect on LAI. Panel (b) shows the results from the method of calibrating the model using the YGP to reduce the water holding capacity of the soil (SOLYGP). Panel (a) pearson correlation is 0.208, Panel (b) correlation is 0.467.

Although improving the representation of soil moisture improves the correlation between rainfall and crop yield, the method used to do so does not improve crop model skill. This is because the calibration method was replaced and so LAI was no longer optimized. Since there is a larger correlation between LAI and crop yield (in comparison to soil moisture and crop yield) this leads to a decrease in model performance. Therefore, an independent method of improving soil moisture parameterization whilst keeping the existing method of calibration is required. A method to to this explored in appendix 9. In this appendices, soil moisture characteristics are predicted using machine learning rather than the existing pedo-transfer method. These additional results show that machine learning has the po-

tential to improve crop modelling through improved parameterization of the soil moisture balance routine.

5.5 Discussion

The results of this chapter present implications for GLAM sensitivity to different climatological yield drivers and calibration. In particular, the role of temperature and rainfall. The yield gap parameter is also shown to have a significant effect on the relationship between rainfall and predicted yield, more so than coordinates used as an analogous proxy in machine learning models. This has implications for the representation of the relationship between rainfall and predicted yield in a changing climate. These aspects of model behaviour were analysed by using machine learning models as bench-marks. Comparing the models in this way shows the potential for model improvements such as increased sensitivity to rainfall, as well as providing a comparison to show the large relative effect of the yield gap parameter. The results also show that different machine learning models may give different relative importance to input variables whilst achieving the same model performance.

5.5.1 Mechanistic knowledge for machine learning and equifinality

The results in section 5.4.2 demonstrate that machine learning models can achieve good model skill for entirely different reasons. Although both the support vector machine and random forest model achieved high correlation coefficients in South Africa, the Principal components which were found to be most beneficial to model performance varied. Since Principal components were used as inputs rather than correlated variables, this demonstrates entirely different information may be most influential for model performance depending on the model architecture used. To expand on this, and further test this assertion, future work could use of a machine learning model ensemble. Each model in the ensemble could be used to 'vote' upon which set of variables or mode of variation in the most im-

portant, with weighting by model importance. If the voting is consistent across datasets then this could provide an answer as to which mechanistic inputs are most valuable to the machine learning models. However, careful choice would have to be made to select which model frameworks would populate the ensemble to ensure a lack of bias against each specific model structure, for example, some models structures such as random forest and gradient boosting algorithms are similar and so would be more likely to give similar answers. This idea is similar to the study by Rudin (2019) who grouped machine learning models into similar model structures and determined feature importance for different groups of models.

The results of this chapter broadly agree with the conclusions of Lischeid et al. (2022) who also demonstrated that similar model performance can be achieved using alternative predictors. This study by Lischeid et al. (2022) also compared the random forest and support vector machine models to come to this conclusion. The use of principal components as predictors in this chapter allows the additional statement to be added to the conclusion that even with no redundancy between predictors (in which the same or similar information is provided to the model by multiple predictors) models will still achieve similar results with different predictors.

The reasons behind the results of the feature importance analysis are likely due to the problem of equifinality for models in general, but in this case machine learning models. Equifinality is the principle that the same outcome can result from different potential causes (Beven 2006). In this case, very similar model performance can result from different principal component inputs. Since the principal component inputs are transformed from components of weather variability, it may not be surprising that different components of weather variability can lead to very similar predictions of crop yield.

However, differences between the importance placed on different features may also be due

to differences in model structure. A potential hypothesis is that features may differ in importance depending on the quantity of data used for training. If so, this may lead to differences in feature importance, as support vector machines use only a small subset of the training data to fit the model function and make predictions (determined by the support vectors). Hence, if the importance of variables is different for the data used to construct the support vectors from the entire dataset, then this may lead to differences in feature importance.

5.5.2 Drivers of observed and modelled yield responses to variability in climate and effects of calibration

Machine learning models are more sensitive to climate variability than the GLAM model. Furthermore, GLAM is more sensitive to temperature variability than rainfall variability. Hence, the correlation between rainfall and yield for machine learning models is more similar to that of observed rainfall and yield. This is also positively correlated with model performance. This leads to the conclusion that the GLAM model should be more sensitive to rainfall variability in an effort to improve model performance.

The YGP parameter was shown to have a significant effect on both the correlation between rainfall and simulated yield within space, and the inter-annual correlation between rainfall and simulated yield. For spatial correlations this is less surprising as it would be expected that mean yields should correlate in space with rainfall magnitudes and so by proxy, the yield gap parameter improves the correlation with rainfall. However it is more surprising that the YGP could have a significant effect on the inter-annual correlation with rainfall as the YGP parameter is static throughout time. However, a possible explanation for this is that the yield gap parameter may affect the bias variance trade off over time between observed and predicted yield. That is, that as the YGP parameter corrects predictions to better capture mean yields, the calibration may reduce the variance of the yield predictions.

Hence, this may affect the correlation with rainfall. These strong effects of the YGP have significant implications for future projections made using a crop model calibrated using an empirical correction factor. For spatial variability, the significant effect of the yield gap parameter results in the correlation between rainfall and crop yield to be based more on location in space, and so less on the actual processes which govern the relationship between rainfall and yield such as evapotranspiration. Therefore, given that rainfall patterns are known to change, and are expected to do so under climate change (Kruger & Nxumalo 2017, Fauchereau et al. 2003, Dunning et al. 2018, Chapman et al. 2020), the resulting yield projections will be in error. This problem could also be described as over-fitting, as the model is over-fit to the current spatial pattern of yield and rainfall. By contrast, machine learning models were less affected by the removal of the latitude and longitude input features, which could be described as analogues for the YGP. This implies that the models better represent the effect of rainfall on yield regardless of location, and so are more likely to generalize and better represent the relationship between rainfall and yield. This over-fitting problem may also affect the ability to predict climate impacts from the effects of El Niño, also due to changing rainfall patterns from climate oscillations (Dore 2005, Kruger 1999). Therefore, calibration should take account of the distribution of El Niño years.

5.5.3 significance of calibration effects for other crop models

These results may be significant for other crop models which also use calibration to correct for spatial differences in yield (most if not all crop models do this to some extent). Ultimately, this problem is an issue of appropriate complexity. With greater parameter interaction, calibration may have unintended effects. Therefore, future work should aim to determine how varying complexity of models may affect inter-annual relationships between yield and weather depending on degree of calibration. GLAM is a relatively simple model (in comparison to some other crop models such as APSIM (Keating et al. 2003)),

and so such effects may be more pronounced in more complex models.

5.5.4 Further work required to improve crop model performance

The results of this chapter show that rainfall effects in GLAM are mediated by the effects of the YGP, and likely need improving. In environments in which rainfall is a significant driver of crop yields this is likely to be beneficial for model performance. Therefore, future crop model development needs to focus on the evaluation of the correlation between rainfall and simulated yield under different environmental conditions. Since it is unlikely that a larger effect of duration is the cause of the weaker correlation with rainfall, this may be due to the representation of the soil moisture characteristics. If the soil moisture is inaccurate, this may cause a reduction in the correlation between simulated yield and rainfall.

Although duration was shown to not have a mediating effect on the correlation between simulated yield and rainfall, simulated crop duration was particularly poorly correlated with observed crop yield, indicating that the importance of duration on yield simulations may be too great. If importance of duration for crop yield predictions is to be reduced, fixing duration is unlikely to be a permanent solution. For such a solution to be devised, either effect of temperature on duration could be reduced or effect of duration on yield could be reduced. Reducing the effect of duration on yield could involve altering the rate of change in growth depending on growth stage. Conversely, the effect of photo-period on crop growth could be increased.

Improving the accuracy of the soil water balance was shown to provide an improvement in the correlation between simulated yield and rainfall. However this did not necessarily lead to an improvement in model performance. It is argued that this is because the method to improve the accuracy of the soil water balance (through calibration) replaced the default calibration method which affects the simulated leaf area index. This default method is more strongly correlated with the output yield and so optimizing based on

this variable will have a more prominent effect on model performance against observed yield. Therefore, a different method should be used to improve the prediction of soil moisture parameters. Appendix 9 shows that machine learning can be used to improve the prediction of soil moisture parameters. Therefore, further work is required to integrate the machine learning method of soil moisture parameter predictions into the GLAM model to test for improvements in model skill as well as improvements in the simulated yield-rainfall correlation. As well as improvements in overall performance, the integration of such a method would improve the spatial applicability of the GLAM model. Currently, pedo-transfer functions used by crop models are used outside of the spatial ranges in which they were developed, this results in poor model performance in certain regions. For example, the Saxton & Rawls (2006) function was not developed for high clay soils (clay % above 60) or soils with high organic matter (OM) content (OM above 8%). This results in poor performance in regions with high clay soils such as Brazil (Demattê et al. 2019). Therefore, greater value is added to this potential further work by testing the new method for both general performance and in regions with high clay soils which may be better simulated by integrating machine learning methods.

A potential hypothesis which arises from the results in this chapter, is that the importance of rainfall, temperature and crop duration for the determination of crop yield may differ depending on spatial scale. Therefore, crop models which are developed and parameterized at the field scale (Challinor et al. 2018) may be biased against relationships which are more important at the regional scale. Figure 5.36 revealed little relationship in the observed data used between crop yield and the predicted crop duration by GLAM. Furthermore, rainfall correlated strongest with observed yield (especially in South Africa and Tanzania). Some compelling evidence for this is discussed by Iizumi et al. (2014c), who found that parameters representing the yield response to temperature showed a dependency on spatial scale. If it can be determined that drivers of crop yield variability

vary in importance at different spatial scales, a potential solution may be that crop models need a specific regional scale set of parameterizations or calibrations which place greater emphasis on model processes shown by sensitivity analysis and evaluation against regional scale observations to be more important.

The role of calibration must also be assessed, especially if the calibrated model is to be used for future projections. The yield gap parameter is shown to have a significant effect in some cases on the relationship between rainfall and crop yield predictions, as discussed in the previous section this may lead to model over-fitting. This result lends credence to the approach of some global studies (e.g. Jägermeyr et al. (2021)) which do not use absolute yields to determine impacts of climate change and instead use relative measures of yield decline. This ensures that the analysis is not biased towards the current spatial variation in climate or crop yield relationships. An interesting future study may be to assess the impact of calibration on yield - rainfall relationships for a suite of gridded crop models. The purpose of which would be to determine if over-fitting in crop models due to calibration occurs across an ensemble of models.

5.5.5 The role of machine learning for crop model improvement

This chapter demonstrates multiple ways in which machine learning can be used for crop model improvement. The main focus is bench-marking crop model performance to diagnose the conditions which are most detrimental to model performance and the relationships between input and output variables. Fundamentally, bench-marking demonstrates that poor crop model performance is not due to lack of sufficient data. In using this knowledge, it can be determined which model processes require improvement based on the comparison. Secondly, also presented (in appendix 9) is a method to improve prediction of soil moisture characteristics. This is presented as an example of how machine learning may replace a model process which is already based on a statistical correlation, and so

process knowledge is not lost by replacement with a method which is usually known for being less interpret-able than a process model.

Bench-marking is shown as a way to understand if sensitivity to climate information is required to increase, and which drivers require increased sensitivity. The combination of increased model sensitivity to a particular climatic predictor and improved model performance indicate that an increase in the sensitivity to said predictor may also improve model performance. To assess the conditions under which machine learning improves upon the performance of a mechanistic model, and then to assess the modelled and predicted relationships between the climate and crop yields can help to understand which agro-climatic relationships are most important to replicate most effectively. Of course, due the spatial heterogeneity of limiting conditions which affect crops (Sacks et al. 2010), as well as effective management and climate stressors, (van Bussel et al. 2015a, Teixeira et al. 2013) the relationship between climate and observed crop yield will vary in space and so will require different bench-marking assessments depending on the dataset.

Furthermore, a method which has potential to improve process based modelling is demonstrated in appendix 9. It is demonstrated that soil moisture characteristics can be better predicted using machine learning than empirical correlations used by crop models currently at the regional scale. Further work should be undertaken to understand the effect of incorporating machine learning to predict soil moisture characteristics on the overall model performance and correlation with rainfall. If correlation is improved through this method, it may enable an increased correlation with rainfall and better generalization of the rainfall relationship with yield.

A further use for machine learning in crop model improvement may be to downscale crop yield predictions. Predictions from GLAM at the country scale (1 set of predictions for the entire country) are shown by comparing this study with Jennings et al. (2022) to

have better model skill than gridded predictions. Furthermore, calibration, used to adjust crop yield estimates to sub-country gridded observations, can affect the relationship with rainfall, potentially leading to over-fitting. The correlation between machine learning predictions and rainfall is shown in this chapter to be less affected by spatial location. Therefore, using machine learning to produce a spatial distribution of crop yield estimates, down-scaled from country level crop yield estimates may produce results less prone to over-fitting to the current spatial distribution of rainfall. Using machine learning to downscale crop model predictions is shown in Folberth et al. (2019) to produce predictions of high performance.

5.5.6 Limitations and uncertainties associated with the yield dataset

It is demonstrated in section 5.4.3.1 that there are a large number of grid cells in which positive correlations exist between solar radiation, temperature and observed crop yield. subsequent negative correlations between rainfall and observed crop yield are also identified. Although in some circumstances solar radiation may have a positive correlation with yield (for instance, in regions which typically have temperatures below the optimum growing temperature for maize (Prasad et al. 2017, Sacks et al. 2010)) This is unlikely to be the case in dry regions of South Africa. Furthermore, although flooding due to excess rainfall can have detrimental impacts on crop yields, rainfall is generally associated with a positive correlation with crop yield.

A possible cause of this is that yield estimates in the GDHY dataset are derived from NPP (Net primary productivity). In turn, NPP is determined by Photo-synthetically active radiation (Iizumi et al. 2014b). This will likely have a positive correlation with solar radiation which in turn has a negative correlation with rainfall. However, this limitation does not have significant impact on the conclusions of this chapter as the focus is placed upon the spatial relationship between climate and crop yield to be considered when bench-

marking the models, which is consistent with expected relationships between temperature, rainfall and yield.

One source of uncertainty which requires addressing is the question of whether the yield data may be more sensitive to weather than that of true observed yield data and so could be the reason why ML models trained on this data are more sensitive to rainfall effects than GLAM. To mitigate against this potentiality, results can be compared to the results found in chapter 3. In chapter 3, GLAM model performance was significantly worse for a year in which excess rainfall lead to increases in crop yields. This effect was not captured by GLAM, and so this result therefore backs up the finding using the GDHY data in this chapter that GLAM is not sensitive enough to rainfall effects which leads to underestimation of yields.

5.5.7 Potential reasons for poor model performance in Zambia

Across both ML models and GLAM, model performance in Zambia is poor. Since both model types performed poorly, it is more likely that poor model performance is due to poor correlation between input climate variables and observed crop yield. Furthermore, also due to similarity in performance, the reason is unlikely to be due to lack of representation of important processes for yield, or over-parameterization. GLAM incorporates more information than the machine learning models, such as predicted crop variety, thermal time requirements of maize, potential and water limited evapotranspiration, vapour pressure deficit, and soil moisture information. In contrast, machine learning inputs are much simpler, namely, minimum and maximum daily temperature averaged over each growing season, rainfall, solar radiation, and dry day and hot day threshold indicators. For both model types to perform poorly requires that adding more information (from GLAM) would be unlikely to improve the machine learning methods, and simplifying the crop model parameterization is unlikely to enable stronger performance either.

Observed correlations between rainfall, temperature and crop yield across space are shown to be poor in section 5.4.3. This is most likely the reason why models performed poorly in Zambia. If a second dataset becomes available for the country this would be most beneficial for confirming if this is the reason for poor model performance.

5.5.8 Recommendations for GLAM crop modelling

New insights useful for GLAM crop model development are shown in this chapter. Firstly, it is shown that correlation between simulated yield and rainfall is too insensitive in some cases. This is likely due to inaccuracies of the prediction of soil moisture characteristics. Future GLAM simulations should use new methods to predict soil moisture characteristics to enable an improved correlation between rainfall and crop yield. It is shown that machine learning methods can outperform pedo-transfer correlations for the prediction of soil moisture characteristics, not just here, but studies from the literature have also shown that machine learning is useful for this purpose (Gunarathna et al. 2019, Amanabadi et al. 2019, Pham et al. 2023, Lamorski et al. 2008). Therefore machine learning should be integrated into the GLAM crop model by using ML methods to predict soil moisture characteristics (RLL, DUL, SAT). The importance of more accurately predicting soil moisture characteristics cannot be overstated, particularly due to the effects of such predictions on the rainfall-simulated yield correlation, and the large uncertainties associated with soil data (Folberth et al. 2016).

