DIAGNOSTICS OF THE THEORETICAL UNDERPINNING OF THE SOCIO-HYDROLOGICAL MODEL IN "WATER EFFICIENCY IN SUSTAINABLE COTTON-BASED PRODUCTION SYSTEMS" PROJECT IN MAHARASHTRA, INDIA

D. Djohan

TU Delft

Diagnostics of the Theoretical Underpinning of the Socio-hydrological Model in "Water Efficiency in Sustainable Cotton-based Production Systems" Project in Maharashtra, India

Evaluation of model performance and the quantification of errors using Monte Carlo sampling, GLUE, linear regressions, linear PCA, and kernel PCA

by

Dennis Djohan

to obtain the degree of Master of Science

at the Delft University of Technology,

defended publicly on April 29, 2021

Student number:	4872983	
Project duration:	March 5, 2020 – April 29, 2021	
Thesis committee:	Dr. S. Pande,	TU Delft, chair
	Dr.Ir. E. Abraham,	TU Delft, supervisor
	Dr.Ir. S. de Vries,	TU Delft, supervisor
	MSc. C.E.M. Luger,	TU Delft, supervisor

Abstract

Background: This research is part of the project "Water Efficiency in Sustainable Cotton-based Production Systems" between Solidaridad Asia and TU Delft. The project aims to increase the livelihood of smallholder farmers in the Maharashtra, India through. A socio-hydrological (SH) model is used extensively in this research as an evaluation tool. However, the baseline research indicates that the lack of stress mechanics in the SH model used in the intervention might cause inaccuracies in yield estimation. Furthermore, it has never been validated at a farmer level before. This research aims to implement the stress mechanics in the SH model, evaluate the overall performance of the model in terms of predicting crop yield, identify potential sources of errors, and give recommendations for future studies.

Method: The research uses iterative top-down approach due to the large study area and the varied nature of the 308 farmers surveyed. The research implements the water and temperature stress mechanics based on Food and Agriculture Organization's (FAO) AquaCrop framework. For the performance evaluation process, this research uses four model scores namely Nash-Sutcliffe (NS), log of NS, Mean Absolute Error (MAE), and the coefficient of determination (r^2). Furthermore, the sources of uncertainties are divided into two categories namely lack of knowledge (i.e. generated by parameter, input, observation, and structural errors) and variability (i.e. generated by climate variations). The lack of knowledge uncertainty is investigated using Monte Carlo Sampling calibration and the Generalized Likelihood Uncertainty Estimation (GLUE) concept is used to obtain the uncertainty intervals of the model. Going further, the errors are divided into residual and structural error. The latter is explained and quantified through a structural error model using a combination of qualitative analysis, Principal Component Analysis, multiple linear regressions, and projection of the data into kernel space. Then residuals between the model yield + structural error vs. the observed yield is thought to be explained by the residual errors. Lastly, the effects of climate variations to the stability of the model are evaluated.

Results and Discussions: After the initial calibration the model has NS value of -0.343 to -0.996, NS_{log} of -0.655 to -1.91, MAE of 447.1 to 553.2 kg/ha, and r^2 of 0.003 to 0.008. Because of the poor performance of the model, the uncertainty intervals from GLUE is not enough to capture the total errors of the model. However, after adjustment using the structural error model, the model scores become NS = 0.83, $NS_{log} = 0.56$, MAE = 149 kg/ha, and $r^2 = 0.859$. The adjusted yield calculation has a residual error as Gaussian distribution with $\sigma = 150$ kg/ha. The qualitative analysis identified several factors that contribute to the errors viz. farmers' capital, irrigation behavior, and crop production process such as canopy cover growth. Lastly, there is no major instability found through the bootstrap analysis.

Conclusions: The physical model is not performing well, especially when it is calculating yield for individual farmers over a large study area. However, the structural error model can adjust the yield prediction so that it is close to the observed yields. This indicates the poor performance is likely to be caused by the prevalence of structural errors in the model instead of the uncertainties regarding parameters, input, or observation values. Therefore, it is recommended for future research to address this first. This can be done by further study and incorporation of more crop production processes, soil water simulation, and exploratory interviews to identify patterns and more factors that can influence the errors.

Acknowledgement

First of all, would like to thank Saket for being the main supervisor for my thesis. He was very patient throughout the project even though I was slow at the start and needed to change my topic halfway in. He also helped massively in synthesizing the results of my research and teaching me to be more critical of my results.

Next, I would like to thank the Christianne, and the rest of the team (Faiz, Ludo, and Sukrati) for all the weekly meetings to discuss ideas, suggestions, results, and coding problems we are facing. The constant communication also really helped to lessen the effects of isolation due to COVID.

Then, I would like to thank Edo and Sierd, who were very helpful even though they joined the committee towards the end of the project. Their comments were valuable in improving the results of my thesis and pushed me in the right direction to get that extra mile.

I also would like to thank Jente for all the help at the start in preparation for the project, even though in the end we could not go to India, which is too bad!

Furthermore, I would like to thank my family and friends. They are always there to listen to my issues regarding the project and at the same time helps me to take my mind off of it so I do not get overwhelmed. This thesis would not be possible without their constant support.

Dennis Djohan
Delft, April 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Agricultural crisis in Maharashtra

India is the largest cotton producer in the world, where the majority of their production comes from Maharashtra (Lalitha et al., 2009). While the percentage of agriculture in India's GDP has been declining from 56% in 1951 to 25% in 2001, 58% of the total number of workers are still dependent on the agricultural sector (Mishra, 2006). However, drought is a major challenge and about one-sixth of India is deemed to be drought prone area (Udmale et al., 2014b). Between 1901 to 2010 about 17% were drought years, and they severely affect the agriculture and economy of the country despite drought mitigation measures (Kumar et al., 2013). The majority of the farmers do not have the means to mitigate the drought impact due to lack of water saving technologies and use of flood irrigation during drought (Udmale et al., 2014b).

Small farmers are often forced to borrow money, not just for agriculture but for their daily consumption (Sravanth and Sundaram, 2019). However, development banks are becoming less inclined to lend money towards farmers. The percentage of total bank loans towards the agricultural sector is halved in early 2000s from the previous figure of 12-20% in 1994 (Merriott, 2016). This decline does not reflect the shrinking of agricultural sector in India's GDP and there is an increase in the volume for non-bank loans with higher interest rates (Merriott, 2016), i.e. the farmers are turning to other sources of loans. The most common loan source is private money lenders (28.4%), followed by relatives (22.93%), and lastly the development bank (3.94%) (Behere and Behere, 2008). The high interest rate that the small farmers are forced to take often led to a perpetual debt traps that end in suicides. This is made worse by the fact that there is a change in demographics in the agricultural sector that showed increase of small farmers. Since 1960s, the number of farm holdings increased from 48.9 to 115.6 million, however, composition wise only the smallholder farmers (0-2 ha of land) increased from 51% to 62%, while the number of medium to large farmers is declining (Mishra, 2006). This means that more farmers are vulnerable to water scarcity and extreme weather conditions.

Government have passed debt relief bills in the past to reduce the farmers burdens (RBI, 2009; Kerala Legislative Assembly, 2012). However, multiple studies have shown that debt relief are not effective in combating the crisis and come with detrimental effects. Following debt relief, there are observed behavioral changes to the borrowers such as no increased investment or productivity, increasing default rates, and longer repayment rate (Kanz, 2012; Giné and Kanz, 2014; De and Tantri, 2014). These behavioral changes contribute to the further decline of credit access for the farmers as banks are becoming less likely to give out loans (De and Tantri, 2014). Moreover, debt reliefs are mostly reactionary and are believed to not be the solution (Sravanth and Sundaram, 2019). Qualitative exploration has shown that there are more factors other than debt that contributes to the crisis and high suicide rates viz. addictions, environmental problems, stress, poor irrigation, poor crop price, increased seed cost, diminishing investment, etc (Dongre and Deshmukh, 2012; Sud, 2009). To address the environmental stress and create employment opportunities, the government implemented The National Rural Employment Guarantee Act. Through this, the government pay the people to build ponds, dykes, and other drought mitigation infrastructures (Udmale et al., 2014a). However, the majority of the farmers are not satisfied with the scheme due to insufficient labor wages (Udmale et al., 2014a).

1.2 Interventions by Solidaridad and TU Delft

”Water Efficiency in Sustainable Cotton-based Production Systems” project is a 5-year project of Solidaridad Asia in collaboration with TU Delft, where they aim to improve the livelihood of smallholder cotton farmers. They plan to achieve this goal through several means such as physical interventions (e.g. reservoirs), financial (e.g. fixed crop prices), education, and promotions. Currently, the study is performed in the state of Maharashtra in India. It focuses on four districts in the north-eastern region (Vidarbha) of Maharashtra, namely Nagpur, Amravati, Wardha, and Yavatmal.

A baseline study (Hatch et al., 2019) was conducted in 2019 by a team from TU Delft to assess the hydrological resources and socio-economic characteristics of 308 farmers in three of the four study areas i.e. Amravati, Wardha, and Yavatmal. Fig. 1.1 shows the study area and the locations of the interviewed farmers. They did an extensive survey regarding farming practices and financial characteristics. Climate data, land use, and soil characteristics of the area are used alongside the survey data to model the performance of the farmers in terms of crop yield and capital generation. The team used a socio-hydrological (SH) model from (Pande and Savenije, 2016) as a basis for their assessment. In the past, this SH model was also used by den Besten (2016) to estimate the crop yield in the state level.

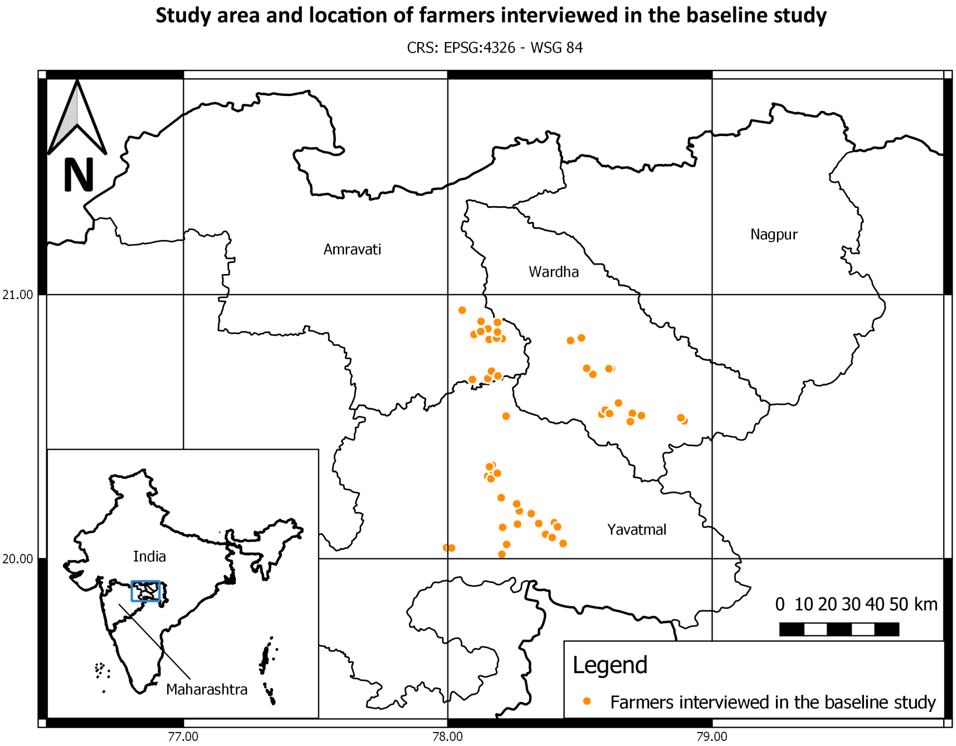


Figure 1.1: Map of the study area and farmers location. The project currently focuses at the three districts: Yavatmal, Amravati, and Wardha. These districts are located in the Vidarbha region, north east of Maharashtra in India. The points indicates the location of the 308 farmers interviewed. Made in QGIS (QGIS Development Team, 2021).

1.3 Knowledge gap

The chapter 1.1 noted the shortcomings of the past interventions and other approaches to increase the farmers' livelihood mentioned in chapter 1.2 are explored. Thus, the importance of the SH model as an intervention tool increases, e.g., for evaluating the cost benefit of building new reservoirs. However, currently there are some limitations to the SH model. The baseline noted that in the current model, the water only starts to affect the crop yield calculation when it is depleted fully and does not take into account the stress from a water deficit, thus overestimating the yield. Furthermore, while the water balance calculation is daily, it only adjust the crop yield annually. This means that the model cannot be used for a crop yield evaluation in the mid of planting season by the farmers or institutions as an adaptation tool.

Furthermore, there were no prior validation for the SH model at the farmers' level, only at the district and state level. For this, the baseline provides new information regarding the observed yield for the validation of the 308 farmers. Further, a QGIS Soil and Water Assessment Tool (QSWAT) model was used to calculate the recharge of the irrigation ponds (Janssen, 2020) and the results are incorporated into the SH model calculation. It is unknown how the model will perform with all these additional information.

1.4 Research objectives

First, this research aims to address the lack of stress mechanics and improve the foundation of the current model for future uses in the project. This is done by implementing daily stress mechanics based on FAO's AquaCrop model framework. AquaCrop model is able to simulate crop yield in several simplified steps namely canopy cover growth → transpiration → above-ground biomass → crop yield (Vanuytrecht et al., 2014). The stress mechanics are able to adjust the crop yield based on water and temperature conditions during growing season. More details of these will be explained in chapter 2.4. The addition of the stress mechanics hopes to reduce the inaccuracies suggested by the baseline study. Moreover, by changing the framework of the code to do daily water and yield calculations it will create a more robust foundation for future additions to the model.

Second, once the stress mechanics are implemented, the new SH model will be calibrated and its performance will be evaluated by comparing the calculated yield with the observed yield. The total error will be quantified, and the potential factors that might contribute to the discrepancies are identified.

The output of this research will be an adjusted model and a report that outlines the shortcomings of the model alongside the recommendations to address them. It can be used by Solidaridad and the TU Delft team for future studies and improve their intervention strategies.

1.5 Research questions

To achieve the mentioned research objectives the research aims to answer the question:

> "What is the overall performance of the model in calculating the cotton yield in the Maharashtra region with the addition of the stress mechanics?"

Alongside the main research question there are sub research questions:

- How do the uncertainties from the model and external variations affect the calculated crop yield?
- Which factors are most influential towards the error between the predicted crop yield and observed yield?

Chapter 2

Methodology

This chapter will first describe the characteristics of the study area, followed by the general approach and conceptual framework of the research. Next the overview of the SH model is discussed alongside with the details of the implementation of stress mechanics. Lastly, the evaluation methods are discussed.

2.1 Description of the study area

The baseline from Hatch et al. (2019) has produced a considerable information regarding the properties of the area and can be used for more details. However, a short summary of the site is provided here. As mentioned previously the study is located in the Vidarbha region of Maharashtra. The region is covered with black soil due to volcanic activities (Hatch et al., 2019). This black soil consists of around 54% clay, 32.5% silt, and 13.5% sand (Katti, 1979). This type of soil is suitable for growing cotton due to its high moisture retention and the relatively high temperature (Janssen, 2020). This region is influenced by the summer monsoon. Vidarbha has between 50-60 rainy days with an average rainfall of 150 - 200 mm (Ratna, 2012). Mean temperature of the area ranges from 29°C to 43°C during the day and 13°C to 28°C at night (Meteoblue, 2021). The 308 farmers sample consists mostly of small to medium farmers (0.5 to 10 ha), however, there are several large farmers (>10 ha). The cotton area ranges from 0.5 to 35 ha with a mean of 4.7 ha and a standard deviation of 4.1 ha. On average, each farmer produces about 1,500 kg of cotton per ha. 238 of these farmers (77% of total farmers) are in debt and 107 (45% of farmers in debt) have loan interest higher than 10%. This indicates similar situation discussed in chapter 1.1.

2.2 Research approach and conceptual framework

Fig. 2.1 shows the iterative research framework that is followed in order to answer the research questions. To make reading this report easier, its structure also follows this iterative framework. The introduction (chapter 1) can be seen as the context analysis from previous study. The methodology (chapter 2) can be seen as the designing and implementation in the model. The results (chapter 3) are the outcome of the evaluation. Lastly, the conclusion and recommendations (chapter 4) are the result of context analysis that can be used for future studies.

First, the findings from the baseline report can be considered as the context analysis regarding the limitation of the model i.e. absence of stress mechanics and lack of validation at farmer level. The stress mechanics code is designed and implemented in the water balance and yield calculation. In the latest iteration, these implementations include root zone growth, canopy growth, water stress, temperature stress, and irrigation calculation. Section 2.3 and 2.4 discusses the basis of the SH model and the implementation of the stress mechanics.

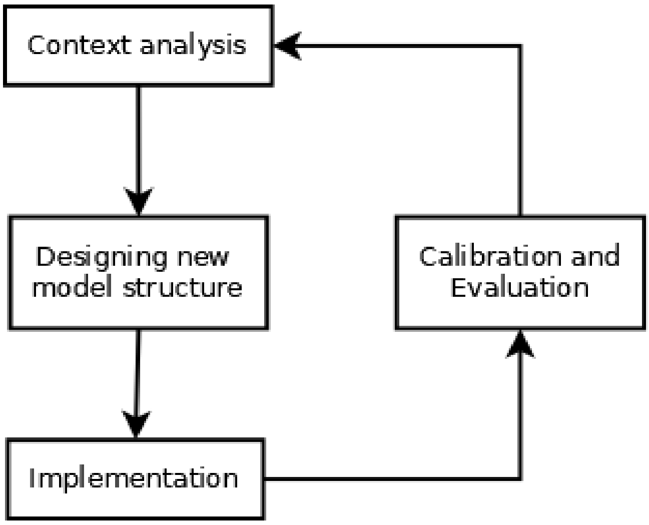


Figure 2.1: The approach used in this research. It is an iterative top down method. The implementations of the stress mechanics are adjusted multiple times according to the result of sub-research question 1 and 2 in the evaluation. In the latest iteration, the evaluation results are presented as a context analysis for future studies.

Fig. 2.2 provides more details on the research framework and breaks down the evaluation process. First, the newly added mechanics in the model is analyzed and its parameters calibrated using Monte Carlo Sampling (MCS). This achieves two things, first, as a sensitivity analysis to check which added parameters are impactful to the performance of the model. When a parameter is deemed not influential i.e. similar performance score across its value range, it can be removed so as to not add unnecessary complexity to the model. This also limits the number of parameters that are used for the calibration so as to reduce the number of iterations needed to find the pareto optimal solutions. Second, once a set of high impact parameters is found, it is run in the simulation again to be calibrated. After that the set of optimal parameters (θ) can be obtained from the best performing iterations. This is done through the Generalized Likelihood Uncertainty Estimation (GLUE) concept and the parameter set is used to find the uncertainty interval of the model predictions.

The external variations to the climate data such as precipitation, evaporation, and irrigation are also considered to test the stability of the model performance. This is done by running the model through bootstrap climate data.

Combined, the results from MCS, GLUE, and bootstrap analysis are used to answer the first sub-research question regarding the uncertainties of the model.

Furthermore, the mean of the predicted yields ($M(\theta, I)$) is then compared with the observed error (O) to obtain the total error. The total error is separated into two. The first is residual error (ϵ_r) which is a function of parameter errors (ϵ_θ), input errors (ϵ_I), and observation errors (ϵ_o). The second is structural error (ϵ_s), which is a function of factors ϕ that are correlated to the structural error. The factors ϕ are obtained by doing PCA and multi linear regressions with various variables. The structural error model itself is obtained by doing PCA in the kernel space (i.e. highly nonlinear space) and using linear regressions in the kernel space. The qualitative analysis is done alongside this to explore the potential ways these factors affect the error generation.

The results from this are used to answer the second sub-research question regarding the most influential factors towards the error.

The loop is repeated multiple times and the implementations are adjusted according to the each evaluation. The implementation discussed in this chapter and the results presented in chapter 3 are of the latest iteration. Finally, using the findings from the previously mentioned analysis, the performance of the model is evaluated to answer the main research question.

More information regarding the uncertainties and errors can be found in section 2.5.

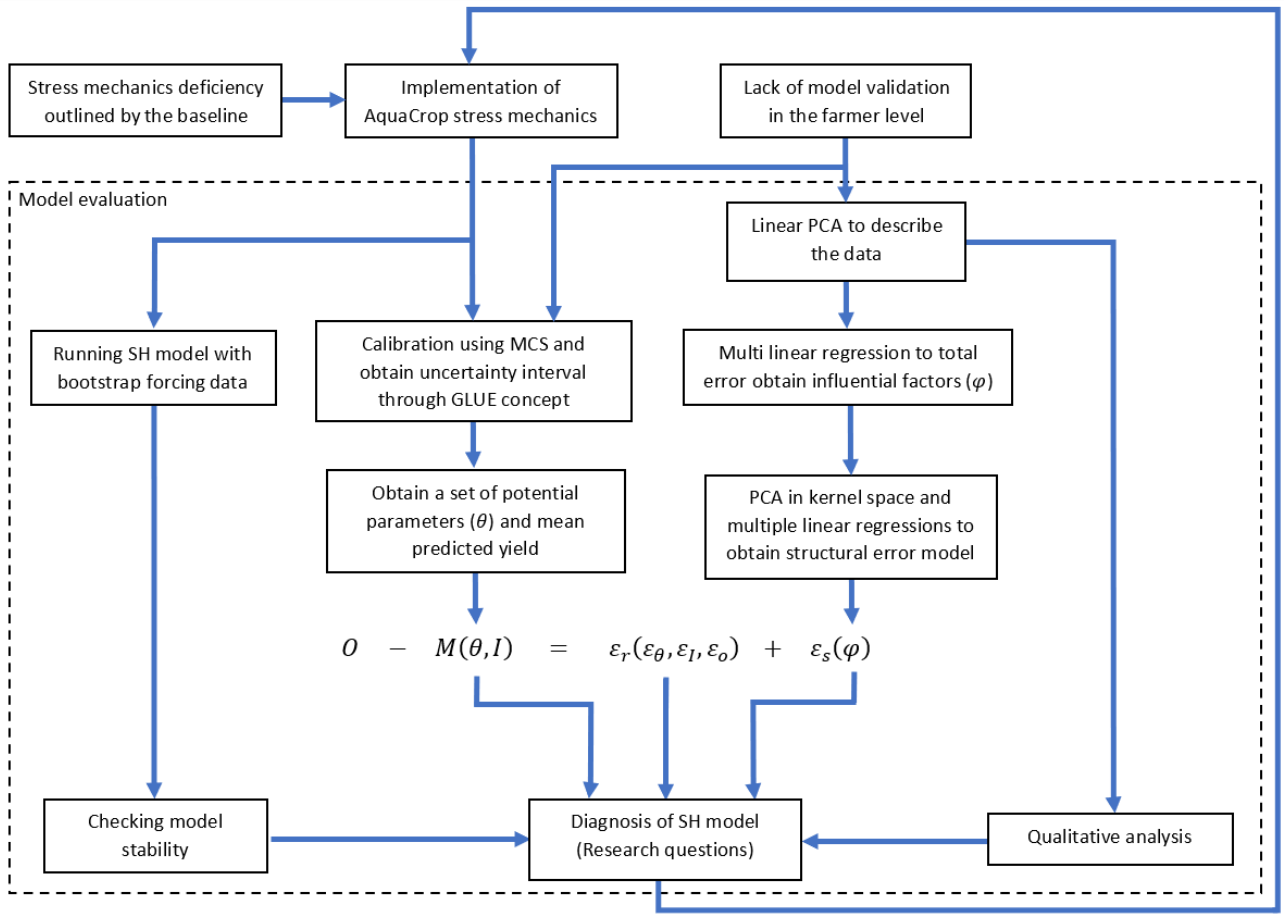

Stress mechanics deficiency outlined by the baseline
Implementation of AquaCrop stress mechanics
Lack of model validation in the farmer level
Model evaluation
Linear PCA to describe the data
Running SH model with bootstrap forcing data
Calibration using MCS and obtain uncertainty interval through GLUE concept
Multi linear regression to total error obtain influential factors (φ)
Obtain a set of potential parameters (θ) and mean predicted yield
PCA in kernel space and multiple linear regressions to obtain structural error model
$O - M(\theta, I) = \varepsilon_r(\varepsilon_\theta, \varepsilon_I, \varepsilon_o) + \varepsilon_s(\varphi)$
Checking model stability
Diagnosis of SH model (Research questions)
Qualitative analysis

Figure 2.2: Conceptual framework. The research is driven by the lack of stress mechanics and model validation. AquaCrop stress mechanics are implemented and the evaluation of the model performance is done.

2.3 Schematics of the base socio-hydrology model

The sociohydrological model framework used in the baseline study is based on the model by Pande and Savenije (2016) as shown in fig. 2.3. These dynamics are driven mainly by climactic variability. The model provide insight to the system dynamics of smallholder farmers through six assets, namely water storage capacity, capital, livestock, soil fertility, grazing access, and labor. This research is focused on the climate variability, soil moisture, and crop production and attempts to improve the feedback response from an annual to a daily calculation. The soil fertility factor and labor availability factor are sometimes adjusted as well due to their direct influence to the crop production. The other parts of the code that calculates the livestock, capital, expenditure, income, soil fertility are left as the original.

2.4 Implementation of the stress mechanics

This section will go into the details on the changes in the soil moisture and crop production part of the model. The stress mechanics use the framework of FAO's AquaCrop and their parameters as a basis for the hydrology, water stress, and yield calculation. More detailed information regarding the processes and the parameters can be seen in the manual by FAO (2012); Raes et al. (2012). An overview of all parameter values used in the model can also be seen in table A.1 in the appendix.

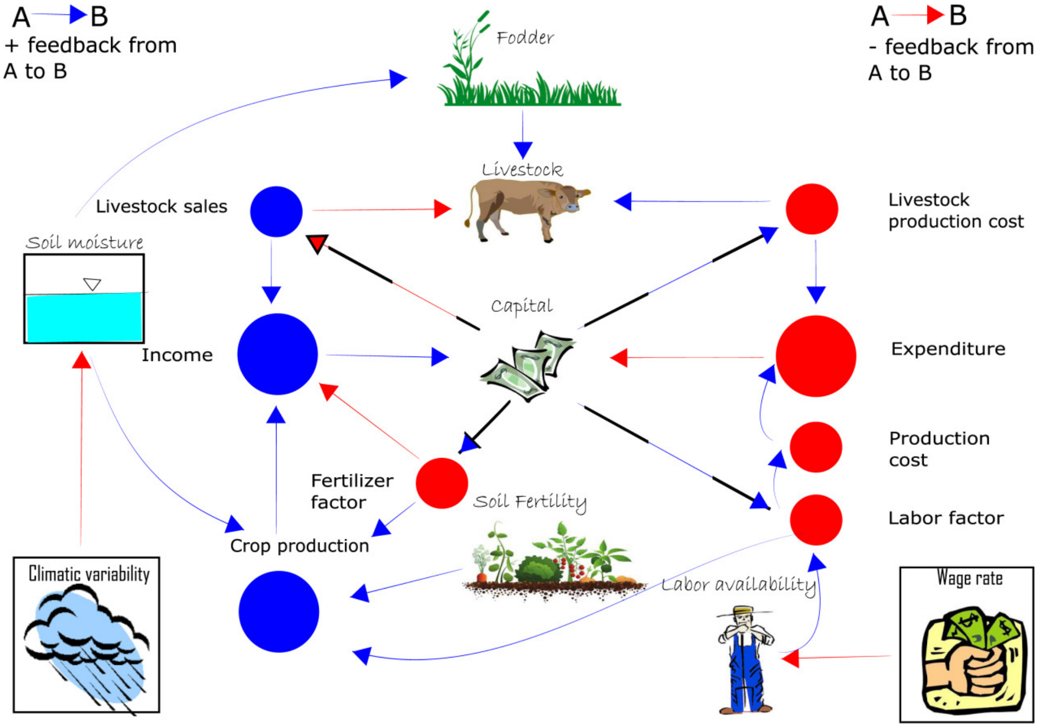


Figure 2.3: Framework of the base SH model obtained from Pande and Savenije (2016). The model captures the dynamics of the smallholder farmers through their assets (state variables). These include, water storage, grazing area, livestock, capital, soil fertility, and labor. This research will focus on climate variability, soil moisture, and crop production part of the model.

Figure 2.4 shows the overview of the stress mechanics from the input climate data to the yield output. The climate forcings used are irrigation, precipitation, reference evapotranspiration, and temperature. Information on these data is included in section 2.4.1. Two stresses are implemented namely water and temperature stress. The water stress is affecting crop growth via the root zone and canopy cover and the temperature stress is affecting the daily biomass production. There is a positive feedback loop present in soil moisture → water stress → canopy growth → soil evaporation → soil moisture. The increased canopy cover will reduce soil evaporation which will increase water available for plant growth. Then, there is negative feedback loop in soil moisture → water stress → canopy growth → transpiration → soil moisture. The increased canopy cover will increase transpiration demand, thus reducing soil moisture and inhibit canopy growth. The transpiration demand met is converted into biomass production using normalized crop water productivity (CWP). Then at the end of the season the biomass is adjusted into the crop yield using the harvest index (HI), fertilizer, and labor factor. Each step is discussed in more details in their own respective sections (2.4.1 - 2.4.9).

2.4.1 Data description

Gridded daily temperature and precipitation (P [mm/d]) data used are developed by India Meteorological Department (IMD) (Srivastava et al., 2009; Pai et al., 2014). The resolution for precipitation data is 0.25 x 0.25 degrees. While the max, min, and mean temperature ($T_{\max}, T_{\min}, T$ respectively [°C]) data are in 1 x 1 degrees resolution. The daily data is available from 1975 to 2019. The reference evaporation (ETc) is estimated using Hargreaves Equation from the temperature data (George H. Hargreaves and Zohrab A. Samani, 1985) (See Eq. 2.1). For the water equivalent of extraterrestrial radiation (R_a [mm/d]), average monthly value at 20° N latitude are used (Samani, 2000) - see appendix A.2 for detailed values. On evaluation, the temperature data generally have $RMSE$ value of < 0.5°C over most parts of India (Srivastava et al., 2009). The IMD4 precipitation data used in this model are compared to IMD1, IMD2, IMD3, APHRO, and IMD_OP. The comparison between the data give

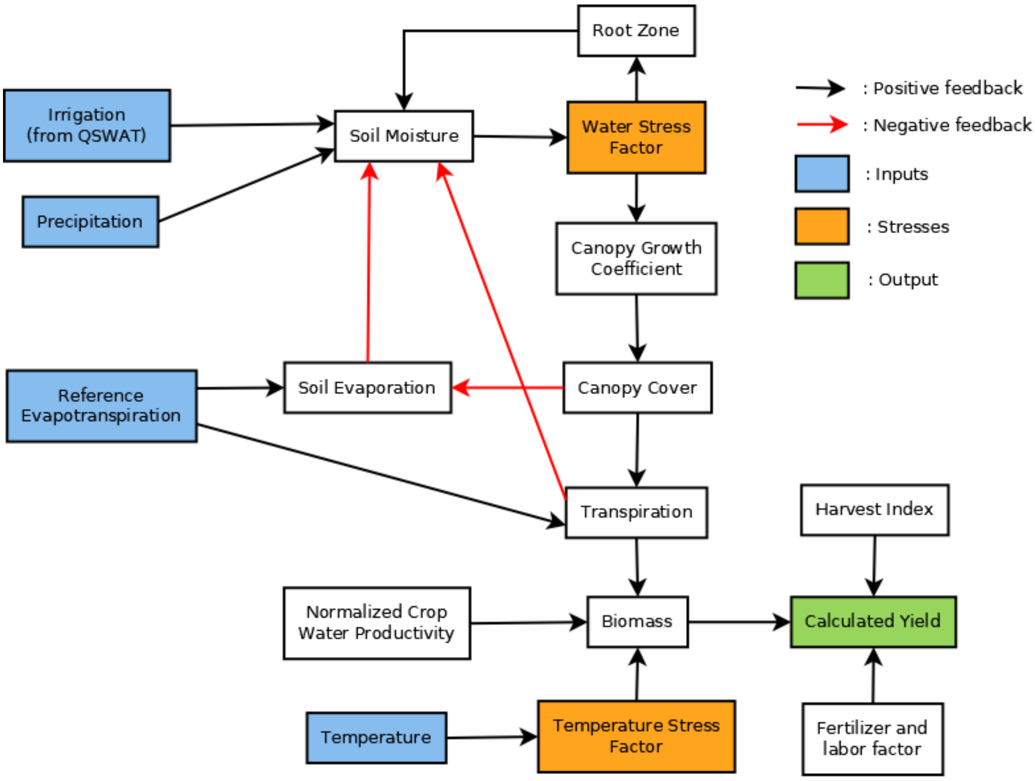


Figure 2.4: Overview of the stress mechanics processes. Going from the left to the right of the diagram is the process from the input climate data through the crop production to produce the calculated yield. The four inputs are irrigation, precipitation, ET_c, and temperature. The two stresses implemented are water and temperature stress. The processes from soil moisture → canopy cover → ET → biomass → yield are largely based on FAO's AquaCrop model. The details are discussed in section 2.4.2 to 2.4.9.

$RMSE$ of 0.11 to 0.43 mm/day and bias of -0.16 to 0.42 mm/day in annual comparison. It must also be noted that the data resolution are low and they might not represent the farm condition very well.

$$\mathrm{ET_c} = 0.0023 R_\mathrm{a} (T + 17.8) \sqrt{T_\mathrm{max} - T_\mathrm{min}} \tag{2.1}$$

The irrigation data comes from the QSWAT model. The thesis of Janssen (2020) provides details on the implementation of the QSWAT model. In summary, the model uses digital elevation, land use, soil type, and weather data to calculate the reservoir inflows at particular intake points. However, the inflow data is only available from 1979 to 2014. In order to obtain the approximate irrigation data for the year 2014-2019, the total precipitation in the region is compared across the dataset to find the most similar years. It is determined by comparing every year's data and finding the combination with the lowest mean absolute error (MAE). Fig. 2.5 shows the most similar year for the 2014-2019. Each year's pair irrigation data is used as a substitute for the missing data. Overall, they are all very similar with the exception of 2015 and 2008 pair, however, this is deemed to be acceptable. Because of the computational limitations, the irrigation data provided by the QSWAT model is monthly. For this, the available water is re-sampled into a daily data with an equal amount every day. Currently, there are only 351 and 358 considered intake points in Ghatanji and Hinhanghat respectively. The limitation for the QSWAT model output is that it does not take into account upstream water intake therefore the inflow produced is likely overestimated. In the SH model, each farmer is assigned the closest intake point. Furthermore, if multiple farmers are using the same intake point, the allocation of available water is divided evenly for the farmers for every time step. This is because the model is run for each individual farmer for the duration of the simulation period and currently it does not allow interaction between the farmers water intake i.e. the model does not know if farmer A can take 100% of the inflow if farmer B does not fill their pond at a particular timestep, thus it is always allocated between the two e.g. 50% for each farmer in a case of two farmers.

In addition to the climate data, various information from the baseline survey is used to supplement the calculations viz. maximum soil depth, type of irrigation used.

2.4.2 Root zone expansion and soil moisture

The effective root zone calculation ($Z_{eff}[mm]$) starts at the planting date. In between the plant date and the 90% of crop emergence (t_{cco}) the Z_{eff} is equal to the minimum effective root zone (Z_n [mm]) unless limited by the maximum effective depth Z_{max}. The Z_{max} is limited by the soil depth (d_{soil} [mm]) or the maximum effective root zone of the plant $Z_{max,plant}$ [mm] - See Eq. 2.3. When $t \geq \frac{t_{cco}}{2}$ the Z_{eff} is calculated using Eq. 2.2. Where Z_o [mm] is the initial value of the root zone, which is usually half of Z_n - see eq. 2.4. The root zone is growing over the duration period of t_x [days] with a shape factor of n [-]. The fig. 2.6 shows the overview of the root zone growth over the planting season.

$$Z_{eff} = \begin{cases} min\left(max\left(Z_o + (Z_{max} - Z_o)\sqrt[n]{\frac{t-\frac{t_{cco}}{2}}{t_x-\frac{t_{cco}}{2}}}, Z_n\right), Z_{max}\right), & \text{if } t \geq \frac{t_{cco}}{2} \\ min(Z_n, Z_{max}), & \text{if } 0 \leq t \leq \frac{t_{cco}}{2} \end{cases} \tag{2.2}$$

$$Z_{max} = min(Z_{max,plant}, d_{soil}) \tag{2.3}$$

$$Z_o = \frac{Z_n}{2} \tag{2.4}$$

2.4.3 Soil water stress

The total water available (TAW [mm]) for the plant changes over time following the evolution of root zone throughout the planting season as seen in Fig. 2.6. The TAW is calculated from the difference between water content at the field capacity (SM_{fc} [mm]) and the wilting point (SM_{wp} [mm]). The SM_{fc} and SM_{wp} are calculated by multiplying the Z_{eff} with the soil field capacity (FC [-]) and wilting point (WP [-]) respectively.

$$SM_{fc} = Z_{eff}FC \tag{2.5}$$

$$SM_{wp} = Z_{eff}WP \tag{2.6}$$

$$TAW = SM_{fc} - SM_{wp} \tag{2.7}$$

There are two thresholds that affects the calculation of the water stress factor ($K_s[-]$). p_{up} [-] signifies the fraction of TAW where the growth of plant starts being inhibited and p_{low} [-] is the fraction of TAW where plant growth stops completely. $Dr_{up/low}$ [mm] are these thresholds expressed as soil moisture in the root zone - see Eq. 2.8. The depletion (D_e [mm]) of water in the soil is calculated by subtracting the the soil moisture at a particular timestep (SM [mm]) from the maximum soil moisture capacity at Z_{max} ($SM_{fc,max}[mm]$) - See Eq. 2.9.

$$Dr_{up/low} = p_{up/low}TAW \tag{2.8}$$

$$De = SM_{fc,max} - SM \tag{2.9}$$

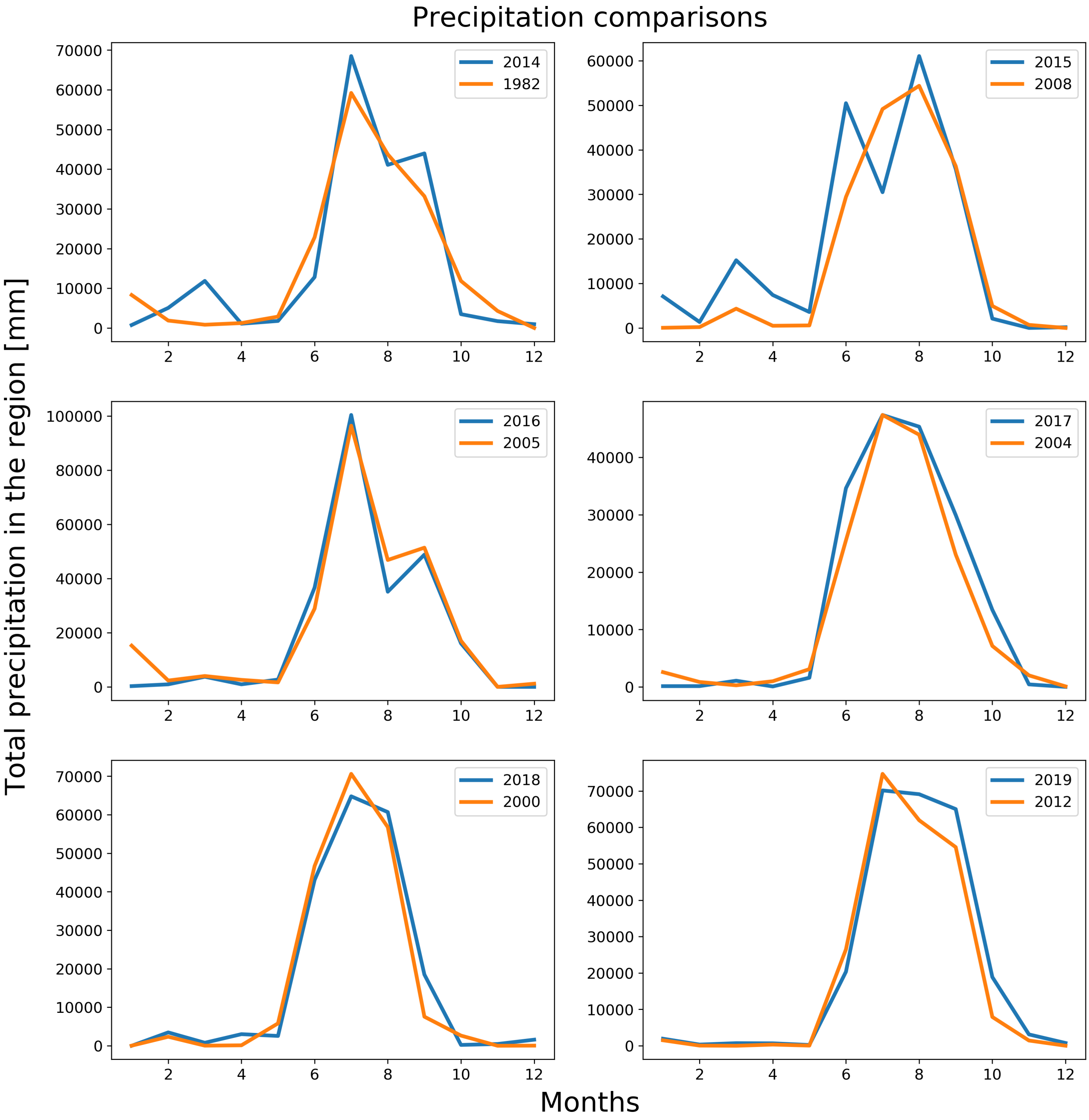


Figure 2.5: The comparison of total precipitation for year 2015 to 2019 across the whole dataset. The metric used to determine the closest year is the mean absolute error. By selecting two years with the most similar precipitation distribution, it is hoped that the irrigation data can be used to substitute for the missing years. 2014 is closest with 1982, 2015 with 2008, 2016 with 2005, 2017 with 2004, 2018 with 2000, and 2019 with 2012.