Future work which makes use of long term climate projections should use model simulations which do not calibrate the yield gap parameter (at least as a control). This is to check for spatial over-fitting potentially introduced by the calibration of the YGP. This is because if rainfall patterns significantly change (Dore 2005, Kruger 1999) this will affect the observed spatial relationship between weather and crop yield therefore meaning potential errors resulting from calibration.

Furthermore, the effect of duration on yield should be assessed at the regional scale. Although it was found that duration does not significantly effect the relationship between simulated yield and rainfall, the effect of duration on simulated yield may be too great in comparison to observed yield at the regional scale. Crop duration data is very difficult to obtain at the regional scale. However, a future potential study could involve the use of surveys to poll farmers as to the crop durations typically experienced in different regions. This data could then be used to assess regional scale duration sensitivity.

5.6 Conclusions

The bench-marking comparisons between GLAM and machine learning under different climatic conditions reveal that, due to the larger sensitivity of machine learning to spatial variations in climate, particularly rainfall, machine learning is often able to better predict crop yield. Machine learning methods do benefit from process knowledge outputs from crop models, however it is difficult to ascertain which specific variables from crop models are most important for machine learning. This is particularly due to the fact that different model architectures place importance on entirely different variables, even when input variables are orthogonal in their modes of variation. This could be due to differences in model structure, however, equally, the problem of equifinality may be a cause. To improve crop models, machine learning has many applications where data quantity allows. However, to apply machine learning, process understanding gained is often difficult to interpret. Furthermore, conflicting results from different architectures could lead to spurious conclusions about reality if only a single model architecture is used. For this reason, it is recommended that any approach which seeks to gain process understanding from machine learning should use an ensemble of model architectures.

Information from bench-marking is able to identify targets to improve crop models. Specifically, this chapter showed that the the effects of rainfall on simulated crop yield is requires

improvement. Therefore, this raises the question of whether crop models are too sensitive to the effects of temperature and duration on yield. Through the sensitivity test described and undertaken in section 5.4.6 it is shown that although the effect of duration on simulated yield is likely too strong, this does not significantly affect the relationship between simulated yield and rainfall. A likely improvement to crop model skill may therefore be an improvement in the accuracy of the simulated soil water balance, appendix 9 proves that this is possible with machine learning, it is also shown that improved prediction of soil moisture will improve the correlation between simulated yield and rainfall (shown in Figure 5.38). Further work is required to integrate machine learning into GLAM for the prediction of soil moisture characteristics in order to demonstrate the value of doing so across a range of soil conditions.

The potential of the bench-marking approach relies on the model performance of the machine learning methods, which in turn relies on the quantity and quality of observed data. In some countries (e.g. Zambia), lack of a large quantity of high quality data reduces the effectiveness of the machine learning models and hence the ability of the models to provide useful benchmarks. Any further research leading from this chapter should seek to acquire a larger dataset of high quality observed data.

Although this chapter is an analysis of the GLAM crop model against machine learning, the results have implications for all crop models which use thermal time to drive crop growth and so show a strong correlation between duration and yield. Models which show such relationships should be evaluated against machine learning models in rainfall driven environments to assess whether the effect of duration on crop yield is over-emphasised. Furthermore, the relationship between simulated crop yield and rainfall in other crop models should be evaluated against machine learning in a similar manner. This can be used to assess whether other crop models also show similar insensitivity to rainfall in some environments. The results also have implications for crop model calibration, as it is shown

that calibration against observed yields affects the relationship with rainfall which may lead to over-fitting.

5.6.1 Novel contributions from this chapter

From this chapter, the following contributions are made to the scientific literature:

- Machine learning is presented as a method to benchmark a crop model in order to learn how to improve crop model parameterization and results
- Through bench-marking with machine learning, it is shown that the GLAM model is not sensitive enough to the effects of rainfall on crop yield, this may be the case for other models as well
- It is shown that different machine learning architectures (e.g. Random forest models and support vector machines) can provide very similar performance at crop yield prediction but using entirely different variables
- YGP Calibration is shown to have a significant impact on simulated crop yield-weather relationships
- A new method to predict soil moisture characteristics using machine learning is presented and a comparison is made against the methods used to predict the same variables within process based crop models (analysis found in appendix 9).

This chapter demonstrates that through bench-marking of crop yield estimates from GLAM and machine learning models, the appropriate level of importance to place on the effects of rainfall and temperature can be determined. Machine learning models are consistently more sensitive to climate variations than the GLAM crop model, and so therefore, if increased sensitivity to a particular driver of yield such as rainfall is beneficial for machine learning model performance, then this should also be beneficial for the crop

model. Generally, the machine learning models are quicker and easier to apply than the crop model, and so this brings an additional benefit as a simple test to understand if the crop model should be more sensitive to a particular driver.

Although (Lischeid et al. 2022) have also shown that random forest models and support vector machines can give similar model performance when using different predictors, this chapter further builds upon this finding. By using Principal components as inputs to the feature importance algorithm, it is demonstrated that machine learning models can predict crop yield with similar degrees of skill using different predictors which are uncorrelated. By doing this, the possible explanation that correlated features may cause these differences can be disregarded. Therefore, a new recommendation can be made that any new insight into the factors which are important for crop yields from machine learning must take into account the differences between ML models.

Furthermore, in showing that machine learning has the potential to outperform soil moisture pedo-transfer functions, used to predict soil moisture characteristics by crop models, it should be recommended that crop models should use machine learning to predict these parameters rather than empirical pedo-transfer functions. Further weight is added to this statement because pedo-transfer functions use empirical coefficients which are not based on theory or knowledge and so do not offer insight which cannot be provided by machine learning.

6 Discussion

The results and discussion of each of the chapters in this thesis present common themes which are explored in this discussion. Machine learning is presented as an alternative to process based crop modelling. In this respect, ML methods are leveraged to understand the strengths and weaknesses of both approaches, and to present methods to improve crop modelling using the GLAM crop model. ML methods are further explored to also better understand model sensitivity to potential errors in climate data. The results in this thesis are useful for 2 key reasons. Firstly, as a case study look into how machine learning methods compare, and may be used to further improve crop modelling methodologies. Secondly, a comparison of machine learning methods, the strengths and weaknesses of different methods, and the sensitivity of machine learning methods is shown. Broadly, many of the results in this thesis build upon the few recent studies which use process based crop models with machine learning (Feng et al. 2019, Shahhosseini et al. 2021, Leng & Hall 2020, Droutsas et al. 2019). Feng et al. (2019), Shahhosseini et al. (2019) both showed that process based information can improve the prediction of machine learning models, however, here it is shown that the way that ML methods use the process information may depend on model structure. Furthermore, it is most useful to account for correlations between variables when undertaking a permutation feature importance if variables are highly correlated. This is to account for bias against highly correlated variables, which become less important due to cross correlation between them (Molnar 2022). A comparison between crop modelling and machine learning was made by (Leng & Hall 2020), however this thesis builds upon this comparison by showing the value of comparing across climatic conditions, and models to reveal both the strengths and weakness of both approaches, and how to improve crop modelling. The application of the method from Watson et al. (2015) to introduce climatological errors into machine learning (in chapter 4 is a novel idea. In doing this, this chapter aims to further build upon studies such as Leng & Hall

(2020), Park & Lee (2021), Jung et al. (2021), Zhang et al. (2021) which use machine learning to address the potential impacts of climate change, and studies such as Wolanin et al. (2020), Delerce et al. (2016), Shahhosseini et al. (2019) which compare the results of different machine learning methods using climate input data. Ultimately, the results of the chapter should be used to inform future researchers of the varying sensitivity of machine learning methods to changes and errors in climate input data across different models and temporal scales. This has wider relevance outside of the field of crop modelling to other fields also investigating the potential impacts of climate change.

6.1 Machine learning and mechanistic modelling: collaboration versus competition

Overwhelmingly, this thesis demonstrates that neither machine learning or crop models are infallible. Usefully however, model performance differs for different reasons. In chapter 3 it was demonstrated that although machine learning models performed poorly in hot conditions, the GLAM crop model performed well by comparison, and all machine learning models outperformed GLAM when rainfall lead to bumper crop yields in 2007. In the case of 2003, poor machine learning model performance was due to the small number of years in the training data in which high temperatures and associated low crop yields were present. Conversely, the reason for poor GLAM model performance in 2007 may have been discovered from the analysis in chapter 5. In this chapter, GLAM model performance is driven more strongly by temperature (and duration) than correlation with rainfall. By contrast, ML model performance is more strongly driven by rainfall. In some circumstances, this may lead to greater overall performance. This lead to the analysis in 9 and Figure 5.39 showing that crop model performance may be improved by integrating ML to predict soil moisture properties. However, as Figure 5.39 also demonstrates, improvements in model performance may only be made if rainfall is a limiting factor to crop growth. This will depend heavily on the region of evaluation, many regions may exhibit other limiting factors

such as nutrient limitations, temperature, or pests, which would limit the effectiveness of improving the soil moisture balance routine.

Although in many cases, ML models outperform that of process based models (Leng & Hall 2020, Beven 2023, Feng et al. 2019) (and this thesis). contrary to speculation (Nearing et al. 2021, Maryasin et al. 2018) it is unlikely that process knowledge will be replaced by machine learning. This is because process based models provide a level of interpretability unmatched by ML. ML models can deliver similar results for very different reasons (Lischeid et al. 2022), (and chapter 4 of this thesis). Although this is also a problem in process based model calibration (Beven 2006, Wallach et al. 2021), it can also be more easily solved because parameters are more explicitly tied to reality (e.g. soil moisture or number of days to flowering rather than being simply a weight assigned to a node of a neural network). Therefore, for process based models, this problem (referred to as equifinality) can be addressed using methods such as GLUE (Schulz et al. 1999) with the appropriate data available to determine uncertainty bounds of parameter values. This can be less easily achieved if parameters are not physically based, empirical, or mechanistic in some way.

Given the predictive strength of ML whilst also considering the improved interpretability of process based modelling, ML should therefore be used with process based models rather than as a replacement for them. This thesis shows methods to do this, including prediction of model sub-processes and bench-marking. A full set of recommendations for the joint use of the two approaches is summarised in section 7.6. Other work in the literature has also shown that joint use of approaches is a promising avenue for future research (Droutsas et al. 2022, Feng et al. 2019, 2022, Zhao et al. 2022).

6.2 Machine Learning and extreme events: Can knowledge supplement data?

Throughout this thesis, climatological machine learning inputs are supplemented by knowledge to varying degrees. In chapter 3 and 4 the number of days above 32 °C is used as an input for machine learning models. This temperature threshold is taken from (Hawkins & Sutton 2012) and is used to represent a threshold above which heat stress may affect maize crops. This proved to be a valuable indicator for model performance, however it did not go for enough to help machine learning models predict the effects of the 2003 European heat wave. Extreme events such as this were better predicted using the GLAM crop model. Furthermore, crop failures in chapter 4 were also better predicted with the GLAM crop model. These 2 combined results indicate that crop modelling can achieve greater performance than ML models when predicting very out of sample extreme events. However, not all ML model training configurations and set ups were tested. For example, if ML models were trained with bias towards extremes (through sampling), this may improve results. ML models trained as classifiers to predict crop failures with sampling biased towards extremes should be tested as part of future work.

The presence of extreme events such as heat waves and droughts and their effects creates imbalanced datasets, which present more challenge for ML models. Due to the nature of 'relative' crop failures, they are significantly less in number than non-crop failures. Although there are numerous methods for addressing the problem of imbalanced datasets such as minority under-sampling, or dataset normalization by sampling evenly across a cumulative distribution (see 3.5) methods may not be effective as under-sampling will reduce the number of data points available for model training and oversampling may lead to over-fitting. If available, the best option is always to collect more data, however this may not always be possible.

Chapter 5 demonstrated that data from the GLAM crop model may improve the predictions of machine learning models, however improvements will depend on crop model performance. It follows that value offered to machine learning from the GLAM crop model will depend on the performance of GLAM. In the chapter, the hybrid models improved machine learning model performance in South Africa, however in the other three countries, in some cases (especially in the case of the SVM model) augmenting the machine learning models with GLAM outputs was detrimental to model performance. This is likely the case due to misspecification. For example, if biomass estimates from GLAM are in error in some way, they will contribute to machine learning model error. If more GLAM model outputs are included as ML input features, which all are in error to a certain extent, they will contribute more to overall model error. If GLAM model performance is poor (such as in Zambia), adding a large number of variables from GLAM to the machine learning models will only increase error overall.

Chapter 3 argued that embedded process knowledge from GLAM is important for the prediction of extreme events. This is the most likely explanation for the superior model performance for prediction of the effects of the 2003 heat wave (because the yield anomaly is captured regardless of YGP calibration extent - see 8). Furthermore, chapter 4 found that GLAM was more able to predict crop failures in both datasets than the ML methods. Similarly, chapter 5 showed that process knowledge from GLAM is useful for machine learning however the use of GLAM model data may depend on the structure of the ML methods used. As discussed this could be due to support vector machines using less data to construct decision surfaces than tree based models. An interesting concept which this alludes to is the idea that process knowledge may vary in importance for different quantities of data. A potential future study which would seek to answer this question could reduce data quantities in a systematic manner similar to chapter 3 but include a method of assessing the importance of different variables at each step in the data reduction process.

A possible outcome could be that temperature related variables become more important as data is reduced. Temperature variables were found to be more important by the support vector machine in chapter 5 which implies this may be the case. Mechanistically, this could be because typical variation in rainfall data may be larger than that of temperature data. Therefore, this could mean that more rainfall data is required before a signal emerges between the relationship with crop yield.

Other studies have shown that integration of process knowledge and ML can produce more accurate prediction of extremes if appropriate feature engineering is used (Feng et al. 2019, 2022, Zhao et al. 2022). Feng et al. (2019) has shown that extreme event indicators (such as heat days and aridity indices) calculated at different crop growth stages estimated from a process based model can achieve greater performance for extreme events. However, as shown in chapter 5, added value from process based models as inputs to ML models very much depends on the model performance of the process based model. This will also depend on the dataset used, as chapter 4 showed, different datasets affect ML sensitivity to extreme events to different magnitudes. A comprehensive assessment of how coupling ML methods with process based model outputs is therefore needed across multiple datasets to truly understand the effectiveness of process model outputs for ML models.

6.3 Model sensitivity: crop models and machine learning

Section 4 showed that machine learning methods are typically more sensitive to potential uncertainty from climate models when compared to the results of (Watson et al. 2015). This is to be expected given the flexible nature of machine learning models (i.e. model parameters are tuned to a particular dataset, previous process knowledge of the system is not used to constrain the model parameters). Increased sensitivity of machine learning in comparison to process based models was also a theme in chapter 5. In this case, machine learning models were found to be more sensitive to climatological drivers of crop yield

than the GLAM crop model. This result was key to the idea that machine learning can be used to assess the sensitivity of crop model - climate relationships, with machine learning providing a benchmark to assess appropriate level of sensitivity. One potential uncertainty raised in this chapter is that the yield data, being not strictly observed, could lead to over sensitive models. Although this is a possibility, this analysis still provided insight into how to improve process based crop model results, and findings from the analysis were corroborated by earlier findings in chapter 3.

Fundamentally, the differences in sensitivity between crop models and machine learning is one of the reasons why the two methodologies must be treated very differently when it comes to interpretation of results. Crop models will produce predictions based on relationships which are not always represented in the observed data. For example, as in chapter 5 the relationship between yield and crop duration was not the same as that of the observed data, furthermore, the effect of rainfall was underestimated. This is likely because relationships between climate and crop yields differ from country to country (as demonstrated in chapter 5) Therefore, crop models which are parameterized to either function in a different region, or more broadly, may be miss-specified when focusing on an individual region. Uniqueness of place is a concept relating to this addressed by Beven (2000). Although the paper discusses this concept in the context of hydrological modelling, crop - climate relationships also may experience regionalization effects (regionalization being that processes may be best represented by region specific parameters). However, Iizumi et al. (2014c) have also showed that optimum parameter values can differ with spatial scale, particularly those related to heat mechanisms on yield. A potential future study may be to use machine learning to better understand how optimum crop model parameters may change with changing spatial scale.

On the other hand, ML models may be misled by unrepresentative relationships in observed data. This is a challenge, and a limitation to the bench-marking study methodology. In

this circumstance, process based models which incorporate validated embedded knowledge from other experiments become especially valuable. An ideal solution to this problem is to use multiple datasets when bench-marking models which incorporate information from a range of sources and data types. For example, future studies which incorporate both yields obtained using remote sensing indices, regional scale census statistics, and field scale observed yields collected in field would provide the most comprehensive dataset for bench-marking ML and crop models.

6.4 Potential model performance for future climate projections

Although all chapters in this thesis focus on application of models in the current climate (meaning it is assumed that there is no overall trend or change in the climate system from the beginning to the end of the datasets used). Statistical models as well as process based crop models are used in many cases for prediction of impacts resulting from future scenarios of climate change(Jägermeyr et al. 2021, Rosenzweig et al. 2014, Lobell & Burke 2010). Chapter 4 demonstrates that machine learning models are more sensitive than crop models to changes in the distribution of input climate data. This is especially relevant and important for climate impact projections. As Porter & Gawith (1999) have discussed, climate change is postulated to cause shifts in the distribution of climate, leading to an increased frequency of hot temperatures, and importantly, more hotter temperatures outside of the historical record. This presents a challenge for machine learning as new (or novel) climates will not be present in the training data and so the impacts of climate change may be better suited to process based models for this reason.

Chapter 4 provides some further answers to this question. In particular, uncertainty, (represented by perturbations) tends to effect ML models more so than process based models, with neural networks often being the most sensitive. Therefore, prediction of future crop yields may be most uncertain when using neural networks (in comparison

to other ML methods). This likely due to different levels of complexity in the models. Appropriate complexity is the idea that the greater the number of parameters required to simulated a process, the greater the uncertainty arising from the interactions between them (Challinor et al. 2009). Neural networks have the most parameters of the models tested. Since increasing parameters does not equal decreases in uncertainty (and in fact can have the opposite effect) (Schulz et al. 1999), this could mean that neural networks can be too complex for crop yield prediction given data limitations often present at the regional scale (Müller et al. 2017). This is in agreement with Guo et al. (2014) who found that more complex time series models can have poorer accuracy than models trained on spatial relationships.

One possible angle to explore in future work is the question of whether spatial variability in climate could be used as an analogue for future climate change, hence addressing the problem of future climate projections being outside of the historical record and so potentially unsuitable for prediction by machine learning methods. Such a methodology would have to use process based crop model projections of impacts for training and model evaluation and utilise a perfect model approach to due the absence of future testing data. Such a study will provide further evaluation of the efficacy of machine learning for the prediction of future climate impacts. Guo et al. (2014) have presented evidence that although ML models trained on spatial data outperform that of time series models, high accuracy is limited only to crop yields within normal ranges. Therefore performance will likely be limited in novel climates (Williams & Jackson 2007).

6.5 The role of the Yield gap parameter for analysis of GLAM model results

Model calibration is shown to have a significant effect on crop model performance in each of the chapters. However, when analysing the correlations between modelled yield and

rainfall in chapter 5 it was revealed that the yield gap parameter also has a very strong effect on the correlation between predicted yield and rainfall. This is interesting as the yield gap parameter is not explicitly designed to affect how the model represents the relationship between rainfall and yield. However, this result may not be surprising as the yield gap parameter improves the spatial prediction of yield, and observed yield is likely to be higher where rainfall is higher and vice versa. However, The presence of such a substantial effect of the yield gap parameter makes any interpretation of the reasons for model skill much more difficult as changing parameters which would be expected to have a more significant effect on the rainfall, yield correlation such as transpiration efficiency or the threshold for soil water stress have comparatively little effect against changing the yield gap parameter.

Furthermore, as shown in supplementary analysis (section 8) YGP calibration affects the variance of model predictions, potentially to the detriment of reproducing realistic variability in yield. This result should be noted for future simulations using the GLAM model. In particular, the importance of reproducing absolute yields, as opposed to yield variability and anomalies should be determined before calibration. If reproducing anomalies or relative yields is more important than providing an absolute value of yields, it could be argued that calibration may be detrimental to model results.

6.6 Is machine learning the future of climate impacts modelling?

This thesis demonstrates that although machine learning offers far greater general model performance for prediction of impacts of climate variability on crops throughout a range of conditions, it can be difficult to predict extreme events and extremities of variability. Fundamentally, this is due to the limitation of crop yield data upon the coverage of the training dataset. Crop yield and soils data of sufficient quality are hard to obtain at the regional scale (Müller et al. 2017, Moriondo et al. 2007, Kasampalis et al. 2018).

One possible solution to this may be to derive crop yield estimates from the normalized difference vegetation index (NDVI) obtained through satellite imagery. Studies have show that crop yields can be effectively predicted using NDVI (Mkhabela et al. 2011, Labus et al. 2002, Wall et al. 2008, Moriondo et al. 2007). Labus et al. (2002) developed a multiple linear regression model to estimate wheat yield from NDVI parameters and region specific coefficients, using this method yield was predicted with an R^2 of 0.753. Much further work has also demonstrated the ability to predict crop yields using NDVI and other vegetation indices using remote sensing (Son et al. 2014, Abdul-Jabbar et al. 2023). However, coefficients from yield estimation models from NDVI can be region specific, and so do not generalize well across environments (Moriondo et al. 2007, Labus et al. 2002). Furthermore, uncertainties are often associated with remotely sensed data sets which prevent them from being truly observed data (for example, the static harvested areas used for crop yield mapping to construct the Iizumi & Sakai (2020), Iizumi et al. (2014b) dataset used as part of chapters 4 and 5. Therefore, although machine learning and remote sensing can be used to increase the data availability for crop yield estimation, this is not a perfect solution due to the uncertainty between remotely sensed data sets and ground truth observed crop yield data.

Process knowledge will always be indispensable for climate impacts modelling. This is due to the greater interpretability of process based models. Both this thesis and Lischeid et al. (2022) showed that machine learning predictions of crop yields can make use of input variables in different ways depending on the model architecture used. If this problem is due to equifinality, this problem may be more difficult to solve for machine learning models. For process based models, Equifinality can be addressed more easily because model parameters are tied to tangible environmental factors such as soil moisture or rate of accumulation of biomass. Therefore, the importance of these parameters can be assessed empirically to determine the appropriate level of influence they should have on model outputs. However

for machine learning, internal parameters are not necessarily based on empirically based environmental factors and so this cannot be done.

Potential future uses of machine learning of vast, some of which were discussed in section 5.5.5 as well as this discussion section. The key advantage of machine learning is the ability to learn relationships implicit in data which may not be captured by existing methods. As shown in this thesis, this can be used to inform model improvement to great effect. Calibration is also a key area where machine learning may be used to further improve crop modelling methodologies. Where parameters are not explicitly known, optimization using machine learning may provide greater model accuracy using machine learning. Machine learning may also be used to downscale crop model predictions (Folberth et al. 2019). Results from this thesis show that this may be especially beneficial because machine learning is less prone to spatial over-fitting of climate relationships with crop yield. A further future use of machine learning may be to reconcile the difference between crop models through emulation. Currently many different types of crop models exist (e.g. GLAM (Challinor et al. 2004), DSSAT (Hoogenboom et al. 2019), LPJML, (Schaphoff et al. 2018), Aquacrop (Raes et al. 2009), APSIM (Keating et al. 2003), and JULES-crop (Osborne et al. 2007, 2015) among many others), however there can be significant differences between crop model predictions (Jägermeyr et al. 2021, Asseng et al. 2013, Bassu et al. 2014). A potential future avenue of work could be to use machine learning methods to classify differences between different crop model structures, then the interpretation of the resultant classification models could be used to understand why crop model predictions differ. This would require cooperation between modelling groups (potentially as part of the Agricultural model inter-comparison project) in order to better interpret the results of the machine learning meta-models.

7 Conclusions

The conclusions of this thesis result from the results and discussion from each of the chapters containing original work. Therefore, each chapter is summarised here before bringing this information together to conclude common threads from each chapter first discussed in the discussion section (chapter 6). Further to this, a summary of advantages and disadvantages of process based and machine learning approaches is described in section 7.5. This leads into the discussion of recommendations for future combined uses of machine learning and process based crop models (section 7.6) and other future work before final concluding remarks.

7.1 Chapter 3 Summary and novel contributions to the scientific literature

Chapter 3 asked how much data is required for machine learning to outperform a process based crop model for the case study country of France using sub-national crop yield statistic data and re-analysis climate data. The study used 5 machine learning models, namely, Random forest, gradient boosting, k-nearest neighbours, support vector machine, and a neural network. These 5 models were compared to predictions from the GLAM crop model (Challinor et al. 2004) originally published by (Watson et al. 2015). The models were trained and evaluated using the same train/test split as GLAM to compare to the existing model configuration from the previous study. Furthermore, the number of years and subsequent number of data points used to train the machine learning models was reduced in order to ask how much data is required for machine learning to outperform process based model predictions.

The results from this chapter found that although neural networks had the greatest model performance overall of the machine learning methods, the effect of reducing data on model

predictions had greatest negative effect on model performance. Random forest, gradient boosting and k-nearest neighbours models offered similar performance however the effect of reducing data did not have as strong an effect. Conversely, although GLAM crop model skill was worse overall in comparison to machine learning methods, the model was more easily able to capture the effects of extreme events such as the 2003 European heat wave, which produced large negative crop yield anomalies. This improved model performance is likely the result of embedded process knowledge, meaning that extreme heat parameterizations improved model skill for extreme heat events. Although machine learning methods were unable to predict the effects of the 2003 heat wave anomaly, models were able to recognize that high rainfall in 2007 lead to increased crop yields. The contrasting conditions resulting in the disparity between model performance lead to the construction of the meta-model to determine which model methodology (GLAM or ML) is best to use depending on climate conditions. Machine learning is shown to provide improved performance in all circumstances except very hot conditions with high solar radiation. The disparity of performance across conditions lead to the work in chapter 5 which went a step further to ask how to use the comparisons of the two methods to improve process based crop modelling.

The novelty of this chapter comes from the comparison of machine learning with a crop model within data limited conditions, and showing how different climate conditions and extreme events may affect model performance differently between methods. Leng & Hall (2020) have made a comparison between machine learning methods and an ensemble of crop models although the data set used was larger than that used in this chapter. Furthermore, this study did not show the differences between different machine learning models, as only one framework (random forest) was used. Shahhosseini et al. (2019) has also addressed the question of how data quantity affects the ability of machine learning methods to predict crop yields. This study also reduced the number of years of training data in

5 year increments, however, each 5 year set represented roughly 0.4 million data points, a much larger number than in this chapter (roughly 450 unique crop yield data points). Furthermore, the Shahhosseini et al. (2019) study was a comparison purely between machine learning methods, and did not place the comparison in the context of process based crop model performance.

7.2 Chapter 4 Summary and novel contributions to the scientific literature

Chapter 4 built upon chapter 3 by both looking specifically at the prediction of crop failures, as well as showing the effects of uncertainty from climate input data on the outputs of machine learning methods. This chapter also built upon the previous by using both the data from the previous chapter as well as the GDHY data set (Iizumi et al. 2014b, Iizumi & Sakai 2020), using a subset of the data for South Africa. The research aims of this chapter were: what is machine learning performance for crop yield failure prediction across contrasting environments, and how would uncertainty in climate input data affect machine learning model performance and correct failure prediction rate. Climatological perturbations were introduced to test data used to evaluate the effect of uncertainty on machine learning model performance. The method used to do this was taken from Watson et al. (2015) and was used to compare the effects of uncertainty across different temporal scales, in both temperature and rainfall data. The effects of uncertainty was assessed across the five machine learning methods first used within the previous chapter, with some loose comparisons made to process based model results published by (Watson et al. 2015). Performance metrics were chosen to assess both effects on general model performance as well as that of crop failure prediction. The magnitude of the perturbations was chosen based on uncertainty between climate models.

Model results can be divided into model performance and crop failures, as well as the

effects of the perturbation scheme used to simulate climate model uncertainty. Model performance was shown to be greater among the machine learning models than the GLAM crop model bench-mark comparison. However, similar to the previous chapter, machine learning methods struggled to reproduce crop failures. Which machine learning method produced the most accurate predictions depended on the dataset used. The French dataset favoured the neural network model, as well as the tree based methods (Random forest and gradient boosting), however the South Africa dataset favoured the support vector machine. The uncertainty affected all machine learning methods in the same way, however the magnitude of the errors in the output predictions differed between models. Neural networks in general were found to be most sensitive to uncertainty in weather data. This is likely due to the larger number of model parameters of the model. uncertainty in rainfall and temperature affected the models differently depending on the timescale of the perturbations. Perturbations of inter-annual variability in rainfall affected model performance more than perturbations in mean rainfall, whereas temperature perturbations affected model outputs both for the mean and inter-annual variability of temperature. When comparing the results of the perturbation scheme to that of Watson et al. (2015) machine learning methods were found to be more sensitive to climatological input uncertainty than the process based GLAM crop model.

The novelty of this chapter comes from the assessment of the effects of climate uncertainty and extrapolation on machine learning methods. No such study which compares the effects of climate uncertainty on the predictions from the 5 machine learning methods used for the chapter. This comparison is very useful for crop or climate modellers researching the impacts of climate change and variability and would like to know which models to choose and how sensitive they may be to uncertainty within climate input data. Moreover, Results also show the effects of extrapolation on models, useful for future studies focusing on projection of the effects of extreme events, or forecasting future crop failures.

7.3 Chapter 5 Summary and novel contributions to the scientific literature

Chapter 5 built upon chapter 3 to use the comparison between machine learning and process based crop modelling to understand how the crop model parameters may affect model performance. This method was used to determine potential avenues for crop model improvement. The GLAM crop model was used as a case study process based crop model and so results are especially relevant to those which aim to use GLAM, however, GLAM shares structural similarities with many other process based crop growth models, and so the results also have wider implications for crop modellers who use other models. This chapter addresses the very pertinent question in the current literature which is how can ML aid process based model simulations (Huntingford et al. 2019, Nearing et al. 2021, Lischeid et al. 2022, Zhang et al. 2023). This is addressed by both using ML as a benchmarking tool as well as for prediction of model sub-processes. Many of the joint uses discussed for ML and process based models are summarised in section 7.6.

In this chapter, comparisons are made between ML and process based crop modelling to answer 3 key questions relating to model improvement. The following research questions are addressed in this chapter:

1. How do crop models and ML represent climate-yield relationships differently?
2. How can process based models be used to improve machine learning predictions?
3. Can insights from machine learning be used to inform crop model improvement?

To answer each of these questions, a selection of model comparisons are made under different conditions summarised in Table 5.1. Model comparisons lead to the finding that ML outperforms GLAM for conditions in which rainfall more strongly drove spatial variation in yield. This lead to several tests to determine how to improve the correlation

between rainfall and predicted yield described in section 5.4.6.

The conclusion of these tests revealed that soil moisture characteristics simulated by pedo-transfer functions have a significant effect on the crop yield-rainfall relationship. Without replacing existing calibration methods, the best method to improve the accuracy of such functions is to use machine learning (the efficacy of doing this is explored in appendix 9).

The answer to the second research question is that machine learning predictions can be improved through the use of GLAM process knowledge as inputs. However, this is only the case if the crop model performance is of sufficient skill, if not, process knowledge inputs can be detrimental to model performance. To try to understand which process knowledge inputs are most useful for improving machine learning predictions is more difficult. This is because different machine learning methods will provide different answers as to which sets of variables are most important. This may be due to differences in model structure.

This chapter also discussed the spatial scale of calibration, and how this may affect the relationship between simulated yield and climate variables, particularly rainfall. Comparisons at different spatial scales show that YGP calibration affects simulated yields more than coordinates as inputs for ML models. For this reason, the role of calibration at the regional scale should be explore in more detail, to determine the degree to which models over-fit to current spatial weather patterns.

The novelty of chapter 5 is that machine learning is used as a benchmark comparison tool to understand the agroclimatic conditions under which improvements to the GLAM crop model are most needed. This approach to improve crop models has never been used before, and could be applied to other crop models to better understand where to target crop model improvements in a similar way. The role of the spatial scale of calibration is also assessed against machine learning. The role of spatial scale in calibration has been

addressed by studies such as that of Angulo et al. (2013). In the study, they show that more location specific information leads to better predictions. However, the results from chapter 5 show that calibration may have unintended effects on climatic relationships and so lead to a more accurate representation of the relationship between rainfall and yield, but for the wrong reasons.