The water stress factor (K_s [-]) is determined by the D_e value. The upper and bottom threshold of water stress factor are 0.7 and 0.2 of total water available (TAW) [mm] respectively. The K_s has a value of 1 (no water stress) when the soil moisture is above the upper threshold Dr_{up}. Eq. 2.10 shows the relative water stress depending on the amount of depletion in the soil. Once the soil moisture drops below the Dr_{up} it decreases following a linear line until it reaches 0 (maximum stress) when the soil moisture value is below the bottom threshold Dr_{low}.

$$K_s = \begin{cases} 1, & \text{if } SM_{fc} - De \geq Dr_{up} \\ 1 - \frac{SM_{fc}-De-Dr_{low}}{Dr_{up}-Dr_{low}}, & \text{if } Dr_{low} \leq SM_{fc} - De \leq Dr_{up} \\ 0, & \text{if } SM_{fc} - De \leq Dr_{low} \end{cases} \tag{2.10}$$

Similarly, the root zone growth is affected by the root zone water stress. Eq. 2.11 shows the calculation of root zone water stress factor (K_{sto} [-]) on various soil moisture conditions. K_{sto} starts out at 1 when the soil moisture is above the upper threshold ($Dr_{up,sto}$) [mm] for root zone stress. The upper threshold is expressed as the fraction (p_{sto}) of total water available (TAW) in the root zone - see Eq. 2.12. When this falls below the threshold $Dr_{up,sto}$, the K_{sto} decreases linearly until it reaches 0 when no water is available. Finally, the K_{sto} is used to calculate the adjusted daily root growth ($dz_{eff,adj}$ [mm/d]) - See Eq. 2.13.

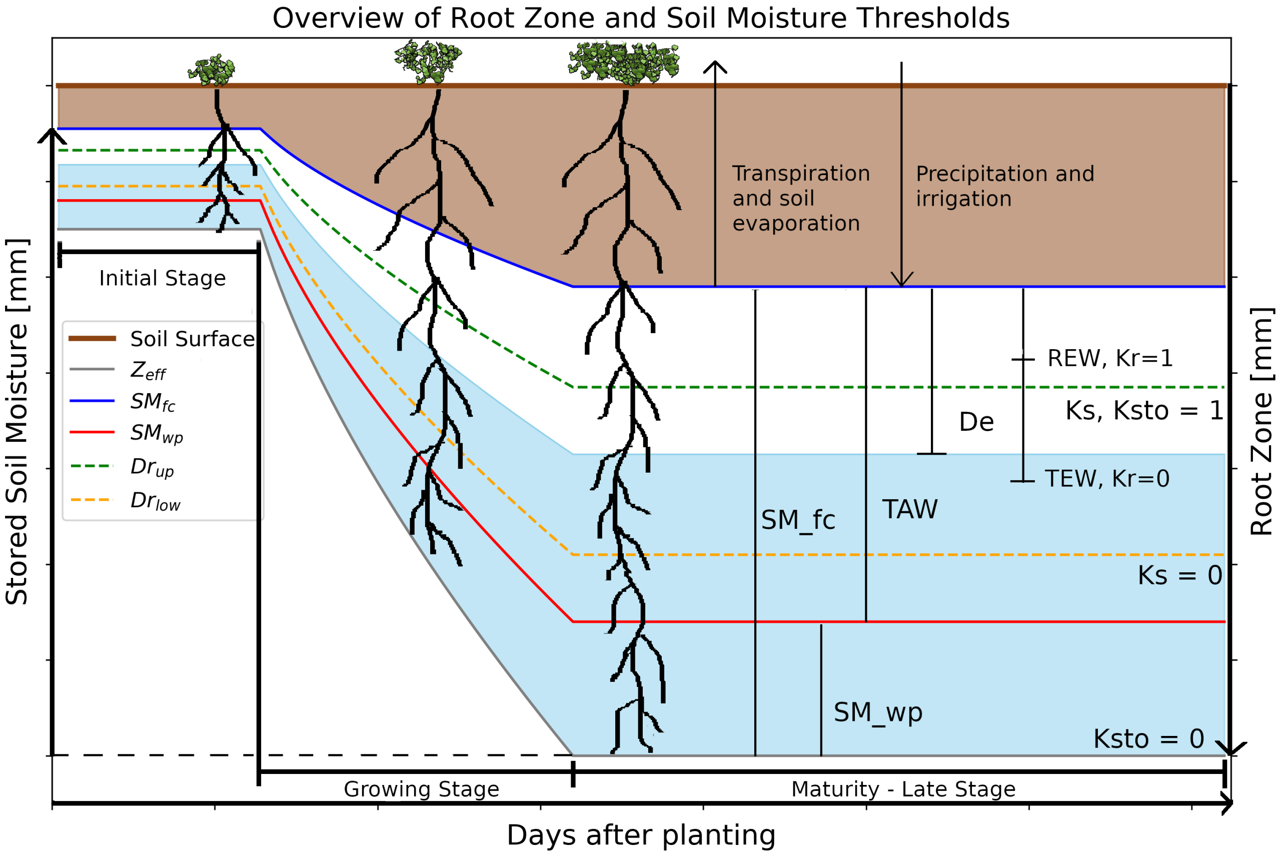


Figure 2.6: Visualization of root zone, stored soil moisture, stress thresholds, and soil evaporation reduction thresholds over the planting season. It is assumed that the soil moisture fills from bottom to up. The soil moisture field capacity at a certain timestep (SM_{fc}), wilting point (SM_{wp}), stress thresholds ($Dr_{up/low}$), and available water (TAW) are determined by the evolution of the root depth (Z_{eff}). The figure also visualizes the change in the stresses (K_s and K_{sto} depending on the depletion (De)

$$Ks_{sto} = \begin{cases} 1, & \text{if } SM_{fc} - De \geq Dr_{up,sto} \\ 1 - \frac{SM_{fc}-De}{Dr_{up,sto}}, & \text{if } 0 \leq SM_{fc} - De \leq Dr_{up,sto} \\ 0, & \text{if } SM_{fc} - De = 0 \end{cases} \tag{2.11}$$

$$Dr_{up,sto} = p_{sto}TAW \tag{2.12}$$

$$dz_{eff,adj} = dz_{eff} \times Ks_{sto} \tag{2.13}$$

2.4.4 Canopy growth

During planting season the growth of the plant is measured by Canopy cover (CC) [-]. It is the fraction of soil covered by the leaves of the crop by looking top down on an area. The prevalence of water stress in the system can cause a reduction of canopy growth. Hence canopy growth coefficient (CGC [-]) needs to be adjusted using the K_s value - See Eq. 2.14.

$$CGC_{adj} = K_s(CGC) \tag{2.14}$$

Equation 2.15 calculates the change in canopy cover $(\frac{dCC}{dt})$ $[s^{-1}]$. The initial value for the canopy cover CC_o is the value of CC when 90% of the sprouts appear. The calculation for $(\frac{dCC}{dt})$ also starts at this time (time t = 0 at t_{CCo} [day]). The equation shows the two phases of the growing stage. The first stage is when the CC has not reached the midway point of the maximum canopy cover (CC_x [-]). The canopy growth then slows down at the second stage as it reaches the CC_x. It stays at the CC_x until maturity and starts to decline when senescence is triggered towards the end of the season. During senescence the CC declines at the rate of canopy decline coefficient (CDC [-]) for every day after the trigger date ($t_{sen} = 0$ at the start of senescence). The change in CC during senescence can be seen in fig. 2.16. The overview of the canopy growth can be seen in Fig. 2.7. The canopy cover is assumed to go to zero after the harvesting season.

$$\frac{dCC}{dt} = \begin{cases} CC_0(CGC_{adj})e^{t(CGC_{adj})}, & \text{if } CC \leq 0.5(CC_x) \\ 0.25\left(\frac{CC_x^2}{CC_0}\right)e^{-(t(CGC_{adj}))}CGC_{adj}, & \text{if } CC > 0.5(CC_x) \end{cases} \tag{2.15}$$

$$\frac{dCC}{dt} = -0.05\,(CDC)\,e^{\frac{CDC(t_{sen})}{CC}} \tag{2.16}$$

2.4.5 Soil evaporation and transpiration

The model uses the dual crop coefficient to calculate the soil evaporation and crop transpiration from the reference evapotranspiration (ET_c). First, the basal crop coefficient (K_{cb} [-]) is used to calculate the crop transpiration and the soil evaporation coefficient (K_e [-]) for the soil evaporation. Adding both coefficients yields the combined coefficient $K_{c,max}$. Fig. 2.8 shows the K_{cb} and $K_{c,max}$ evolution throughout the year. The K_{cb} is divided into several stages. The initial stage (L_{ini}) lasts for 30 days, starting from planting date until 10% of the ground is covered by the crop. Next it enters the development phase L_{dev} as the crop grows, where the basal crop coefficient increases linearly from $K_{cb,ini}$ to $K_{cb,mid}$. Then it stays there during the mid stage L_{mid}. When the crop reaches maturity it enters the late stage L_{late} where the basal crop coefficient drops to $K_{cb,end}$ at harvest. The actual transpiration is then calculated using Eq. 2.17 (Vanuytrecht et al., 2014). It is proportional to the CC value, the K_{cb} value, and the water stress to account for water available in the soil.

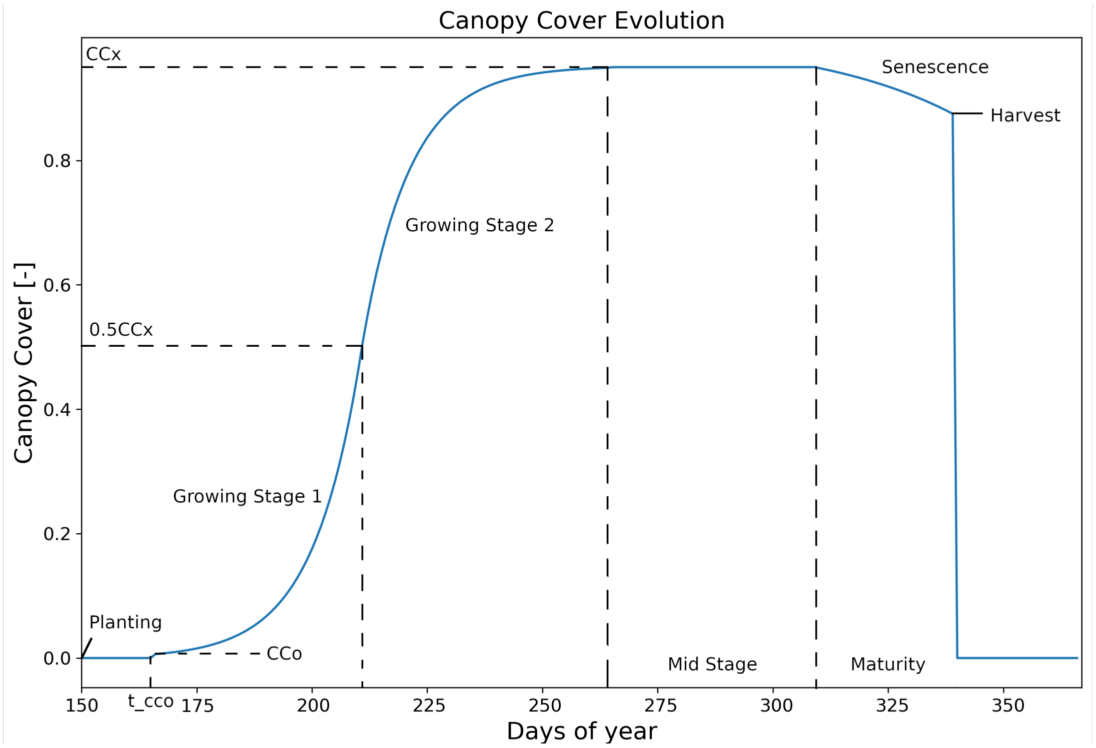


Figure 2.7: Visualization of canopy cover over the planting season. It is split into four stages. In the first growing stage, the CC grows rapidly until it reaches the second growing stage at $0.5CC_x$ where it slows down as it reaches the CC_x. In mid stage the CC stays relatively constant. During maturity the senescence is triggered and CC falls until the harvesting date.

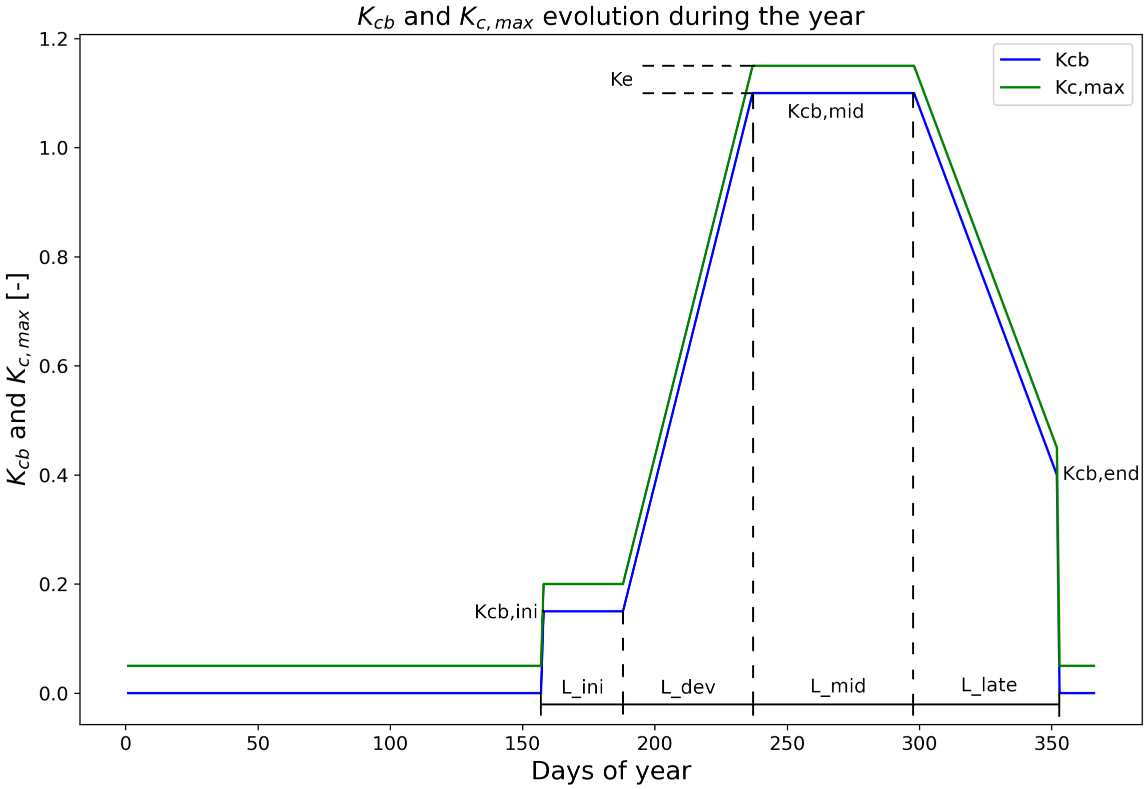


Figure 2.8: K_{cb} and $K_{c,max}$ evolution during the year for cotton (Allen et al., 1998). The dual crop coefficient is divided into four stages. This follows the plant growth. The basal crop coefficient (K_{cb}) increases linearly from the initial stage (L_{ini}) to the mid stage (L_{mid}) during development stage (L_{dev}). It drops to linearly in the late stage (L_{late}). The soil evaporation coefficient (K_e) stays constant at 0.05.

Currently, the value of the $K_{c,max}$ is assumed to be constant at $K_{cb} + 0.05$ (a maximum value for a completely wet surface (Allen et al., 1998)). This is due to the model not taking into account surface roughness/aerodynamics, relative humidity, and the height of the plant.

In order to calculate the soil evaporation (E_s [mm]), Eq. 2.18 is used (Allen et al., 1998). It is inversely proportional to the amount of CC as it covers more ground during the season. Moreover, the soil evaporation depends on the water depletion from the ground surface (Allen et al., 2005). When the surface is wet or the depletion is below the readily evaporable water (REW [mm]) the soil evaporation reduction coefficient (K_r [-]) is equal to 1. In this situation, energy is the limiting factor. As the depletion increases and goes past the REW the K_r decreases linearly until it reaches the total evaporable water (TEW [mm]) threshold, where it is equal to 0 and soil evaporation is inhibited completely - See Fig. 2.6 and Eq. 2.19. Currently, the calculation does not take into account the amount of wetted surface area from precipitation or irrigation.

$$T_p = K_s(CC)K_{cb}ET_c \tag{2.17}$$

$$E_s = K_r(1 - CC)K_eET_c \tag{2.18}$$

$$K_r = \begin{cases} 1, & \text{if } De \leq REW \\ \frac{TEW-De}{TEW-REW}, & \text{if } REW \leq De \leq TEW \\ 0, & \text{if } De \geq TEW \end{cases} \tag{2.19}$$

The total evapotranspiration (ET_a [mm]) is calculated by adding T_p and E_a - Eq. 2.20.

$$ET_a = T_p + E_s \tag{2.20}$$

2.4.6 Irrigation

The irrigation supplied (Irr [mm]) is calculated using Eq. 2.21. Irrigation is triggered when soil moisture goes below 50% of the total water available (TAW). The volume for irrigation depends on the De over the irrigated area (A_{res} [m^2]) and its irrigation system efficiency (c_{irr} [-]). The irrigation system is separated into drip, sprinkler, and furrow. Each of these has its own associated efficiency. The information for the A_{res} and c_{irr} for each farmer are obtained from the survey. This volume is limited by the current water content in the reservoir S_{res} [mm m^2]. The change of water volume in the reservoir is calculated using Eq. 2.22. The inflow from the intake point (I_{res} [mm m^2]) is obtained from the QSWAT data. Another flux considered is the open water evaporation (E_o [mm m^2]), which is calculated in Eq. 2.23. For the maximum capacity of the reservoir ($S_{res,max}$ [mm m^2]) and the water surface area of the reservoir (A_{res} [m^2]), it is assumed that irrigating farmers have a well with a diameter of 3m and depth of 10m.

$$Irr = min\left(S_{res}, max\left((De)\left(\frac{A_{irr}}{c_{irr}}\right), 0\right)\right) \tag{2.21}$$

$$S_{res} = max\left(0, min\left(S_{res-1} + I_{res} - E_o - Irr, S_{res,max}\right)\right) \tag{2.22}$$

$$E_o = ET_c(A_{res}) \tag{2.23}$$

2.4.7 Soil moisture

The soil moisture (SM) for every timestep is then calculated using Eq. 2.24 by adding all the fluxes such as precipitation (P), irrigation (Irr), and total evapotranspiration (ET_a) up to the maximum field capacity $SM_{fc,max}$.

$$SM_{i+1} = max\left(SM_i + P + Irr - ET_a, SM_{fc,max}\right) \tag{2.24}$$

2.4.8 Biomass and water adjusted yield calculation

The daily biomass increase ($m\left[\frac{g}{m^2 day}\right]$) is calculated using Eq. 2.25 and the total biomass ($B\left[\frac{g}{m^2 year}\right]$) can be obtained by summing the m during the planting season. The ratio of the transpiration being met (T_a) and the reference evaporation (ET_c) is multiplied with normalized crop water productivity ($CWP\left[\frac{g}{m^2 mm}\right]$) to obtain the amount of biomass produced per unit area per mm of water transpired.

$$m = K_T(CWP)\frac{T_a}{ET_c} \tag{2.25}$$

$$B = \sum(m) \tag{2.26}$$

Another factor that affects the biomass value is the temperature stress. If the temperature is above or equal to the upper threshold (T_{up} [°C]) the temperature stress coefficient (K_T [-]) is equal to 1 and does not affect the biomass production. However, when it goes below the T_{up} the K_T is decreasing along a logistic function until it reaches $K_T = 0$ at the bottom temperature threshold (T_{bot} [°C]). T_{up} and T_{bot} are 20 °C and 0 °C respectively. Eq. 2.27 shows the calculation for temperature stress and Eq. 2.28 shows the shape parameter for the equation.

$$K_T = \begin{cases} 1, & \text{if } T \geq T_{up} \\ \frac{1}{1+999e^{k(T-T_{bot})}}, & \text{if } T_{bot} \leq T \leq T_{up} \\ 0, & \text{if } T \leq T_{bot} \end{cases} \tag{2.27}$$

$$k = ln\left(\frac{\frac{0.001}{999}}{T_{up}-T_{bot}}\right) \tag{2.28}$$

The water adjusted cotton yield (Y_o)is then obtained from the total biomass (B) at harvest by multiplying the it with a harvest index (HI [-]) value of cotton.

$$Y_o = HI(B) \tag{2.29}$$

2.4.9 Fertilizer and Labour Factor

Currently, the soil fertility is not activated in the model. Therefore the crop yield is only adjusted based on how much Nitrogen (N) from fertilizer is applied (N_{app} [kg N/y]) vs. the maximum applicable fertilizer (N_{max} [kg N/y]) for the crop area (A_{crop} [ha]). Additional Nitrogen from manure (N_{manure}) is also considered. The relative fertilizer used (N_{rel} [-]) can be seen in Eq. 2.30. The fertilizer factor (F_N [-]) is then calculated using Eq. 2.31. The minimum yield is (Y_{min}) is assumed to be 375 kg/ha. The maximum yield Y_{max} is calculated assuming all evapotranspiration demand is met i.e. $T_a = ET_c$ and $K_T = 1$ throughout the planting season - see Eq. 2.32.

$$N_{rel} = \frac{N_{app} \times N_{manure}}{N_{max} \times A_{crop}} \tag{2.30}$$

$$F_N = \frac{Y_{min}}{Y_{max}} + N_{rel}\left(1 - \frac{Y_{min}}{Y_{max}}\right) \tag{2.31}$$

$$Y_{max} = HI\sum(CWP) \tag{2.32}$$

The labor on the farm can vary depending on the wage rate, and the number of the labor determine the crop output of the farm. The ratio (L_{rel} [-]) of available labor (L_{av} [person]) and the maximum labor (L_{max} [person]) is used to calculate the labor factor (F_{lab} [-]) - See Eq. 2.33 and 2.34. The shape of the labor factor ($f_{shape,lab}$ [-]) is negative as the first few missing laborers do not impact the yield as much as the later.

$$L_{rel} = \frac{L_{av}}{L_{max}} \tag{2.33}$$

$$F_{lab} = \frac{e^{L_{rel} \times f_{shape,lab}} - 1}{e^{f_{shape,lab}} - 1} \tag{2.34}$$

The F_N and F_{lab} are used in adjusting the crop yield in Eq. 2.35. Currently this adjustment is calculated on an annual basis.

$$Y' = Y_o(F_N)(F_{lab}) \tag{2.35}$$

2.5 Model Evaluation

In order to compare the model performance, the Nash-Sutcliffe (NS) coefficient is used - see Eq. 2.36 (Nash and Sutcliffe, 1970). The numerator represents the difference between the model calculated yield (Y') and the observed yield (Y) and the denominator is the difference between the observed yield and the mean of the observed yield ($\bar{Y}$). The NS value 1 means the model is representing the real yield perfectly. A value of 0 means that the model has the same performance compared to using the mean yield as a prediction i.e. using no model at all. Anything below 0 is worse than using the mean yield. The log of NS coefficient (NS_{log}) is also used in this research - see Eq. 2.37. Another metric used for the performance evaluation is the Mean Absolute Error (MAE) - see Eq. 2.38. Lastly, the coefficient of determination r^2 value is also used. Equation 2.39 shows the calculation of r^2, where Y_{pred} is the predicted value from the linear regression. These metrics were used on several studies to evaluate model performances (Toumi et al., 2016; Paredes et al., 2015; Mkhabela and Bullock, 2012) and they are used in the calibration process - see section 2.5.2.

$$NS = 1 - \frac{\sum_{i=1}^{n}(Y' - Y)^2}{\sum_{i=1}^{n}(Y - \bar{Y})^2} \tag{2.36}$$

$$NS_{log} = 1 - \frac{\sum_{i=1}^{n}(log(Y') - log(Y))^2}{\sum_{i=1}^{n}(log(Y) - log(\bar{Y}))^2} \tag{2.37}$$

$$MAE = \frac{\sum_{i=1}^{n}|Y' - Y|}{n} \tag{2.38}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{pred} - Y)^2}{\sum_{i=1}^{n}(Y - \bar{Y})^2} \tag{2.39}$$

2.5.1 Defining uncertainties topology

The error topology that is used in this research can be seen in fig. 2.9. It identifies the parts in which the errors are influencing. The uncertainty is divided into two categories, first is due to lack of knowledge and second is variability. Lack of knowledge uncertainties represents the imperfection of our knowledge of the system and variability represents the inherent variation in the system (Rotmans et al., 2003). The lack of knowledge category is split into four errors viz. input, parameter, structural, and observation errors. The lack of knowledge includes measurement errors from the input and observed data. It also takes into account the parameter and structural errors of the model. Parameter errors deals with the uncertainties of the parameters values while structural errors looks into the deficiencies in the model processes and extent (gray box in figure). Recall from section 2.2 that the total error (yellow box in figure) is split into two: the structural error (ϵ_s), and the residual error (ϵ_r), which is the combination of the input, parameter, and observation errors. The second source of uncertainty is due to variability, which comes from the external variations such as climate forcings.

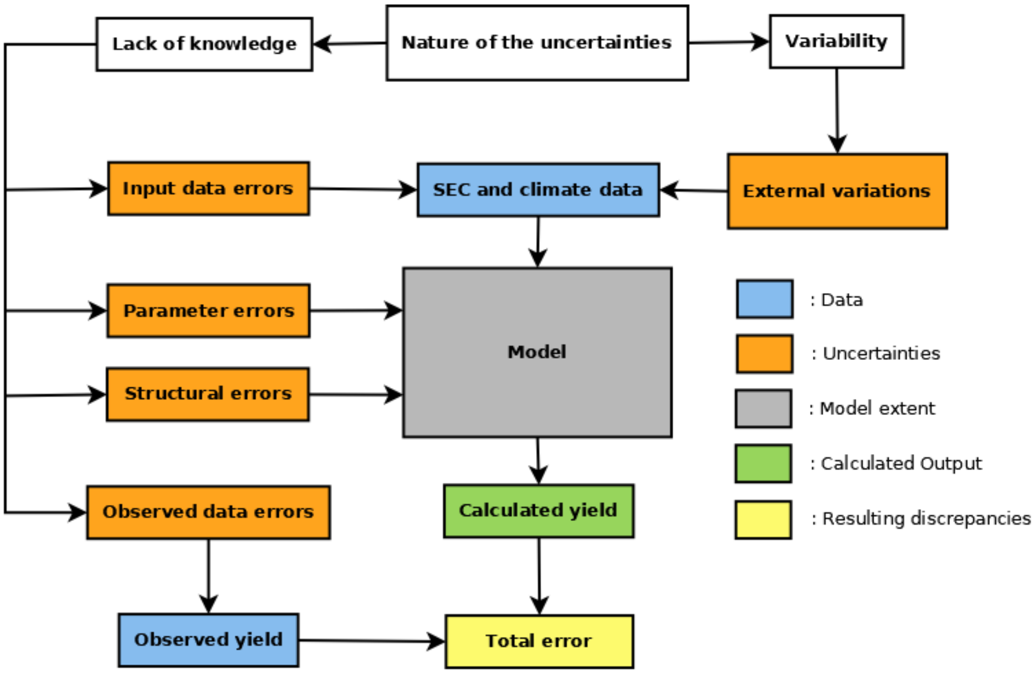


Figure 2.9: Uncertainties topology used in this research, it identifies the parts of the process being influenced by the errors. The errors (orange box) are used to evaluate the model performance. The uncertainties are divided into two categories: lack of knowledge and variability.

Error descriptions

Input and observation errors are associated with the used data in this project. The data are described in section 2.4.1. The resolution for the precipitation and temperature can contribute to the error as they might not represent the true condition at a farmers' field. Similarly, the assumption used for the irrigation inflow at year 2014-2019 can contribute to this input error since they do not represent the true flow.

The observed crop yield obtained in the survey in 2019 was the previous crop yield in 2018. The time between the harvest and the survey might contribute to the inaccuracies due to forgetfulness. There is also a distinct increments of the yield that can be observed from the survey answers as people round their yield to the 0.5 quintals/acre (around 120 kg/ha). This increments amount to around 8% of the observed mean yield, which is quite substantial. These input and observation errors are not to be studied in depth in this research, however, they are acknowledged when discussing the results.

Parameter errors are the errors caused by the difference between the used and true parameter values. The addition of the stress mechanics introduces a lot of parameters into the model. Discrepancies for each of these parameter from their true value contribute to the parameter error of the model. Appendix A.1 contains the input parameters used in the model.

The structural errors are errors due to incorrect representation of a process in the model or the lack thereof. It is often caused by the definition of the model extent and assumptions. For example, assumed linear shape of water stress function, lack of groundwater fluxes, inactive soil fertility code, etc.

2.5.2 Calibration and uncertainty interval

Given the nature of variations in the parameter values across the study region and the input uncertainties mentioned in the previous section, Monte Carlo Sampling (MCS) is used for to calibrate the model and find the uncertainty interval. MCS method takes into account both the natural variability of the parameter values and the measurement errors (Gardner and O'neill, 1981). The MCS aims to find the approximate range of parameter values that are the most representative of the study area. This is done by selecting a set of parameters and their range of values (see Appendix A.1 for detailed values) and doing a simulation for large iterations with a random parameter values selected each time. The Generalized Likelihood Uncertainty (GLUE) introduces the concept of many feasible combination of parameter values that have similar likelihood in describing the system (Beven and Binley, 1992). Furthermore, different objective functions can give different information about the system. If a parameter set performs well on one objective function it does not mean that it will perform as well on the rest i.e. pareto optimal solution. Therefore, the optimal parameter sets selected must be above the performance thresholds for all four objective functions mentioned at the start of chapter 2.5.1. Unfortunately, determining the thresholds is subjective and can be seen as a shortcoming for this method. To mitigate the subjectivity, they are selected in such a way that ensures that at least 50% of the observed yield lies in between the uncertainty interval of the calculated yield (however, it must also be noted that the 50% threshold itself is subjectively selected). Based on this, the threshold values used in this experiment are: $NS \geq -1.0$, $NS_{log} \geq -2.0$, $MAE \leq 600$, and $r^2 \geq 0.003$.

Moreover, MCS is computationally expensive due to a large number of parameters present and the complexity of the model. To minimize this limitation, only a small subset of parameters are calibrated at any time. For example, in the current iteration of the model only HI, t_{cco}, and labor factor shape are calibrated. These parameters are chosen based on a combination expert judgment and prior MCS analysis. For example, some parameters are found to not be influential to the crop calculation so they are just assigned a particular value. Some parameters are also discarded from the model when it is deemed not influential in order to maintain the complexity of the model down (e.g. the addition of convex stress function introduces shape parameter and when it was deemed to be not influential in improving the model score it is discarded and the model is reverted to use linear function).

The set of optimal parameters are then used to obtain the uncertainty interval of the predicted crop yield. From this the mean yield will be used to evaluate the total error.

2.5.3 Structural Error Model

A structural error model ($\epsilon_s(\phi)$) is presented in an attempt to capture a portion of the total error and adjust the predicted yield (Beven, 2005; Kennedy and Hagan, 2001). To do this, this research utilized linear PCA and regression of survey variables, PCA in kernel space, and context analysis. In this research, context analysis refers to the qualitative analysis of the context at the model extent or boundaries to identify missing processes.

Linear Principal Component Analysis and Regression

First, linear PCA is used to reduce the dimensionality of the factors tested by doing basis transformation of the data (Wold et al., 1987). Reducing the dimension of the factors and identifying their principal components (PCs) helps to identify patterns/trend and describe the data more easily. The factors being tested include socio-economic variables from the baseline survey such as children help, cotton area, cotton yield, crop price, seeds cost, pesticide cost, fertilizer cost, and fertilizer usage. Other variables tested include soil depth, latitude, longitude, precipitation, evaporation, irrigation totalling 14 variables. The eigenvalue used as a cutoff for selecting the retained PC is 1. This is so that the PC can at least explain the same amount of variance compared to the variables being measured. Furthermore, after the basis transformation, a varimax (orthogonal) rotation is done to adjust the loading factors (Brown, 2009). This process rotates the principal components so that one variable will not be explained by multiple PCs and makes a clearer distinction between the PCs. This is useful for the evaluation during the qualitative analysis.

The obtained principal components are used in a multiple linear regressions with the total error to evaluate their first order effects. Both the PCA and the linear regression is done in SPSS Statistics software (IBM Corp, 2019). Once the influential factors are identified, they are discussed in a qualitative analysis to find the potential ways these parameters contribute to the errors. They are also used in the Kernel PCA to model the structural error.

PCA in Kernel Space

While the components obtained from linear PCA is useful to evaluate the first order effects of the variables, it may be insufficient to capture of the underlying manifold of the data structure that can be non-linear. Thus, to explore the non linearity, these variables are mapped to a higher dimensional space using kernel functions where they are linearly separable (Schölkopf et al., 1998; Raschka and Mirjalili, 2019). Then PCA can be done in this new kernel space by maximizing the variance for each component and minimizing their covariance. To do this, the KPCA command in the scikit-sklearn plugin on python is used (Pedregosa et al., 2011). There are various kernels that can be used with this plugin, i.e., sigmoid, radial basis function (RBF), cosine, and polynomial (degree 2 to 5). Cross validation is done in order to select the best kernel function. First, the available data is split into 75% training data and 25% test data. Then, the training data are used to find the eigenvectors or the transformation "model" using the KPCA command. Following this, the test data are projected to the kernel space using the KPCA "model". The resulting PCs in the kernel space is then used in a multiple linear regression model with the total error. The explained variance will be used to select which PCs will be used in the regression. The PCs with highest variance are kept until there is one that brings the cumulative variance above 90%. The proportion of explained variance (R^2) is calculated using equation 2.40, where $\sum_{i=1}^{k} \lambda_i$ is the cumulative variance, and $\sum_{i=1}^{n} \lambda_i$ is the total eigenvalue for the total number of non-zero components (n). The resulting model from the regression ($\varepsilon_s(\varphi)$) is used to predict the structural error.

$$R^2 = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{2.40}$$

Residual error

The residuals between the predicted total error and the observed total error is attributed to the residual error. The distribution of the residual errors is deemed to be caused by uncertainties with regards to the input, parameters, and observation values.

2.5.4 External variability and model stability

The MCS method is also used to randomly sample the climate data (with replacement for every iteration) over the recorded period as a bootstrap analysis. The MCS can create a bootstrap climate data that mimics the structure of the climate data to test the stability of the model (Royston and Sauerbrei, 2009). As an example of the sampling, a precipitation data for the first day of the year (DOY) can be randomly sampled from every year of the current data where DOY=1. This repeats for every DOY for the simulation time. This method assumes stationarity i.e. trends due to climate change are not considered so there is no time dependence. In order to reduce unwanted variability a high number of iteration is needed (Hesterberg, 2011). Considering the size of our climate data (daily data over 308 locations over 45 years) an iteration size of 1,000 is used. The bootstrap method is used on the precipitation and evaporation data.