7.4 Common threads across chapters

Common themes and techniques unite chapters 3, 4 and 5, chiefly among which is the theme of model comparison between machine learning and process based crop modelling. Chapter 3 compared machine learning and crop modelling to answer how much data is required for machine learning to outperform process based crop modelling. Subsequently chapter 5 used the experience and knowledge gained from chapter 3 to ask how to improve process modelling using machine learning. Chapter 4 also took the results from chapter 3 and focused instead (in part) on comparing machine learning model performance for crop failure prediction. This aspect of chapter 4 arose from the result in chapter 3 in which machine learning methods performed poorly at predicting the effects of the 2003 heat wave.

Greater sensitivity of machine learning methods to climate variability is also a common thread of chapters 4 and 5. This idea is first explored in 4 by comparing the effects of climatological uncertainty on different machine learning architectures. This chapter found that machine learning sensitivity to the effects of temperature and rainfall uncertainty differs by model, environment and temporal scale. The finding that sensitivity to input uncertainty affects model architectures differently lead to the decision in chapter 5 to compare both a support vector machine and random forest model to the results from GLAM. The support vector machine was chosen because the model structure is sufficiently different from the random forest model as well as the result in chapter 4 that showed that

the random forest and support vector machine were 2 of the best performing models in South Africa.

A common thread throughout this thesis is the value of process knowledge for predictions. The results in chapter 3 indicated that embedded process knowledge from the GLAM crop model allowed the model to better predict the effects of the 2003 heat wave. Chapter 5 further explored this idea by adding process knowledge as inputs to the machine learning methods. This had mixed effects on model performance. Where GLAM model performance was of sufficient skill, adding information from GLAM to the machine learning models such as planting dates and biomass estimates proved useful, however if GLAM model skill was much worse than machine learning model skill, then adding GLAM information as inputs to the models was detrimental to machine learning model performance. Process knowledge will always be indispensable for the prediction of climate impacts, regardless of which methods are used. Even if machine learning methods are used over process based crop models, machine learning methods still require a degree of feature engineering to create meaningful inputs to gain the best model predictions. questions arise during the feature engineering process which require process knowledge to answer such as how to determine the growing season dates, how to better capture the effects of extreme events, and what are the most important input features for the model. In particular, features which involved the use of heat stress temperature thresholds were found to be useful such as the 32 degree temperature threshold.

The effects of extreme events is also a common theme across chapters. Due to the data limitations for the prediction of extreme events, predicting the effects of extreme events will always be a challenge for modelling in general, but especially machine learning methods which do not use process knowledge to constrain model parameters. Techniques such as minority over-sampling which normalize the dataset to ensure the same proportion of extreme event instances as other instances can reduce the overall explanatory power of

the model due to the fewer number of data points or introduce bias. Some minority over sampling methods such as those discussed by Chawla (2009) may be more useful, and so could be explored in potential future work.

Using machine learning to gain process understanding with the view to improving process based crop models is an overall aim of this thesis. Through doing this, it is shown that GLAM underestimates the effects of rainfall. Looking back to chapter 3 this may be the reason why GLAM performs poorly in the the year 2007 when high rainfall caused large increases in crop yields. Further work should explore the role of the soil water balance routine to improve crop model correlations with rainfall where rainfall is a key driver of observed crop yield.

7.5 Summary: advantages and disadvantages of ML and process based crop models

Throughout this thesis machine learning and crop modelling is compared to assess the strengths and weaknesses of the two approaches. In summary, Table 3.6 from chapter 3 is revisited here (Table 7.1) using lessons learned from this thesis to inform the advantages and disadvantages of the approaches. The two key advantages of process based crop modelling is interpretability and improved extrapolation ability. Interpretability is addressed in section 6.6, however it is shown in practice to be useful in section 5.4.6. The strength of crop modelling to predict outside of the historic range of data found in the training dataset is addressed in sections 3.4.1 and 6.4. ML advantages and disadvantages are different from that of process based models. Key to chapter 5 is the flexibility of ML. Because ML model parameters depend on the training data, in the circumstances in which machine learning outperformed GLAM, poor GLAM performance was not due to poor or insufficient data. Therefore ML can be a very useful bench-marking tool.

Table 7.1:
Comparative advantages and disadvantages of machine learning and crop modelling

Method	Advantages	Disadvantages
Crop modelling	Interpret-able (section 5.4.6, also discussed in section 6.6)	Better calibration requires more site specific information (Figure 4.11 and section 5.4.4)
	Process knowledge improves extrapolation (chapters 3 and 4, Figure 3.6)	Calibration can be subjective (addressed in section 2.4)
Machine learning	improved general performance (chapters 3 and 5)	Can be more difficult to predict the effects of extremes (Figures 3.6, 4.17, 4.18)
	Improved sensitivity to weather variability (section 5.4.3)	Equifinality can lead to difficult to interpret feature importance (section 5.4.2)
	Flexible model structure (chapter 5)	Parameter interaction can cause varying behaviour between different model architectures (Figure 4.20 and section 5.4.2)

7.6 Recommendations for use of ML with process based crop models

This work shows multiple avenues in which ML could be used with process based crop models. Table 7.2 summarises multiple directions for the combined use of models discussed in this thesis. ML has many uses (not inclusive of those discussed here) which can be used for great benefit for crop modelling. Here, it is shown that comparisons with ML can be used to identify conditions for appropriate model selection, as well as conditions under which crop model performance requires improvement. ML can also be used within a process based model to improve sub-processes. GLAM is especially well suited to this (as shown by (Droutsas et al. 2022)) because semi-empirical parameterizations used by GLAM rely on data and so can be better optimized using ML. As discussed, improvements in ML over crop model simulations show that poor crop model performance is not due to data limitations, conversely however, improved crop model performance over ML indicates

deficiencies in the dataset. Analysis from this thesis into the spatial scale of calibration also indicated that ML would be well suited to down-scaling crop model simulations (as shown by Folberth et al. (2019)). Process knowledge from crop models can also be useful for ML in some circumstances, however more research is needed to identify the appropriate knowledge required by ML for best improvements in performance.

Table 7.2:
Recommendations for future ML and process based model usage

Usage	relevant section/Figure
Identify appropriate conditions for model selection	Figure 3.11
Identify mechanistic crop model deficiencies	Figures 5.34, 5.35
Incorporation of Process based knowledge into ML	Figure 5.16, section 5.4.2
Incorporation of ML to improve crop model subprocess parameterization	Appendix 9
Machine learning to downscale crop model simulations	Figure 5.33

7.7 Further work

Throughout this thesis, many avenues of future work are presented, as there is a great deal of new ideas and questions which arise from the results. Firstly, chapter 5 showed that feature importance of machine learning methods can vary significantly depending on the model chosen. A potential future study to determine feature importance using an ensemble of machine learning methods is mentioned. Fisher et al. (2019) have addressed this problem in a general machine learning context and have proposed that machine learning methods should be grouped based on a measure of similarity between models called the model class reliance (MCR), which gives a more comprehensive description of model feature importance accounting for different model structures providing different descriptions of feature importance. The application of the approach presented in Fisher et al. (2019) along with the accounting of correlations between highly correlated variables undertaken in this thesis is most likely to provide the best account of the value of process model

understanding for machine learning.

In Both chapters 3 and 4 machine learning methods presented poorly predict the effects of extreme events. This relates to the idea of learning from imbalanced data explored by Chawla et al. (2004), Chawla (2009) in which many sampling techniques for addressing the problem of imbalanced data are addressed. In chapter 3 a method of majority under-sampling was explored in an attempt to improve the predictions of the low crop yield values by reducing bias towards larger values, however this line of analysis did not achieve improved performance. A future study should aim to further address machine learning prediction of crop failures whilst also introducing synthetic generation of data to determine if this will improve predictions. This may be a very interesting and useful study as oversampling techniques which make use of synthetic data can be very powerful tools for reducing machine learning bias and improving predictions of minority classes (Chawla et al. 2002, Fernández et al. 2018).

Similarly, although statistical models of generating synthetic data to improve predictions could be very successful, for crop-climate modelling, it may also be a very successful solution to use crop models to generate new data for machine learning. In chapter 3 it is mentioned that a potential idea may be a 2 way pipeline which would integrate machine learning and crop modelling to generate synthetic data for each other. In this hypothetical study, climates in which GLAM the GLAM crop model performs better than machine learning could be used to generate synthetic data to improve machine learning. In chapter 5 countries which GLAM performed adequately improved machine learning predictions lending credence to this idea. Studies such as Folberth et al. (2019), Shahhosseini et al. (2019), Feng et al. (2019) have also trained machine learning methods using synthetic data to great effect also implying the potential for such a method. The other part of the 2 way pipeline (using machine learning predictions to improve GLAM) was explored in chapter 5. The potential of a 2 way pipeline could be further explored by using ML produced down-

scaled predictions to calibrate the YGP parameter. This would involve using machine learning to downscale crop yield simulations (similar to methods used by Folberth et al. (2019)), then using the downscaled simulations to calibrate the YGP parameter. This could be useful for then allowing estimates of biomass and leaf area index calibrated to finer spatial scales.

In appendix 9 it is also shown that machine learning can be used to improve the prediction of soil moisture characteristics relative to existing empirical methods used by many crop models. Further work is required to integrate the machine learning based predictions of soil moisture characteristics into crop models to understand if this will improve the correlation between rainfall and yield predictions. If so, this will make a compelling case for the use of machine learning to predict such variables instead of the most commonly used method at the regional scale which is to use pedo-transfer functions.

Chapter 4 showed that rainfall and temperature uncertainty can have different magnitudes of effects on machine learning outputs depending on the model structure used. perturbations were added to the input of the test data set instead of the training dataset to simulate the potential differences which may arise from differences in unseen data relative to the training data, such as future climate change scenarios or forecasting future crop failures. Therefore, this was deemed a more interesting methodological set up, (as well as enabling a comparison to the results of Watson et al. (2015)). Therefore, a future study which uses this information further could be to take a climate model ensemble (such as the CMIP6 model ensemble used by (Jägermeyr et al. 2021)) and compare yield predictions using weather data from each climate model in the ensemble against present day climate information and crop yields. Such an analysis would show the relative uncertainty of machine learning and climate models in comparison to the AgMIP crop model ensemble presented in Jägermeyr et al. (2021).

7.8 Concluding remarks

In this thesis, machine learning methods are compared to crop modelling to understand how to improve model predictions. Although the first chapter took an adversarial approach, to pit the modelling approaches against each other, Regional scale prediction most benefits from the combined use of both approaches. this is for 2 key reasons. Firstly, data limitations imposed by the availability of crop yield data can limit the ability of machine learning methods to predict the effects of extreme events. Secondly, process based crop models are more suitable for interpretation than machine learning methods. This is because it is much easier to address equifinality if model parameters are explicitly tied to measurably empirical processes. However, the benefit machine learning offers to regional scale crop yield prediction and climate impacts assessment is undeniable. Although machine learning may offer greater predictive performance than process based crop modelling, an even greater benefit is the use of machine learning to provide a comparative test for process based crop models therefore allowing the bench-marking of model processes. Comparison against machine learning shows that crop model performance is not limited by data availability. Therefore, this comparison can be used to further understanding of how to better represent the reality of crop-climate relationships.

More broadly, this thesis is a comparison of existing understanding of physical processes with the capability of machine learning to represent the same processes. As such, the material in this thesis is of wider interest to those who wish to compare process understanding with machine learning. Ultimately, although this thesis compares predictive performance between knowledge based and data driven models, the two modelling approaches were shown to have very different results and behaviour. Furthermore, the interpretability of process based methods is invaluable for understanding and model improvement. Therefore, although many comparisons are made between process models, existing techniques and machine learning, both in this thesis and the wider literature in other fields (Kim

et al. 2021, Perol et al. 2018, Semmler et al. 2021, Leng & Hall 2020, Maryasin et al. 2018, Brecht & Bihlo 2023). Machine learning should be used to further human understanding rather than replace existing process knowledge.

References

Abdi, H. & Williams, L. J. (2010), 'Principal component analysis', *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433–459.

Abdul-Jabbar, T., Ziboon, A. & Albayati, M. (2023), Crop yield estimation using different remote sensing data: literature review, *in* 'IOP Conference Series: Earth and Environmental Science', Vol. 1129, IOP Publishing, p. 012004.

Abram, N. J., Henley, B. J., Sen Gupta, A., Lippmann, T. J., Clarke, H., Dowdy, A. J., Sharples, J. J., Nolan, R. H., Zhang, T., Wooster, M. J. et al. (2021), 'Connections of climate change and variability to large and extreme forest fires in southeast australia', *Communications Earth & Environment* **2**(1), 1–17.

Adhikari, U. & Nejadhashemi, A. P. (2016), 'Impacts of climate change on water resources in Malawi', *Journal of Hydrologic Engineering* **21**(11), 05016026.

Ahmed, S. A., Diffenbaugh, N. S., Hertel, T. W., Lobell, D. B., Ramankutty, N., Rios, A. R. & Rowhani, P. (2011), 'Climate volatility and poverty vulnerability in Tanzania', *Global Environmental Change* **21**(1), 46–55.

Amadu, F. O., McNamara, P. E. & Miller, D. C. (2020), 'Yield effects of climate-smart agriculture aid investment in southern Malawi', *Food Policy* **92**, 101869.

Amanabadi, S., Vazirinia, M., Vereecken, H., Vakilian, K. A. & Mohammadi, M. (2019), 'Comparative study of statistical, numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data', *Eurasian Soil Science* **52**, 1555–1571.

Angulo, C., Rötter, R., Lock, R., Enders, A., Fronzek, S. & Ewert, F. (2013), 'Implication of crop model calibration strategies for assessing regional impacts of climate change in Europe', *Agricultural and Forest Meteorology* **170**, 32–46.

Aranguren, M., Castellón, A. & Aizpurua, A. (2020), 'Wheat yield estimation with ndvi values using a proximal sensing tool', *Remote Sensing* **12**(17), 2749.

Asfaw, D., Black, E., Brown, M., Nicklin, K. J., Otu-Larbi, F., Pinnington, E., Challinor, A., Maidment, R. & Quaife, T. (2018), 'Tamsat-alert v1: A new framework for agricultural decision support', *Geoscientific Model Development* **11**(6), 2353–2371.

Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., Kimball, B. A., Ottman, M. J., Wall, G., White, J. W. et al. (2015), 'Rising temperatures reduce global wheat production', *Nature climate change* **5**(2), 143–147.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rötter, R. P., Cammarano, D. et al. (2013), 'Uncertainty in simulating wheat yields under climate change', *Nature climate change* **3**(9), 827–832.

Asseng, S., Martre, P., Maiorano, A., Rötter, R. P., O'Leary, G. J., Fitzgerald, G. J., Girousse, C., Motzo, R., Giunta, F., Babar, M. A. et al. (2019), 'Climate change impact and adaptation for wheat protein', *Global change biology* **25**(1), 155–173.

Atkinson, D. & Porter, J. R. (1996), 'Temperature, plant development and crop yields', *Trends in Plant Science* **1**(4), 119–124.

Baldos, U. L. C. & Hertel, T. W. (2014), 'Global food security in 2050: the role of agricultural productivity and climate change', *Australian Journal of Agricultural and Resource Economics* **58**(4), 554–570.

Bandara, J. S. & Cai, Y. (2014), 'The impact of climate change on food crop productivity, food prices and food security in south asia', *Economic Analysis and Policy* **44**(4), 451–465.

Barlow, K., Christy, B., O'Leary, G., Riffkin, P. & Nuttall, J. (2015), 'Simulating the

impact of extreme heat and frost events on wheat crop production: A review', *Field crops research* **171**, 109–119.

Basso, B., Cammarano, D. & Carfagna, E. (2013), Review of crop yield forecasting methods and early warning systems, *in* 'Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy', Vol. 18, p. 19.

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., Rosenzweig, C., Ruane, A. C., Adam, M., Baron, C. et al. (2014), 'How do various maize crop models vary in their responses to climate change factors?', *Global change biology* **20**(7), 2301–2320.

Bengio, Y., Simard, P. & Frasconi, P. (1994), 'Learning long-term dependencies with gradient descent is difficult', *IEEE transactions on neural networks* **5**(2), 157–166.

Bergamaschi, H., Costa, S. M. S. d., Wheeler, T. R. & Challinor, A. J. (2013), 'Simulating maize yield in sub-tropical conditions of southern Brazil using GLAM model', *Pesquisa Agropecuária Brasileira* **48**, 132–140.