On the other hand, the irrigation data is dependent on the precipitation and evaporation data. Therefore, instead of using random sampling, the irrigation data is adjusted based on the bootstrap data used. The sum of bootstrap precipitation (P_{boot} [mm]) and evaporation (E_{boot} [mm]) over the planting season is compared with the sum of original data. Eq.2.41 shows the calculation of the irrigation factor (F_{irr} [-]) from the ratio of precipitation and evaporation. The evaporation is inversely proportional to the irrigation inflow. The resulting F_{irr} is used to adjust the original irrigation inflow see Eq. 2.42.

$$F_{irr} = \frac{1}{2}\left(\frac{\Sigma P}{\Sigma P_{boot}} + \frac{\Sigma E_{boot}}{\Sigma E}\right) \tag{2.41}$$

$$Irr_{adj} = F_{irr} \times (Irr) \tag{2.42}$$

Chapter 3

Results and Discussions

This chapter presents the findings from the research alongside with the discussions regarding the implications and the limitations. First, the model's overall performance is described. It is then followed by the uncertainties from the parameters and variability. Lastly, the results from the structural error analysis will show the potential locations of the errors.

3.1 Performance of the Model

Figure 3.1 and 3.2 shows the overview of the evolution for various variables. The model was run from the year 2010 to warm up the model. When looking at these figures it seems that the added stress processes are responding very well to the forcings. Fig. 3.1 shows that the evolution of soil moisture closely follows the rainfall events and the transpiration. In turn, the water stress function follows the soil moisture closely. Because there is no water stress at the start of the planting season, the canopy cover growth also seem to be unhindered at the start. As the soil moisture and water stress drops the canopy cover also starts to experience senescence. It can also be observed that the transpiration slowly drops to 0 as the soil moisture reaches the wilting point. Fig. 3.2 shows the evolution of the transpiration and soil evaporation fluxes in a clearer scaling. Alongside it, the daily biomass growth can be seen to follow the transpiration movement closely. These two graphs show that there is no major fundamental issues to the additions of the stress processes.

3.1.1 Crop yield prediction of the SH model

This section will discuss the results of the crop yield calculation by the model. After calibration, the average crop yield and their uncertainty intervals are calculated using the pareto optimal range of parameter sets. How to determine the threshold of optimal parameters are discussed in chapter 2.5.2. The threshold values used in this experiment are: $NS \geq -1.0$, $NS_{log} \geq -2.0$, $MAE \leq 600$, and $r^2 \geq 0.003$. In total, 1,557 parameter sets out of the 10,000 simulations were used to obtain the range of uncertainty intervals. 54.2% of the 308 farmers have the y=x line falls within the uncertainty intervals.

Figure 3.3 shows the plot of calculated yield against the observed yield. It appears that there is a cap to the model output around the 2,000 kg/ha limit. This can be explained by the inability of the calibration to account for these high yields, suggesting that there are structural errors within the model - see chapter 3.1.3. In other words, the calibration can only squeeze the data points into the mean yield as it has the least errors mathematically. Because of this, the highest discrepancies come from the high observed yields.

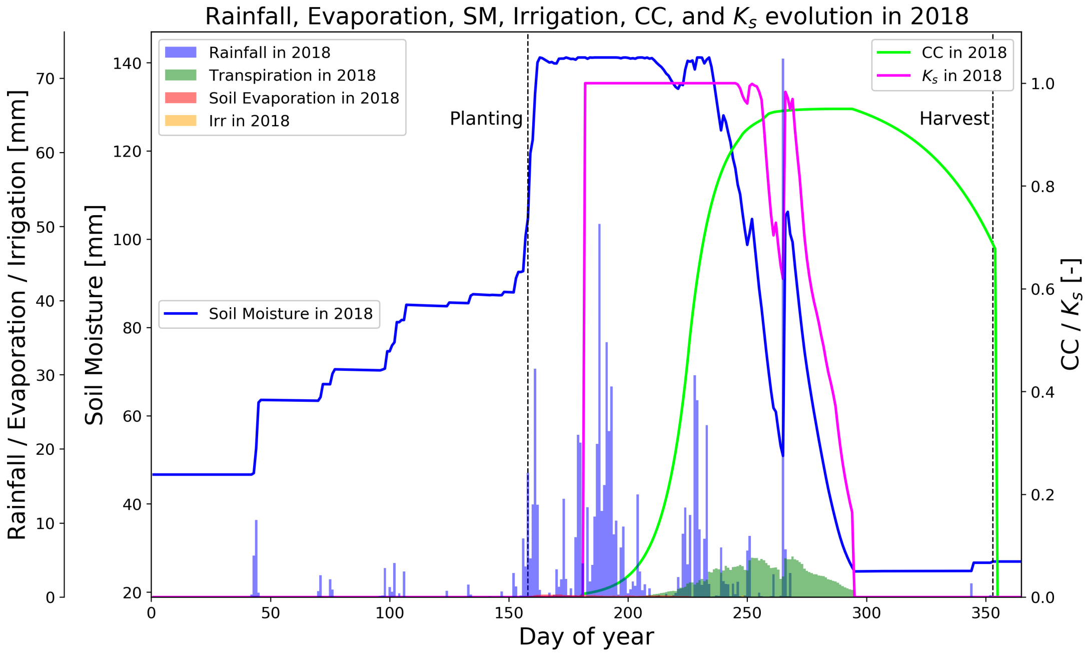


Figure 3.1: Seasonal evolution of soil moisture, water stress factor (Ks), and canopy cover (CC) due to rainfall, transpiration, soil evaporation, and irrigation. Soil moisture increases and decreases during the rainy season and peak of planting season respectively. The K_s starts to decrease when there the water is depleting. The observed CC evolution is also following the conditions very well. It grows rapidly during the time when water is abundant and senescence is triggered when water is depleted. The irrigation is hardly visible due to the small storage capacity of the new pond.

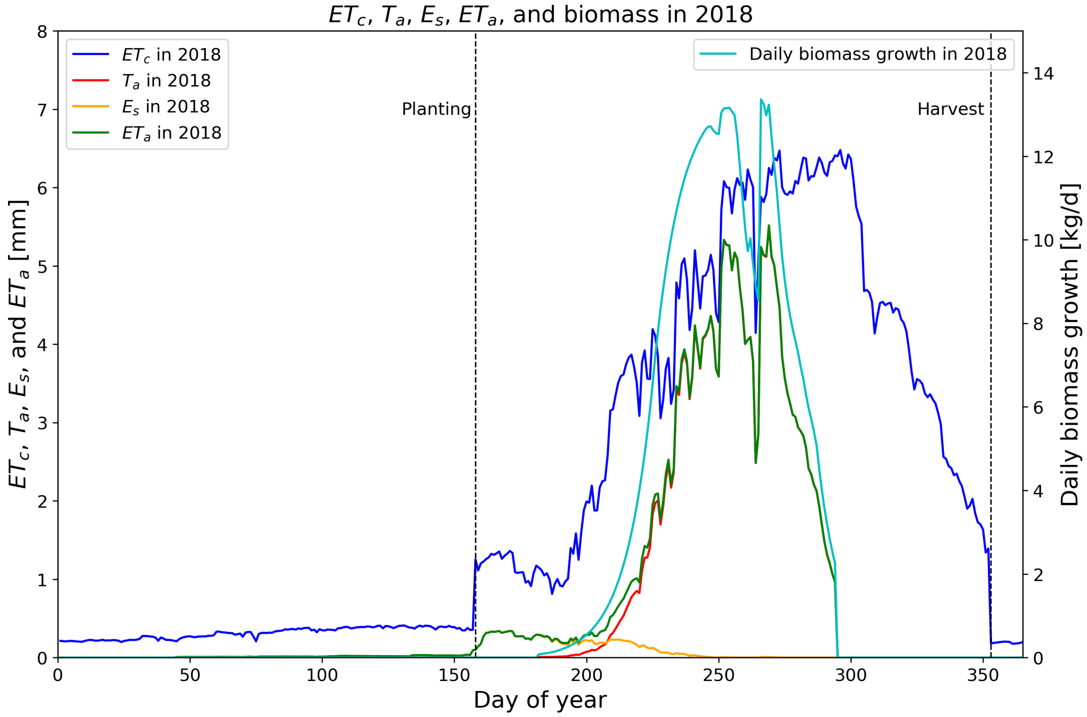


Figure 3.2: Seasonal evolution of transpiration, soil evaporation, and biomass. The actual evaportranspiration (ET_a) is seen to follow the reference (ET_c). Then it drops around day 300, mainly due to the lack of soil moisture. The daily biomass produced follows the transpiration closely.

The three parameters that are selected for the calibration are HI, t_{cco}, and $f_{shape/lab}$. From prior sensitivity analysis through MCS these three parameters are consistently influential to the objective functions. HI is influential due to its direct influence towards the yield calculation in the end i.e. the various parameters calculated during the crop production can be compensated by adjusting the value of HI. Similarly the $f_{shape/lab}$ have a direct influence and can compensate for the uncertainties in the labour availability calculation. The only other parameter from the yield calculation is the t_{cco}, this is likely due to its influence in determining whether the canopy cover will reach its maximum before the rainfall ends. The table 3.1 shows the ranges of objective function scores obtained from the pareto optimal parameter sets. All of the objective function scores are poor. The NS and NS_{log} values suggest that the model performs worse than using the average observed yield as a prediction. The MAE represents around 30% error from the observed yield. Lastly, from the r^2 values the calculated yield does not seem to be correlated to the observed yield. For more details on the variation of objective functions with the parameter values see Appendix C.

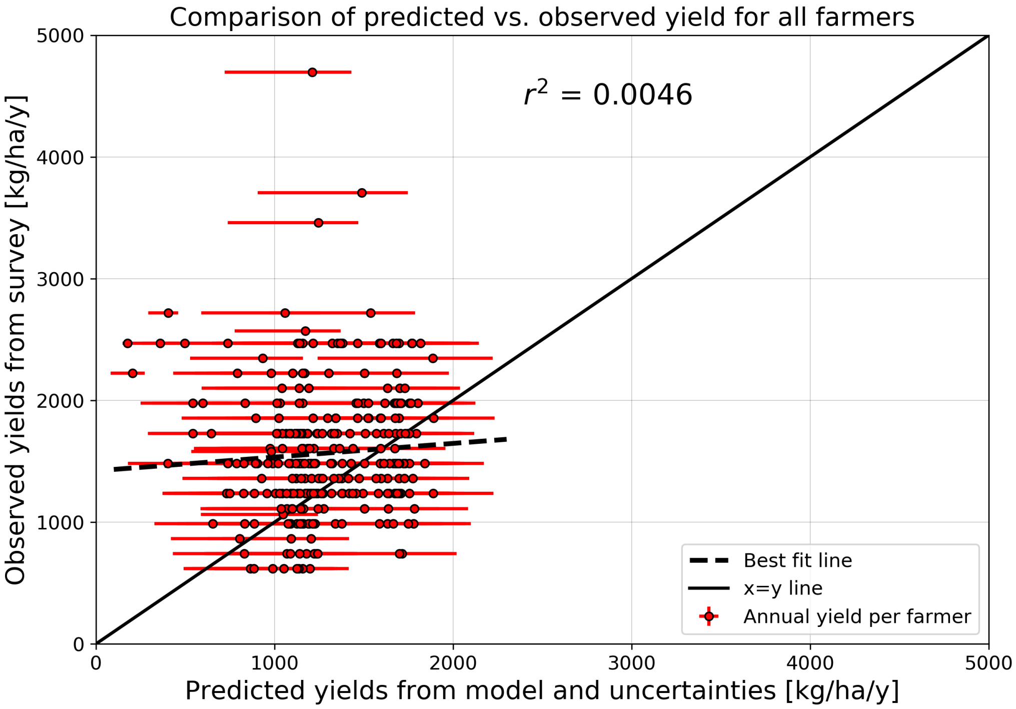


Figure 3.3: Yield comparison between the model and observed data. The model calculation is unable to simulate the high crop yield as it clusters around the 2,000 kg/ha. Further, the r^2 value is very low at 0.0046 and indicate that the model results are not correlated to the observed data.

Table 3.1: Ranges of calibrated parameters and objective functions. The poor values for the objective functions indicate poor performance of the model.

	HI [-]	t_{cco} [days]	$f_{shape,lab}$ [-]	NS [-]	NS_{log} [-]	MAE [kg/ha]	r^2 [-]
Max	0.4	7	-6.457	-0.343	-0.655	553.2	0.008
Min	0.25	30	-9.995	-0.996	-1.91	447.1	0.003

Comparison to other studies

Prior research that used the SH model i.e. the baseline by Hatch et al. (2019) and the report by den Besten (2016) showed reasonable inference between observed and model results at the district and state level respectively. Additionally, table 3.2 shows the overview of the statistics of the performances of various studies that uses AquaCrop model. For the most part, these studies give out adequate results

Table 3.2: Overview of statistics of FAO's AquaCrop model performance from various studies.

	Experimental setup Location and timeframe	Crop	NS or NS_{log}	MAE or $RMSE$	r^2
Paredes et al. (2015)	Daxing, North China Plain Irrigation Experiment Station 2008-2011 around 400 mm in June-Oct	Soybean	-0.47 to 0.82 NS for SM	<7.3% $RMSE$ for CC 302 kg/ha $RMSE$ for yield	0.22 to 0.86 for SM
Tan et al. (2018)	Xinjiang, China Irrigation Experimental Station 2012,2013,2015,2016 4-5 fields per year around 950 mm rainfall in April-Oct	Cotton	-	644 to 2,207 kg/ha $RMSE$ for yield	0.86 to 0.98 for CC 0.00 to 0.96 for SM 0.01 to 0.96 for yield
Toumi et al. (2016)	40 km East of Marrakech, Morocco 2002-2004 Irrigated zone R3 (experimental site) 6 validated fields 190-250 mm rainfall annually	Wheat	0.73 to 0.99 for CC -2.62 to 0.55 for SM	3.63 to 16.79% $RMSE$ for CC 14-28% of SM 100 kg/ha $RMSE$ for yield	0.83 to 0.99 for CC 0.65 to 0.90 for SM 0.98 for yield
Mkhabela and Bullock (2012)	Saskatchewan, Canada 2003-2006 4-5 fields 514-542 mm rainfall annually	Wheat	-	610.5 kg/ha MAE for yield (17%) 40.5 mm MAE for SM	0.90 for SM 0.66 for yield

even though there are some poor performances such as the NS score for SM, r^2 that is close to 0 in couple studies, and the high $RMSE$ for cotton yield. Looking at the experimental setups, the majority of these studies are performed in a more controlled environment and once calibrated, the model is validated on select fields. Paredes et al. (2015) found that the model is not suitable for irrigation scheduling when not using the dual crop coefficient approach and raises an important point to properly partitioning the ET_c to simulate soil waster properly. Tan et al. (2018) has a high variation of r^2 value. This is mainly due to a very poor model performance in the year 2015 and the calculation of soil water, salinity, and HI. It is observed, that the studies performed in an area with high precipitation are performing less well than the opposite counterparts. While it can mean that the model is not able to simulate water conditions on high precipitation it can also be due to the crop selection. As the validation for cotton is poorer compared to the rest. It must be noted however, that the sample size of these studies is small.

So far, the SH model has only been used to and validated to district and state level. On the other hand, various studies of AquaCrop model has been used on a small scale fields. However, it appears that the SH model does not perform very well when it attempted to do calculations of a micro level (farmers) at a macro scale (at 3 districts). It is also known that FAO's AquaCrop have difficulties in upscaling to a regional scale due to the spatial heterogeneity (Han et al., 2020). The adequate results from the previous studies with the SH model can be attributed to the averaging the yield results to the district and state level, because the NS score of the model is close to 0 which indicates that it can perform similarly compared to the average yield. Recommendations to the calibration and mentioned flaws is included in chapter 4.1.

3.1.2 External variability effects on model stability

Fig. 3.4, shows the comparison of the observed and model yield using bootstrap data. The parameter values used for the bootstrap analysis is the mean of the calibrated parameter sets. The spread of the data points follows the patterns seen in fig. 3.3 as it clusters around the 2,000 kg.

Furthermore, it appears that there is no major instability in the model under various external conditions, as the uncertainty intervals are not too far from the data points. The data points with small uncertainty intervals mostly come from the farmers with irrigation system, which is supported by the idea that farmers with irrigation system are less affected by climate variation and vice versa. From this results it seems that variability it is not an important factor to the poor performance.

3.1.3 Structural error model

Linear Principal Component Analysis

14 variables were selected to be tested against the total error. Using SPSS Statistics software's Dimension Reduction function, these variables are reduced to 5 principal components (PC). The represented eigenvalues and variance for each PC can be seen in table 3.3. Combined, these PCs account for 63.8% of the variance, it is deemed acceptable considering the number of parameters analyzed. Table 3.4 gives the list of variables tested. It also provides the reduced principal components (PC 1-5) and their correlation against each variable. To select which variables a PC is representing, a correlation of ≥ 0.5 or ≤ -0.5 is considered. The perceived variables for each component can be seen as a light gray cell down of the PC column in table 3.4. PC1 and PC2 can be categorized as the location and water components, PC3 is the cost related components, PC4 is the SEC components, and PC5 is the fertilizer component. PC5 amounts to 7.3% of the variance but it contains only one variable so fertilizer usage can be more influential than expected.

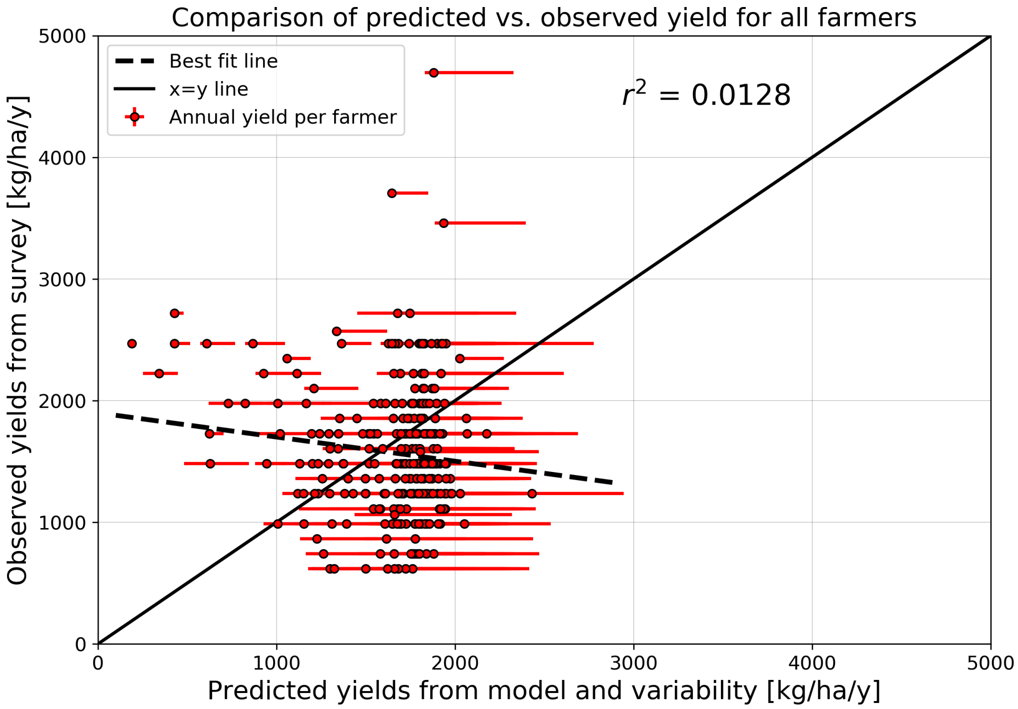


Figure 3.4: Yield comparison between the model and observed data when bootstrap climate data is used. The result is similar to the previous comparison in that it fails to simulate the high crop yield and clusters around the 2,000 kg/ha. The data points with low uncertainty intervals mostly come from the irrigating farmers, and vice versa. Overall, the climate variation does not seem to impact the stability of the model.

Table 3.3: The explained variance for each of the component. The 5 components account for 63.8% of the variance. Both the initial and rotated values are presented, but the rotated components are used in this research.

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.359	23.995	23.995	2.864	20.458	20.458
2	1.861	13.293	37.287	2.255	16.108	36.565
3	1.613	11.524	48.811	1.608	11.484	48.049
4	1.097	7.834	56.645	1.183	8.453	56.502
5	1.008	7.203	63.848	1.028	7.346	63.848

Linear regression of principal components with the total error

The first principal components PC1 represents the perceived variation of evaporation, irrigation, and latitude against the yield difference. Fig. 3.5 shows that the total error is proportional to the PC1. Table 3.4 shows negative correlation of evaporation and irrigation against the PC1, thus the farmers with less evaporation and the need for irrigation tend to be underestimated by the model. Similarly the higher latitude farmers are seen to be underestimated as well. In fig. 3.6 PC2 shows a stronger correlation with the total error. The model tend to underestimate at a higher precipitation level, which is consistent with the studies comparison - see section 3.1.1. It also suggest a strong correlation by the longitude variable. Both of these, supports the correlation seen in PC1 with regards to the evaporation, irrigation, and location. Another variable in PC2 is the soil depth, it is underestimating at a lower soil depth, this might be due to the inaccurate representation of the soil surface simulation as the model SM is operating on a one bucket system.

PC3 is relatively uncorrelated as seen in fig. 3.7. The perceived costs and prices do not impact the yield difference. The SH model assumes that even in negative capital conditions, the farmers can pro-

Table 3.4: Rotated component matrix, this table indicates which variables are assigned to which components. The light gray cell helps to visualize this. The correlation thresholds for a variable to be assigned are ≥ 0.5 or ≤ -0.5

	Component				
	1	2	3	4	5
Children help	-0.167	-0.099	-0.024	0.698	0.111
Cotton area	0.111	0.340	-0.038	0.623	-0.345
Cotton yield	0.440	0.194	0.002	-0.305	-0.247
Crop price	-0.062	-0.020	0.589	-0.085	0.007
Seeds cost	0.060	-0.127	-0.095	-0.077	0.370
Pesticide cost	0.163	-0.100	0.729	0.026	-0.159
Fertilizer cost	-0.053	0.056	0.826	0.049	0.107
Fertilizer amount	0.059	0.296	0.145	0.152	0.783
Soil depth	0.541	-0.620	-0.003	-0.149	0.127
Latitude	0.906	-0.233	-0.018	-0.172	0.116
Longitude	0.068	0.894	-0.033	-0.083	0.049
Precipitation	-0.203	0.852	-0.046	0.040	-0.003
Evaporation	-0.910	0.099	0.056	0.171	-0.129
Irrigation	-0.776	0.021	-0.087	-0.294	0.007

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser
Rotation converged in 7 iterations.

ceed to the following year and obtain the necessary resources for their farming activities - see Pande and Savenije (2016) for details on the socioeconomic aspect of the model. The change in capital does not change the seed and fertilizer available to the farmer and that means the yield do not change and total error stays the same. PC4 contains cotton area and children help, and it has a weak correlation - see fig. 3.8. This is consistent with the findings on the sensitivity analysis of the labor factor shape. Lastly, the PC5 contains only the fertilizer amount used. Even though it amounts for the least variance of the 5, it only comes from one variable. Fig. 3.9 indicates that the current influence of fertility amount is very strong. The low fertilizer amount can mean that the soil is fertile and does not need additional Nitrogen. It is seen that the model heavily underestimates in this condition. This suggest that the inactive soil fertility process of the model can play an important part in improving the model prediction as it will bring up the fertilizer factor and thus the yield. Eq. 3.1 shows the linear model using the significant components PC 1, 2, 4, and 5. Combined they account for 0.288 of the variance of the total error.

$$\epsilon_{s,linear} = 215.6PC2 - 210.4PC5 + 112.0PC1 - 71.6PC4 \tag{3.1}$$

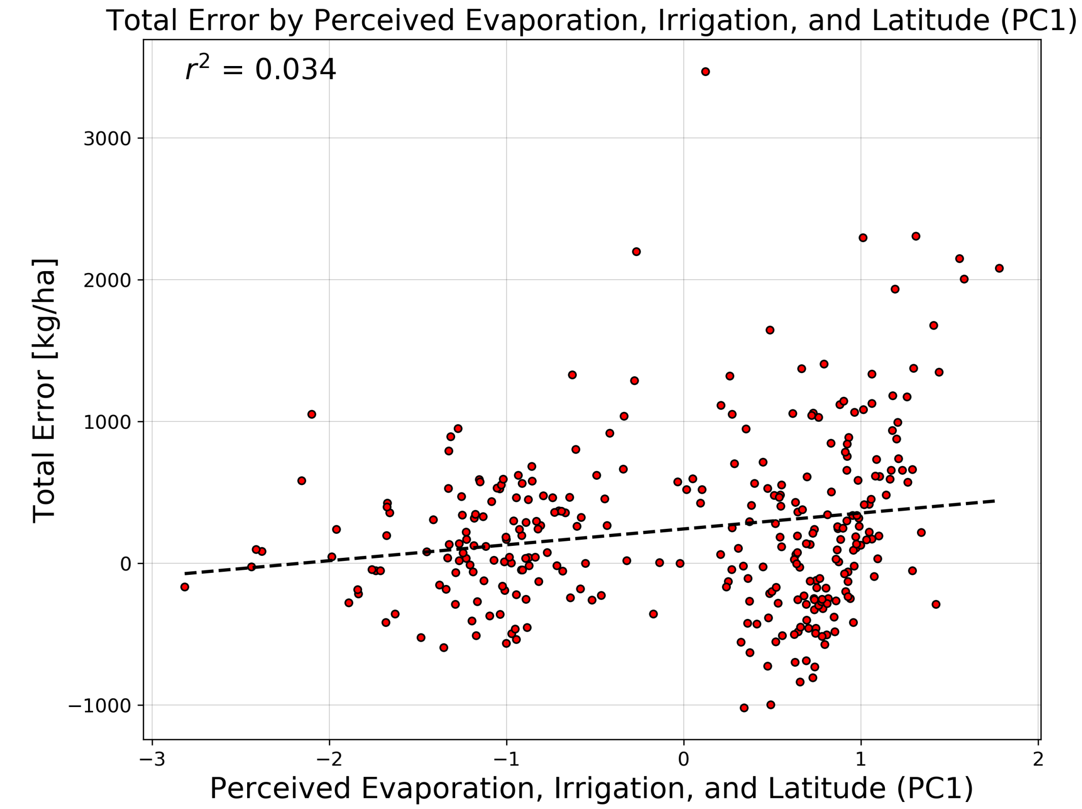


Figure 3.5: Regression result between the PC1 with the total error.

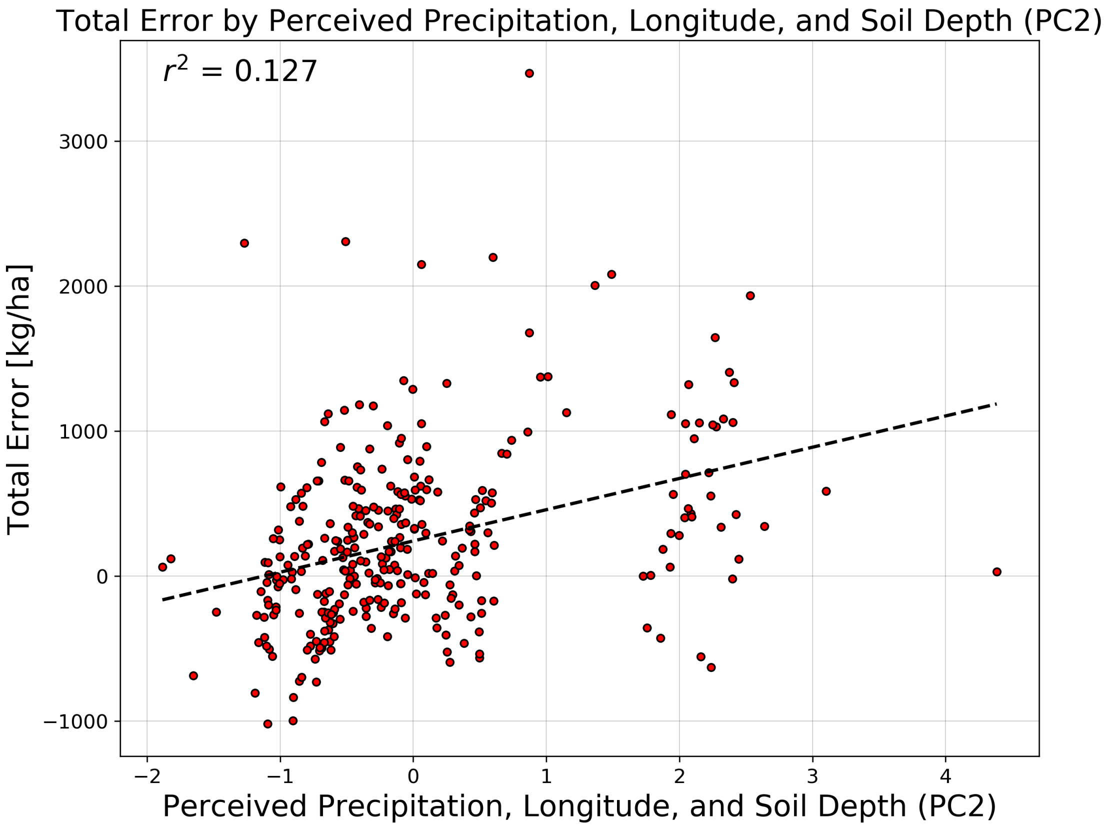


Figure 3.6: Regression result between the PC2 with the total error.

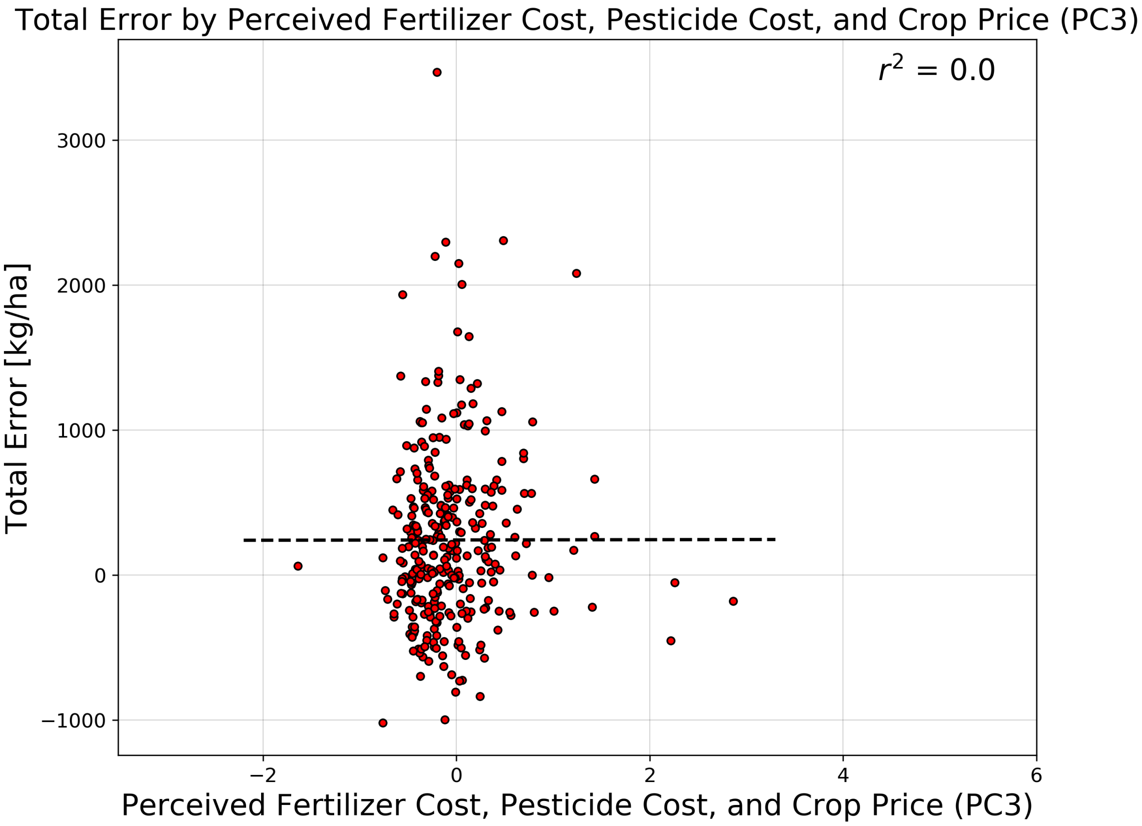


Figure 3.7: Regression result between the PC3 with the total error.

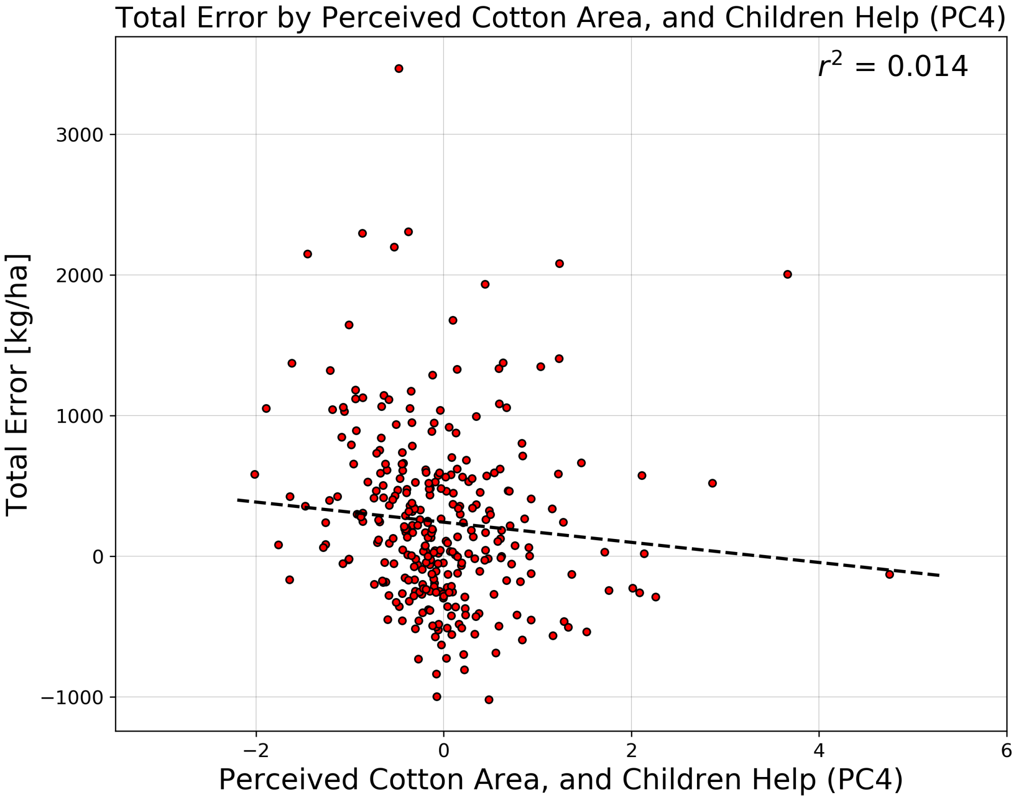


Figure 3.8: Regression result between the PC4 with the total error.

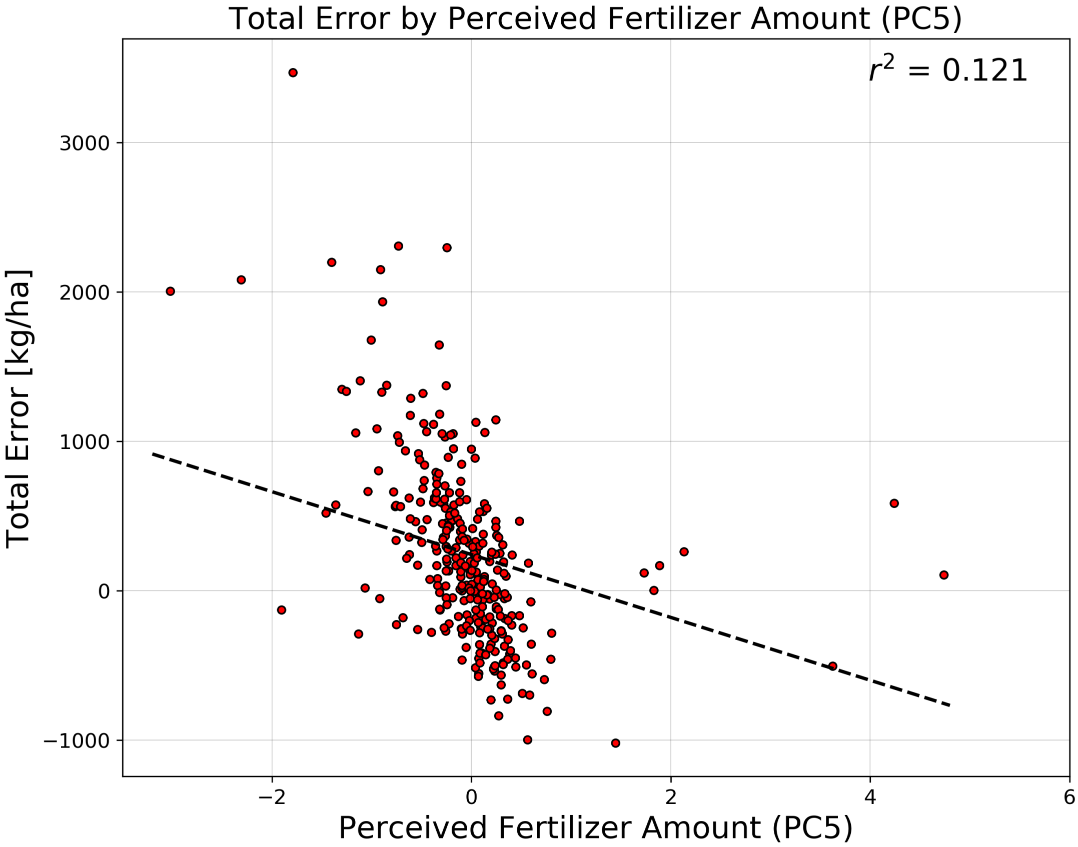


Figure 3.9: Regression result between the PC5 with the total error.

3.1.4 PCA in Kernel Space and Regression with the Total Error

Cross validation result and the error structure model

Table 3.5 gives an overview to the performance of various kernel models on both train and test data. The polynomial kernels perform poorly compared to the non-polynomial kernels. Only degree 2 polynomials are seen to give an acceptable results. However, score drops between the train and test data are observed, which can indicate overfitting making it not suitable to be used. On the other hand, the cosine, RBF, and sigmoid kernels perform similarly well and there is no major jump between the train and test data on all the scores. In this research, sigmoid kernel is used as it has the best values in all scores. However, it must be noted that the other two kernels can be good substitutes especially during future testing and when more data are acquired. More detailed information regarding the cross validation such as the retained components, plots of test data, and summary of regression model in kernel space for various kernels can be seen in Appendix E.

The comparison of the prediction of the total error using the regression model and the observed total error is seen in figure 3.10. Overall, the predicted total error match the observed error well with perhaps a slight deviation at higher error values. The difference between the predicted and observed error is attributed to the residual error (ϵ_r) and its distribution can be seen in the histogram at figure 3.11. The figure also includes the assumed Gaussian distribution of the residual error based on the histogram with $\sigma = 150$ kg/ha. This assumed distribution can be used as a priori distribution in future research e.g. calibration. For now it is used to visualize the uncertainty interval of the yield prediction.