Beven, K. (2006), 'A manifesto for the equifinality thesis', *Journal of hydrology* **320**(1-2), 18–36.

Beven, K. (2023), 'Benchmarking hydrological models for an uncertain future', *Hydrological Processes* p. e14882.

Beven, K. J. (2000), 'Uniqueness of place and process representations in hydrological modelling', *Hydrology and earth system sciences* **4**(2), 203–213.

Bhatia, N. & Vandana (2010), 'Survey of nearest neighbor techniques', *CoRR* **abs/1007.0085**.

Bhavsar, H. & Ganatra, A. (2012), 'A comparative study of training algorithms for supervised machine learning', *International Journal of Soft Computing and Engineering (IJSCE)* **2**(4), 2231–2307.

Birch, C., Vos, J. & Van der Putten, P. (2003), 'Plant development and leaf area production in contrasting cultivars of maize grown in a cool temperate environment in the field', *European Journal of Agronomy* **19**(2), 173–188.

Black, E., Blackburn, M., Harrison, G., Hoskins, B. & Methven, J. (2004), 'Factors contributing to the summer 2003 European heatwave', *Weather* **59**(8), 217–223.

Bowden, C., Foster, T. & Parkes, B. (2023), 'Identifying links between monsoon variability and rice production in India through machine learning', *Scientific Reports* **13**(1), 2446.

Brecht, R. & Bihlo, A. (2023), 'Towards replacing precipitation ensemble predictions systems using machine learning', *arXiv preprint arXiv:2304.10251* .

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L. & Peng, B. (2019), 'Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches', *Agricultural and Forest Meteorology* **274**, 144–159.

Caparas, M., Zobel, Z., Castanho, A. D. & Schwalm, C. R. (2021), 'Increasing risks of crop failure and water scarcity in global breadbaskets by 2030', *Environmental Research Letters* **16**(10), 104013.

Carleton, T. A. & Hsiang, S. M. (2016), 'Social and economic impacts of climate', *Science* **353**(6304), aad9837.

Castelli, M., Clemente, F. M., Popovič, A., Silva, S. & Vanneschi, L. (2020), 'A machine learning approach to predict air quality in california', *Complexity* **2020**.

Castelvecchi, D. (2016), 'Can we open the black box of AI?', *Nature News* **538**(7623), 20.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & Lopez, A. (2020), 'A comprehensive survey on support vector machine classification: Applications, challenges and trends', *Neurocomputing* **408**, 189–215.

Challinor, A. J., Ewert, F., Arnold, S., Simelton, E. & Fraser, E. (2009), 'Crops and climate change: progress, trends, and challenges in simulating impacts and informing adaptation', *Journal of experimental botany* **60**(10), 2775–2789.

Challinor, A. J., Koehler, A.-K., Ramirez-Villegas, J., Whitfield, S. & Das, B. (2016a), 'Current warming will reduce yields unless maize breeding and seed systems adapt immediately', *Nature Climate Change* **6**(10), 954–958.

Challinor, A. J., Müller, C., Asseng, S., Deva, C., Nicklin, K. J., Wallach, D., Vanuytrecht, E., Whitfield, S., Ramirez-Villegas, J. & Koehler, A.-K. (2018), 'Improving the use of crop models for risk assessment and climate change adaptation', *Agricultural systems* **159**, 296–306.

Challinor, A. J., Parkes, B. & Ramirez-Villegas, J. (2015), 'Crop yield response to climate change varies with cropping intensity', *Global change biology* **21**(4), 1679–1688.

Challinor, A. J., Simelton, E. S., Fraser, E. D., Hemming, D. & Collins, M. (2010), 'Increased crop failure due to climate change: assessing adaptation options using models and socio-economic data for wheat in China', *Environmental Research Letters* **5**(3), 034012.

Challinor, A. J., Watson, J., Lobell, D. B., Howden, S., Smith, D. & Chhetri, N. (2014), 'A meta-analysis of crop yield under climate change and adaptation', *Nature Climate Change* **4**(4), 287–291.

Challinor, A., Slingo, J., Wheeler, T. & Doblas-Reyes, F. (2016b), 'Probabilistic simulations of crop yield over western India using the demeter seasonal hindcast ensembles', *Tellus A: Dynamic Meteorology and Oceanography* **57**(3), 498–512.

Challinor, A., Wheeler, T., Craufurd, P., Slingo, J. & Grimes, D. (2004), 'Design and optimisation of a large-area process-based model for annual crops', *Agricultural and forest meteorology* **124**(1-2), 99–120.

Challinor, A., Wheeler, T., Slingo, J. & Hemming, D. (2005), 'Quantification of physical and biological uncertainty in the simulation of the yield of a tropical crop using present-day and doubled CO_2 climates', *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1463), 2085–2094.

Chapman, S., E Birch, C., Pope, E., Sallu, S., Bradshaw, C., Davie, J. & H Marsham, J. (2020), 'Impact of climate change on crop suitability in sub-saharan Africa in parameterized and convection-permitting regional climate models', *Environmental Research Letters* **15**(9), 094086.

Chawla, N. V. (2009), 'Data mining for imbalanced datasets: An overview', *Data mining and knowledge discovery handbook* pp. 875–886.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* **16**, 321–357.

Chawla, N. V., Japkowicz, N. & Kotcz, A. (2004), 'Special issue on learning from imbalanced data sets', *ACM SIGKDD explorations newsletter* **6**(1), 1–6.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A. et al. (2005), 'Europe-wide reduction in primary productivity caused by the heat and drought in 2003', *Nature* **437**(7058), 529–533.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**(368), 829–836.

COGECA, C. (2003), 'Assessment of the impact of the heat wave and drought of the summer 2003 on agriculture and forestry', *Committee of Agricultural Organisations in the European Union and General Committee for Agricultural Cooperation in the European Union, Brussels, Belgium* .

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Cover, T. & Hart, P. (1967), 'Nearest neighbor pattern classification', *IEEE transactions on information theory* **13**(1), 21–27.

d'Amour, C. B., Wenz, L., Kalkuhl, M., Steckel, J. C. & Creutzig, F. (2016), 'Teleconnected food supply shocks', *Environmental Research Letters* **11**(3), 035007.

Deka, P. C. et al. (2014), 'Support vector machine applications in the field of hydrology: a review', *Applied soft computing* **19**, 372–386.

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Patiño, V. H., Garcés Varón, G. & Jiménez, D. (2016), 'Assessing weather-yield relationships in rice at local scale using data mining approaches', *PloS one* **11**(8), e0161620.

Demattê, J. A., Dotto, A. C., Paiva, A. F., Sato, M. V., Dalmolin, R. S., Maria do Socorro, B., da Silva, E. B., Nanni, M. R., ten Caten, A., Noronha, N. C. et al. (2019), 'The Brazilian soil spectral library (bssl): A general view, application and challenges', *Geoderma* **354**, 113793.

Deva, C. R., Urban, M. O., Challinor, A. J., Falloon, P. & Svitákova, L. (2020), 'Enhanced leaf cooling is a pathway to heat tolerance in common bean', *Frontiers in plant science* **11**, 19.

Devereux, S. (2002), 'The Malawi famine of 2002'.

Devereux, S. (2009), 'Why does famine persist in Africa?', *Food security* **1**, 25–35.

Di Luca, A., Pitman, A. J. & de Elía, R. (2020), 'Decomposing temperature extremes errors in CMIP5 and CMIP6 models', *Geophysical Research Letters* **47**(14), e2020GL088031.

Diez-Sierra, J. & Del Jesus, M. (2020), 'Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods', *Journal of Hydrology* **586**, 124789.

Dong, T. & Dong, W. (2021), 'Evaluation of extreme precipitation over Asia in CMIP6 models', *Climate Dynamics* **57**(7-8), 1751–1769.

Dore, M. H. (2005), 'Climate change and changes in global precipitation patterns: what do we know?', *Environment international* **31**(8), 1167–1181.

Dorion, S., Lalonde, S. & Saini, H. S. (1996), 'Induction of Male Sterility in Wheat by Meiotic-Stage Water Deficit Is Preceded by a Decline in Invertase Activity and Changes in Carbohydrate Metabolism in Anthers', *Plant Physiology* **111**(1), 137–145.

Droutsas, I., Challinor, A. J., Deva, C. R. & Wang, E. (2022), 'Integration of machine learning into process-based modelling to improve simulation of complex crop responses', *in silico Plants* **4**(2).

Droutsas, I., Challinor, A., Swiderski, M. & Semenov, M. (2019), 'New modelling technique for improving crop model performance - application to the GLAM model', *Environmental Modelling & Software* **118**, 187–200.

Dunning, C. M., Black, E. & Allan, R. P. (2018), 'Later wet seasons with more intense rainfall over Africa under future climate change', *Journal of Climate* **31**(23), 9719–9738.

Durand, J.-L., Delusca, K., Boote, K., Lizaso, J., Manderscheid, R., Weigel, H. J., Ruane, A. C., Rosenzweig, C., Jones, J., Ahuja, L. et al. (2018), 'How accurately do maize crop

models simulate the interactions of atmospheric co2 concentration levels with limited water supply on water use and yield?', *European Journal of Agronomy* **100**, 67–75.

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K., Büchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T. et al. (2015), 'The global gridded crop model intercomparison: data and modeling protocols for phase 1 (v1. 0)', *Geoscientific Model Development (Online)* **8**(2).

Ellis, F. & Manda, E. (2012), 'Seasonal food crises and policy responses: A narrative account of three food security crises in Malawi', *World Development* **40**(7), 1407–1417.

Ewert, F., Rötter, R., Bindi, M., Webber, H., Trnka, M., Kersebaum, K., Olesen, J., van Ittersum, M., Janssen, S., Rivington, M., Semenov, M., Wallach, D., Porter, J., Stewart, D., Verhagen, J., Gaiser, T., Palosuo, T., Tao, F., Nendel, C., Roggero, P., Bartošová, L. & Asseng, S. (2015), 'Crop modelling for integrated assessment of risk to food production from climate change', *Environmental Modelling & Software* **72**, 287–303.

Ewert, F., van Ittersum, M. K., Heckelei, T., Therond, O., Bezlepkina, I. & Andersen, E. (2011), 'Scale changes and model linking methods for integrated assessment of agri-environmental systems', *Agriculture, Ecosystems & Environment* **142**(1), 6–17. Scaling methods in integrated assessment of agricultural systems.

Falconnier, G. N., Corbeels, M., Boote, K. J., Affholder, F., Adam, M., MacCarthy, D. S., Ruane, A. C., Nendel, C., Whitbread, A. M., Justes, É. et al. (2020), 'Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-saharan Africa', *Global change biology* **26**(10), 5942–5964.

Fan, X., Duan, Q., Shen, C., Wu, Y. & Xing, C. (2020), 'Global surface air temperatures in CMIP6: historical performance and future changes', *Environmental Research Letters* **15**(10), 104056.

FAOSTAT (2022), 'Crop production, yield, harvested area (global - national - annual) - faostat'.

URL: *https://data.apps.fao.org/catalog/dataset/crop-production-yield-harvested-area-global-national-annual-faostat/resource/45e995e9-1021-4288-abed-ab933d3a3c01*

Farooq, M., Hussain, M. & Siddique, K. H. (2014), 'Drought stress in wheat during flowering and grain-filling periods', *Critical reviews in plant sciences* **33**(4), 331–349.

Fauchereau, N., Trzaska, S., Rouault, M. & Richard, Y. (2003), 'Rainfall variability and changes in southern Africa during the 20th century in the global warming context', *Natural hazards* **29**(2), 139.

Feng, P., Wang, B., Liu, D. L., Waters, C. & Yu, Q. (2019), 'Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern australia', *Agricultural and Forest Meteorology* **275**, 100–113.

Feng, P., Wang, B., Liu, D. L., Yu, Q. & Hu, K. (2022), 'Coupling machine learning with APSIM model improves the evaluation of climate extremes impact on wheat yield in australia', *Modeling Processes and Their Interactions in Cropping Systems: Challenges for the 21st Century* pp. 251–275.

Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V. (2018), 'Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary', *Journal of Artificial Intelligence research* **61**, 863–905.

FEWS NET (2023), 'FEWS NET Famine Early Warning Systems Network'.

URL: *https://fews.net/*

Feynman, R. P. (1966), 'Symmetry in physical laws', *The Physics Teacher* **4**(4), 161–174.

Fischer, K. & Hajdu, F. (2015), 'Does raising maize yields lead to poverty reduction? a case study of the massive food production programme in South Africa', *Land Use Policy* **46**, 304–313.

Fisher, A., Rudin, C. & Dominici, F. (2019), 'All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.', *J. Mach. Learn. Res.* **20**(177), 1–81.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N. & Obersteiner, M. (2019), 'Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning', *Agricultural and Forest Meteorology* **264**, 1–15.

Folberth, C., Skalskỳ, R., Moltchanova, E., Balkovič, J., Azevedo, L. B., Obersteiner, M. & Van Der Velde, M. (2016), 'Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations', *Nature communications* **7**(1), 11872.

Foster, T. & Brozović, N. (2018), 'Simulating crop-water production functions using crop growth models to support water policy assessments', *Ecological Economics* **152**, 9–21. **URL:** *https://www.sciencedirect.com/science/article/pii/S0921800917314313*

Franke, J. A., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Balkovic, J., Ciais, P., Dury, M., Falloon, P. D., Folberth, C. et al. (2020), 'The GGCMI phase 2 experiment: global gridded crop model simulations under uniform changes in co 2, temperature, water, and nitrogen levels (protocol version 1.0)', *Geoscientific Model Development* **13**(5), 2315–2336.

Freund, Y., Schapire, R. E. et al. (1996), Experiments with a new boosting algorithm, *in* 'ICML', Vol. 96, Citeseer, pp. 148–156.

Gaupp, F., Hall, J., Hochrainer-Stigler, S. & Dadson, S. (2020), 'Changing risks of simultaneous global breadbasket failure', *Nature Climate Change* **10**(1), 54–57.

Giannini, A., Biasutti, M., Held, I. M. & Sobel, A. H. (2008), 'A global perspective on African climate', *Climatic Change* **90**(4), 359–383.

Goulart, H., Van Der Wiel, K., Folberth, C., Balkovic, J. & Van Den Hurk, B. (2021), 'Storylines of weather-induced crop failure events under climate change', *Earth System Dynamics* **12**(4), 1503–1527.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K. & Kumari, M. (2019), 'Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils', *Water* **11**(9), 1940.

Guo, W. W., Xue, H. et al. (2014), 'Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models', *Mathematical problems in Engineering* **2014**.

Hachigonta, S. & Reason, C. (2006), 'Interannual variability in dry and wet spell characteristics over Zambia', *Climate Research* **32**(1), 49–62.

Hachigonta, S., Reason, C. & Tadross, M. (2008), 'An analysis of onset date and rainy season duration over Zambia', *Theoretical and applied climatology* **91**, 229–243.

Haile, M. G., Wossen, T., Tesfaye, K. & von Braun, J. (2017), 'Impact of climate change, weather extremes, and price risk on global food supply', *Economics of Disasters and Climate Change* **1**, 55–75.

Haley, P. & Soloway, D. (1992), Extrapolation limitations of multilayer feedforward neural networks, *in* '[Proceedings 1992] IJCNN International Joint Conference on Neural Networks', Vol. 4, pp. 25–30 vol.4.

Hansen, J. & Jones, J. (2000), 'Scaling-up crop models for climate variability applications', *Agricultural Systems* **65**(1), 43–72.

Hawkins, E., Fricker, T. E., Challinor, A. J., Ferro, C. A., Ho, C. K. & Osborne, T. M. (2013), 'Increasing influence of heat stress on french maize yields from the 1960s to the 2030s', *Global change biology* **19**(3), 937–947.

Hawkins, E. & Sutton, R. (2012), 'Time of emergence of climate signals', *Geophysical Research Letters* **39**(1).

Haylock, M. R., Hofstra, N., Tank, A. M. G. K., Klok, E. J., Jones, P. D. & New, M. (2008), 'A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006', *Journal of Geophysical Research: Atmospheres* **113**(D20).

He, X., Chaney, N. W., Schleiss, M. & Sheffield, J. (2016), 'Spatial downscaling of precipitation using adaptable random forests', *Water resources research* **52**(10), 8217–8237.

Heinicke, S., Frieler, K., Jägermeyr, J. & Mengel, M. (2022), 'Global gridded crop models underestimate yield responses to droughts and heatwaves', *Environmental Research Letters* **17**(4), 044026.