Table 3.5: Performance comparison of various kernel functions on train and test data. The kernels are evaluated using MAE, NS, and r^2. The sigmoid kernel has the best performance (light grey colored row) in all scores. Thus, sigmoid kernel is chosen to map the variables onto the kernel space.

	MAE [kg/ha]		NS [-]		r^2 [-]	
	Train	Test	Train	Test	Train	Test
Poly, deg: 2	178	228	0.82	0.67	0.820	0.680
Poly, deg: 3	428	470	0.088	-0.084	0.084	-0.009
Poly, deg: 4	438	469	0.050	-0.055	0.046	-0.013
Poly, deg: 5	440	468	0.042	-0.047	0.037	-0.013
Cosine	185	184	0.80	0.79	0.792	0.792
RBF	166	183	0.85	0.79	0.837	0.789
Sigmoid	139	167	0.89	0.82	0.887	0.823

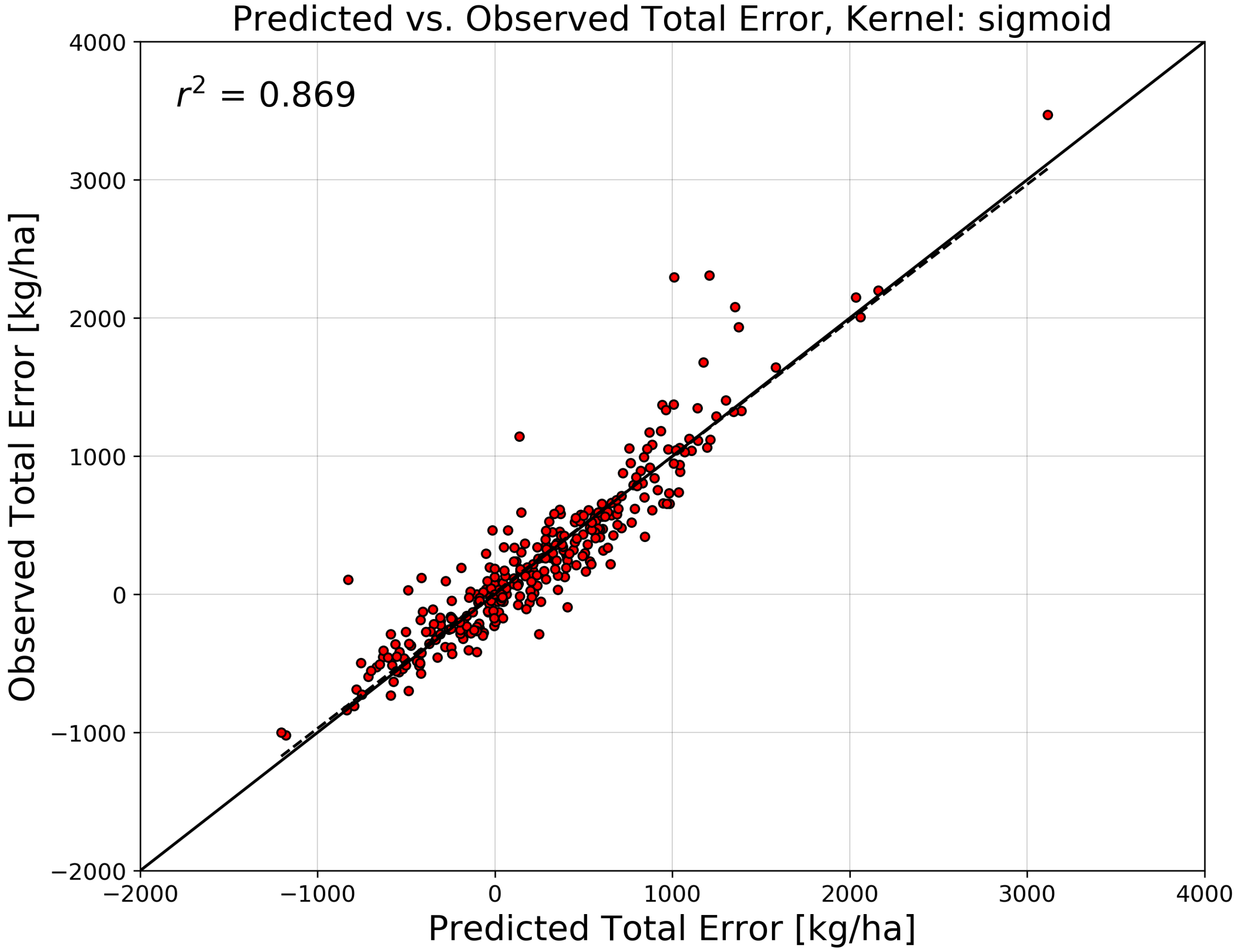


Figure 3.10: Prediction of total error using the structural error model. It is shown that the structural error model is able to capture around 87% of the variance of the total error.

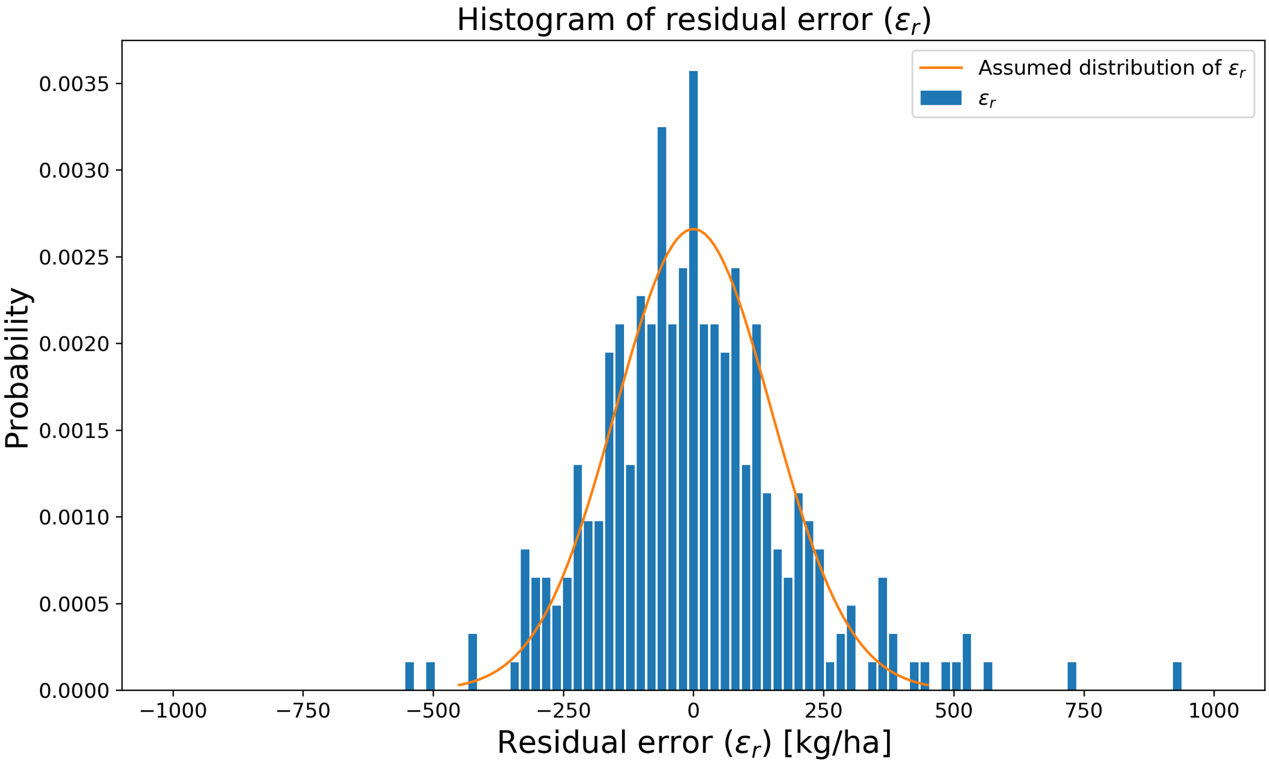


Figure 3.11: Distribution of ϵ_r and its assumed gaussian distribution (with $\sigma = 150$ [kg/ha]). This is assumed to be the uncertainty interval of the predictive model.

Adjusted prediction of cotton yield

The calculated yield from the physical model is adjusted with the structural error model for a new prediction of the cotton yield. The new model prediction is substantially better than the physical model. Table 3.6 shows the comparison between the physical and the combined model. The adjustment using structural error model improves the prediction in all four scores. This improvement can also be seen in figure 3.1.4. The predicted yield follows the observed yield very well even at high yield values. The uncertainty interval from the ϵ_r can be seen as a yellow are around the fit line of the predicted yield. Even though the model still tends to underestimate at higher yield, the y=x line is still within the uncertainty interval. The large difference in performance between the prediction from the SH model and the SH + structural error model indicates the prevalence of structural error in the SH model.

Table 3.6: Comparison of model scores between physical vs. combined (physical + structural error) model. It sees major improvement in all four scores.

	MAE [kg/ha]	NS [-]	NS_{log} [-]	r^2 [-]
SH model	489	-0.624	-1.145	0.005
SH + structural error model	149	0.83	0.56	0.859

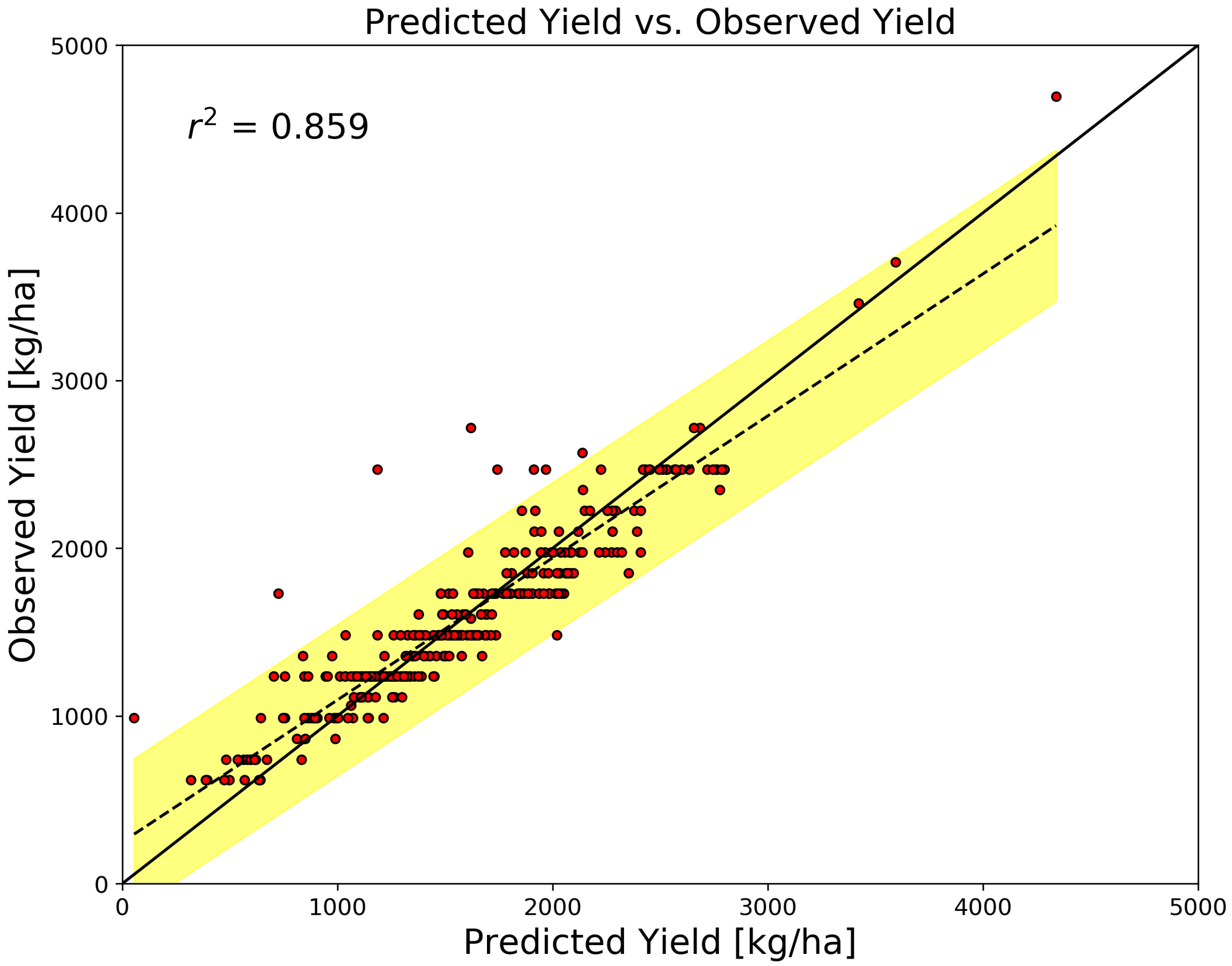


Figure 3.12: Predicted vs. observed cotton yield by the physical and structural error model. The yellow area surrounding the fit line is the uncertainty interval obtained from the assumed distribution of ϵ_r

3.2 Qualitative analysis

Following the evaluation of the model performance and structural error analysis, a qualitative analysis is performed to identify the potential ways the factors can contribute to the errors.

Capital of farmers

The results from the regression suggest that latitude and longitude of the farmers are correlated with the total error. Fig. 3.13 shows the location of the farmers and their yield difference. Fig. 3.14 shows the mean capital of farmers and standard deviation. There is a reasonable inference that can be drawn when comparing the two figures. Amravati and Yavatmal are seen to have a lower capital compared to Wardha. Similarly, the yield difference are more likely to overestimate in Amravati and Yavatmal compared to Wardha. This is also supported by the regression in fig 3.5 and 3.6. One might argue that richer farmers have better ways to mitigate climate impact, more capital to afford fertilizer, have access to better tools, etc, that are not captured very well by the model.

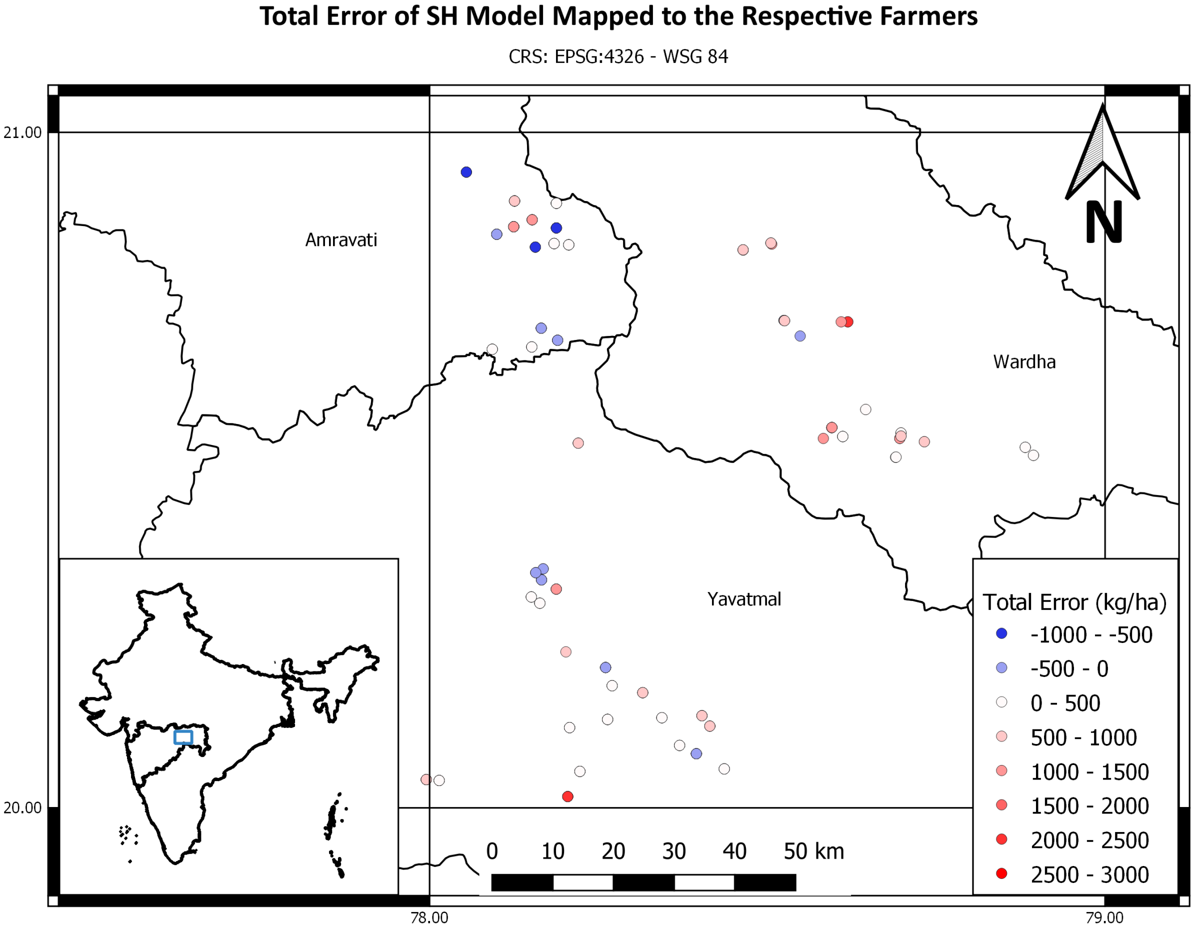


Figure 3.13: Map of total error for the farmers. It is seen that Amravati and Yavatmal are more likely to have overestimated yields compared to Wardha. It is also noted that the underestimations have larger magnitude compared to the underestimation. Made in QGIS (QGIS Development Team, 2021).

Irrigation process in the model

The yield comparison from fig 3.3 and 3.4 shows that the farmers with high crop yield are not captured very well by the model. Figure 3.15 and 3.16 show similar plots, with the exception that only rainfed

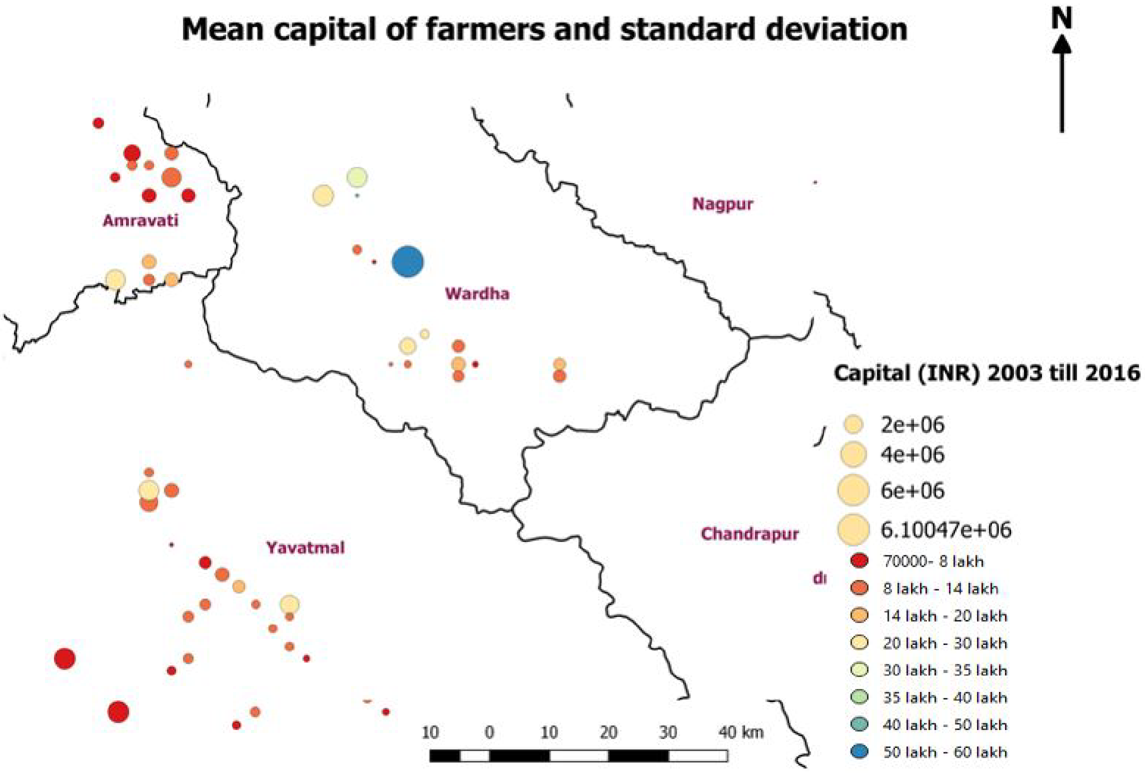


Figure 3.14: Mean capital of farmers found during the baseline study. Obtained from Hatch et al. (2019). Amravati and Yavatmal are seen to have poorer farmers compared to Wardha. An inference can be made between the distribution of capital and the total error.

farmers data are used. Both plots show a stronger correlation compared to their previous counterparts. The r^2 value of the model yield using the optimal parameter sets improve from 0.0046 to 0.1703, whole vs. rainfed data respectively. Similarly, the r^2 for the model yield using the bootstrap data improve from 0.0128 to 0.0797, whole vs. rainfed data respectively. This significant difference in performance indicates that irrigation mechanics should be investigated.

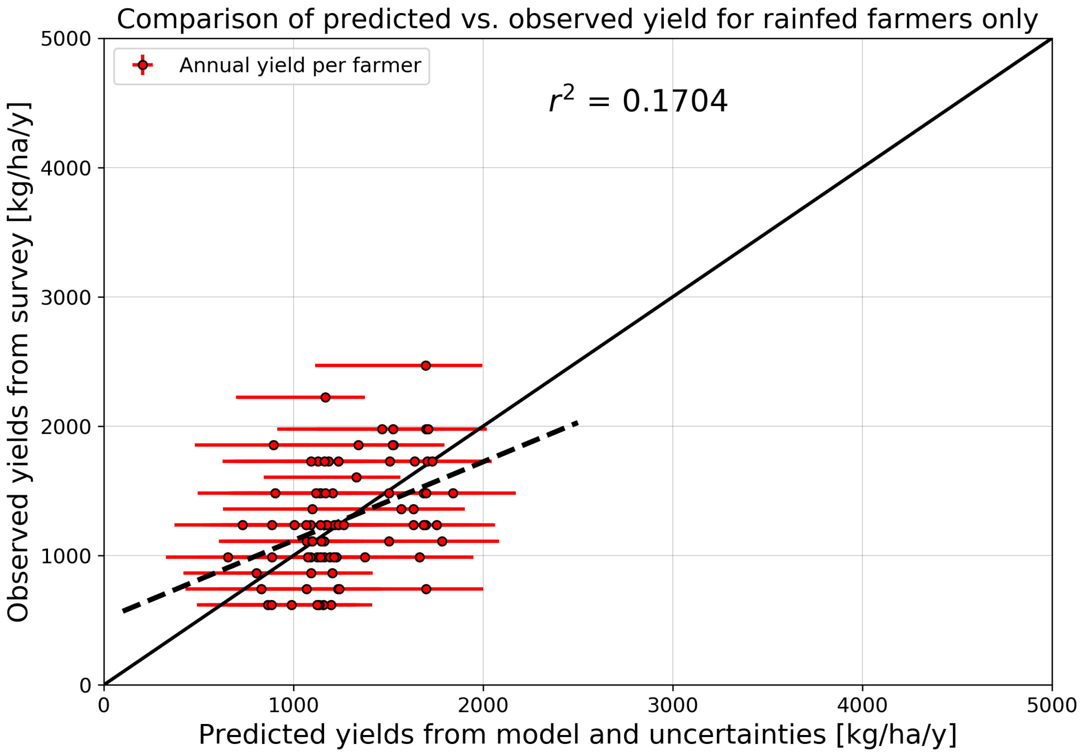


Figure 3.15: Yields using optimal parameter sets

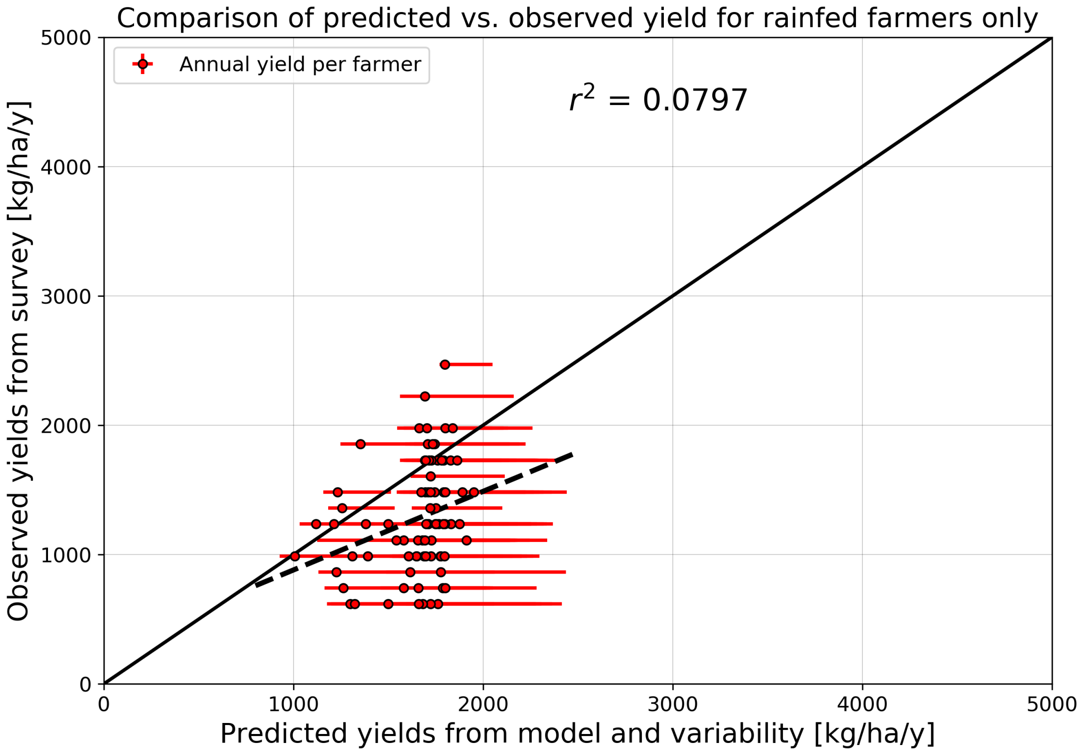


Figure 3.16: Model yield comparison against the observed data when only considering rainfed farmers with uncertainty interval that comes from the climate data variability.

Validation against CC and SM

In this experiment, the model is only calibrated against the crop yield. Here, the canopy cover and the soil moisture will be compared to the NDVI reading and GLEAM data to check whether they are performing well.

First, the canopy cover data will be compared to the NDVI reading from the Sentinel 2 data. The reading is obtain using Google Earth Engine (GEE) (Gorelick et al., 2017). Sample of the script used can be seen in appendix F. In total, 7 readings were taken, i.e. the maximum value of NDVI each month over the planting season from June to December. The NDVI value is then compared with the calculated CC from the model. It was found that most of the coordinates associated with the farmers are the location of their house instead of their farm. To work around this, the coordinates of the farmers are mapped out using Google Earth Pro software. The average CC of farmers in a given village is compared to the 5 NDVI readings in the surrounding area that resembles cotton fields. Another factor that determines the selection is the temporal variation of the field during the 2018 planting season, the relatively green fields between June and December are selected. In the end, there are 5 locations that are used for the comparison viz. Wardha, Ghatanji, Yavatmal, Amravati, and Hinhanghat. Fig. 3.17 shows the comparison of CC and NDVI in Wardha. Tenreiro et al. (2021) shows that the relationship between CC and NDVI for industrial crop is very close to 1:1. Thus, the model is over predicting the CC by about 0.2 at the peak around September and October. Similarly, the NDVI maps (fig. 3.19) shows that the canopy cover in the general region peaks in September. Overall the timing of the growth is correct, however the maximum canopy cover needs to be lowered. This is one aspect of the model that can be adjusted in future research to reduce the structural error.

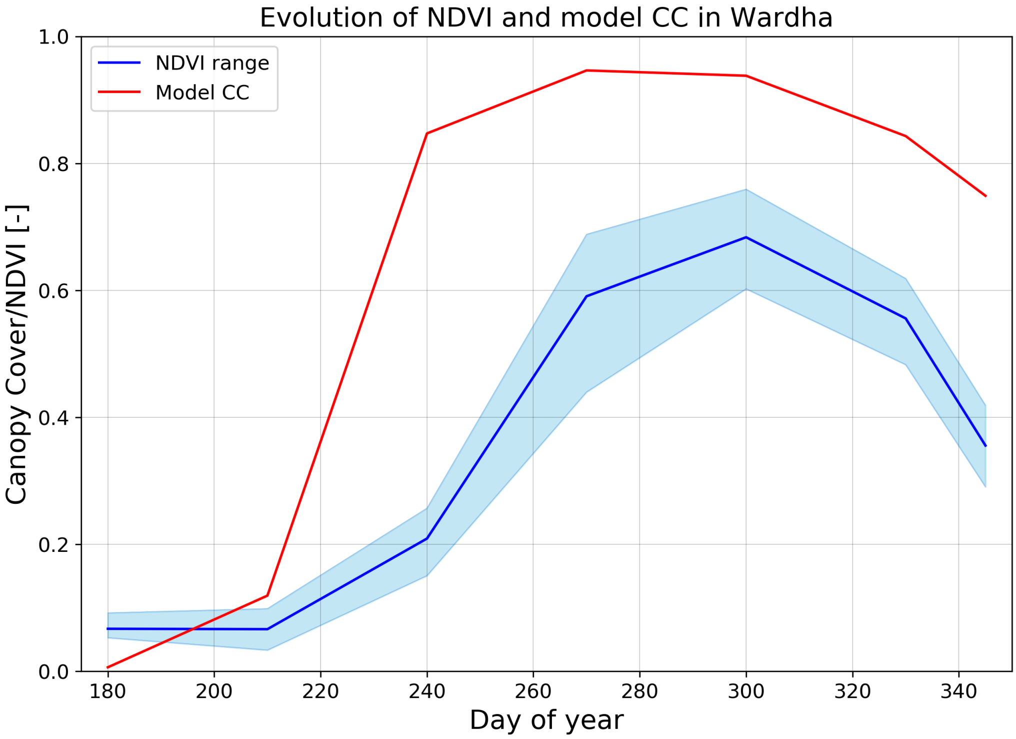


Figure 3.17: Model canopy cover vs. NDVI in Wardha

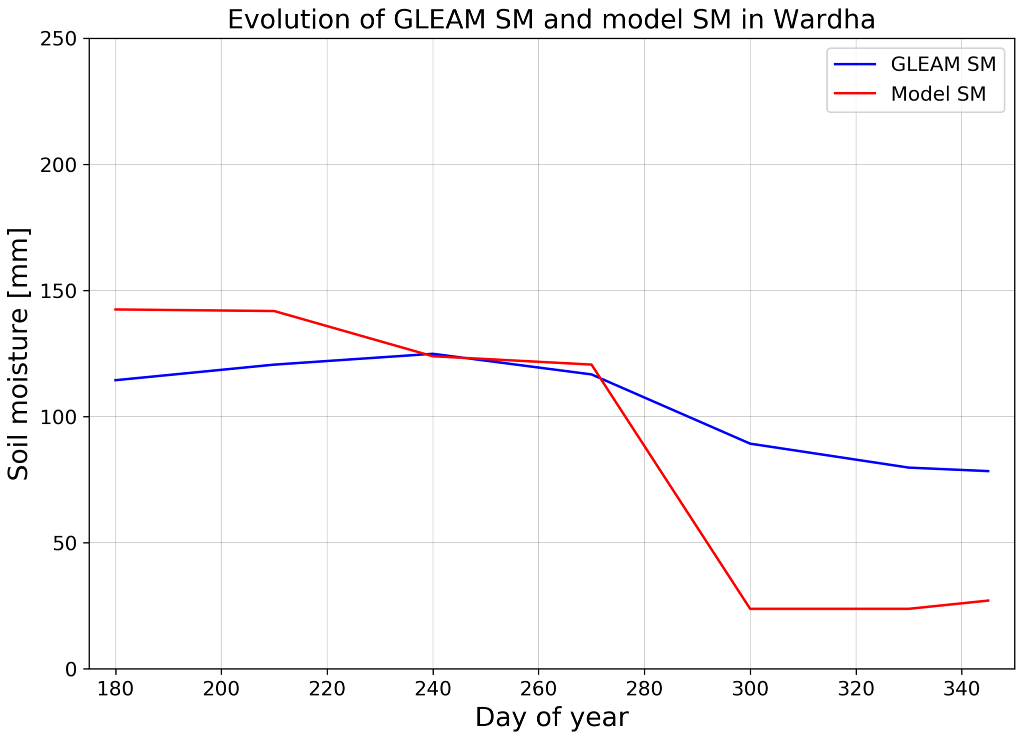


Figure 3.18: Model soil moisture vs. GLEAM in Wardha

Secondly, the soil moisture from the model is compared with the GLEAM v3 root zone soil moisture data (Martens et al., 2017; Miralles et al., 2011). The GLEAM data used is daily with a resolution of 0.25 x 0.25 degrees. The soil moisture is compared at the day of year: 180, 210, 240, 270, 300, 330, 345 following the planting season. Figure 3.18 showcased these SM comparison. It seems that both of them are giving similar SM calculation at the start of the season. Towards the end of the season the SH model predicts less SM compared to GLEAM. Both the CC and SM performance can be considered reasonable, though future validation using site data will provide a better comparison. The complete comparison for all location can be seen in appendix G.

3.2.1 Summary of model evaluation and error analysis

From the model evaluation, it is found that the SH model by itself is performing poorly in terms of crop yield prediction. However, when adjusted with the structural error model, the predicted yield is close to the observed yield. This indicates a strong effect of structural error towards the total error. From the linear regression of the PCA components and the qualitative analysis this structural error can be explained by the capital, fertilizer, irrigation behavior, and crop production process (CC and SM).

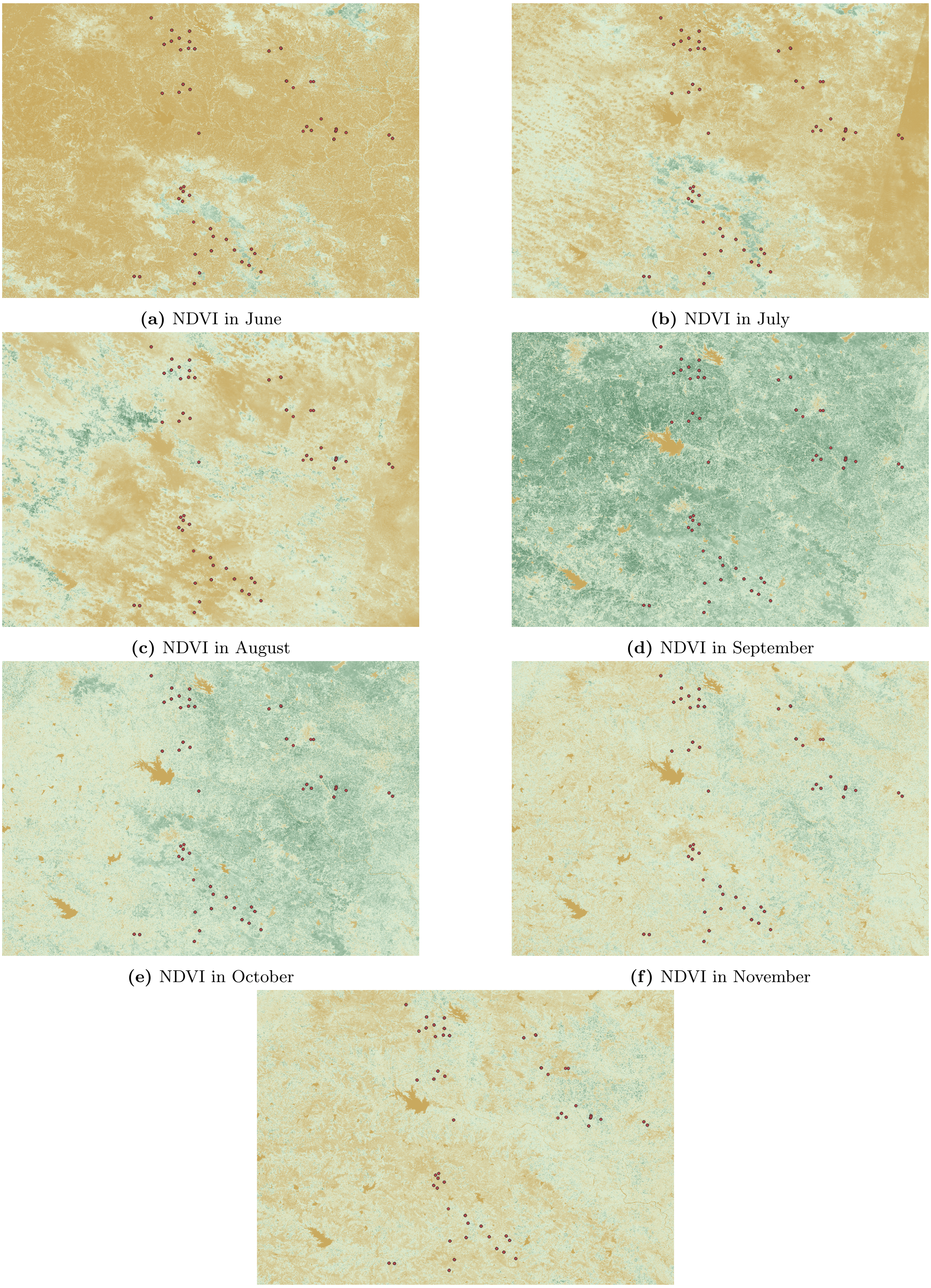
3.3 Limitation, Restrictions, and Assumptions

The implemented FAO's AquaCrop framework is still a more simplified version of the original. Thus the yield calculation might not be the most accurate. There are various processes that can be implemented to potentially improve the yield prediction. For example, the soil fertility stress and its interaction with water productivity, the soil salinity stress and movement of salt, evolving harvest index throughout the season based on various conditions, the root water extraction rate, etc. The assumptions being made in the process also contributes to the inaccuracies, for example, the rainfall interception is assumed to be 0, the water inputs and outputs are assumed to move immediately (not limited by a certain rate), etc.

Using MCS for calibration is computationally expensive and there are more efficient method. Another method considered at the start of this research is Bayesian approach. It is a more efficient approach compared to MCS (Oakley and O'Hagan, 2004) given the size and the complexity of the model. It updates the prior probability density functions (PDF) with its posterior PDF for every iteration until it is close to the observed data (Muehleisen and Bergerson, 2016). Furthermore, it prevents overfitting of the data because this method try to maximize the likelihood of the model output is statistically consistent with the observed data instead of minimizing the error (Muehleisen and Bergerson, 2016). This is especially important given the calibration results discussed in chapter 3.1. However, in the early version of the model there were erratic changes in the model results when the inputs and parameters are changed, so it was deemed not appropriate to do Bayesian approach (Oakley and O'Hagan, 2004). Moreover, the variance of the model output is found to be heteroscedastic and would not fit for the usage of Bayesian inference - See appendix B.1 for details. It is noted however, that when the model performs relatively well in the future and more efficient calibration method is needed, this approach can be considered.

In order to model the structural error, it is assumed that the total error (ϵ_{tot}) can be separated to the residual error (ϵ_r) and structural error (ϵ_s). In reality it is difficult, if not impossible, to disaggregate the various errors (Beven, 2005). However, the method presented here is deemed good enough for the purpose of this research, i.e., a rough first order approximation of the magnitude of the structural error to evaluate the its prevalence.

With regards to the survey data, the farmers coordinates are of their house instead of the field. It poses a problem when trying to validate the CC with the NDVI since the exact location is unknown. Further, there are multiple farmers with the same coordinate, most likely due to rounding of the decimals. For future studies it is best to obtain the field coordinate and make sure that it is not rounded since it makes distinguishing between farmers difficult.



(a) NDVI in June

(b) NDVI in July

(c) NDVI in August

(d) NDVI in September

(e) NDVI in October

(f) NDVI in November

(g) NDVI in December

Figure 3.19: Evolution of NDVI in the study region over the planting season from June to December 2018. The raster data is taken from Sentinel 2 from GEE (Gorelick et al., 2017). The points represent the farmers' locations.

Chapter 4

Conclusion and Recommendations

To answer the first sub research question “How do the uncertainties from the model and external variations affect the calculated crop yield?”. From the calibration results of the SH model, the uncertainties is shown to be very high. This is due to the model being unable to calibrate the higher yield values and it must compensate this by adjusting the threshold of objective functions so that the majority of the observed yield falls within the uncertainty interval. However, when adjusted by the error model, the uncertainty interval is considerably smaller. Lastly, the external variations do not affect the model significantly and the model is quite stable.

The aforementioned result indicates that structural errors are prevalent, which brings us to the second research question “Which factors are the most influential towards the error between the predicted crop yield and observed yield?”. The linear PCA and the multiple linear regressions indicate that precipitation, irrigation, evaporation, soil depth, locations, labor, and fertilizer amount are correlated with the total error. Among these, fertilizer amount has the strongest correlation. Upon qualitative analysis, the capital distribution in the area is seen to have a reasonable inference to the errors. The capital difference might be linked to the fertilizer and irrigation usage among the farmers. Moreover, the irrigated vs. rainfed farmers comparison indicate a high influence of irrigation towards the error. Lastly, the validation of CC and SM indicates some discrepancies and that there are improvements that can be done to the simulation of water and crop production mechanics.

To conclude, “What is the overall performance of the model in calculating the cotton yield in the Maharashtra region with the addition of the stress mechanics?”. The SH model by itself is not performing very well. Especially at individual farmer level over a large region. The SH model has NS value of -0.343 to -0.996, NS_{log} of -0.655 to -1.91, MAE of 447.1 to 553.2 kg/ha, and r^2 of 0.003 to 0.008. The additions of the stress mechanics did little to improve this performance. However, when used in combination with the structural error model, the crop yield prediction score improve to NS value of 0.83, NS_{log} of 0.56, MAE of 149 kg/ha, and r^2 value of 0.859. Recommendations to improve the model performance is discussed in the section 4.1.

4.1 Recommendations

Table 4.1: Lists of recommendations for the shortcomings discussed in previous chapters

Shortcomings	Recommendations
Calibration	Currently, the model was only calibrated to the crop yield. Step calibration can be utilized to check the other parameters viz. canopy cover, soil moisture, and soil fertility (when validation data are obtained from future survey). Furthermore, the calibration using MCS takes too much time for a model this complex. When most of the structural errors are addressed the model can be calibrated using more efficient statistical inference. One example (but not limited to) is a bayesian approach as it can reduce the number of iterations to a far smaller number compared to using the MCS method.
Capital, irrigation, behavior	Conduct exploratory interviews or workshops with the stakeholders to find possible patterns between irrigated and rainfed farmers, or farmers with high capital. These patterns then can be used in a survey to obtain quantifiable variables for the model. This survey can include things such as irrigation behavior, technique, weather forecast usage, etc. These variables can either be used to improve the structural error model or to implemented directly in the SH model to reduce the structural error. For example, certain model processes that can be triggered given certain variables or conditions. The appendix H shows some preliminary RANAS questions for irrigation and expenditure cuts that can be used as a basis for the exploratory interviews and surveys.
Soil fertility	Currently, the soil fertility is not activated in the model. Only the fertilizer usage is implemented in the model. Further, the fertilizer factor calculation is very basic with only a linear function to represent the usage. Soil fertility should be activated in order to give a more representative condition in the soil. Moreover, processes such as soil erosion and effects of soil moisture to fertility can be added to improve the model mechanics. Lastly, the fertility adjustment factor is currently calculated annually, this can be reworked to be a daily calculation. This will create a better foundation for future implementations and improve the model as an adaptation tool.
Water balance	Two buckets process can be used to give a more accurate representation of the vertical flow in the soil. Appendix I showcased the preliminary implementation of a two bucket system. Moreover, a horizontal flow could help in capturing the dynamics between crop fields and grazing area.
Crop production	There are processes that can still be added to the crop production code to potentially make it better. For example, the soil salinity stress and movement of salts, evolving harvest index, root extraction rate, etc.
Miscellaneous	- The coordinates obtained during the survey should be the coordinates of the farm instead of the farmers' house as it is more useful to the model i.e. to be used in the validation of CC, SM, and other parameters. - Restructuring the model framework so it runs all farmers at once to allow feedback response between them (especially important for irrigation water intake). - Implementation of a changing K_e coefficient e.g. due to wetted surface. - A look into the system dynamics of the area to investigate the relationships between influential variables.

Bibliography

Allen, R. G., Pereira, L. S., and Raes, D. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56 Table of Contents. Technical report, FAO - Food and Agriculture Organization of the United Nations.

Allen, R. G., Pruitt, W. O., Raes, D., Smith, M., and Pereira, L. S. (2005). Estimating Evaporation from Bare Soil and the Crop Coefficient for the Initial Period Using Common Soils Information. *Journal of Irrigation and Drainage Engineering*, 131(1):14–23.

Andrade, L., O'Malley, K., Hynds, P., O'Neill, E., and O'Dwyer, J. (2019). Assessment of two behavioural models (HBM and RANAS) for predicting health behaviours in response to environmental threats: Surface water flooding as a source of groundwater contamination and subsequent waterborne infection in the Republic of Ireland. *Science of the Total Environment*, 685:1019–1029.

Behere, P. and Behere, A. (2008). Farmers' suicide in Vidarbha region of Maharashtra state: A myth or reality? *Indian Journal of Psychiatry*, 50(2):124.

Beven, K. (2005). On the concept of model structural error. *Water Science and Technology*, pages 167–175.

Beven, K. and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298.

Brown, J. D. (2009). Choosing the Right Type of Rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter*, 13(November):20–25.

Contzen, N., Mosler, H.-J. (2015). Methodological sheets #1-6: The RANAS approach to systematic behavior change; The RANAS model of behaviour change; The RANAS behavioural factors; The RANAS behaviour change techniques; Doer/non-doer analysis to specify the critical behavioural factors; &. *National Ambient Air Quality Monitoring*, 7(1):829–843.

De, S. and Tantri, P. (2014). Borrowing Culture and Debt Relief: Evidence from a Policy Experiment. *Asian Finance Association (AsianFA) 2014 Conference Paper.*

den Besten, N. (2016). The socio-hydrology of smallholders in Marathwada, Maharashtra state [India]. Technical Report August, TU Delft.

Dongre, A. R. and Deshmukh, P. R. (2012). Farmers' suicides in the Vidarbha region of Maharashtra, India: A qualitative exploration of their causes. *Journal of Injury and Violence Research*, 4(1):2–7.

FAO (2012). AquaCrop: Reference Manual. Technical Report June, FAO, Land and Water Division.

Gardner, R. H. and O'neill, R. V. (1981). Parameter Uncertainty and Model Predictions: A Review of Monte Carlo Results. *Journal of Chemical Information and Modeling*, 53(9):1689–1699.

George H. Hargreaves and Zohrab A. Samani (1985). Reference Crop Evapotranspiration from Temperature. *Applied Engineering in Agriculture*, 1(2):96–99.

Giné, X. and Kanz, M. (2014). The Economic Effects of a Borrower Bailout Evidence from an Emerging Market. *The Review of Financial Studies*, 31(5).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Han, C., Zhang, B., Chen, H., Liu, Y., and Wei, Z. (2020). Novel approach of upscaling the FAO AquaCrop model into regional scale by using distributed crop parameters derived from remote sensing data. *Agricultural Water Management*, 240(January):106288.

Hatch, N., Raghunathan, P., Willaard, T., Van, C., Mohammed, W., Saket, Y., Contributors, P., Adank, K., Klessens, T., Daniel, D., Ekka, A., Steele-Dunne, S., Tyagi, A., Nagi, A., and Pastore, P. (2019). Baseline Study Water Efficiency in Sustainable Cotton-based Production Systems in Maharashtra, India Principal Investigator Saket Pande. Technical report, TU Delft.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):497–526.

IBM Corp (2019). IBM SPSS Statistics for Windows.

Janssen, J. M. (2020). Estimating new reservoir locations with the use of a hydrological model for small holder cotton farmers in Maharashtra , India . Estimating new reservoir locations with the use of a hydrological model for small holder farmers in Maharashtra , India . Technical report, TU Delft.

Kanz, M. (2012). What Does Debt Relief Do for Development ? Evidence from India ' s Bailout Program for Highly-Indebted Rural Households. *American Economic Journal: Applied Economics*, 8(4).

Katti, R. K. (1979). Search for Solutions To Problems in Black Cotton Soils. *Indian Geotechnical Journal*, 9(1):1–82.

Kennedy, M. C. and Hagan, A. O. (2001). Bayesian calibration of computer models. *Statistical Methodology*, 63(3):425–464.

Kerala Legislative Assembly (2012). THE KERALA FARMERS' DEBT RELIEF COMMISSION (AMENDMENT) BILL, 2012.

Kumar, K. N., Rajeevan, M., Pai, D. S., Srivastava, A. K., and Preethi, B. (2013). On the observed variability of monsoon droughts over India. *Weather and Climate Extremes*, 1:42–50.

Lalitha, N., Ramaswami, B., and Viswanathan, P. K. (2009). India's experience with Bt cotton: Case studies from Gujarat and Maharashtra. In *Biotechnology and agricultural development: Transgenic cotton, rural institutions and resource-poor farmers*, pages 135–167. Routledge, Abingdon.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, 10(5):1903–1925.

Merriott, D. (2016). Factors associated with the farmer suicide crisis in India. *Journal of Epidemiology and Global Health*, 6(4):217–227.

Meteoblue (2021). Climate Yavatmal.

Miralles, D. G., Holmes, T. R., De Jeu, R. A., Gash, J. H., Meesters, A. G., and Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, 15(2):453–469.

Mishra, S. (2006). Suicide of Farmers in Maharashtra. Technical report, Indira Gandhi Institute of Development Research.

Mkhabela, M. S. and Bullock, P. R. (2012). Performance of the FAO AquaCrop model for wheat grain yield and soil moisture simulation in Western Canada. *Agricultural Water Management*, 110:16–24.

Muehleisen, R. T. and Bergerson, J. (2016). Bayesian Calibration-What, Why And How. *4th International High Performance Buildings Conference at Purdue.*

Nash, J. E. and Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 66(3):751–769.

Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., and Mukhopadyay, B. (2014). Development of a new high spatial resolution (0 . 25 ° × 0 . 25 °) Long Period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region data sets of different spatial resolutions and time period. *MAUSAM*, 1(1):1–18.

Pande, S. and Savenije, H. H. G. (2016). A sociohydrological model for smallholder farmers in Maharashtra, India. *Water Resources Research*, 52(3):1923–1947.

Paredes, P., Wei, Z., Liu, Y., Xu, D., Xin, Y., Zhang, B., and Pereira, L. S. (2015). Performance assessment of the FAO AquaCrop model for soil water , soil evaporation , biomass and yield of soybeans in North China Plain. *Agricultural Water Management*, 152:57–71.

Pedregosa, F., Weiss, R., and Brucher, M. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

QGIS Development Team (2021). QGIS Geographic Information System.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2012). AquaCrop reference manual. Annex 1. Crop parameters. *Reference Manual of AQUACROP*, pages 2–69.

Raschka, S. and Mirjalili, V. (2019). *Python Machine Learning*. Packt Publishing, 3 edition.

Ratna, S. B. (2012). Summer monsoon rainfall variability over Maharashtra, India. *Pure and Applied Geophysics*, 169(1-2):259–273.

RBI (2009). Agricultural Debt Waiver and Debt Relief Scheme, 2008.

Rotmans, J., Sluijs, J. P. V. A. N. D. E. R., Asselt, M. B. A. V. A. N., Janssen, P., and Krauss, M. P. K. V. O. N. (2003). A Conceptual Basis for Uncertainty Management. *Integrated Assessment*, 00(0).

Royston, P. and Sauerbrei, W. (2009). Bootstrap assessment of the stability of multivariable models. *Stata Journal*, 9(4):547–570.

Samani, Z. (2000). Estimating Solar Radiation and Evapotranspiration Using Minimum Climatological Data. *Journal of Irrigation and Drainage Engineering*, 126(4):265–267.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5).

Sravanth, K. R. S. and Sundaram, N. (2019). Agricultural Crisis and Farmers Suicides in India. *International Journal of Innovative Technology and Exploring Engineering*, 8(11).

Srivastava, A. K., Rajeevan, M., and Kshirsagar, S. R. (2009). Development of a high resolution daily gridded temperature data set (1969 – 2005) for the Indian region. *Atmospheric Science Letters*, 10(4):249–254.

Sud, S. (2009). *The changing profile of Indian agriculture*. Business Standard Books, New Delhi.

Tan, S., Wang, Q., Zhang, J., Chen, Y., Shan, Y., and Xu, D. (2018). Performance of AquaCrop model for cotton growth simulation under film-mulched drip irrigation in southern Xinjiang , China. *Agricultural Water Management*, 196:99–113.

Tenreiro, T. R., García-Vila, M., Gómez, J. A., Jiménez-Berni, J. A., and Fereres, E. (2021). Using NDVI for the assessment of canopy cover in agricultural crops within modelling research. *Computers and Electronics in Agriculture*, 182(November 2020).

Toumi, J., Er-raki, S., Ezzahar, J., Khabba, S., Jarlan, L., and Chehbouni, A. (2016). Performance assessment of AquaCrop model for estimating evapotranspiration , soil water content and grain yield of winter wheat in Tensift Al Haouz (Morocco): Application to irrigation management. *Agricultural Water Management*, 163:219–235.

Udmale, P., Ichikawa, Y., Manandhar, S., Ishidaira, H., and Kiem, A. S. (2014a). Farmers ' perception of drought impacts , local adaptation and administrative mitigation measures in Maharashtra. *International Journal of Disaster Risk Reduction*, 10:250–269.

Udmale, P. D., Ichikawa, Y., Kiem, A. S., and Panda, S. N. (2014b). Drought Impacts and Adaptation Strategies for Agriculture and Rural Livelihood in the Maharashtra State of India. *The Open Agriculture Journal*, 8:41–47.

Vanuytrecht, E., Raes, D., Steduto, P., Hsiao, T. C., Fereres, E., Heng, L. K., Garcia Vila, M., and Mejias Moreno, P. (2014). AquaCrop: FAO's crop water productivity and yield response model. *Environmental Modelling and Software*, 62:351–360.

Wold, S., Esbensen, K. I. M., and Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52.

Appendices

Appendix A

Input Parameter

A.1 Input Parameters Used in the Model

Table A.1 shows the description, ranges, and the value used in the model calculation.

Parameter	Description	Ranges	Used
$Z_{max,plant}$	Maximum effective root depth of cotton [mm]	1000-1700(1)	1400
Z_n	Minimum effective root depth of cotton [mm]	300(2)	300
n	Shape factor of root zone growth [-]	1.5(2)	1.5
t_x	Duration of root growth [days] estimated from the duration from planting to reaching CC_x (from the model)	-	100
p_{sto}	Threshold until root zone stress (fraction of TAW) [-]	0.65(1)	0.65
FC	Field capacity [-]	0.3-0.5	0.42
WP	Wilting point [-]	0.1-0.3	0.17
CC_o	Initial canopy cover when 90% seedlings emerges [-]. 5-7 cm2/plant and there is around 60k-150k plants/ha gives the range(2)	0.003-0.0105	0.0063
t_{cco}	Time to reach initial canopy cover [days]	7-30	7-30
CC_x	Maximum canopy cover [-]. For cotton, it is almost entirely covered(2)	0.9-0.99	0.95
CDC	Canopy decline coefficient [/day], it is a fraction of canopy cover decline per day. Range taken from FAO annex(2), converted from fraction per GDD to fraction per day based on average GDD per day in Maharashtra	0.03-0.045	0.03
CGC	Canopy growing coefficient [/day], it is a fraction of canopy cover growth per day. Range taken from FAO annex(2), converted from fraction per GDD to fraction per day based on average GDD per day in Maharashtra	0.09-0.12	0.097(3)
p_{up}	Upper threshold for water stress [-]	0.7(2)	0.7
p_{low}	Lower threshold for water stress [-]	0.2(2)	0.2
REW	Readily evaporable water [mm]	8-12(4)	10
TEW	Total evaporable water [mm]	25-38(4)	35
T_{up}	Upper threshold for temperature stress effect to biomass production [°C]	0.1(2)	0.1
T_{bot}	Bottom threshold for temperature stress effect to biomass production [°C]	20(2)	20
CWP	Crop water productivity value normalized to evapotranspiration $[g/m^2/mm]$	15(2)	15

HI	Harvest index [-]	25-40(2)	0.25-0.40
$f_{shape,lab}$	Shape of labor factor function	-	-6.457 to - 9.995

Table A.1: Description of parameters in the model and their values (1)(Allen et al., 1998), (2)(FAO, 2012), (3)(den Besten, 2016),(Allen et al., 2005)(4)

A.2 Ra values

Table A.2 shows the extraterrestrial radiation (Ra) values used in the Hargreaves equation to estimate the potential evapotranspiration.

Month	Ra [mm/d]
1	11.2
2	12.7
3	14.4
4	15.6
5	16.3
6	16.4
7	16.3
8	15.9
9	14.8
10	13.3
11	11.6
12	10.7

Table A.2: Monthly average extraterrestrial radiation (Ra) values at 20°N latitude, used in the Hargreaves equation. (Samani, 2000)

Appendix B

Additional Graphs

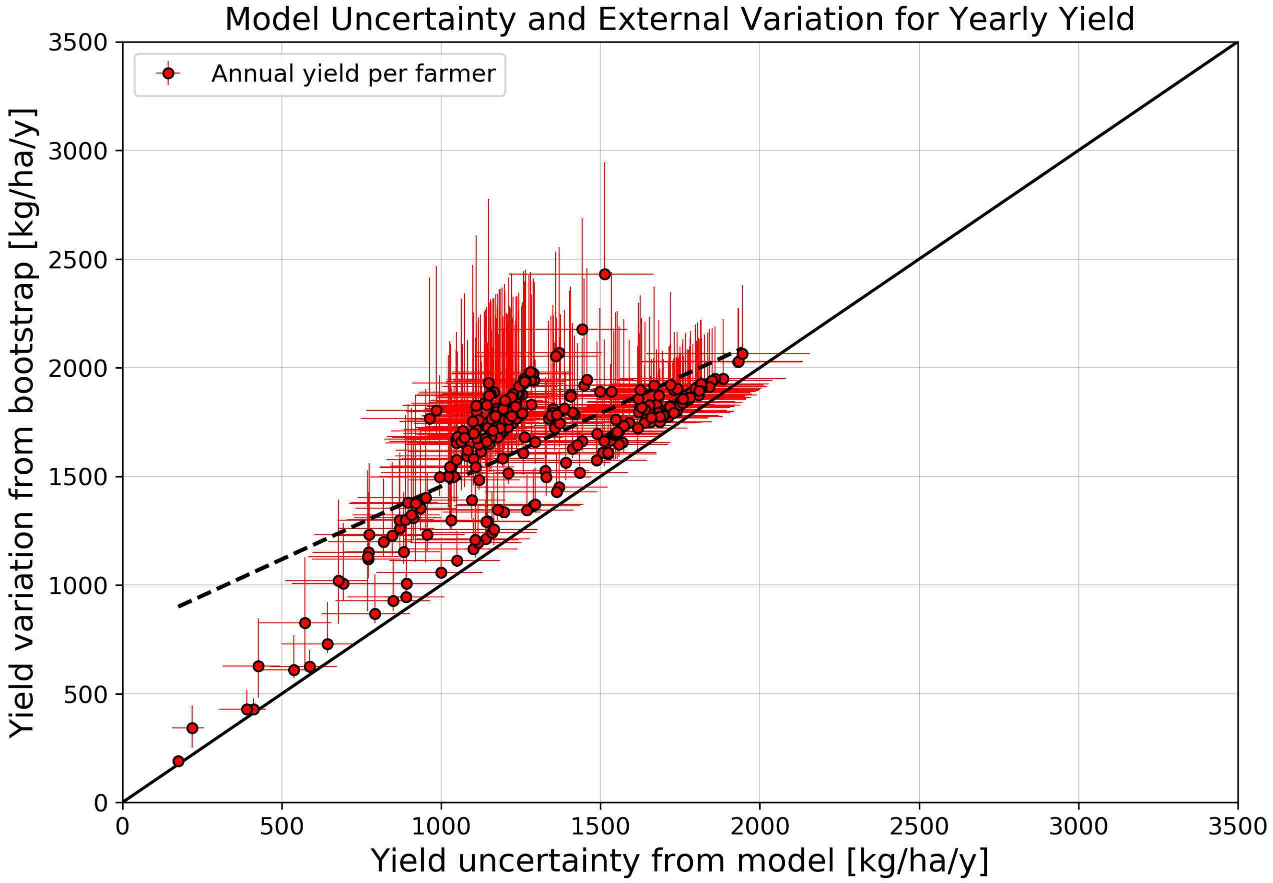


Figure B.1: Scatter plot of calculated yield from the MCS and bootstrap iterations, it show the heteroscedasticity of the variance.

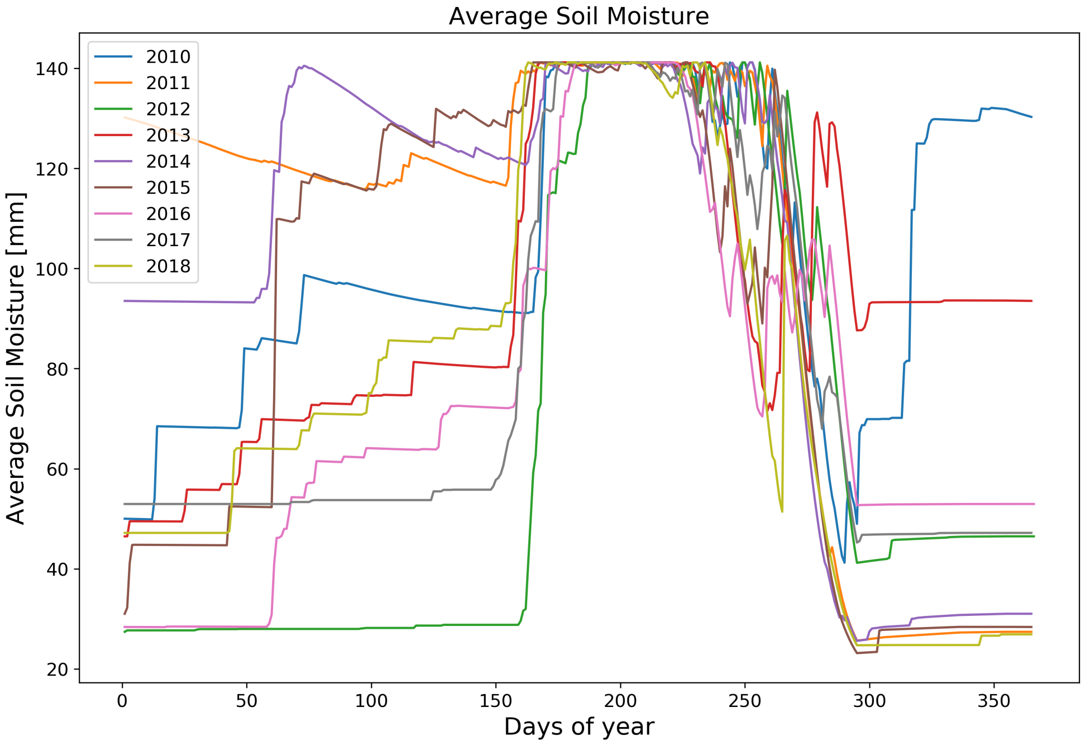


Figure B.2: Evolution of soil moisture over various years

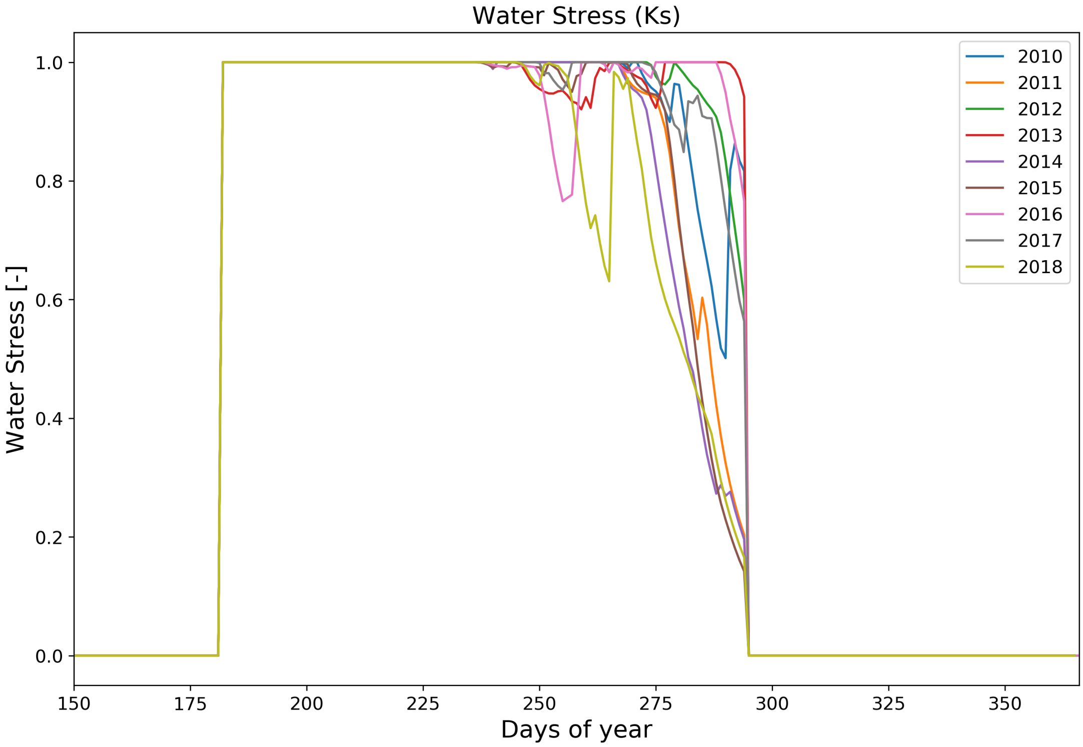


Figure B.3: Evolution of water stress over various years, it highlights the absence of water stress at the start of the season (value of 1 indicates no water stress, vice versa). Water stress occurs towards the end of the season. Value outside of the planting season is not calculated.

Appendix C

MCS Results

C.1 Objective Function Variation with Parameter Values

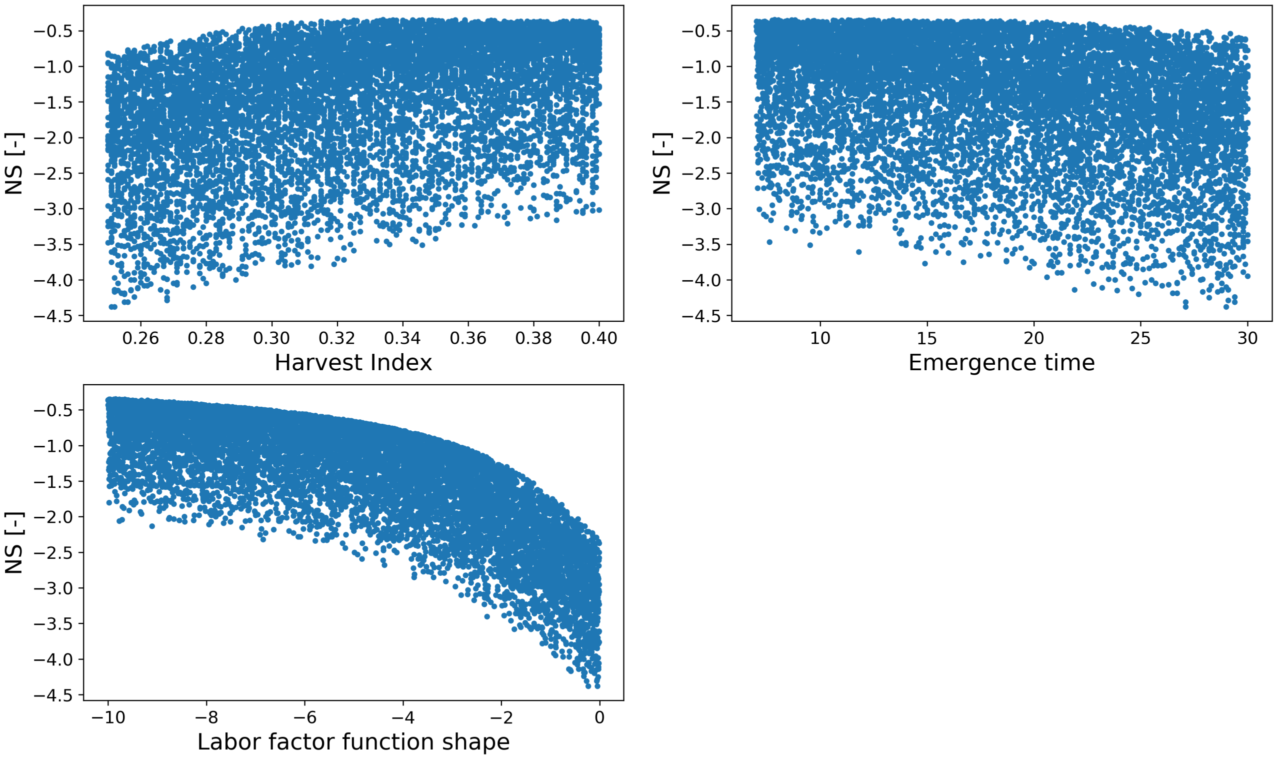


Figure C.1: Nash-Sutcliffe variation with parameter values

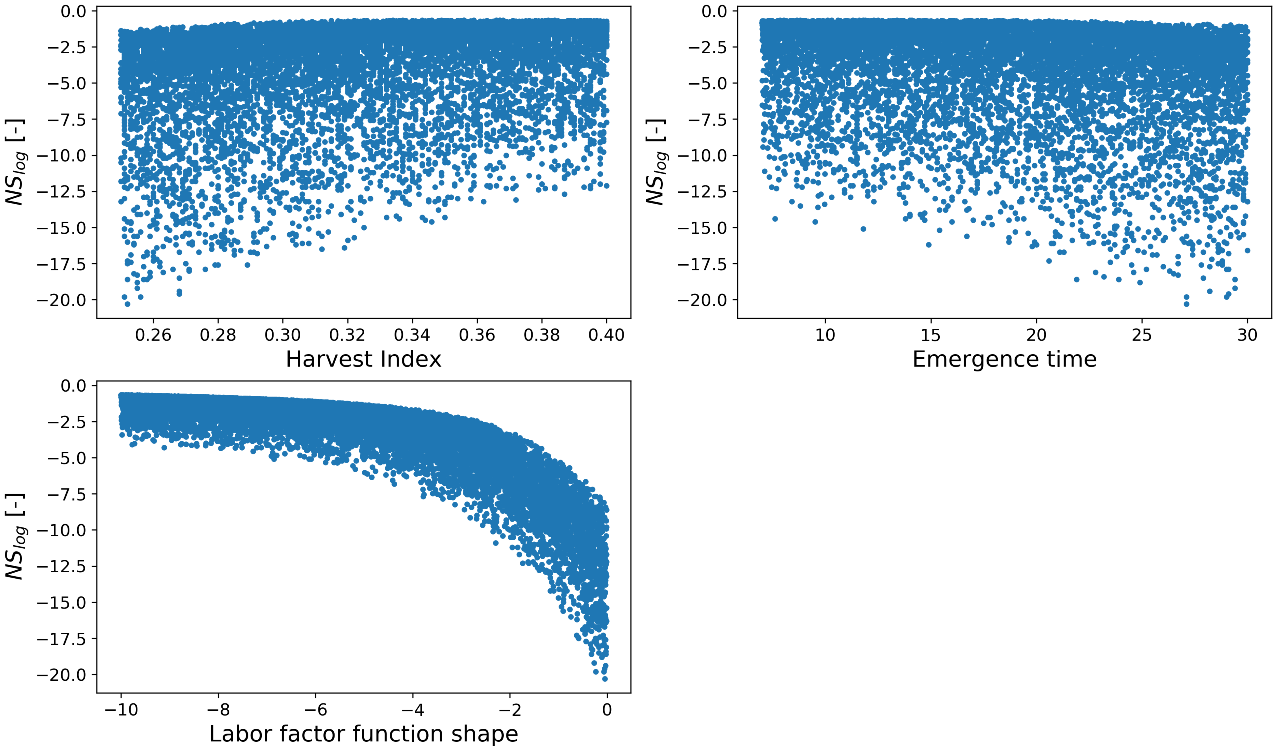


Figure C.2: Log of Nash-Sutcliffe variation with parameter values

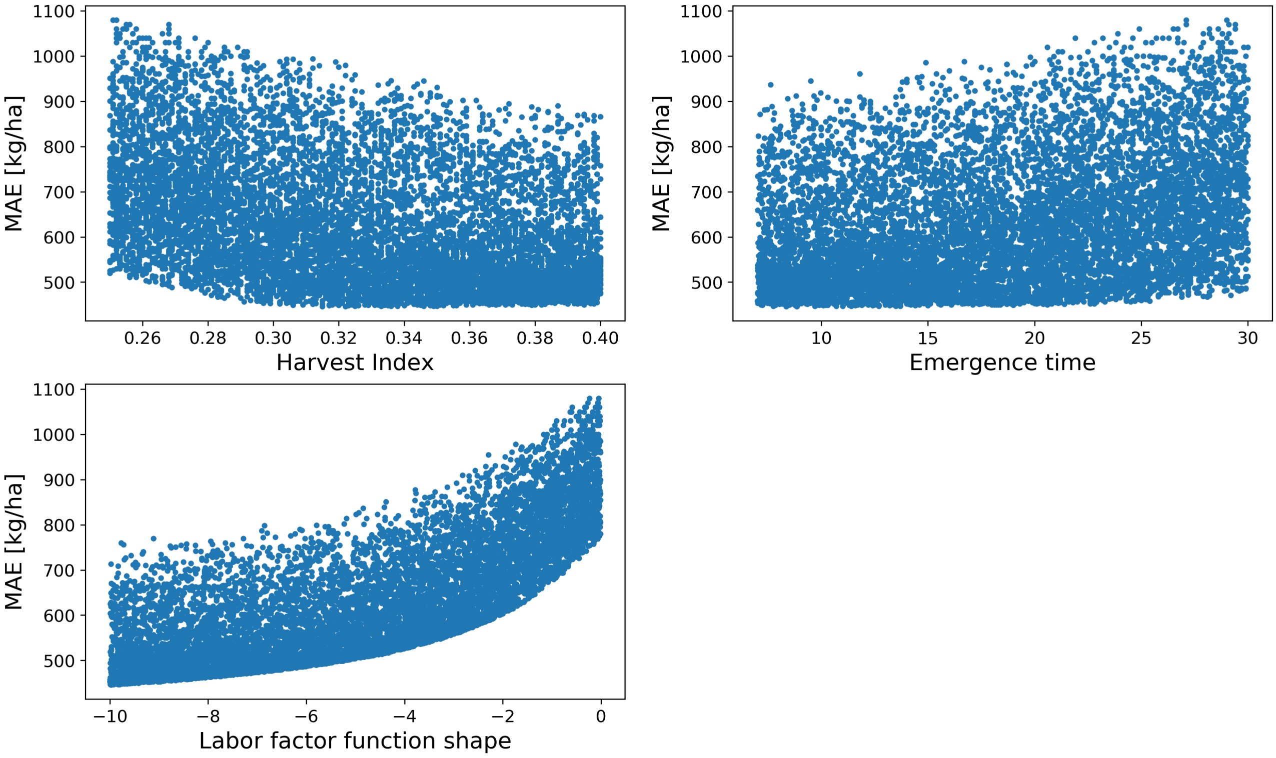


Figure C.3: MAE variation with parameter values

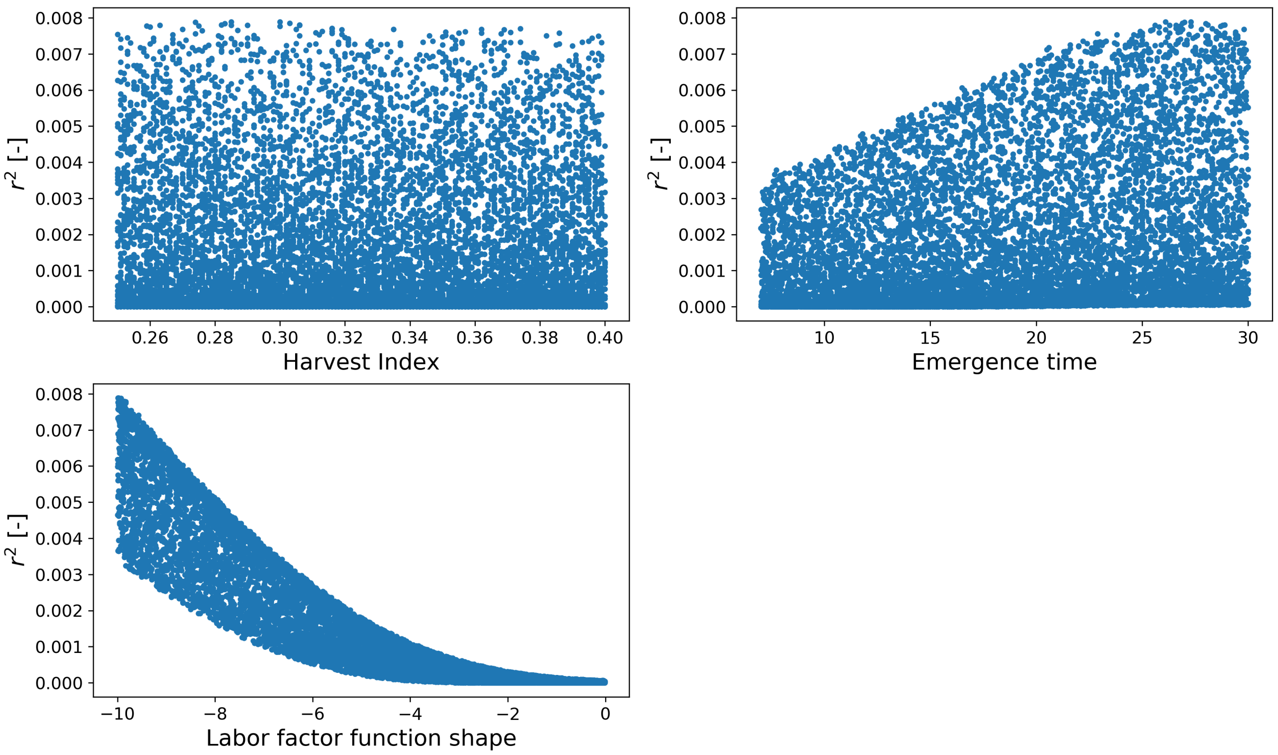


Figure C.4: r^2 variation with parameter values

Appendix D

PCA Additional Materials

This appendix contains additional information regarding the PCA. Table D.1 shows the sample testing, the KMO value of 0.628 indicates that the sample used is suitable for the dimension reduction analysis. The scree plot fig. D.1 and the table D.2 showcases the variance for each component. Lastly, fig. D.2 shows the component plot.