Hendrycks, D., Lee, K. & Mazeika, M. (2019), Using pre-training can improve model robustness and uncertainty, *in* 'International Conference on Machine Learning', PMLR, pp. 2712–2721.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hoffman, A. L., Kemanian, A. R. & Forest, C. E. (2018), 'Analysis of climate signals in the crop yield record of sub-saharan Africa', *Global change biology* **24**(1), 143–157.

Hoffmann, H., Baranowski, P., Krzyszczak, J., Zubik, M., Sławiński, C., Gaiser, T. & Ewert, F. (2017), 'Temporal properties of spatially aggregated meteorological time series', *Agricultural and Forest Meteorology* **234-235**, 247–257.

Hoffmann, H., Zhao, G., Asseng, S., Bindi, M., Biernath, C., Constantin, J., Coucheney, E., Dechow, R., Doro, L., Eckersten, H. et al. (2016), 'Impact of spatial soil and climate input data aggregation on regional yield simulations', *PloS one* **11**(4), e0151782.

Hofmann, T., Schölkopf, B. & Smola, A. J. (2008), 'Kernel methods in machine learning', *The annals of statistics* **36**(3), 1171–1220.

Hoogenboom, G., Porter, C. H., Boote, K. J., Shelia, V., Wilkens, P. W., Singh, U., White, J. W., Asseng, S., Lizaso, J. I., Moreno, L. P. et al. (2019), The DSSAT crop modeling ecosystem, *in* 'Advances in crop modelling for a sustainable agriculture', Burleigh Dodds Science Publishing, pp. 173–216.

Hornik, K. (1991), 'Approximation capabilities of multilayer feedforward networks', *Neural networks* **4**(2), 251–257.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T. & Yang, H. (2019), 'Machine learning and Artificial Intelligence to aid climate change research and preparedness', *Environmental Research Letters* **14**(12), 124007.

Iizumi, T., Luo, J.-J., Challinor, A., Sakurai, G., Yokozawa, M., Sakuma, H., Brown, M. & Yamagata, T. (2014a), 'Impacts of el niño southern oscillation on the global yields of major crops', *Nature communications* **5**(1), 1–7.

Iizumi, T. & Sakai, T. (2020), 'The global dataset of historical yields for major crops 1981–2016', *Scientific Data* **7**(1), 1–7.

Iizumi, T., Sakuma, H., Yokozawa, M., Luo, J.-J., Challinor, A. J., Brown, M. E., Sakurai, G. & Yamagata, T. (2013), 'Prediction of seasonal climate-induced variations in global food production', *Nature climate change* **3**(10), 904–908.

Iizumi, T., Tanaka, Y., Sakurai, G., Ishigooka, Y. & Yokozawa, M. (2014c), 'Dependency

of parameter values of a crop model on the spatial scale of simulation', *Journal of Advances in Modeling Earth Systems* **6**(3), 527–540.

Iizumi, T., Yokozawa, M., Sakurai, G., Travasso, M. I., Romanenkov, V., Oettli, P., Newby, T., Ishigooka, Y. & Furuya, J. (2014b), 'Historical changes in global yields: major cereal and legume crops from 1982 to 2006', *Global ecology and biogeography* **23**(3), 346–357.

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., Franke, J. A. et al. (2021), 'Climate impacts on global agriculture emerge earlier in new generation of climate and crop models', *Nature Food* **2**(11), 873–885.

Jennings, S. A., Challinor, A. J., Smith, P., Macdiarmid, J. I., Pope, E., Chapman, S., Bradshaw, C., Clark, H., Vetter, S., Fitton, N. et al. (2022), 'A new integrated assessment framework for climate-smart nutrition security in sub-saharan Africa: the integrated future estimator for emissions and diets (ifeed)', *Frontiers in Sustainable Food Systems* p. 278.

Jiménez, D., Cock, J., Satizábal, H. F., Pérez-Uribe, A., Jarvis, A., Van Damme, P. et al. (2009), 'Analysis of Andean Blackberry (rubus glaucus) production models obtained by means of artificial neural networks exploiting information collected by small-scale growers in colombia and publicly available meteorological data', *Computers and electronics in agriculture* **69**(2), 198–208.

Jung, J., Han, H., Kim, K. & Kim, H. S. (2021), 'Machine learning-based small hydropower potential prediction under climate change', *Energies* **14**(12), 3643.

Jury, M. & Mwafulirwa, N. (2002b), 'Climate variability in Malawi, part 1: dry summers, statistical associations and predictability', *International Journal of Climatology: A Journal of the Royal Meteorological Society* **22**(11), 1289–1302.

Jury, M. R. (2002a), 'Economic impacts of climate variability in South Africa and development of resource prediction models', *Journal of Applied Meteorology and Climatology* **41**(1), 46–55.

Kasampalis, D. A., Alexandridis, T. K., Deva, C., Challinor, A., Moshou, D. & Zalidis, G. (2018), 'Contribution of remote sensing on crop models: a review', *Journal of Imaging* **4**(4), 52.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N., Meinke, H., Hochman, Z. et al. (2003), 'An overview of APSIM, a model designed for farming systems simulation', *European journal of agronomy* **18**(3-4), 267–288.

Kelly, T., Foster, T. & Schultz, D. M. (2023), 'Assessing the value of adapting irrigation strategies within the season', *Agricultural Water Management* **275**, 107986.

Kennel, C. F., Briggs, S. & Victor, D. G. (2016), 'Making climate science more relevant', *Science* **354**(6311), 421–422.

Khaki, S. & Wang, L. (2019), 'Crop yield prediction using deep neural networks', *Frontiers in plant science* **10**, 621.

Kijazi, A. & Reason, C. (2005), 'Relationships between intraseasonal rainfall variability of coastal Tanzania and enso', *Theoretical and applied climatology* **82**, 153–176.

Kim, D. & Kaluarachchi, J. (2015), 'Validating fao aquacrop using landsat images and regional crop information', *Agricultural Water Management* **149**, 143–155.

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Gourley, J. J. & Hong, Y. (2021), 'Can Artificial Intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study

of four watersheds with different hydro-climatic regions across the conus', *Journal of Hydrology* **598**, 126423.

Kimball, B. A. (2016), 'Crop responses to elevated CO2 and interactions with H2O, n, and temperature', *Current Opinion in Plant Biology* **31**, 36–43. SI: 31: Physiology and metabolism 2016.

Knusel, B., Zumwald, M., Baumberger, C., Hadorn, G. H., Fischer, E. M. F., Bresch, D. N. & Knutti, R. (2019), 'Applying big data beyond small problems in climate research', *Nature climate change* **9**(3), 196–202.

Kruger, A. (1999), 'The influence of the decadal-scale variability of summer rainfall on the impact of el niño and la niña events in South Africa', *International Journal of Climatology: A Journal of the Royal Meteorological Society* **19**(1), 59–68.

Kruger, A. C. & Nxumalo, M. (2017), 'Historical rainfall trends in South Africa: 1921–2015', *Water SA* **43**(2), 285–297.

Kumari, M. & Soni, S. (2017), 'A review of classification in web usage mining using k-nearest neighbour', *Advances in Computational Sciences and Technology* **10**(5), 1405–1416.

Labus, M., Nielsen, G., Lawrence, R., Engel, R. & Long, D. (2002), 'Wheat yield estimates using multi-temporal NDVI satellite imagery', *International Journal of Remote Sensing* **23**(20), 4169–4180.

Lamorski, K., Pachepsky, Y., Sławiński, C. & Walczak, R. (2008), 'Using support vector machines to develop pedotransfer functions for water retention of soils in Poland', *Soil Science Society of America Journal* **72**(5), 1243–1247.

Lange, S. (2016), 'EartH2Observe, WFDEI and ERA-Interim data merged and bias-corrected for ISIMIP (EWEMBI)'.

Lange, S. (2018), 'Bias correction of surface downwelling longwave and shortwave radiation for the EWEMBI dataset', *Earth System Dynamics* **9**(2), 627–645.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436–444.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. & Jackel, L. (1989), 'Handwritten digit recognition with a back-propagation network', *Advances in neural information processing systems* **2**.

Leng, G. & Hall, J. W. (2020), 'Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models', *Environmental Research Letters* **15**(4), 044027.

Lesk, C., Rowhani, P. & Ramankutty, N. (2016), 'Influence of extreme weather disasters on global crop production', *Nature* **529**(7584), 84–87.

Li, H., Keune, J., Smessaert, F., Nieto, R., Gimeno, L. & Miralles, D. G. (2023), 'Land–atmosphere feedbacks contribute to crop failure in global rainfed breadbaskets', *npj Climate and Atmospheric Science* **6**(1), 51.

Li, Y.-F., Wu, Y., Hernandez-Espinosa, N. & Peña, R. J. (2013), 'Heat and drought stress on durum wheat: Responses of genotypes, yield, and quality parameters', *Journal of Cereal Science* **57**(3), 398–404.

Lischeid, G., Webber, H., Sommer, M., Nendel, C. & Ewert, F. (2022), 'Machine learning in crop yield modelling: A powerful tool, but no surrogate for science', *Agricultural and Forest Meteorology* **312**, 108698.

Liu, P., Choo, K.-K. R., Wang, L. & Huang, F. (2017), 'Svm or deep learning? a comparative study on remote sensing image classification', *Soft Computing* **21**(23), 7053–7065.

Lobell, D. B. & Burke, M. B. (2010), 'On the use of statistical models to predict crop yield responses to climate change', *Agricultural and forest meteorology* **150**(11), 1443–1452.

Lobell, D. B., Burke, M. B., Tebaldi, C., Mastrandrea, M. D., Falcon, W. P. & Naylor, R. L. (2008a), 'Prioritizing climate change adaptation needs for food security in 2030', *Science* **319**(5863), 607–610.

Lobell, D. B., Deines, J. M. & Tommaso, S. D. (2020), 'Changes in the drought sensitivity of us maize yields', *Nature Food* **1**(11), 729–735.

Lobell, D. B. & Field, C. B. (2008b), 'Estimation of the carbon dioxide (co2) fertilization effect using growth rate anomalies of co2 and crop yields since 1961', *Global Change Biology* **14**(1), 39–45.

Lutz, F., Herzfeld, T., Heinke, J., Rolinski, S., Schaphoff, S., Von Bloh, W., Stoorvogel, J. J. & Müller, C. (2019), 'Simulating the effect of tillage practices with the global ecosystem model LPJmL (version 5.0-tillage)', *Geoscientific Model Development* **12**(6), 2419–2440.

Lyu, K., Zhang, X., Church, J. A., Slangen, A. & Hu, J. (2014), 'Time of emergence for regional sea-level change', *Nature Climate Change* **4**(11), 1006–1010.

Mahlalela, P., Blamey, R. & Reason, C. (2019), 'Mechanisms behind early winter rainfall variability in the southwestern cape, South Africa', *Climate Dynamics* **53**, 21–39.

Manoharan, A., Begam, K., Aparow, V. R. & Sooriamoorthy, D. (2022), 'Artificial neural networks, gradient boosting and support vector machines for electric vehicle battery state estimation: A review', *Journal of Energy Storage* **55**, 105384.

Marsland, S. (2011), *Machine learning: an algorithmic perspective*, Chapman and Hall/CRC.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., Boote, K. J., Ruane, A. C., Thorburn, P. J., Cammarano, D. et al. (2015), 'Multimodel ensembles of wheat growth: many models are better than one', *Global change biology* **21**(2), 911–925.

Maryasin, B., Marquetand, P. & Maulide, N. (2018), 'Machine learning for organic synthesis: are robots replacing chemists?', *Angewandte Chemie International Edition* **57**(24), 6978–6980.

Marzano, F. S., Rivolta, G., Coppola, E., Tomassetti, B. & Verdecchia, M. (2007), 'Rainfall nowcasting from multisatellite passive-sensor images using a recurrent neural network', *IEEE Transactions on Geoscience and Remote Sensing* **45**(11), 3800–3812.

Maxwell, D. & Fitzpatrick, M. (2012), 'The 2011 somalia famine: Context, causes, and complications', *Global Food Security* **1**(1), 5–12.

Mendelsohn, R. (2007), 'What causes crop failure?', *Climatic change* **81**(1), 61–70.

Mendez, G. & Lohr, S. (2011), 'Estimating residual variance in random forest regression', *Computational Statistics & Data Analysis* **55**(11), 2937–2950.

Minoli, S., Müller, C., Elliott, J., Ruane, A. C., Jägermeyr, J., Zabel, F., Dury, M., Folberth, C., François, L., Hank, T. et al. (2019), 'Global response patterns of major rainfed crops to adaptation by maintaining current growing periods and irrigation', *Earth's Future* **7**(12), 1464–1480.

Mitchell, J. F., Lowe, J., Wood, R. A. & Vellinga, M. (2006), 'Extreme events due to human-induced climate change', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **364**(1845), 2117–2133.

Mitchell, T. M. (1997), 'Machine learning', *Burr Ridge, IL: McGraw Hill* .

Mkhabela, M., Bullock, P., Raj, S., Wang, S. & Yang, Y. (2011), 'Crop yield forecasting on the canadian prairies using MODIS NDVI data', *Agricultural and Forest Meteorology* **151**(3), 385–393.

Molnar, C. (2022), *Interpretable Machine Learning*, 2 edn.
URL: *https://christophm.github.io/interpretable-ml-book*

Monfreda, C., Ramankutty, N. & Foley, J. A. (2008), 'Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000', *Global biogeochemical cycles* **22**(1).

Monteith, J. L. (1996), 'The quest for balance in crop modeling', *Agronomy Journal* **88**(5), 695–697.

Monteith, J. L. & Moss, C. J. (1977), 'Climate and the efficiency of crop production in britain [and discussion]', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **281**(980), 277–294.

Moriondo, M., Maselli, F. & Bindi, M. (2007), 'A simple model of regional wheat yield based on NDVI data', *European Journal of Agronomy* **26**(3), 266–274.

Msowoya, K., Madani, K., Davtalab, R., Mirchi, A. & Lund, J. R. (2016), 'Climate change impacts on maize production in the warm heart of Africa', *Water Resources Management* **30**, 5299–5312.

Mulenga, B. P., Wineman, A. & Sitko, N. J. (2017), 'Climate trends and farmers' perceptions of climate change in Zambia', *Environmental management* **59**, 291–306.

Müller, C., Cramer, W., Hare, W. L. & Lotze-Campen, H. (2011), 'Climate change risks for african agriculture', *Proceedings of the national academy of sciences* **108**(11), 4313–4315.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S. et al. (2017), 'Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications', *Geoscientific Model Development* **10**(4), 1403–1422.

Nalepa, J. & Kawulok, M. (2019), 'Selecting training sets for support vector machines: a review', *Artificial Intelligence Review* **52**(2), 857–900.

Nazarenko, E., Varkentin, V. & Polyakova, T. (2019), Features of application of machine learning methods for classification of network traffic (features, advantages, disadvantages), *in* '2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)', pp. 1–5.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C. & Gupta, H. V. (2021), 'What role does hydrological science play in the age of machine learning?', *Water Resources Research* **57**(3), e2020WR028091.

Newlands, N. K., Ghahari, A., Gel, Y. R., Lyubchich, V. & Mahdi, T. (2019), Deep learning for improved agricultural risk management, *in* 'HICSS'.

Nguyen, K. T. N., François, B., Balasubramanian, H., Dufour, A. & Brown, C. (2023), 'Prediction of water quality extremes with composite quantile regression neural network', *Environmental Monitoring and Assessment* **195**(2), 284.

Nicklin, K. J. (2013), Seasonal crop yield forecasting in semi-arid West Africa, PhD thesis, University of Leeds.

Nwokolo, S. C., Obiwulu, A. U. & Ogbulezie, J. C. (2023), 'Machine learning and analytical model hybridization to assess the impact of climate change on solar pv energy production', *Physics and Chemistry of the Earth, Parts A/B/C* **130**, 103389.

Olive, D. J. & Olive, D. J. (2017), *Multiple linear regression*, Springer.

Ortiz-Bobea, A., Ault, T. R., Carrillo, C. M., Chambers, R. G. & Lobell, D. B. (2021), 'Anthropogenic climate change has slowed global agricultural productivity growth', *Nature Climate Change* **11**(4), 306–312.

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R. & Wheeler, T. (2015), 'Jules-crop: a parametrisation of crops in the joint uk land environment simulator', *Geoscientific Model Development* **8**(4), 1139–1155.

Osborne, T. M., Lawrence, D. M., Challinor, A. J., Slingo, J. M. & Wheeler, T. R. (2007), 'Development and assessment of a coupled crop–climate model', *Global Change Biology* **13**(1), 169–183.