Table D.1: KMO and Bartlett's Test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.628
Bartlett's Test of Sphericity	Approx. Chi-Square	1741.143
	df	91
	Sig.	0.000

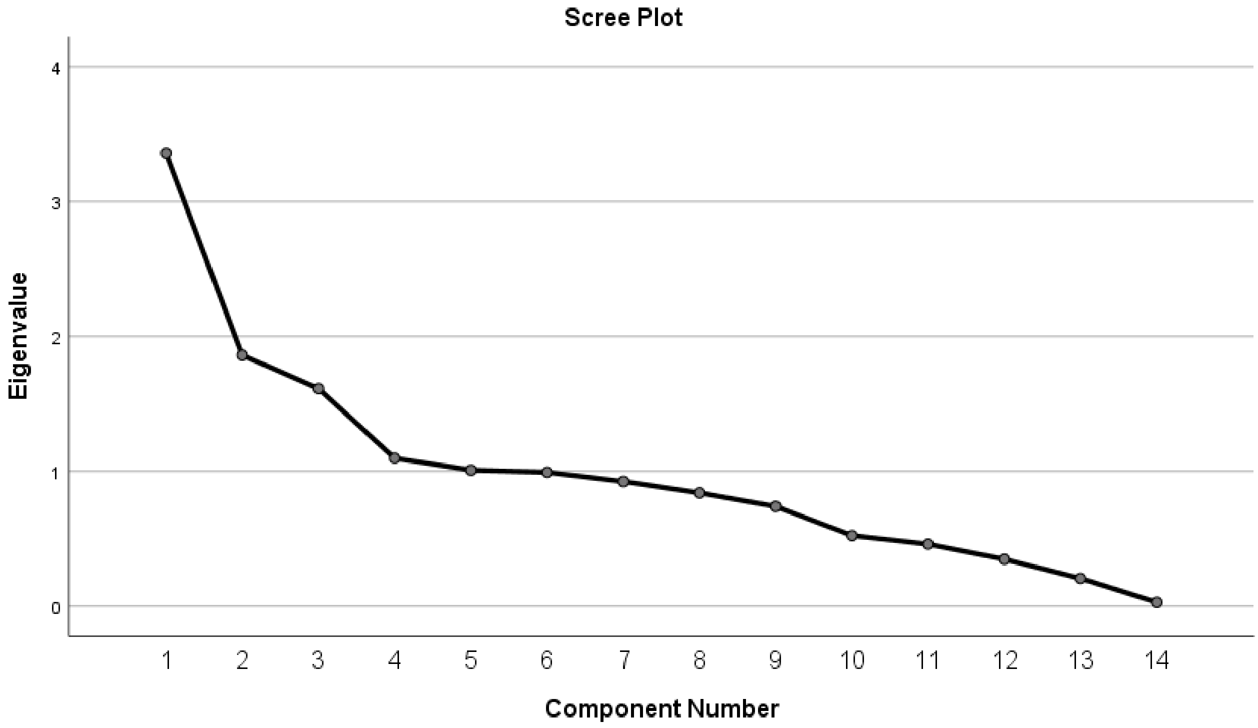


Figure D.1: Scree plot

Table D.2: Total variance explained for all 14 components

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.359	23.995	23.995
2	1.861	13.293	37.287
3	1.613	11.524	48.811
4	1.097	7.834	56.645
5	1.008	7.203	63.848
6	0.992	7.087	70.935
7	0.924	6.602	77.537
8	0.840	6.001	83.538
9	0.740	5.285	88.823
10	0.523	3.739	92.562
11	0.459	3.282	95.843
12	0.348	2.487	98.330
13	0.205	1.467	99.797
14	0.028	0.203	100.000

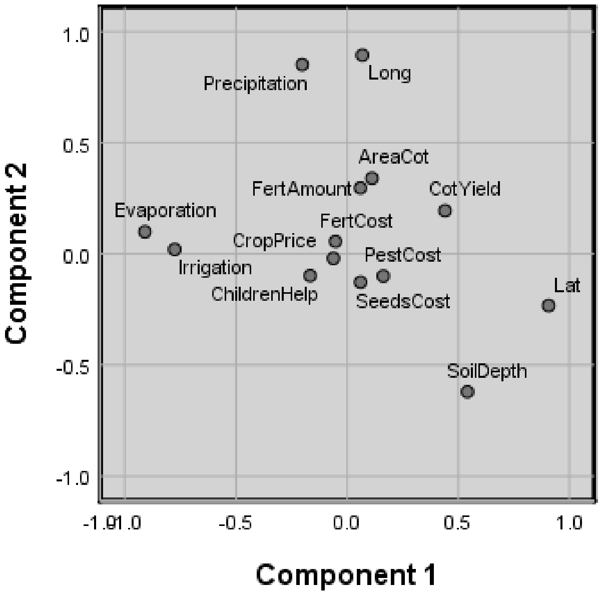


Figure D.2: Component plot

Appendix E

Kernel Functions Performances

This appendix shows the overview of the cross validation result for various kernels used in the Kernel PCA. This include a table of overall performance score with a number of retained components, plots of predicted vs. observed total error, plots of explained variance, and the summary for each regression model.

E.1 Summary of Kernel Performance

Table E.1: Performance comparison of various kernel functions including the number of components

	MAE [kg/ha]		NS [-]		r^2 [-]		Components	
	Train	Test	Train	Test	Train	Test	n total	n retained
Poly, deg: 2	178	228	0.82	0.67	0.820	0.680	147	7
Poly, deg: 3	428	470	0.088	-0.084	0.084	-0.009	229	1
Poly, deg: 4	438	469	0.050	-0.055	0.046	-0.013	230	1
Poly, deg: 5	440	468	0.042	-0.047	0.037	-0.013	229	1
Cosine	185	184	0.80	0.79	0.792	0.792	84	6
RBF	166	183	0.85	0.79	0.837	0.789	230	26
Sigmoid	139	167	0.89	0.82	0.887	0.823	144	7

E.2 Predicted vs Observed Total Error of Test Data

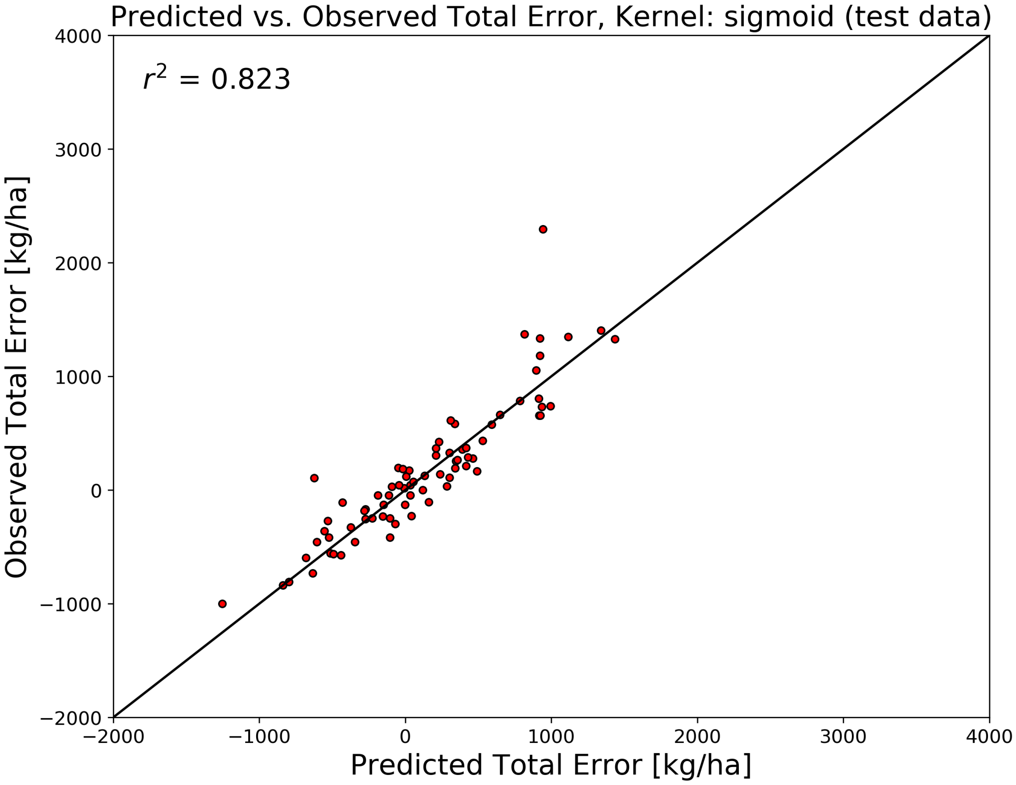


Figure E.1: Predicted vs. observed total error for sigmoid kernel using test data

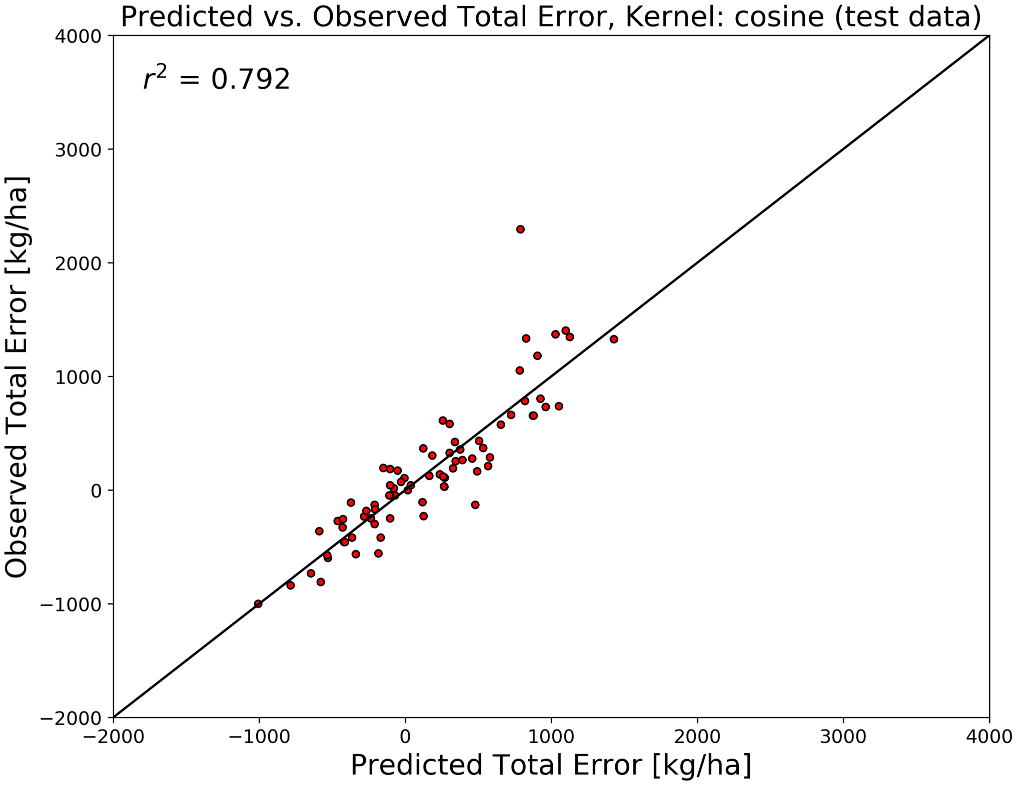


Figure E.2: Predicted vs. observed total error for cosine kernel using test data

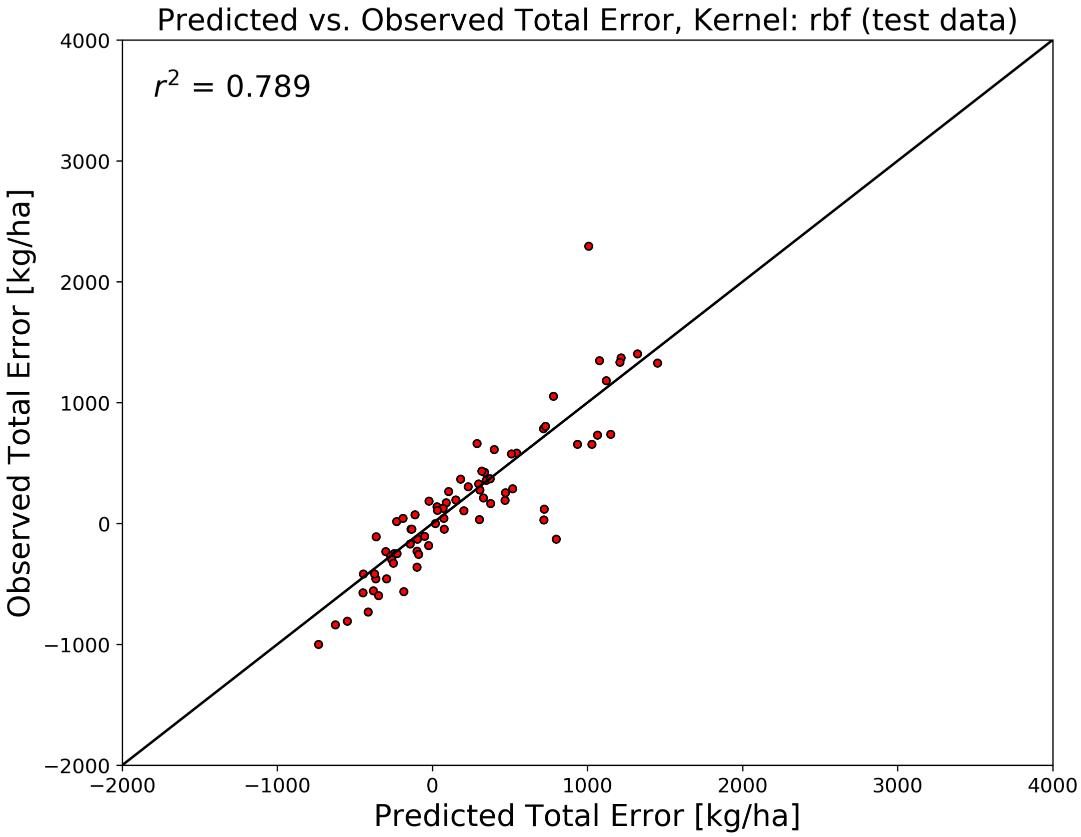


Figure E.3: Predicted vs. observed total error for RBF kernel using test data

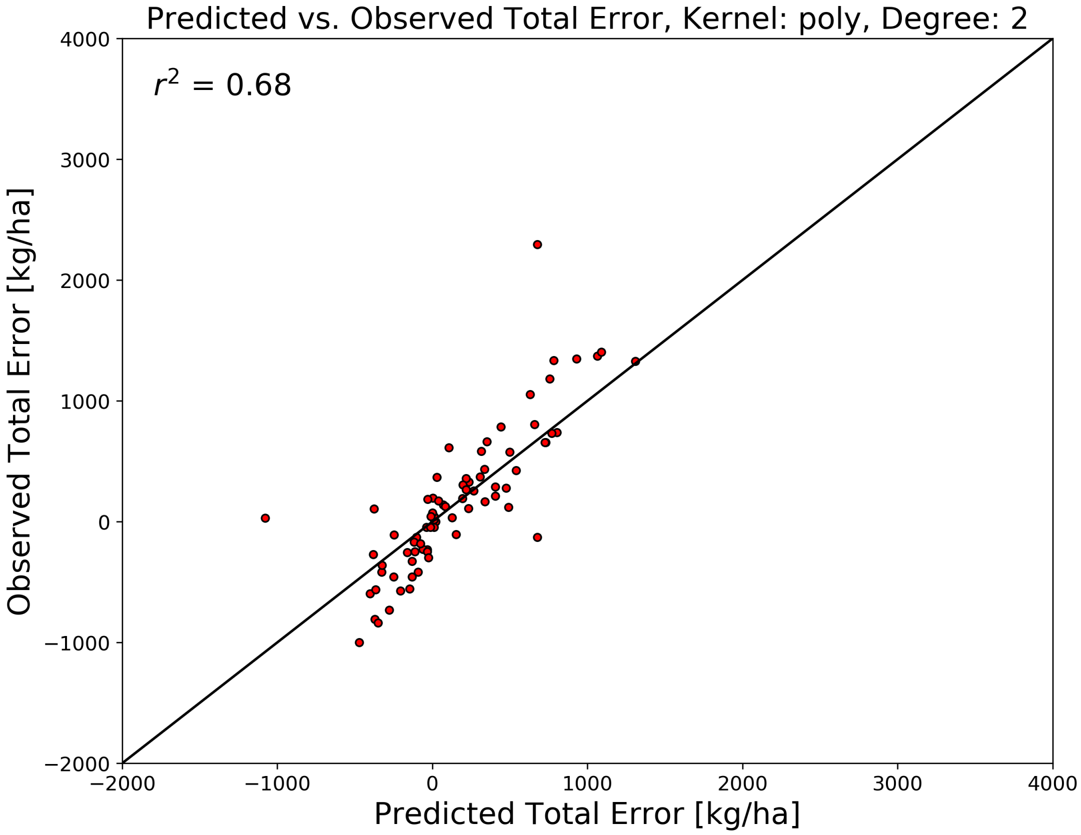


Figure E.4: Predicted vs. observed total error for polynomial degree 2 kernel using test data

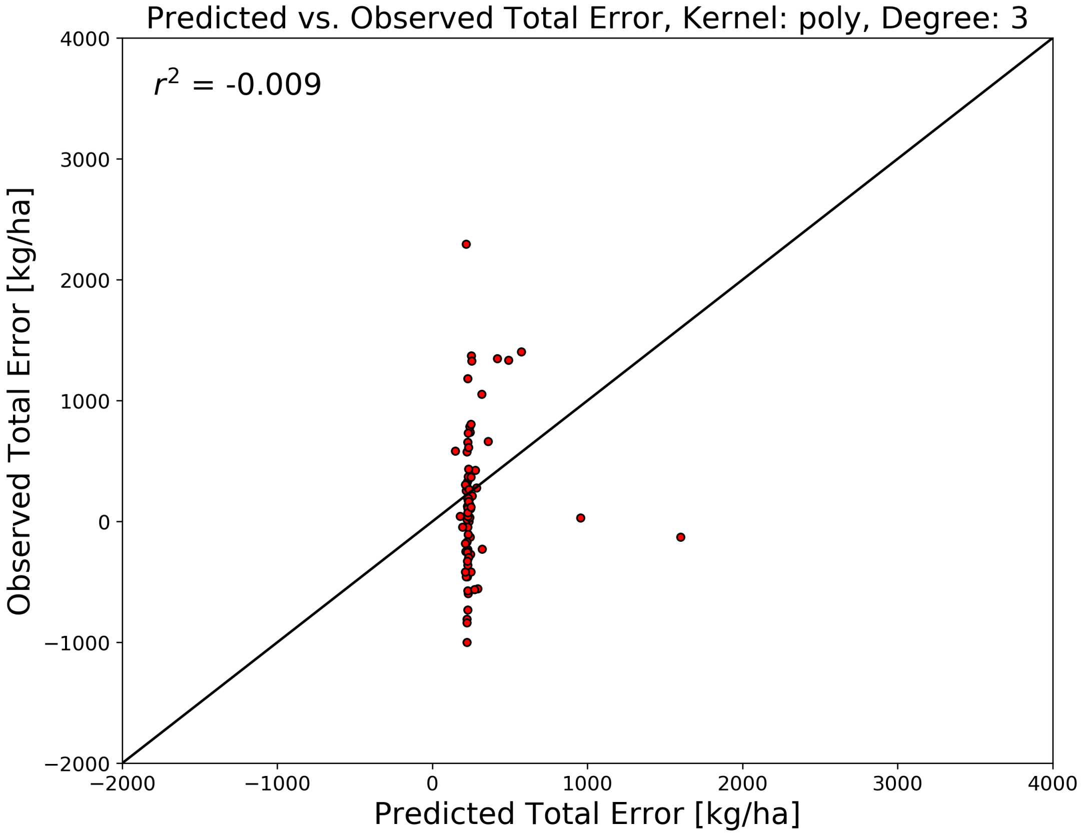


Figure E.5: Predicted vs. observed total error for polynomial degree 3 kernel using test data

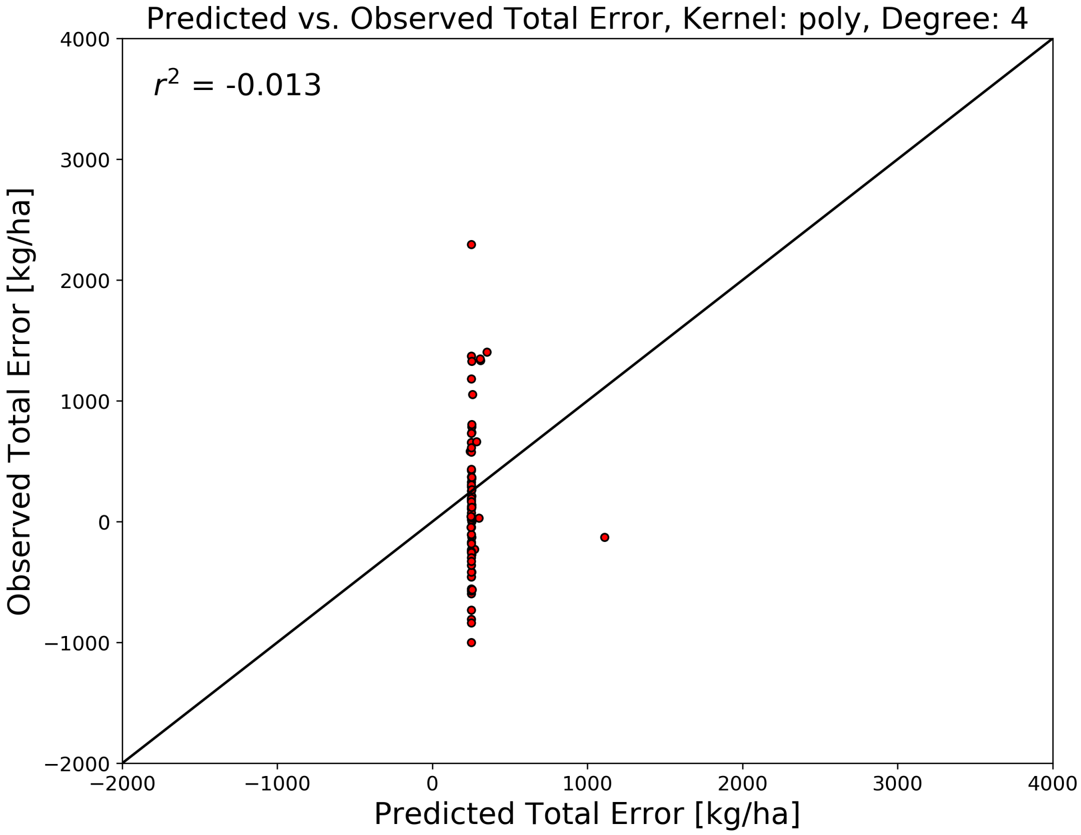


Figure E.6: Predicted vs. observed total error for polynomial degree 4 kernel using test data

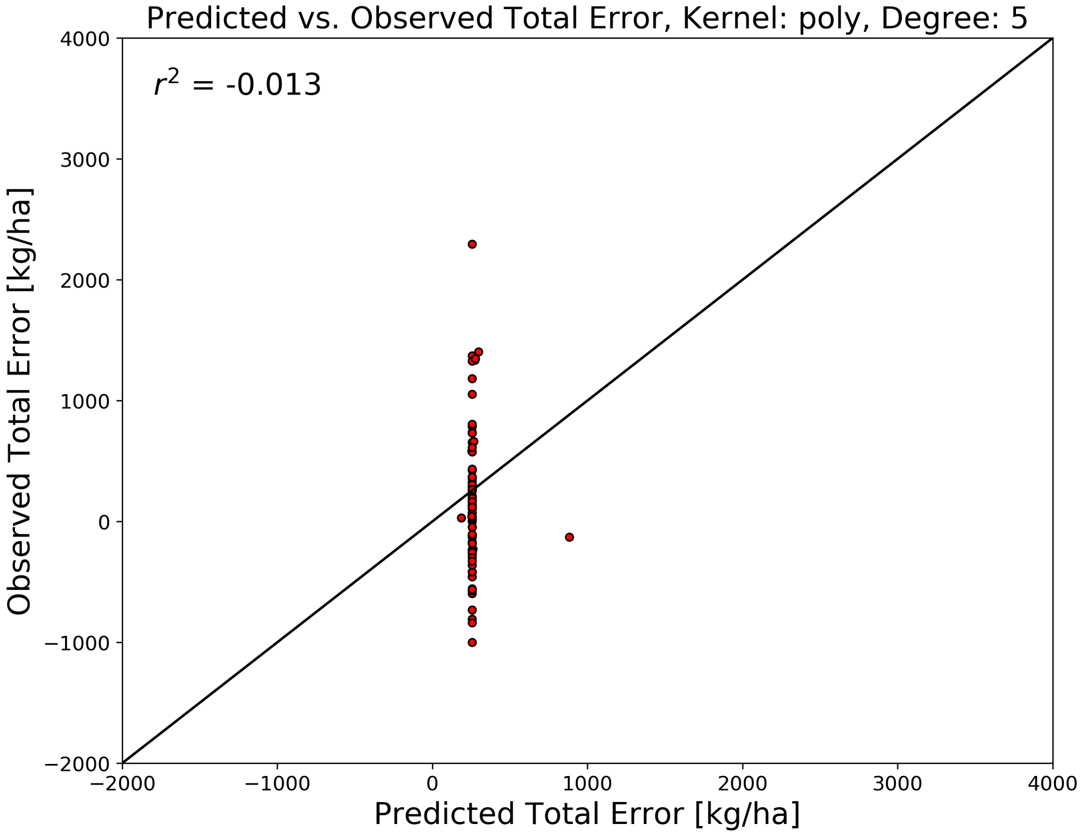


Figure E.7: Predicted vs. observed total error for polynomial degree 5 kernel using test data

E.3 Variance Explained of Components in Kernel Space

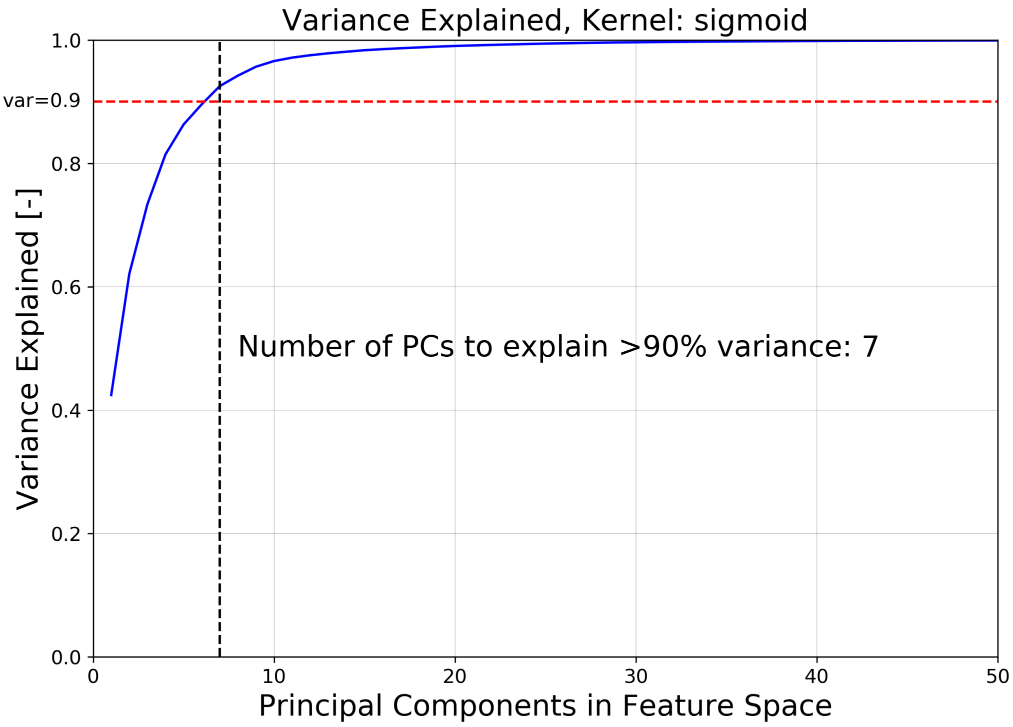


Figure E.8: Variance explained by components in kernel space and the retained components (> 90% variance) for sigmoid kernel

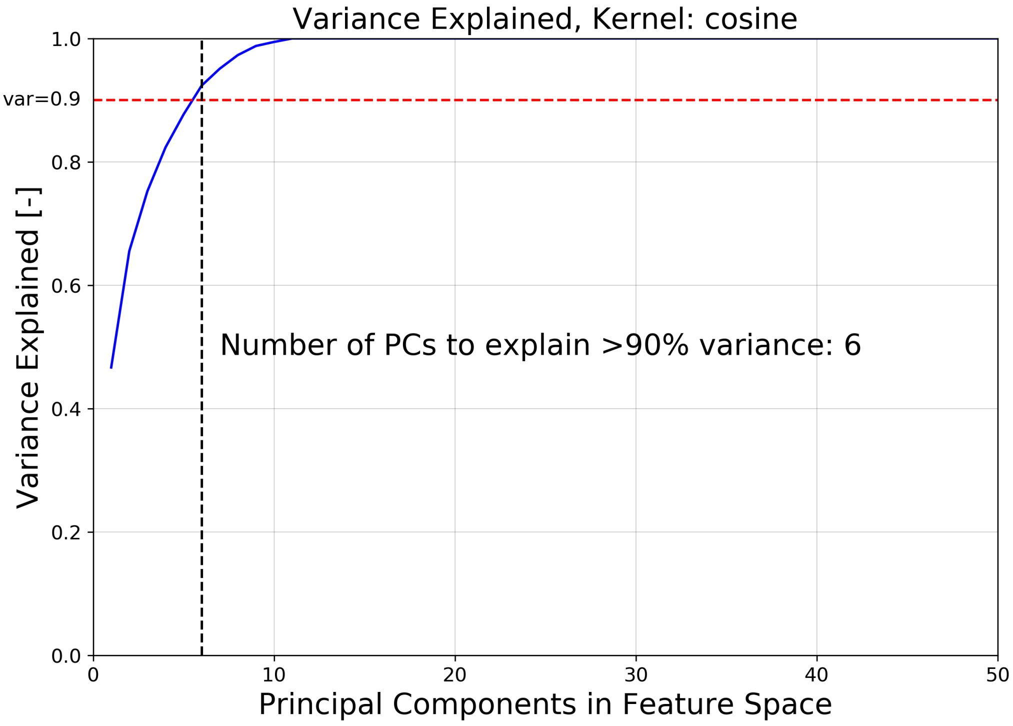


Figure E.9: Variance explained by components in kernel space and the retained components (> 90% variance) for cosine kernel

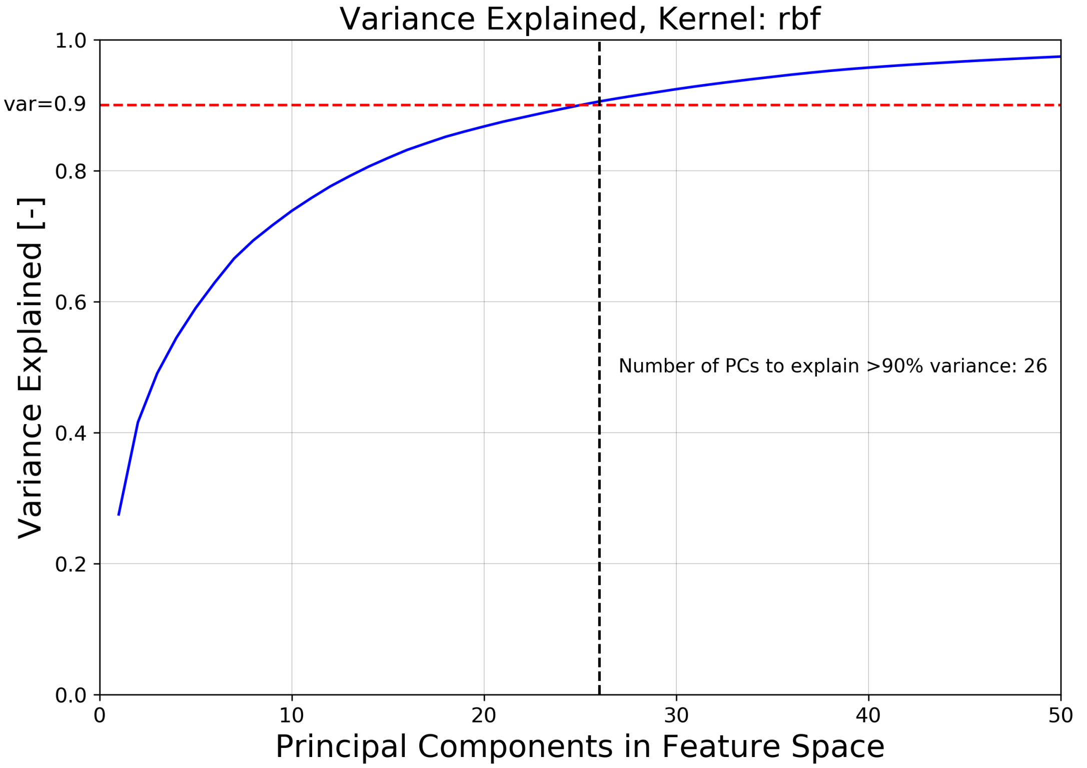


Figure E.10: Variance explained by components in kernel space and the retained components (> 90% variance) for RBF kernel

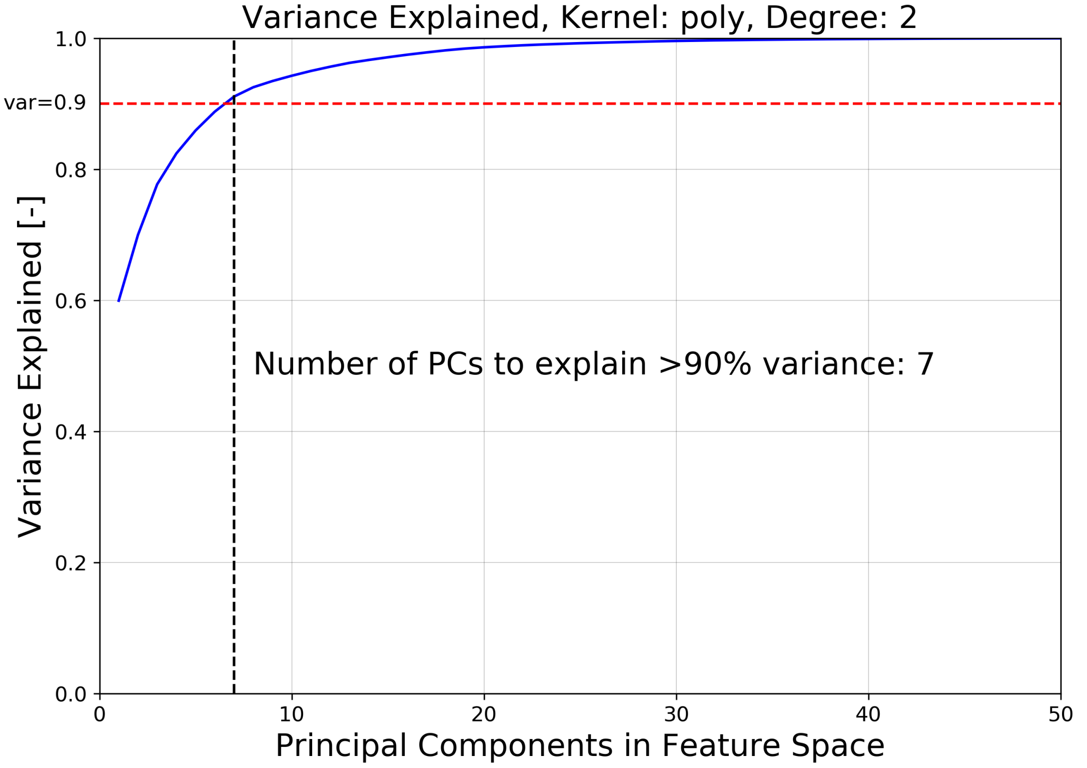


Figure E.11: Variance explained by components in kernel space and the retained components (> 90% variance) for polynomial degree 2 kernel

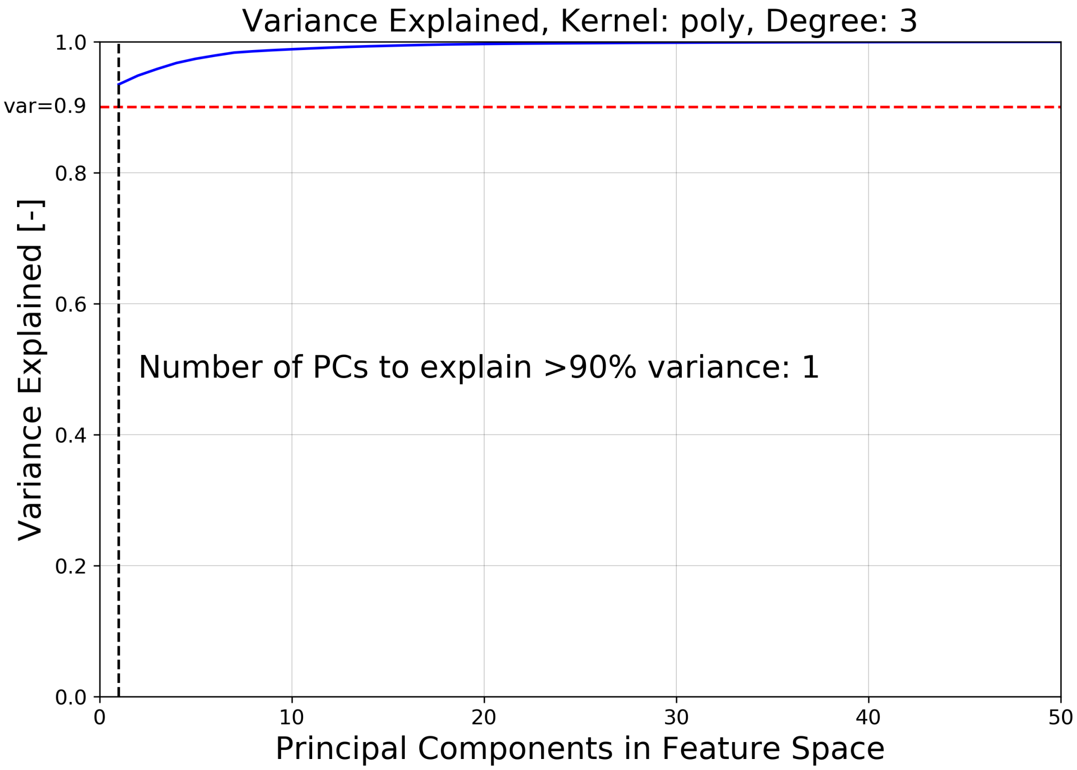


Figure E.12: Variance explained by components in kernel space and the retained components (> 90% variance) for polynomial degree 3 kernel

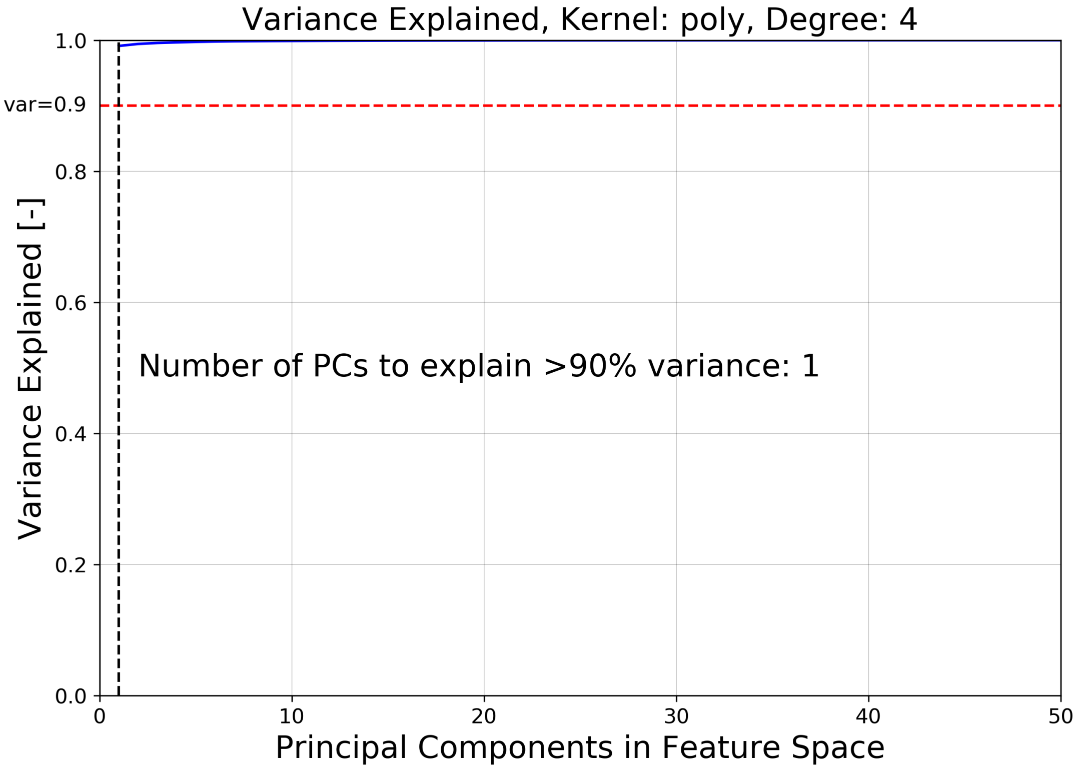


Figure E.13: Variance explained by components in kernel space and the retained components (> 90% variance) for polynomial degree 4 kernel

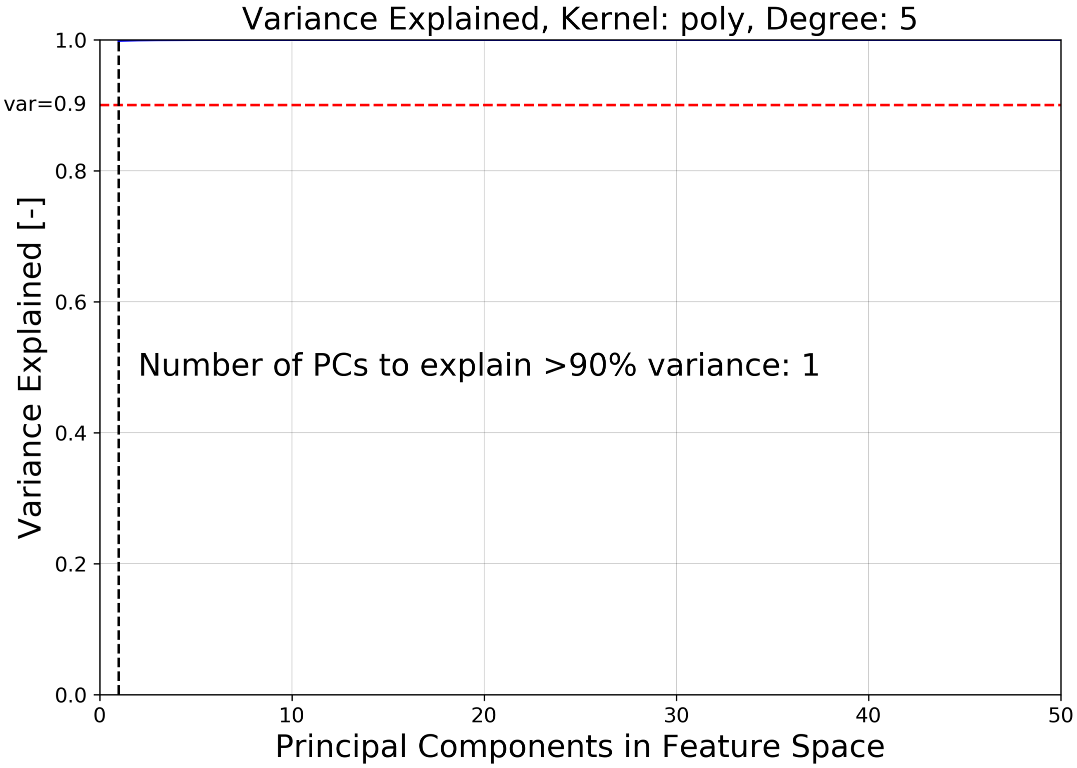
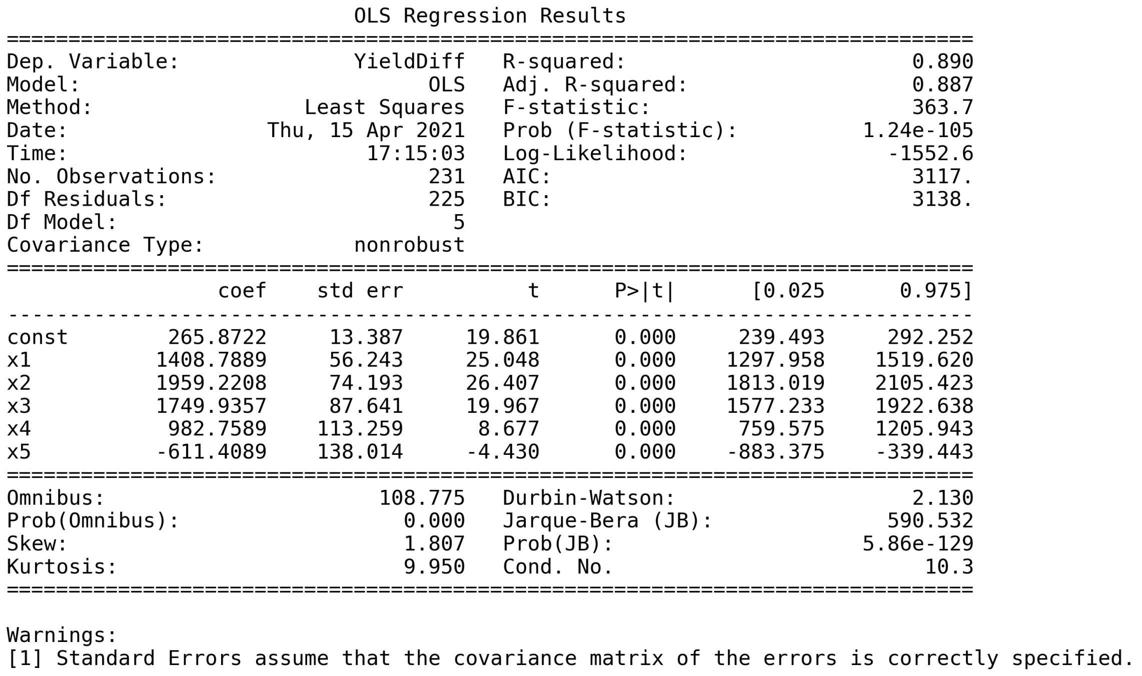


Figure E.14: Variance explained by components in kernel space and the retained components (> 90% variance) for polynomial degree 5 kernel

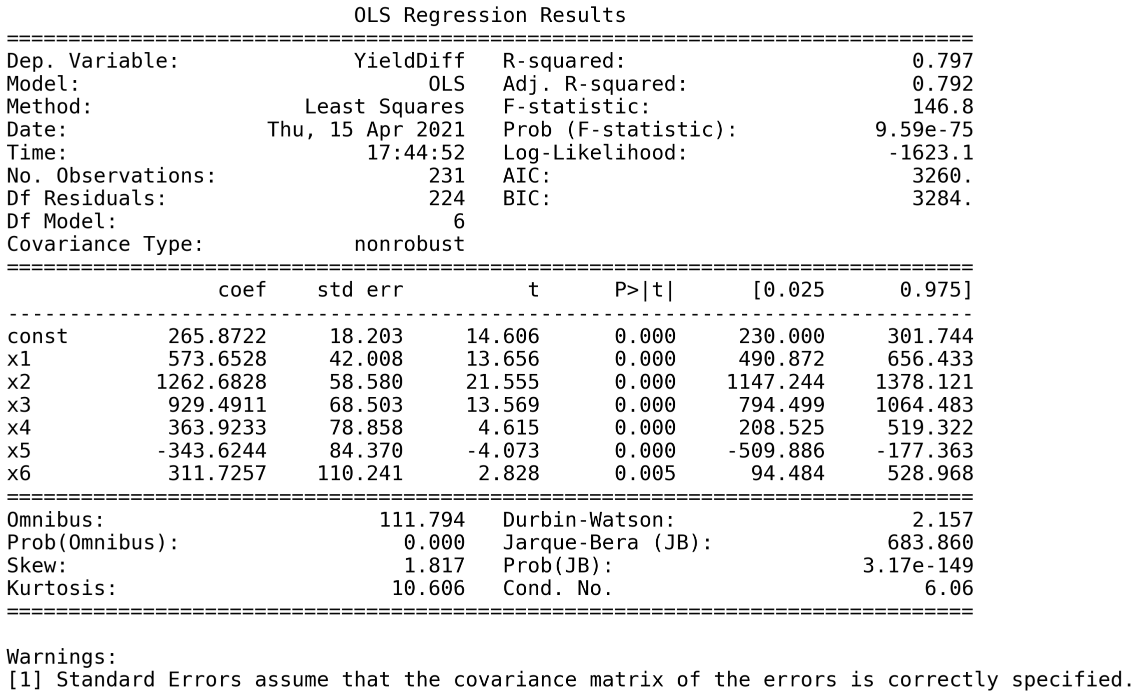
E.4 Summary of Regression Model Using Kernel Space Components



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.890
Model:                            OLS   Adj. R-squared:                  0.887
Method:                 Least Squares   F-statistic:                     363.7
Date:                Thu, 15 Apr 2021   Prob (F-statistic):          1.24e-105
Time:                        17:15:03   Log-Likelihood:                -1552.6
No. Observations:                 231   AIC:                             3117.
Df Residuals:                     225   BIC:                             3138.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     13.387     19.861      0.000     239.493     292.252
x1          1408.7889     56.243     25.048      0.000    1297.958    1519.620
x2          1959.2208     74.193     26.407      0.000    1813.019    2105.423
x3          1749.9357     87.641     19.967      0.000    1577.233    1922.638
x4           982.7589    113.259      8.677      0.000     759.575    1205.943
x5          -611.4089    138.014     -4.430      0.000    -883.375    -339.443
==============================================================================
Omnibus:                      108.775   Durbin-Watson:                   2.130
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              590.532
Skew:                           1.807   Prob(JB):                    5.86e-129
Kurtosis:                       9.950   Cond. No.                         10.3
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

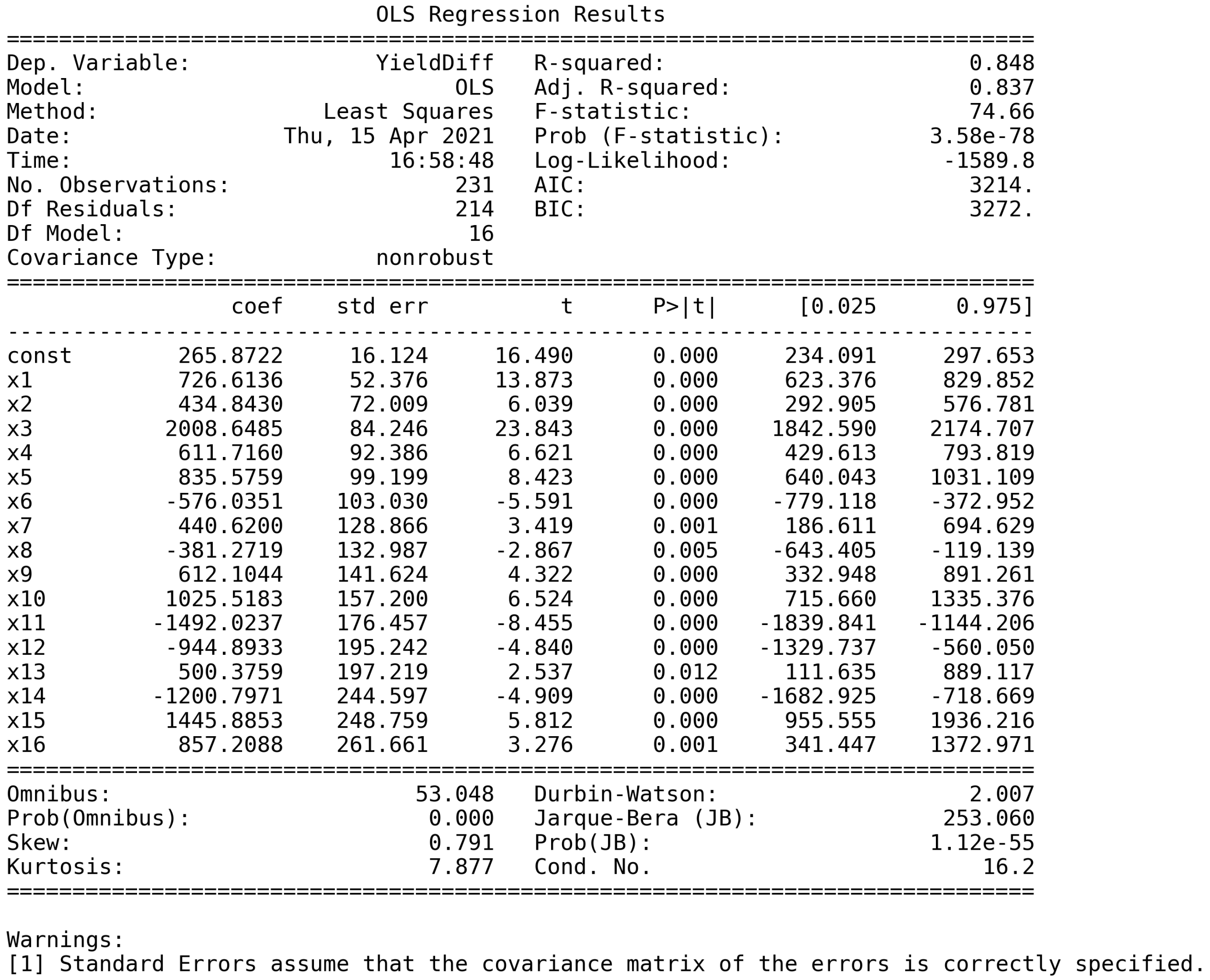
Figure E.15: Summary of regression model for sigmoid kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.797
Model:                            OLS   Adj. R-squared:                  0.792
Method:                 Least Squares   F-statistic:                     146.8
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           9.59e-75
Time:                        17:44:52   Log-Likelihood:                -1623.1
No. Observations:                 231   AIC:                             3260.
Df Residuals:                     224   BIC:                             3284.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     18.203     14.606      0.000     230.000     301.744
x1           573.6528     42.008     13.656      0.000     490.872     656.433
x2          1262.6828     58.580     21.555      0.000    1147.244    1378.121
x3           929.4911     68.503     13.569      0.000     794.499    1064.483
x4           363.9233     78.858      4.615      0.000     208.525     519.322
x5          -343.6244     84.370     -4.073      0.000    -509.886    -177.363
x6           311.7257    110.241      2.828      0.005      94.484     528.968
==============================================================================
Omnibus:                      111.794   Durbin-Watson:                   2.157
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              683.860
Skew:                           1.817   Prob(JB):                    3.17e-149
Kurtosis:                      10.606   Cond. No.                         6.06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

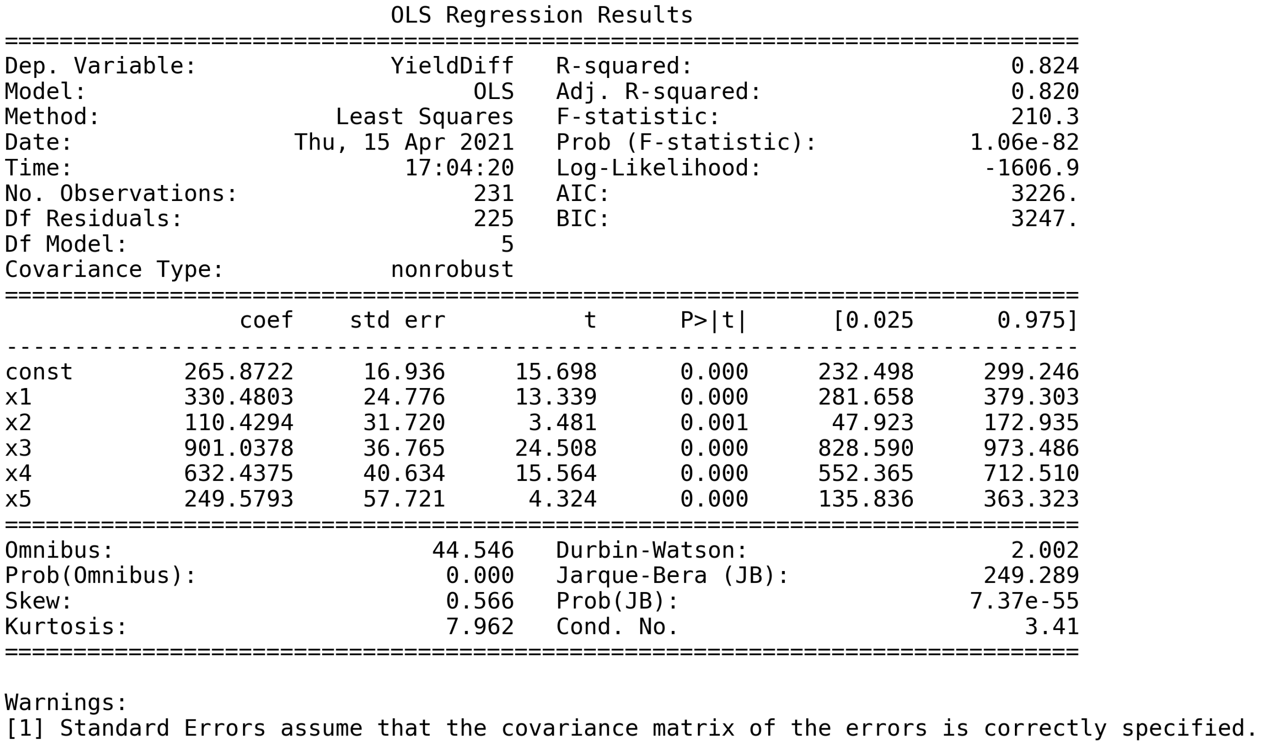
Figure E.16: Summary of regression model for cosine kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.848
Model:                            OLS   Adj. R-squared:                  0.837
Method:                 Least Squares   F-statistic:                     74.66
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           3.58e-78
Time:                        16:58:48   Log-Likelihood:                -1589.8
No. Observations:                 231   AIC:                             3214.
Df Residuals:                     214   BIC:                             3272.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     16.124     16.490      0.000     234.091     297.653
x1           726.6136     52.376     13.873      0.000     623.376     829.852
x2           434.8430     72.009      6.039      0.000     292.905     576.781
x3          2008.6485     84.246     23.843      0.000    1842.590    2174.707
x4           611.7160     92.386      6.621      0.000     429.613     793.819
x5           835.5759     99.199      8.423      0.000     640.043    1031.109
x6          -576.0351    103.030     -5.591      0.000    -779.118    -372.952
x7           440.6200    128.866      3.419      0.001     186.611     694.629
x8          -381.2719    132.987     -2.867      0.005    -643.405    -119.139
x9           612.1044    141.624      4.322      0.000     332.948     891.261
x10         1025.5183    157.200      6.524      0.000     715.660    1335.376
x11        -1492.0237    176.457     -8.455      0.000   -1839.841   -1144.206
x12         -944.8933    195.242     -4.840      0.000   -1329.737    -560.050
x13          500.3759    197.219      2.537      0.012     111.635     889.117
x14        -1200.7971    244.597     -4.909      0.000   -1682.925    -718.669
x15         1445.8853    248.759      5.812      0.000     955.555    1936.216
x16          857.2088    261.661      3.276      0.001     341.447    1372.971
==============================================================================
Omnibus:                       53.048   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              253.060
Skew:                           0.791   Prob(JB):                     1.12e-55
Kurtosis:                       7.877   Cond. No.                         16.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

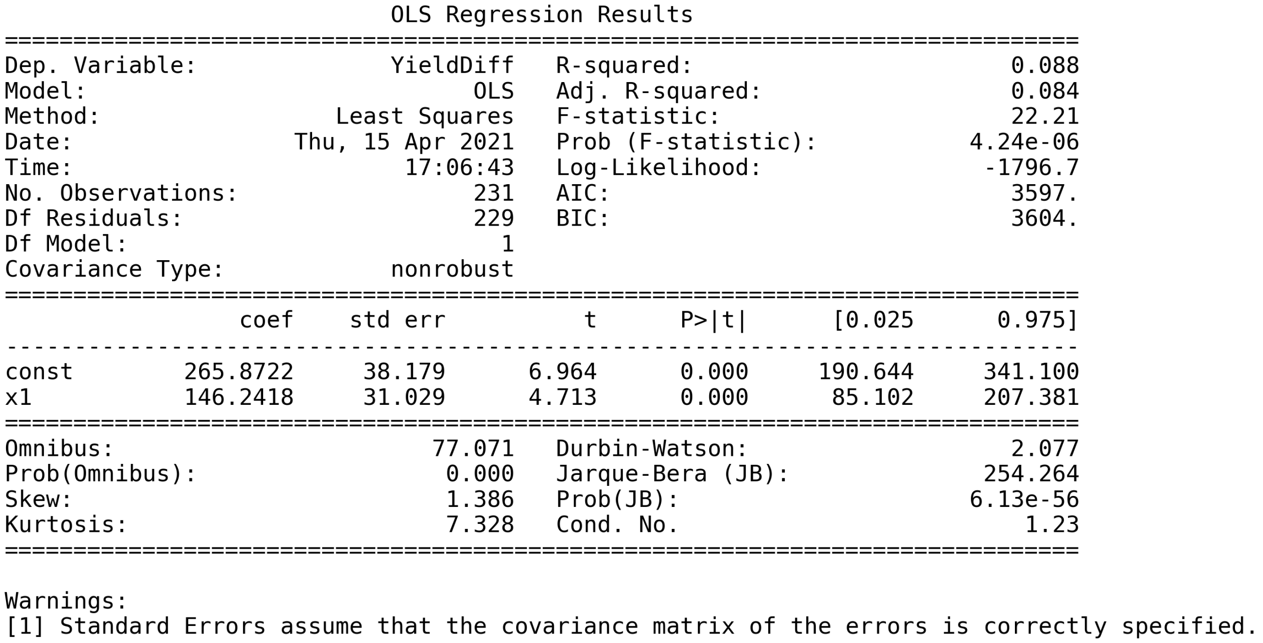
Figure E.17: Summary of regression model for RBF kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.824
Model:                            OLS   Adj. R-squared:                  0.820
Method:                 Least Squares   F-statistic:                     210.3
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           1.06e-82
Time:                        17:04:20   Log-Likelihood:                -1606.9
No. Observations:                 231   AIC:                             3226.
Df Residuals:                     225   BIC:                             3247.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     16.936     15.698      0.000     232.498     299.246
x1           330.4803     24.776     13.339      0.000     281.658     379.303
x2           110.4294     31.720      3.481      0.001      47.923     172.935
x3           901.0378     36.765     24.508      0.000     828.590     973.486
x4           632.4375     40.634     15.564      0.000     552.365     712.510
x5           249.5793     57.721      4.324      0.000     135.836     363.323
==============================================================================
Omnibus:                       44.546   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              249.289
Skew:                           0.566   Prob(JB):                     7.37e-55
Kurtosis:                       7.962   Cond. No.                         3.41
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

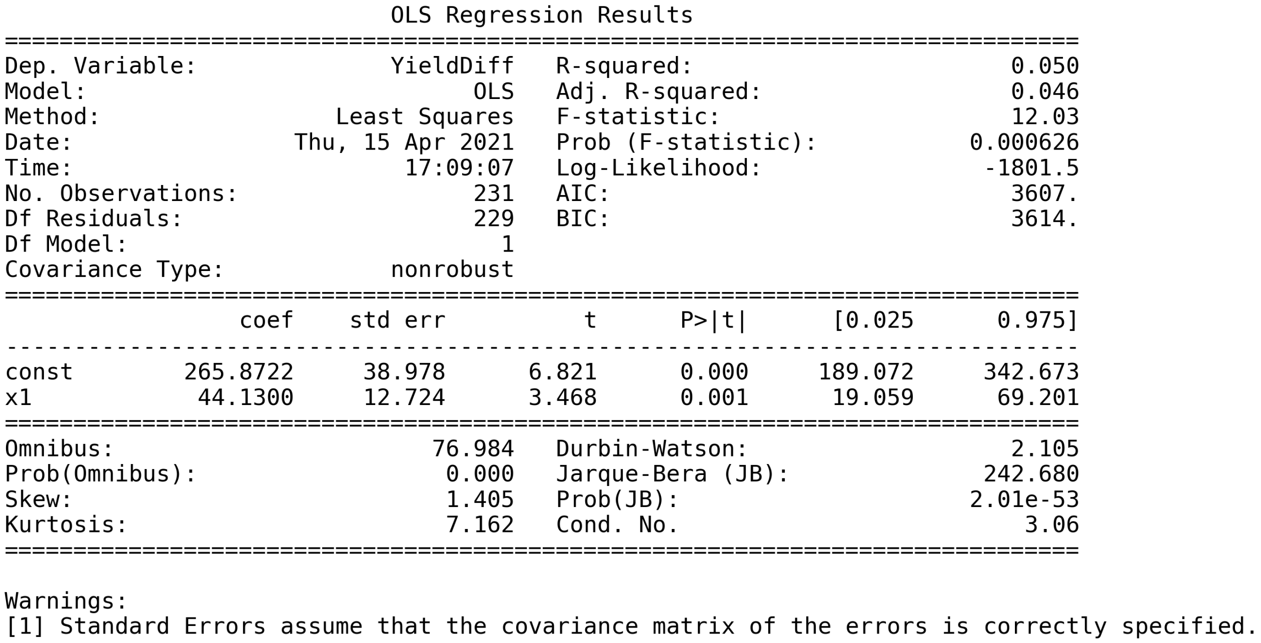
Figure E.18: Summary of regression model for polynomial degree 2 kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.088
Model:                            OLS   Adj. R-squared:                  0.084
Method:                 Least Squares   F-statistic:                     22.21
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           4.24e-06
Time:                        17:06:43   Log-Likelihood:                -1796.7
No. Observations:                 231   AIC:                             3597.
Df Residuals:                     229   BIC:                             3604.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     38.179      6.964      0.000     190.644     341.100
x1           146.2418     31.029      4.713      0.000      85.102     207.381
==============================================================================
Omnibus:                       77.071   Durbin-Watson:                   2.077
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              254.264
Skew:                           1.386   Prob(JB):                     6.13e-56
Kurtosis:                       7.328   Cond. No.                         1.23
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

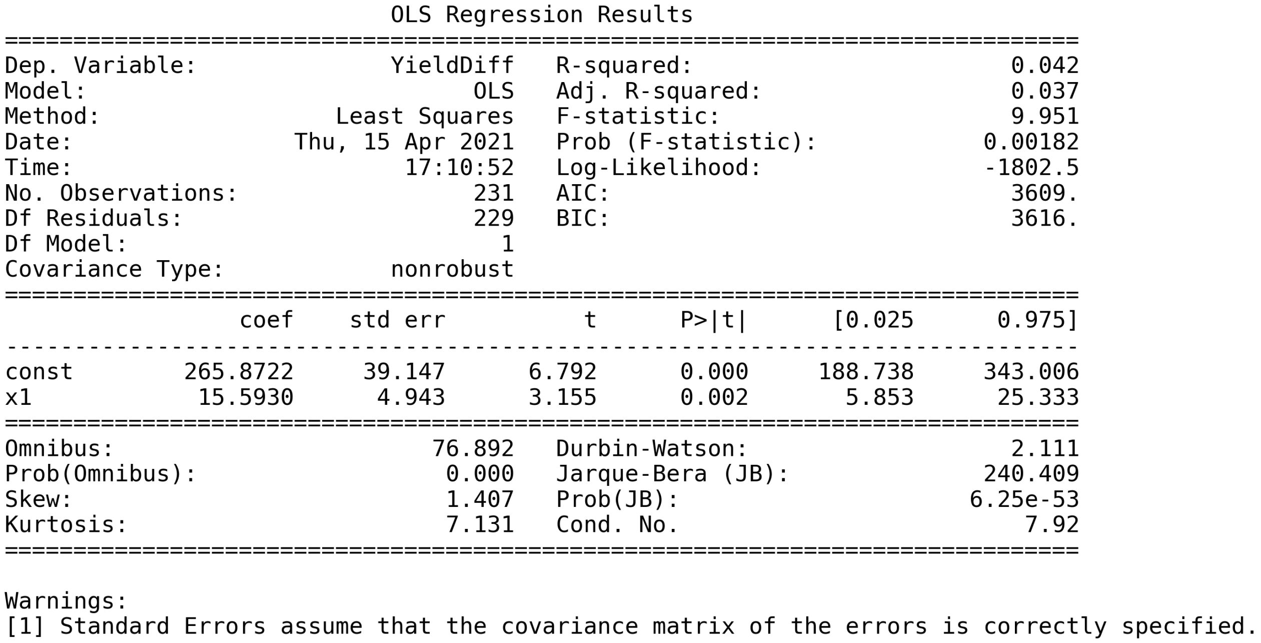
Figure E.19: Summary of regression model for polynomial degree 3 kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                  0.046
Method:                 Least Squares   F-statistic:                     12.03
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           0.000626
Time:                        17:09:07   Log-Likelihood:                -1801.5
No. Observations:                 231   AIC:                             3607.
Df Residuals:                     229   BIC:                             3614.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     38.978      6.821      0.000     189.072     342.673
x1            44.1300     12.724      3.468      0.001      19.059      69.201
==============================================================================
Omnibus:                       76.984   Durbin-Watson:                   2.105
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              242.680
Skew:                           1.405   Prob(JB):                     2.01e-53
Kurtosis:                       7.162   Cond. No.                         3.06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure E.20: Summary of regression model for polynomial degree 4 kernel.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              YieldDiff   R-squared:                       0.042
Model:                            OLS   Adj. R-squared:                  0.037
Method:                 Least Squares   F-statistic:                     9.951
Date:                Thu, 15 Apr 2021   Prob (F-statistic):            0.00182
Time:                        17:10:52   Log-Likelihood:                -1802.5
No. Observations:                 231   AIC:                             3609.
Df Residuals:                     229   BIC:                             3616.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        265.8722     39.147      6.792      0.000     188.738     343.006
x1            15.5930      4.943      3.155      0.002       5.853      25.333
==============================================================================
Omnibus:                       76.892   Durbin-Watson:                   2.111
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              240.409
Skew:                           1.407   Prob(JB):                     6.25e-53
Kurtosis:                       7.131   Cond. No.                         7.92
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure E.21: Summary of regression model for polynomial degree 5 kernel.