Osborne, T., Rose, G. & Wheeler, T. (2013), 'Variation in the global-scale impacts of climate change on crop productivity due to climate model uncertainty and adaptation', *Agricultural and Forest Meteorology* **170**, 183–194.

Ossó, A., Allan, R. P., Hawkins, E., Shaffrey, L. & Maraun, D. (2022), 'Emerging new climate extremes over Europe', *Climate Dynamics* **58**(1), 487–501.

Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., Patil, R. H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F., Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B. & Rötter, R. (2011), 'Simulation of winter wheat yield and its variability in different climates of europe: A comparison of eight crop growth models', *European Journal of Agronomy* **35**(3), 103–114.
URL: *https://www.sciencedirect.com/science/article/pii/S1161030111000542*

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R. & Mouazen, A. (2016), 'Wheat yield prediction using machine learning and advanced sensing techniques', *Computers and Electronics in Agriculture* **121**, 57–65.

Park, S.-J. & Lee, D.-k. (2021), 'Predicting susceptibility to landslides under climate change impacts in metropolitan areas of south korea using machine learning', *Geomatics, Natural Hazards and Risk* **12**(1), 2462–2476.

Pasley, H., Brown, H., Holzworth, D., Whish, J., Bell, L. & Huth, N. (2023), 'How to build a crop model. a review', *Agronomy for Sustainable Development* **43**(1), 1–12.

Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C. & Sanderson, B. M. (2017), 'Precipitation variability increases in a warmer climate', *Scientific reports* **7**(1), 1–9.

Perol, T., Gharbi, M. & Denolle, M. (2018), 'Convolutional neural network for earthquake detection and location', *Science Advances* **4**(2), e1700578.

Pham, H. & Olafsson, S. (2019), 'Bagged ensembles with tunable parameters', *Computational Intelligence* **35**(1), 184–203.

Pham, K., Kim, D., Le, C. V. & Won, J. (2023), 'Machine learning-based pedotransfer functions to predict soil water characteristics curves', *Transportation Geotechnics* p. 101052.

Pielke Jr, R. A. & Landsea, C. N. (1999), 'La nina, el nino, and atlantic hurricane damages in the United States', *Bulletin of the American Meteorological Society* **80**(10), 2027–2034.

Porter, J. R. & Gawith, M. (1999), 'Temperatures and the growth and development of wheat: a review', *European Journal of Agronomy* **10**(1), 23–36.

Porter, J. R. & Semenov, M. A. (2005), 'Crop responses to climatic variation', *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1463), 2021–2035.

Porter, J. R., Xie, L., Challinor, A. J., Cochrane, K., Howden, S. M., Iqbal, M. M., Lobell, D. B. & Travasso, M. I. (2014), 'Food security and food production systems'.

Portmann, F. T., Siebert, S. & Döll, P. (2010), 'Mirca2000—global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling', *Global biogeochemical cycles* **24**(1).

Prasad, A. M., Iverson, L. R. & Liaw, A. (2006), 'Newer classification and regression tree techniques: bagging and random forests for ecological prediction', *Ecosystems* **9**, 181–199.

Prasad, P. V., Bheemanahalli, R. & Jagadish, S. K. (2017), 'Field crops and the fear

of heat stress—opportunities, challenges and future directions', *Field Crops Research* **200**, 114–121.

Prasetya, E. P. & Djamal, E. C. (2019), Rainfall forecasting for the natural disasters preparation using recurrent neural networks, *in* '2019 International Conference on Electrical Engineering and Informatics (ICEEI)', IEEE, pp. 52–57.

Priestley, C. H. B. & Taylor, R. J. (1972), 'On the assessment of surface heat flux and evaporation using large-scale parameters', *Monthly weather review* **100**(2), 81–92.

Proctor, J. (2021), 'Atmospheric opacity has a nonlinear effect on global crop yields', *Nature Food* **2**(3), 166–173.

Proctor, J., Rigden, A., Chan, D. & Huybers, P. (2022), 'More accurate specification of water supply shows its importance for global crop production', *Nature Food* **3**(9), 753–763.

Qiu, J., Wu, Q., Ding, G., Xu, Y. & Feng, S. (2016), 'A survey of machine learning for big data processing', *EURASIP Journal on Advances in Signal Processing* **2016**(1), 1–16.

Radford, A., Metz, L. & Chintala, S. (2015), 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv preprint arXiv:1511.06434* .

Raes, D., Steduto, P., Hsiao, T. C. & Fereres, E. (2009), 'AquaCrop—the fao crop model to simulate yield response to water: Ii. main algorithms and software description', *Agronomy Journal* **101**(3), 438–447.

Rajulapati, C. R., Papalexiou, S. M., Clark, M. P. & Pomeroy, J. W. (2021), 'The perils of regridding: examples using a global precipitation dataset', *Journal of Applied Meteorology and Climatology* **60**(11), 1561–1573.

Ray, D. K., Gerber, J. S., MacDonald, G. K. & West, P. C. (2015), 'Climate variation explains a third of global crop yield variability', *Nature communications* **6**(1), 5989.

Razavi, S., Hannah, D. M., Elshorbagy, A., Kumar, S., Marshall, L., Solomatine, D. P., Dezfuli, A., Sadegh, M. & Famiglietti, J. (2022), 'Coevolution of machine learning and process-based modelling to revolutionize earth and environmental sciences: A perspective', *Hydrological Processes* **36**(6), e14596.

Reason, C. & Rouault, M. (2002), 'Enso-like decadal variability and South African rainfall', *Geophysical Research Letters* **29**(13), 16–1.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. & Prabhat (2019), 'Deep learning and process understanding for data driven earth system science', *Nature* **566**(7743), 195–204.

Reidsma, P., Ewert, F., Lansink, A. O. & Leemans, R. (2010), 'Adaptation to climate change and climate variability in European agriculture: the importance of farm level responses', *European journal of agronomy* **32**(1), 91–102.

Rind, D., Goldberg, R., Hansen, J., Rosenzweig, C. & Ruedy, R. (1990), 'Potential evapotranspiration and the likelihood of future drought', *Journal of Geophysical Research: Atmospheres* **95**(D7), 9983–10004.

Rind, D., Goldberg, R. & Ruedy, R. (1989), 'Change in climate variability in the 21st century', *Climatic change* **14**(1), 5–37.

Robinson, A., Lehmann, J., Barriopedro, D., Rahmstorf, S. & Coumou, D. (2021), 'Increasing heat and rainfall extremes now far outside the historical climate', *npj Climate and Atmospheric Science* **4**(1), 1–4.

Romero, C. C., Hoogenboom, G., Baigorria, G. A., Koo, J., Gijsman, A. J. & Wood, S. (2012), 'Reanalysis of a global soil database for crop and environmental modeling', *Environ. Model. Softw.* **35**(C), 163–170.

Rosenblatt, F. (1961), Principles of neurodynamics. perceptrons and the theory of brain mechanisms, Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T. A. M., Schmid, E., Stehfest, E., Yang, H. & Jones, J. W. (2014), 'Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison', *Proceedings of the National Academy of Sciences* **111**(9), 3268–3273.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., Antle, J. M., Nelson, G. C., Porter, C., Janssen, S. et al. (2013), 'The agricultural model intercomparison and improvement project (agmip): protocols and pilot studies', *Agricultural and Forest Meteorology* **170**, 166–182.

Rowhani, P., Lobell, D. B., Linderman, M. & Ramankutty, N. (2011), 'Climate variability and crop production in Tanzania', *Agricultural and Forest Meteorology* **151**(4), 449–460.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature machine intelligence* **1**(5), 206–215.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), 'Learning representations by back-propagating errors', *nature* **323**(6088), 533–536.

Sacks, W. J., Deryng, D., Foley, J. A. & Ramankutty, N. (2010), 'Crop planting dates: an analysis of global patterns', *Global ecology and biogeography* **19**(5), 607–620.

Sagi, O. & Rokach, L. (2018), 'Ensemble learning: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4), e1249.

Saleem, M. H., Potgieter, J. & Arif, K. M. (2019), 'Plant disease detection and classification by deep learning', *Plants* **8**(11), 468.

Sánchez, B., Rasmussen, A. & Porter, J. R. (2014), 'Temperatures and the growth and development of maize and rice: a review', *Global change biology* **20**(2), 408–417.

Saxton, K. E. & Rawls, W. J. (2006), 'Soil water characteristic estimates by texture and organic matter for hydrologic solutions', *Soil science society of America Journal* **70**(5), 1569–1578.

Saxton, K., Rawls, W., Romberger, J. S. & Papendick, R. (1986), 'Estimating generalized soil-water characteristics from texture', *Soil science society of America Journal* **50**(4), 1031–1036.

Schaphoff, S., von Bloh, W., Rammig, A., Thonicke, K., Biemans, H., Forkel, M., Gerten, D., Heinke, J., Jägermeyr, J., Knauer, J. et al. (2018), 'LPJmL4–a dynamic global vegetation model with managed land–part 1: Model description', *Geoscientific Model Development* **11**(4), 1343–1375.

Schewe, J., Gosling, S. N., Reyer, C., Zhao, F., Ciais, P., Elliott, J., Francois, L., Huber, V., Lotze, H. K., Seneviratne, S. I. et al. (2019), 'State-of-the-art global models underestimate impacts from climate extremes', *Nature communications* **10**(1), 1–14.

Schlenker, W. & Roberts, M. J. (2009), 'Nonlinear temperature effects indicate severe damages to US crop yields under climate change', *Proceedings of the National Academy of sciences* **106**(37), 15594–15598.

Schmidhuber, J. & Tubiello, F. N. (2007), 'Global food security under climate change', *Proceedings of the National Academy of Sciences* **104**(50), 19703–19708.

Schulz, K., Beven, K. & Huwe, B. (1999), 'Equifinality and the problem of robust calibration in nitrogen budget simulations', *Soil Science Society of America Journal* **63**(6), 1934–1941.

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. & Ciampitti, I. A. (2020), 'Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil', *Agricultural and Forest Meteorology* **284**, 107886.

Seidel, S., Palosuo, T., Thorburn, P. & Wallach, D. (2018), 'Towards improved calibration of crop models – where are we now and where should we go?', *European Journal of Agronomy* **94**, 25–35.

Semenov, M. & Porter, J. (1995), 'Climatic variability and the modelling of crop yields', *Agricultural and Forest Meteorology* **73**(3), 265–283. Biospheric Aspects of the Hydrological Cycle.

Semmler, G., Wernly, S., Wernly, B., Mamandipoor, B., Bachmayer, S., Semmler, L., Aigner, E., Datz, C. & Osmani, V. (2021), 'Machine learning models cannot replace screening colonoscopy for the prediction of advanced colorectal adenoma', *Journal of Personalized Medicine* **11**(10), 981.

Shahhosseini, M., Hu, G. & Archontoulis, S. V. (2020), 'Forecasting corn yield with machine learning ensembles', *Frontiers in Plant Science* **11**.

Shahhosseini, M., Hu, G., Huber, I. & Archontoulis, S. V. (2021), 'Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt', *Scientific reports* **11**(1), 1606.

Shahhosseini, M., Martinez-Feria, R. A., Hu, G. & Archontoulis, S. V. (2019), 'Maize yield and nitrate loss prediction with machine learning algorithms', *Environmental Research Letters* **14**(12), 124026.

Siegert, F., Ruecker, G., Hinrichs, A. & Hoffmann, A. (2001), 'Increased damage from fires in logged forests during droughts caused by El $\tilde{N}$ino', *Nature* **414**(6862), 437–440.

Silini, R., Barreiro, M. & Masoller, C. (2021), 'Machine learning prediction of the madden-julian oscillation', *npj Climate and Atmospheric Science* **4**(1), 57.

Son, N., Chen, C., Chen, C., Minh, V. & Trung, N. (2014), 'A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation', *Agricultural and Forest Meteorology* **197**, 52–64.

Stige, L. C., Stave, J., Chan, K.-S., Ciannelli, L., Pettorelli, N., Glantz, M., Herren, H. R. & Stenseth, N. C. (2006), 'The effect of climate variation on agro-pastoral production in Africa', *Proceedings of the National Academy of Sciences* **103**(9), 3049–3053.

Sullivan, M., VanToai, T., Fausey, N., Beuerlein, J., Parkinson, R. & Soboyejo, A. (2001), 'Evaluating on-farm flooding impacts on soybean', *Crop Science* **41**(1), 93–100.

Sutskever, I., Martens, J. & Hinton, G. E. (2011), Generating text with recurrent neural networks, *in* 'Proceedings of the 28th international conference on machine learning (ICML-11)', pp. 1017–1024.

Tallec, T., Béziat, P., Jarosz, N., Rivalland, V. & Ceschia, E. (2013), 'Crops' water use efficiencies in temperate climate: Comparison of stand, ecosystem and agronomical approaches', *Agricultural and Forest Meteorology* **168**, 69–81.

Tariq, M., Ahmad, S., Fahad, S., Abbas, G., Hussain, S., Fatima, Z., Nasim, W., Mubeen, M., ur Rehman, M. H., Khan, M. A. et al. (2018), 'The impact of climate warming and crop management on phenology of sunflower-based cropping systems in Punjab, Pakistan', *Agricultural and Forest Meteorology* **256**, 270–282.

Taunk, K., De, S., Verma, S. & Swetapadma, A. (2019), A brief review of nearest neighbor algorithm for learning and classification, *in* '2019 International Conference on Intelligent Computing and Control Systems (ICCS)', pp. 1255–1260.

Teixeira, E. I., Fischer, G., Van Velthuizen, H., Walter, C. & Ewert, F. (2013), 'Global hot-spots of heat stress on agricultural crops due to climate change', *Agricultural and Forest Meteorology* **170**, 206–215.

Thiery, W., Lange, S., Rogelj, J., Schleussner, C.-F., Gudmundsson, L., Seneviratne, S. I., Andrijevic, M., Frieler, K., Emanuel, K., Geiger, T., Bresch, D. N., Zhao, F., Willner, S. N., Büchner, M., Volkholz, J., Bauer, N., Chang, J., Ciais, P., Dury, M., François, L., Grillakis, M., Gosling, S. N., Hanasaki, N., Hickler, T., Huber, V., Ito, A., Jägermeyr, J., Khabarov, N., Koutroulis, A., Liu, W., Lutz, W., Mengel, M., Müller, C., Ostberg, S., Reyer, C. P. O., Stacke, T. & Wada, Y. (2021), 'Intergenerational inequities in exposure to climate extremes', *Science* **374**(6564), 158–160.

Thornton, P. K., Ericksen, P. J., Herrero, M. & Challinor, A. J. (2014), 'Climate variability and vulnerability to climate change: a review', *Global change biology* **20**(11), 3313–3328.

Torgo, L., Branco, P., Ribeiro, R. P. & Pfahringer, B. (2015), 'Resampling strategies for regression', *Expert Systems* **32**(3), 465–476.

Troy, T. J., Kipgen, C. & Pal, I. (2015), 'The impact of climate extremes and irrigation on US crop yields', *Environmental Research Letters* **10**(5), 054013.

Urban, O., Hlaváčová, M., Klem, K., Novotná, K., Rapantová, B., Smutná, P., Horáková, V., Hlavinka, P., Škarpa, P. & Trnka, M. (2018), 'Combined effects of drought and high temperature on photosynthetic characteristics in four winter wheat genotypes', *Field Crops Research* **223**, 137–149.

van Bussel, L. G., Ewert, F., Zhao, G., Hoffmann, H., Enders, A., Wallach, D., Asseng, S., Baigorria, G. A., Basso, B., Biernath, C., Cammarano, D., Chryssanthacopoulos, J., Constantin, J., Elliott, J., Glotter, M., Heinlein, F., Kersebaum, K.-C., Klein, C., Nendel, C., Priesack, E., Raynal, H., Romero, C. C., Rötter, R. P., Specka, X. &

Tao, F. (2016), 'Spatial sampling of weather data for regional crop yield simulations', *Agricultural and Forest Meteorology* **220**, 101–115.

van Bussel, L. G., Grassini, P., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., de Groot, H., Saito, K., Cassman, K. G. et al. (2015a), 'From field to atlas: Upscaling of location-specific yield gap estimates', *Field Crops Research* **177**, 98–108.

Van Bussel, L., Müller, C., Van Keulen, H., Ewert, F. & Leffelaar, P. (2011), 'The effect of temporal aggregation of weather input data on crop growth models' results', *Agricultural and forest meteorology* **151**(5), 607–619.

van der Velde, M., Tubiello, F. N., Vrieling, A. & Bouraoui, F. (2012), 'Impacts of extreme weather on wheat and maize in France: evaluating regional crop simulations against observed data', *Climatic change* **113**, 751–765.

van der Velde, M., Wriedt, G. & Bouraoui, F. (2010), 'Estimating irrigation use and effects on maize yield during the 2003 heatwave in France', *Agriculture, Ecosystems & Environment* **135**(1), 90–97.

van Ittersum, M. K., Ewert, F., Heckelei, T., Wery, J., Alkan Olsson, J., Andersen, E., Bezlepkina, I., Brouwer, F., Donatelli, M., Flichman, G., Olsson, L., Rizzoli, A. E., van der Wal, T., Wien, J. E. & Wolf, J. (2008), 'Integrated assessment of agricultural systems – a component-based framework for the European union (SEAMLESS)', *Agricultural Systems* **96**(1), 150–165.

Van Klompenburg, T., Kassahun, A. & Catal, C. (2020*a*), 'Crop yield prediction using machine learning: A systematic literature review', *Computers and Electronics in Agriculture* **177**, 105709.

van Klompenburg, T., Kassahun, A. & Catal, C. (2020*b*), 'Crop yield prediction using machine learning: A systematic literature review', *Computers and Electronics in Agriculture* **177**, 105709.

Verschuur, J., Li, S., Wolski, P. & Otto, F. E. (2021), 'Climate change as a driver of food insecurity in the 2007 Lesotho-South Africa drought', *Scientific reports* **11**(1), 1–9.

Vogel, J., Rivoire, P., Deidda, C., Rahimi, L., Sauter, C. A., Tschumi, E., Van Der Wiel, K., Zhang, T. & Zscheischler, J. (2020), 'Identifying meteorological drivers of extreme impacts: an application to simulated crop yields', *Earth System Dynamics Discussions* **2020**, 1–27.

Wall, L., Larocque, D. & Léger, P.-M. (2008), 'The early explanatory power of NDVI in crop yield modelling', *International Journal of Remote Sensing* **29**(8), 2211–2225.

Wallach, D. (2011), 'Crop model calibration: A statistical perspective', *Agronomy Journal* **103**(4), 1144–1151.

Wallach, D., Makowski, D., Jones, J. W. & Brun, F. (2006), *Working with dynamic crop models: evaluation, analysis, parameterization, and applications*, Elsevier.

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N. et al. (2021), 'The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise', *Environmental Modelling & Software* **145**, 105206.

Wang, Y., Bobb, J. F., Papi, B., Wang, Y., Kosheleva, A., Di, Q., Schwartz, J. D. & Dominici, F. (2016), 'Heat stroke admissions during heat waves in 1,916 US counties for the period from 1999 to 2010 and their effect modifiers', *Environmental health* **15**(1), 1–9.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S. & Ahrentzen, S. (2018), 'Random forest based hourly building energy prediction', *Energy and Buildings* **171**, 11–25.

Wang, Z., Zhan, C., Ning, L. & Guo, H. (2021), 'Evaluation of global terrestrial evapotranspiration in CMIP6 models', *Theoretical and Applied Climatology* **143**, 521–531.

Warnatzsch, E. A. & Reay, D. S. (2020), 'Assessing climate change projections and impacts on central Malawi's maize yield: The risk of maladaptation', *Science of The Total Environment* **711**, 134845.

Watson, J., Challinor, A. J., Fricker, T. E. & Ferro, C. A. T. (2015), 'Comparing the effects of calibration and climate errors on a statistical crop model and a process-based crop model', *Climatic change* **132**(1), 93–109.

Webb, T. W., Dulberg, Z., Frankland, S. M., Petrov, A. A., O'Reilly, R. C. & Cohen, J. D. (2023), 'Learning representations that support extrapolation'.

Williams, J. W. & Jackson, S. T. (2007), 'Novel climates, no-analog communities, and ecological surprises', *Frontiers in Ecology and the Environment* **5**(9), 475–482.

Witten, I. H. & Frank, E. (2002), *Data mining: practical machine learning tools and techniques with Java implementations*, ACM New York, NY, USA.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y. & Guanter, L. (2020), 'Estimating and understanding crop yields with explainable deep learning in the Indian wheat belt', *Environmental research letters* **15**(2), 024019.

Woldemeskel, F., Sharma, A., Sivakumar, B. & Mehrotra, R. (2016), 'Quantification of precipitation and temperature uncertainties simulated by cmip3 and cmip5 models', *Journal of Geophysical Research: Atmospheres* **121**(1), 3–17.

Wreford, A. & Adger, W. N. (2010), 'Adaptation in agriculture: historic effects of heat waves and droughts on uk agriculture', *International Journal of Agricultural Sustainability* **8**(4), 278–289.

Xu, L., Crammer, K., Schuurmans, D. et al. (2006), Robust support vector machine training via convex outlier ablation, *in* 'AAAI', Vol. 6, pp. 536–542.

Yang, H., Dobbie, S., Ramirez-Villegas, J., Chen, B., Qiu, S., Ghosh, S. & Challinor, A. (2020), 'South India projected to be susceptible to high future groundnut failure rates for future climate change and geo-engineered scenarios', *Science of The Total Environment* **747**, 141240.

Yang, J., Zhang, J., Wang, Z., Zhu, Q. & Liu, L. (2003), 'Involvement of abscisic acid and cytokinins in the senescence and remobilization of carbon reserves in wheat subjected to water stress during grain filling', *Plant, Cell & Environment* **26**(10), 1621–1631.

Zhang, L., Zhang, Z., Tao, F., Luo, Y., Cao, J., Li, Z., Xie, R. & Li, S. (2021), 'Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning', *Environmental Research Letters* **16**(12), 124043.

Zhang, N., Zhou, X., Kang, M., Hu, B.-G., Heuvelink, E. & Marcelis, L. F. (2023), 'Machine learning versus crop growth models: an ally, not a rival', *AoB Plants* **15**(2), plac061.

Zhang, S. & Chen, J. (2021a), 'Uncertainty in projection of climate extremes: A comparison of CMIP5 and CMIP6', *Journal of Meteorological Research* **35**(4), 646–662.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P. et al. (2017), 'Temperature increase reduces global yields of major crops in four independent estimates', *Proceedings of the National Academy of sciences* **114**(35), 9326–9331.

Zhao, Y., Xiao, D., Bai, H., Tang, J., Liu, D. L., Qi, Y. & Shen, Y. (2022), 'The prediction of wheat yield in the north China plain by coupling crop model with machine learning algorithms', *Agriculture* **13**(1), 99.

Zhu, P., Abramoff, R., Makowski, D. & Ciais, P. (2021b), 'Uncovering the past and future climate drivers of wheat yield shocks in Europe with machine learning', *Earth's Future* **9**(5), e2020EF001815.

Zwiers, F. W. & Kharin, V. V. (1998), 'Changes in the extremes of the climate simulated by CCC GCM2 under CO2 doubling', *Journal of Climate* **11**(9), 2200–2222.

8 Appendix: Comparison between calibrated and uncalibrated (YGP 1) French maize simulations

To test the effect of model calibration on the ability of the GLAM model to capture the 2003 yield anomaly, a calibrated and uncalibrated set of model results are compared in Figure 8.1. The comparison shows that both uncalibrated and calibrated model simulations predict significantly decreased yields in 2003. The main differences between the two simulations is that the uncalibrated model tends to have greater variance in model predictions, including more outliers and a larger inter-quartile range for each year. Calibration of the YGP parameter therefore does not effect the ability of the model to capture the 2003 yield anomaly. The main effect of calibration, other than to reduce the magnitude of yields appears to be a reduction in model variability. In the case of 2003 this effect actually causes the model to underestimate the variability of observed yields in comparison to the uncalibrated simulation which produces a distribution of predictions closer to the observed yield anomaly. These findings indicate a bias - variance trade off effect caused by YGP calibration, meaning although YGP calibration biases yields towards a lower more realistic mean value, variance in yield across space is reduced and potentially becomes less realistic. This identifies a further pathway to future model improvement. Future work should more closely analyse the relationship between YGP calibration and bias variance trade offs across spatial scales.

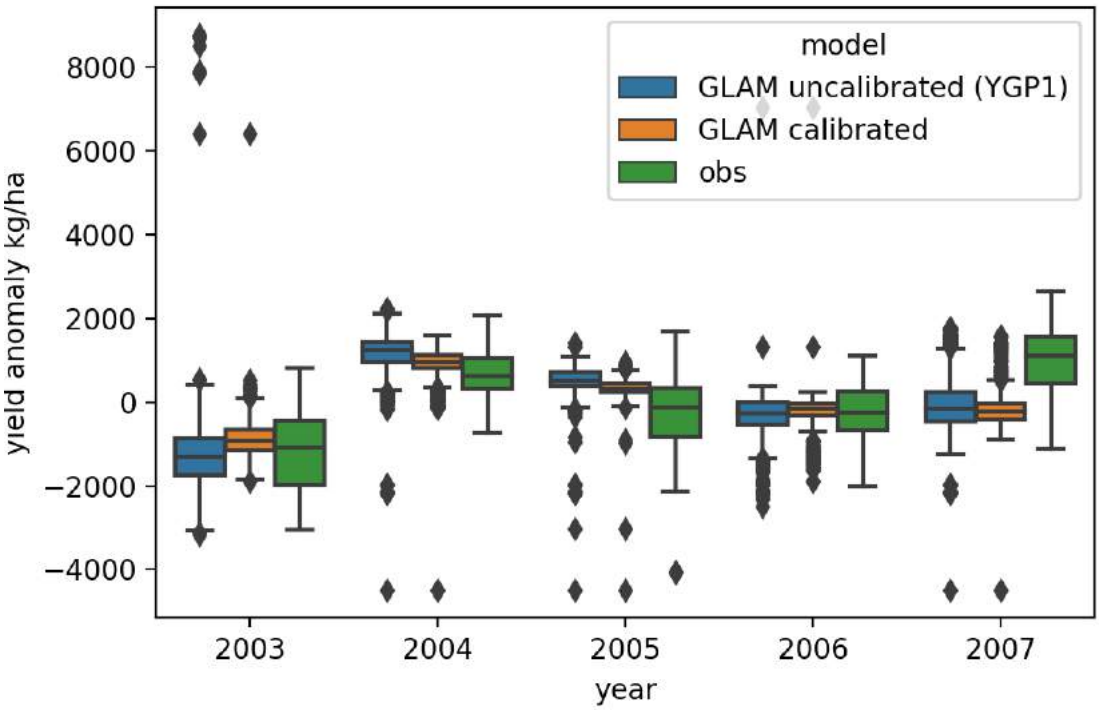

Figure 8.1: Comparison between model simulations between a fully calibrated GLAM model (same as in the main results section of chapter 3, and a simulation in which the yield gap parameter was set to the value of 1 so that it has no effect on simulated leaf area index (and so was left uncalibrated).

9 Appendix: Machine Learning to predict soil moisture parameters

In this section it is demonstrated that machine learning may have the potential to improve crop model predictions through improvements in prediction of the soil moisture water balance. To do this, machine learning (a random forest model) predictions of soil moisture characteristics is compared to predictions of soil moisture characteristics using pedo-transfer functions described in Saxton et al. (1986) and Saxton & Rawls (2006). Pedotransfer functions are sets of empirical correlations used to predict soil moisture char-

acteristics within crop growth models. Pedotransfer functions are described in 2.1.

The machine learning method to predict soil moisture characteristics is benchmarked against existing soil moisture pedotransfer functions in Figure 9.1. This initial test shows that machine learning has the potential to outperform existing empirical methods used in crop models such as GLAM, Aquacrop and LPJML (Jennings et al. 2022, Raes et al. 2009, Lutz et al. 2019). The pedotransfer function methods tend to overestimate larger RLL and DUL values. Furthermore, the random forest model tested greatly improves the RMSE of the prediction over the conventional methods.

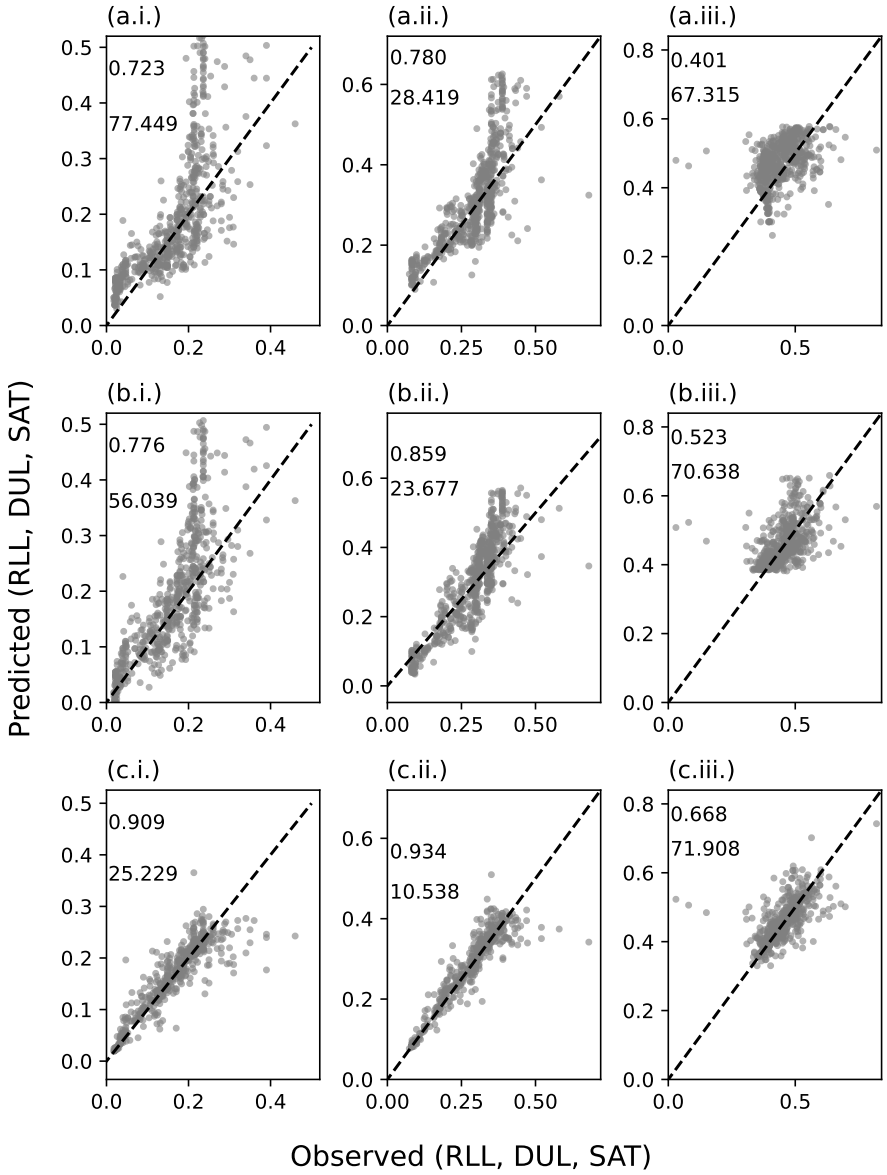

Figure 9.1: Comparison of 3 different methods of predicting soil soil moisture characteristics (RLL-i, DUL-ii and SAT-iii) row (a) corresponds to the pedotransfer function developed in Saxton et al. (1986) and used as part for the simulations in Jennings et al. (2022), row (b) corresponds to the updated pedotransfer function from Saxton & Rawls (2006), row (c) are the results of a random forest machine learning model. Top left corner of each plot shows the correlation coefficient followed by the % RMSE underneath.

This work is related to the outcomes of chapter 5. In the chapter, it was found that more

accurate prediction of soil moisture characteristics would improve correlation between rainfall and simulated crop yield. However, this did not result in improved performance over the standard model calibration (in which optimization of the YGP parameter affects leaf area index), (seen in Figure 5.38). Here it is shown that machine learning has the potential to improve prediction of soil moisture parameters over existing pedotransfer functions. Therefore, it is proposed that for future simulations, instead of trying to calibrate using soil moisture characteristics, the normal method of calibration should be used, and machine learning should also be used to simulate soil moisture parameters.

Soil pedotransfer functions are used across many different crop models and are used as a basis for soil moisture parameterization for many crop modelling studies (e.g. Jägermeyr et al. (2021), Jennings et al. (2022)). Hence, machine learning has the potential to improve soil moisture parameterization, not just for GLAM crop model simulations but for other models as well.