E.5 Python Code Used in Kernel Testing

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 12})
from sklearn.decomposition import PCA, KernelPCA, FactorAnalysis
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
import scipy.stats as stats

#%%get dat
dat=pd.read_excel("KPCA_dat_adj.xlsx", header=0)

X = StandardScaler().fit_transform(dat.iloc[:,0:-3])
y = dat.iloc[:,-3]

#split training and test data
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.25, random_state=1)

# Running KPCA
kernelType="sigmoid"
deg=5
kpca = KernelPCA(kernel=kernelType, fit_inverse_transform=True,n_components=None, degree=deg)
X_train=kpca.fit_transform(X_train)
X_test=kpca.transform(X_test)

#%%
V=kpca.alphas_
D=kpca.lambdas_
D_cumsum=np.cumsum(D)
var_explained=D_cumsum/np.sum(D)
nCount=sum(var_explained <= 0.90)+1

Xplot=np.arange(1,len(var_explained)+1)
Yplot=var_explained
print(len(D))

plt.figure(figsize=(10,7))
plt.plot(Xplot, Yplot, c="blue")
plt.axhline(y=0.90, color='r', linestyle='--')
plt.axvline(x=nCount, color='k', linestyle='--')
plt.xlim(0,50)
plt.ylim(0,1)
plt.grid(color='k', linestyle='-', linewidth=0.1)
plt.text(nCount+1, 0.5,'Number of PCs to explain >90% variance: {}'.format(nCount), ha='left', va='center',fontsize=18)
plt.text(-5, 0.9,'var=0.9', ha='left', va='center')
plt.title("Variance Explained, Kernel: {}" .format(kernelType),fontsize=18)
# plt.title("Variance Explained, Kernel: {}, Degree: {}" .format(kernelType,deg),fontsize=18)
plt.xlabel("Principal Components in Feature Space",fontsize=18)
plt.ylabel("Variance Explained [-]",fontsize=18)
# plt.savefig('Plots/varExplained_{}_deg{}.png'.format(kernelType,deg),dpi=300, bbox_inches = "tight")
plt.savefig('Plots/varExplained_{}.png'.format(kernelType),dpi=300, bbox_inches = "tight")

#%% fitting model with significant pval
X_model=X_train[:,1:nCount+1]
X2 = sm.add_constant(X_model)
est = sm.OLS(y_train, X2)
est2 = est.fit()
print(est2.summary())

#%% Get rid of the model with low significance
X_model=X_train[:,(1,2,3,4,5)]
X2 = sm.add_constant(X_model)
est = sm.OLS(y_train, X2)
est2 = est.fit()
print(est2.summary())

r2=round(est2.rsquared,3)
adj_r2=round(est2.rsquared_adj,3)
f=est2.fvalue
beta=est2.params
pval=est2.pvalues
y_pred_train=est2.predict(X2)

plt.rc('figure', figsize=(8, 5))
plt.text(0.01, 0.05, str(est2.summary()), {'fontsize': 10}, fontproperties = 'monospace') # approach improved by OP -> monospace!
plt.axis('off')
plt.tight_layout()
# plt.savefig('Plots/KPCA_Summary_{}_deg{}.png'.format(kernelType,deg),dpi=300, bbox_inches = "tight")
plt.savefig('Plots/KPCA_Summary_{}.png'.format(kernelType),dpi=300, bbox_inches = "tight")

MAE=np.sum(abs(y_pred_train-y_train))/len(y_pred_train)
NS=1-(np.sum((y_pred_train-y_train)**2)/np.sum((y_train.mean()-y_train)**2))
NS_log=1-(np.sum((np.log10(y_pred_train)-np.log10(y_train))**2)/np.sum((np.log10(y_train.mean())-np.log10(y_train))**2))

print('MAE of {} Kernel PCA Train: {} [kg/ha]' .format(kernelType,MAE))
print('NS of {} Kernel PCA Train: {} [kg/ha]' .format(kernelType,NS))
print('NS log of {} Kernel PCA Train: {} [kg/ha]' .format(kernelType,NS_log))

#predict
X2_test=sm.add_constant(X_test[:,(1,2,3,4,5)])
y_pred=est2.predict(X2_test)

MAE=np.sum(abs(y_pred-y_test))/len(y_pred)
NS=1-(np.sum((y_pred-y_test)**2)/np.sum((y_test.mean()-y_test)**2))
```

```
NS_log=1-(np.sum((np.log10(y_pred)-np.log10(y_test))**2)/np.sum((np.log10(y_test.mean())-np.log10(y_test))**2))

print(’MAE of {} Kernel PCA Test: {} [kg/ha]’ .format(kernelType,MAE))
print(’NS of {} Kernel PCA Test: {} [kg/ha]’ .format(kernelType,NS))
print(’NS log of {} Kernel PCA Test: {} [kg/ha]’ .format(kernelType,NS_log))

#%% plots test data KPCA
plt.figure(figsize=(10,8))
Xplot=y_pred
Yplot=y_test

plt.scatter(Xplot, Yplot, c="red",s=20, edgecolor=’k’)
Xplot2 = sm.add_constant(Xplot)
estPlot = sm.OLS(Yplot, Xplot2)
estPlot2=estPlot.fit()
adj_r2Plot=round(estPlot2.rsquared_adj,3)
plt.text(-1800, 3600,’$r^2$ = {}’.format(adj_r2Plot), ha=’left’, va=’center’,fontsize=18)
plt.plot(np.arange(-2000,4000),np.arange(-2000,4000),color="black")
plt.xlim(-2000,4000)
plt.ylim(-2000,4000)
# plt.title("Predicted vs. Observed Total Error, Kernel: {}, Degree: {}".format(kernelType,deg),fontsize=18)
plt.title("Predicted vs. Observed Total Error, Kernel: {} (test data)" .format(kernelType),fontsize=18)
plt.xlabel("Predicted Total Error [kg/ha]",fontsize=18)
plt.ylabel("Observed Total Error [kg/ha]",fontsize=18)
# plt.savefig(’Plots/KPCA_{}_deg{}.png’.format(kernelType,deg),dpi=300, bbox_inches = "tight")
plt.savefig(’Plots/KPCA_{}.png’.format(kernelType),dpi=300, bbox_inches = "tight")

#%% ################################# KPCA ALL DATA ############################################
#transform all data and estimate them
X_all=kpca.transform(X)
X2_all=sm.add_constant(X_all[:,(1,2,3,4,5)]) #components cumulative to >90% variance and significant in regression model
y_pred=est2.predict(X2_all)

# plots
plt.figure(figsize=(10,8))
Xplot=y_pred
Yplot=y

plt.scatter(Xplot, Yplot, c="red",s=20, edgecolor=’k’)
Xplot2 = sm.add_constant(Xplot)
estPlot = sm.OLS(Yplot, Xplot2)
estPlot2=estPlot.fit()
betaPlot=estPlot2.params
adj_r2Plot=round(estPlot2.rsquared_adj,3)
plt.text(-1800, 3600,’$r^2$ = {}’.format(adj_r2Plot), ha=’left’, va=’center’,fontsize=18)
plt.plot(np.arange(-2000,4000),np.arange(-2000,4000),color="black")
plt.plot(np.unique(Xplot), np.unique(estPlot2.predict(sm.add_constant(Xplot))), color="black",linestyle=’--’)
plt.xlim(-2000,4000)
plt.ylim(-2000,4000)
plt.title("Predicted vs. Observed Total Error, Kernel: {}"  .format(kernelType),fontsize=18)
plt.xlabel("Predicted Total Error [kg/ha]",fontsize=18)
plt.ylabel("Observed Total Error [kg/ha]",fontsize=18)
plt.savefig(’Plots/KPCA_{}_ALLDATA.png’.format(kernelType),dpi=300, bbox_inches = "tight")

MAE=np.sum(abs(y_pred-y))/len(y_pred)
NS=1-(np.sum((y_pred-y)**2)/np.sum((y.mean()-y)**2))
NS_log=1-(np.sum((np.log10(y_pred)-np.log10(y))**2)/np.sum((np.log10(y.mean())-np.log10(y))**2))

print(’MAE of {} Kernel PCA Test: {} [kg/ha]’ .format(kernelType,MAE))
print(’NS of {} Kernel PCA Test: {} [-]’ .format(kernelType,NS))
print(’NS log of {} Kernel PCA Test: {} [-]’ .format(kernelType,NS_log))

#%% plot adjusted yield

Xplot=dat.iloc[:,-2]+y_pred
Yplot=dat.iloc[:,-1]

plt.figure(figsize=(10,8))
plt.fill_between(np.unique(Xplot),np.unique(estPlot2.predict(sm.add_constant(Xplot)))+450, np.unique(estPlot2.predict(sm.
      add_constant(Xplot)))-450, color=’yellow’, alpha=’0.5’)
plt.scatter(Xplot, Yplot, c="red",s=20, edgecolor=’k’)
Xplot2 = sm.add_constant(Xplot)
estPlot = sm.OLS(Yplot, Xplot2)
estPlot2=estPlot.fit()
betaPlot=estPlot2.params
adj_r2Plot=round(estPlot2.rsquared_adj,3)
plt.text(300, 4500,’$r^2$ = {}’.format(adj_r2Plot), ha=’left’, va=’center’,fontsize=18)
# plt.errorbar(Xplot, Yplot,  0, 450, fmt=’r^’,label=’Annual yield per farmer’, elinewidth =0.5,
#               marker=’x’, markersize=’5’,markeredgecolor=’blue’, ecolor=[’red’],barsabove=False)
plt.plot(np.arange(0,5000),np.arange(0,5000),color="black")
plt.plot(np.unique(Xplot), np.unique(estPlot2.predict(sm.add_constant(Xplot))), color="black",linestyle=’--’)
plt.xlim(0,5000)
plt.ylim(0,5000)
plt.title("Predicted Yield vs. Observed Yield",fontsize=18)
plt.xlabel("Predicted Yield [kg/ha]",fontsize=18)
plt.ylabel("Observed Yield [kg/ha]",fontsize=18)
plt.savefig(’Plots/AdjustedYieldUncer.png’,dpi=300, bbox_inches = "tight")

MAE=np.sum(abs(Xplot-Yplot))/len(Xplot)
NS=1-(np.sum((Xplot-Yplot)**2)/np.sum((Yplot.mean()-Yplot)**2))
NS_log=1-(np.sum((np.log10(Xplot)-np.log10(Yplot))**2)/np.sum((np.log10(Yplot.mean())-np.log10(Yplot))**2))

print(’MAE of {} Predicted Yield: {} [kg/ha]’ .format(kernelType,MAE))
print(’NS of {} Predicted Yield: {} [-]’ .format(kernelType,NS))
print(’NS log of {} Predicted Yield: {} [-]’ .format(kernelType,NS_log))

#%% histogram of er
a=dat.iloc[:,-1]-(dat.iloc[:,-2]+y_pred)

```

```
mean = 0
std = 150
# mean,std=norm.fit(a)
x = np.linspace(mean - 3*std, mean + 3*std, 100)

bins = np.linspace(-1000, 1000, 100)
labels=[’$\epsilon_r$’]

plt.figure(figsize=(12,7))
plt.hist(a, bins,rwidth=0.8,label=labels,normed=True)
plt.plot(x, stats.norm.pdf(x, mean, std), label=’Assumed distribution of $\epsilon_r$’)
plt.title(’Histogram of residual error ($\epsilon_r$)’,fontsize=18)
plt.xlabel(’Residual error ($\epsilon_r$) [kg/ha]’,fontsize=18)
plt.ylabel(’Probability’,fontsize=18)
plt.legend()
plt.savefig(’Plots/HistEr.png’,dpi=300, bbox_inches = "tight")
```

Listing E.1: Python script used to perform the evaluation

Appendix F

GEE Script

```
var geometry =
    ee.Geometry.Polygon(
        [[[77.5, 19.5],
          [79, 19.5],
          [79, 21],
          [77.5, 21]]]);

//////////////////////////////////////////////////////////////////////////
// Sentinel 2
//////////////////////////////////////////////////////////////////////////

var s2 = ee.ImageCollection("COPERNICUS/S2")
          .filterDate(’2018-01-01’,’2019-01-01’)
          .filterBounds(geometry);
var images = s2.select([’B2’,’B3’,’B4’,’B8’], [’B’,’G’,’R’,’NIR’]);
var rgb_viz = {min:0, max:2000, bands:[’R’,’G’,’B’]};

function addNDVI(img) {
  var ndvi = img.normalizedDifference([’NIR’, ’R’]).rename(’NDVI’);
  return img.addBands(ndvi);
}

var ndvi_viz4 = {bands:"NDVI", min:0.2, max:1, palette:"000000,00FF00"};
var withNDVI = images.map(addNDVI);
var NDVI_S2 = withNDVI.select([’NDVI’]).max();

Map.addLayer(NDVI_S2, ndvi_viz4, ’NDVI_Sentinel2’);

Export.image.toDrive({image: ee.Image(NDVI_S2),description: "NDVI",scale: 30,region: geometry, maxPixels: 1e10});
```

Listing F.1: GEE script to obtain NDVI data from Sentinel 2

Appendix G

Validation With Gleam and NDVI Data

G.1 Comparison of Model CC and NDVI Data

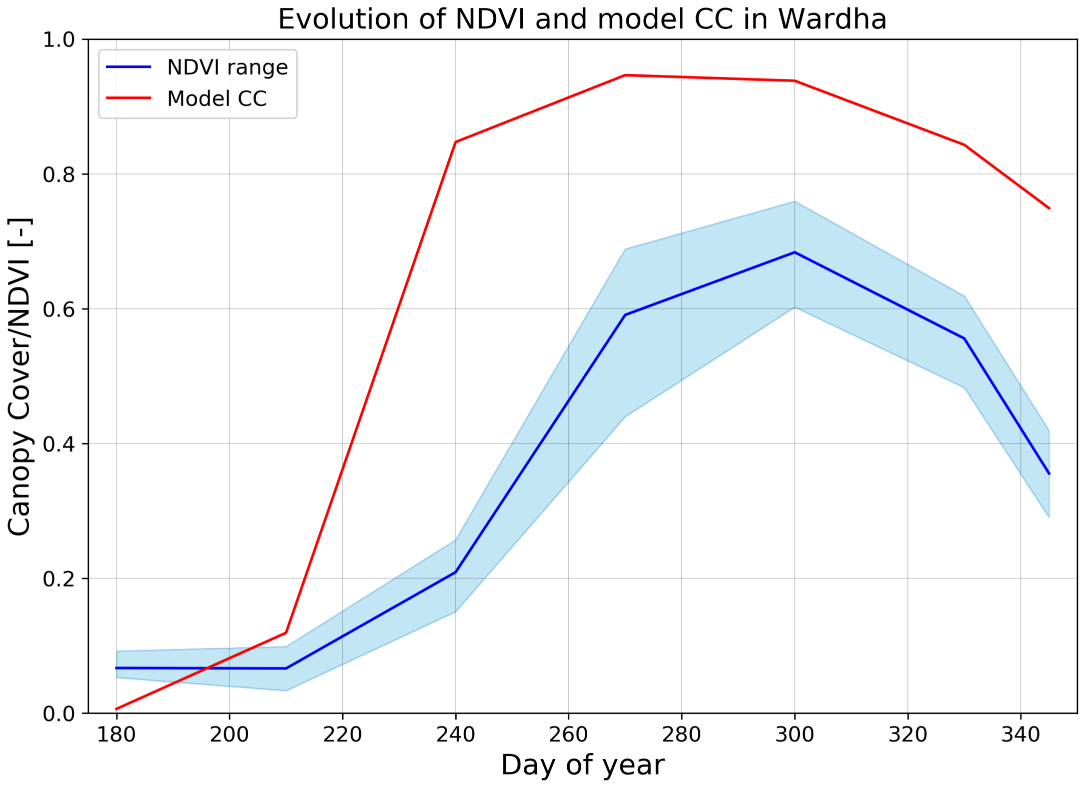


Figure G.1: CC vs NDVI in Wardha

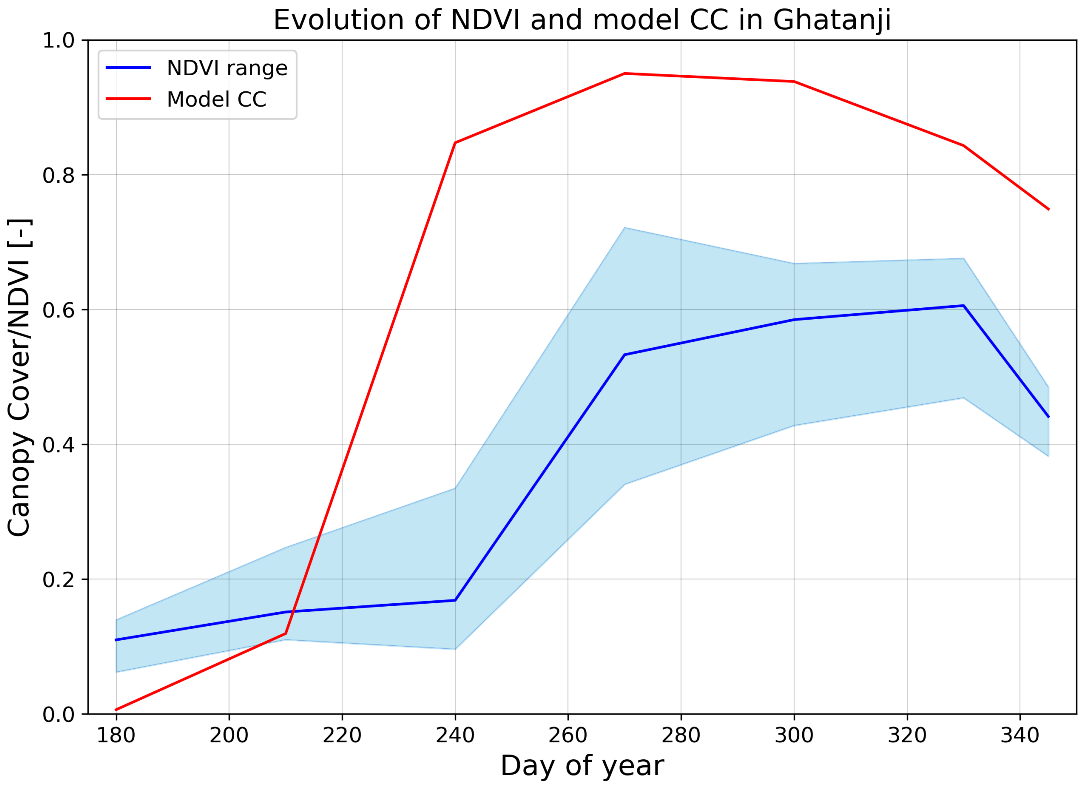


Figure G.2: CC vs NDVI in Ghatanji

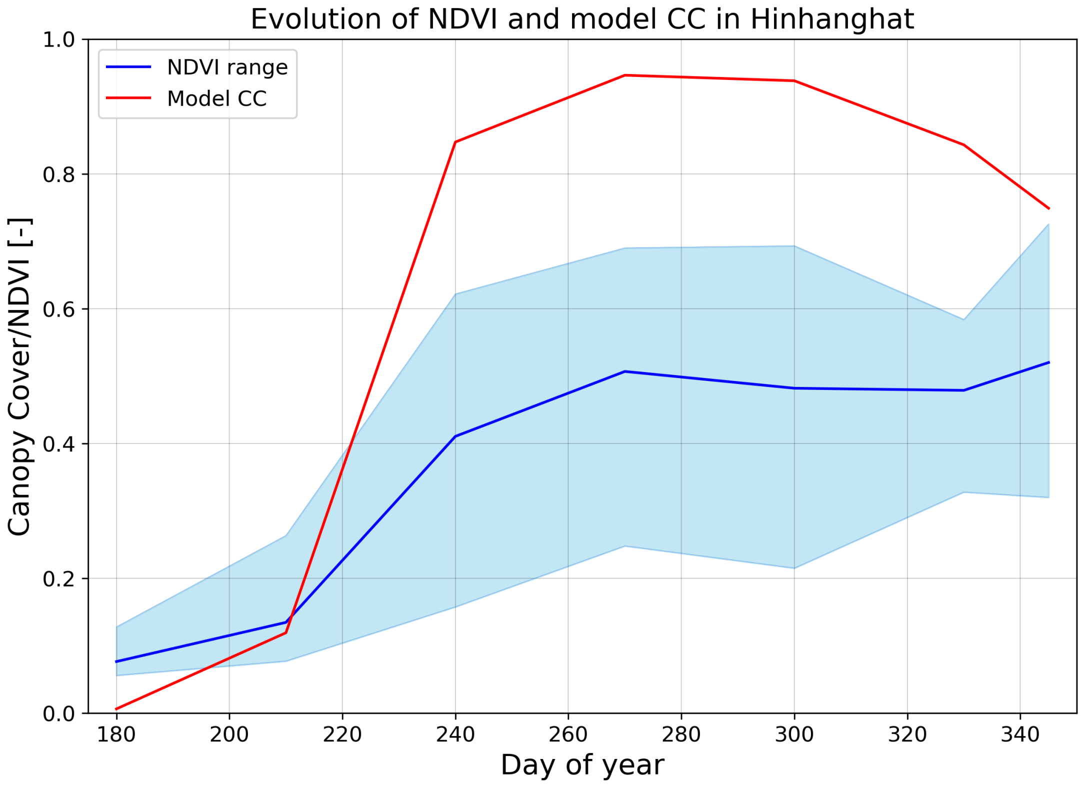


Figure G.3: CC vs NDVI in Hinhanghat

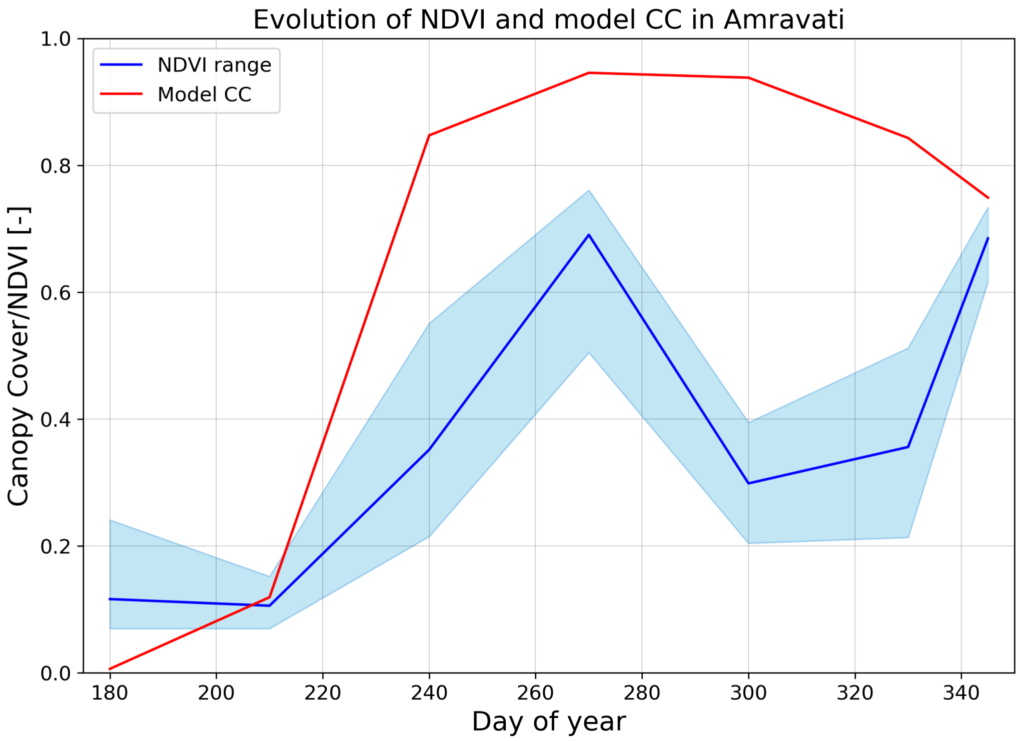


Figure G.4: CC vs NDVI in Amravati

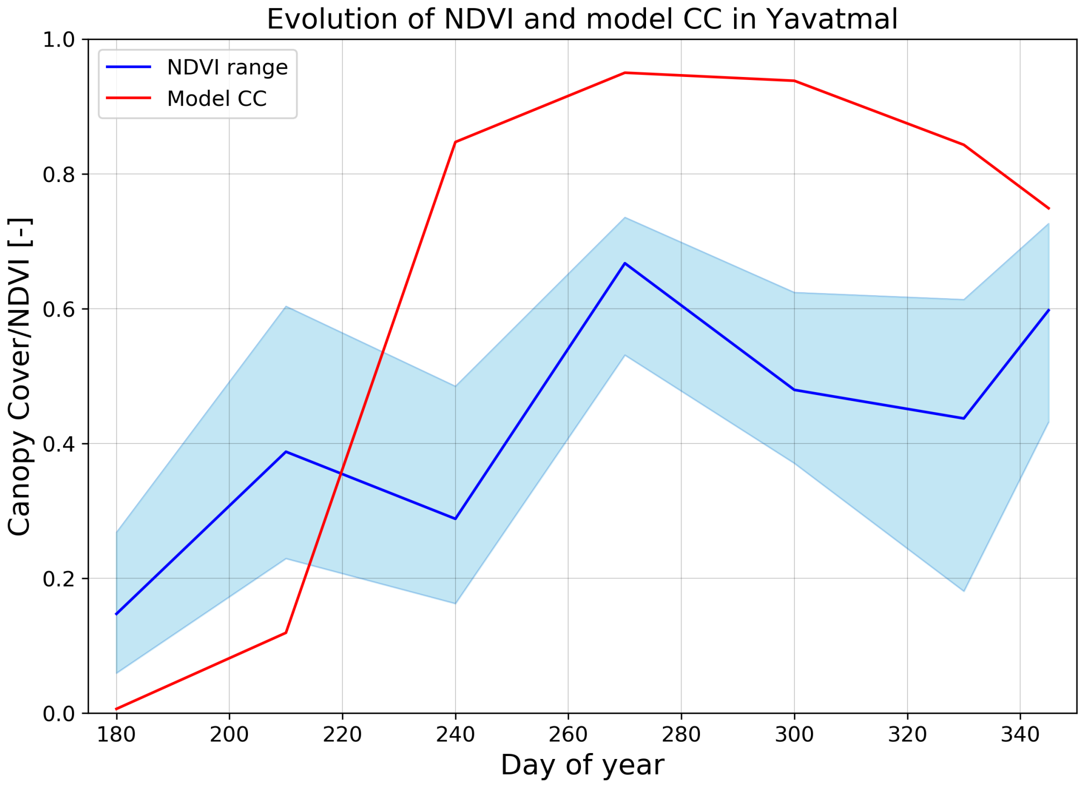


Figure G.5: CC vs NDVI in Yavatmal

G.2 Comparison of Model SM and GLEAM Data

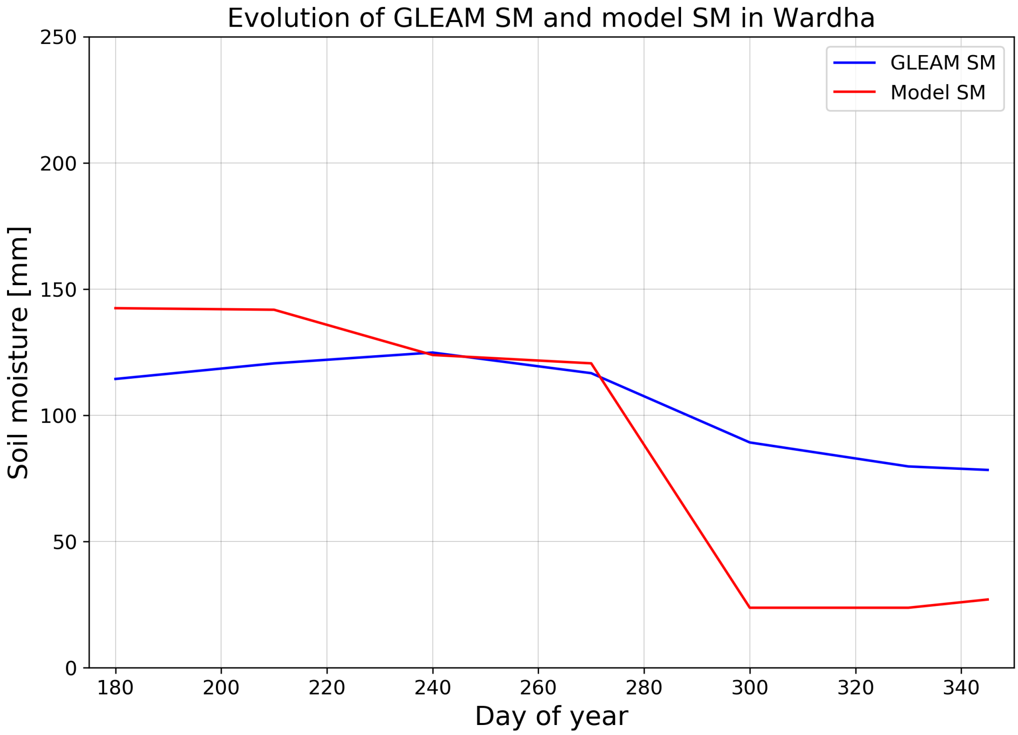


Figure G.6: Model SM vs Gleam data in Wardha

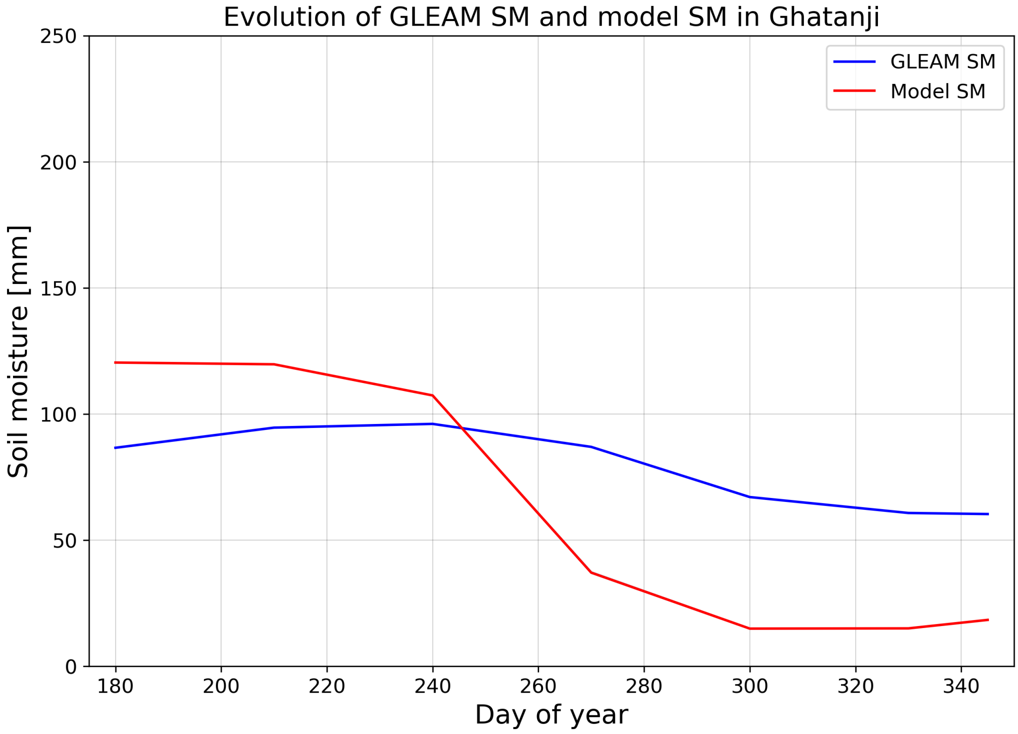


Figure G.7: Model SM vs Gleam data in Ghatanji

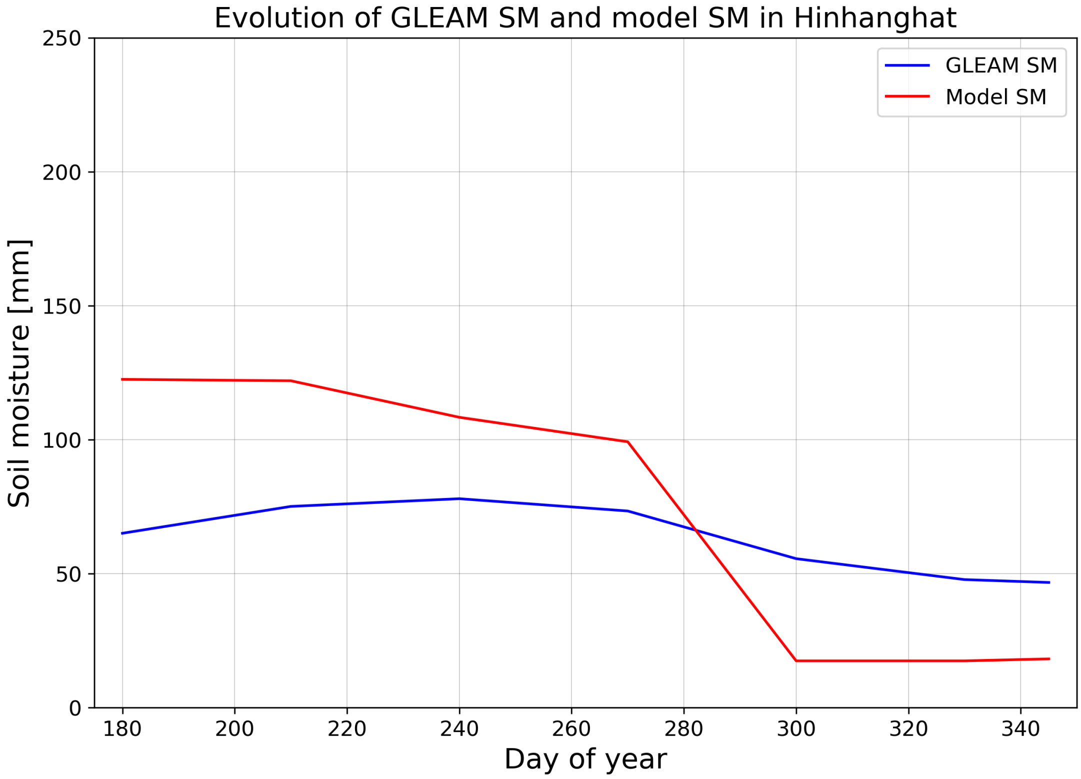


Figure G.8: Model SM vs Gleam data in Hinhanghat

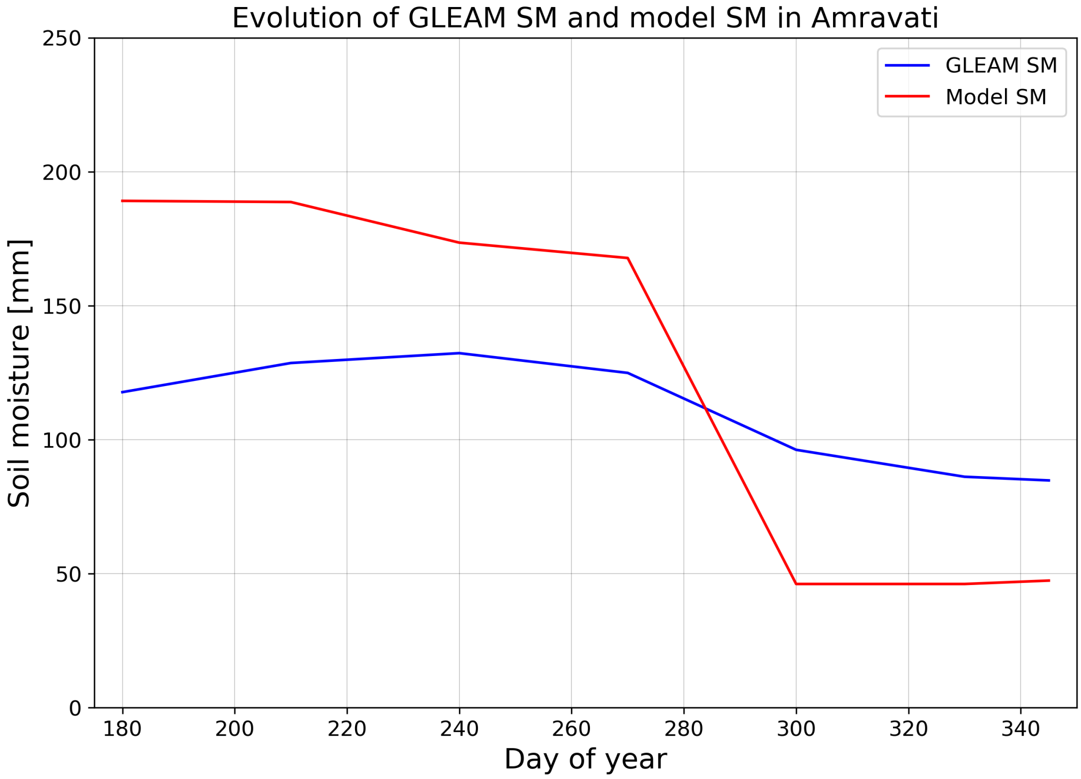


Figure G.9: Model SM vs Gleam data in Amravati

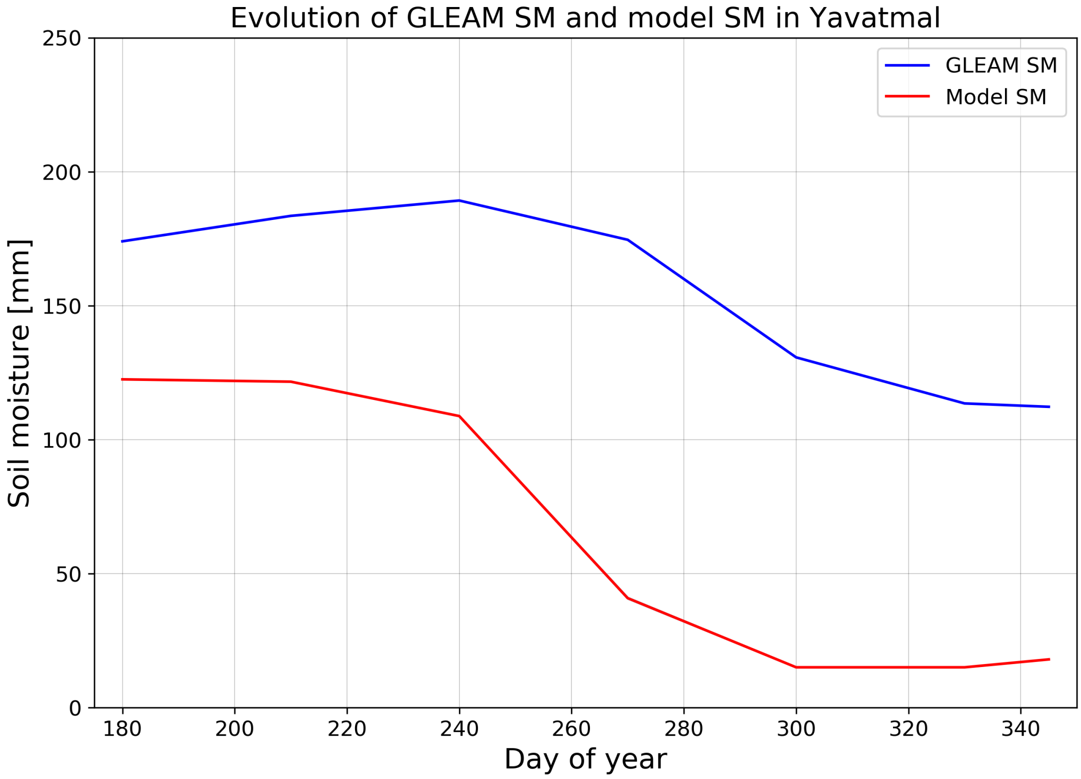


Figure G.10: Model SM vs Gleam data in Yavatmal

Appendix H

RANAS Questions

This appendix lists ideas for the preliminary questions that can be asked when doing the survey. The list include the questions that can be used to improve the mechanics and triggers for the irrigation and the expenditure cut. These questions are by no means complete, it should be treated as a supplementary material when designing the actual questionnaires, and it should be adjusted when exploratory interviews are done. These questions are designed largely based on Andrade et al. (2019) and Contzen, N., Mosler (2015), please refer to those papers for more information about RANAS.

Target group: Smallholder farmers and their dependents

Main points

- How will/did they deal with financial hardship?
  - What triggers them to take action?
  - Expenditure cuts, extra income
- Irrigation usage and behavior
  - Knowledge of proper irrigation usage
  - Information on irrigation techniques depending on what they use

Additional notes:

- Potential Factors – More accurate list should be investigated using exploratory interviews
- Splitting between the do and no doers might provide the difference in responses that will provide insight to the behavior factor
- When doing likert scale, make sure that each score is clearly defined

RANAS Questionnaire

Risk – Knowledge (about risk), vulnerability, severity

Irrigation

- Are you aware of the symptoms of water stress in cotton? (knowledge, list/multiple choice)
- Provide a list of symptoms and ask which one they are familiar with (scoring can give weight to the more important symptoms)
- How many rainless day would cause you to start irrigating your crop? (vulnerability/knowledge, list/multiple choice)
- What would be the impact of X number of rainless days? (severity, Likert scale)

Expenditure cuts

- How much savings are available to you during crisis? (vulnerability)
- How many dependents are in your house? (vulnerability, severity)

Attitude – Belief, feelings

Irrigation

- How much extra effort/time do you think it is to irrigate your crop? (belief, Likert scale)
- How certain are you that irrigation would help improve your crop yield? (belief, Likert scale)
- Do you believe that irrigation would improve the yield of your farm? (belief, Likert scale)
- How do you feel about using the irrigation system? (feelings, Likert scale)

Expenditure cuts

- How much do you prefer off farm work? (belief, Likert scale)
- How much extra effort/time do you need for off farm work? (belief, Likert scale)
- How much extra money do you get from off farm work? (belief, Likert scale)
- How easy is it to sell livestock? (belief, Likert scale)
- How do you feel about selling your livestock? (feelings, Likert scale)

Norm – Others behavior, (dis)approval, personal importance

Irrigation

- How many people you know irrigate their farm? (others behavior, Likert scale)
- How do your families and friends (and helper) think about using irrigation? ((dis)approval, Likert scale)
- How important do you think irrigating your crop is? (personal importance, Likert scale)

Expenditure cuts

- How many people you know do off-farm work? (others behavior, Likert scale)
- How do your families and friends think about off-farm work? ((dis)approval, Likert scale)
- How do your families and friends think about selling livestock? ((dis)approval, Likert scale)
- How strongly do you feel about continuing farm work within a season given that price of crop go below an X amount of INR? (personal importance, Likert scale)
- How strongly do you feel about continuing farm work within a season given that a certain symptom(s) of drought occurs? (personal importance, Likert scale)

Ability – Knowledge, confidence

Irrigation

- Asks them to describe how they irrigate their farm (knowledge, come up with criteria for points/quantification based on the 'correct' method to do it)
- Asks how confident they are about their method (confidence, Likert scale)
- Asks how confident they are about whether they will keep irrigating (confidence, Likert scale)
- Asks how confident they are about fixing the pond if it is broken (confidence, Likert scale)

Expenditure cuts

- Asks them how they determine whether they will do off farm work (knowledge, come up with criteria for quantification)
- Asks them how they decide how many livestock to sell (knowledge, come up with criteria for quantification)

- Asks how confident they are about their method (confidence, Likert scale)
- Asks how confident they are about finding off farm work (confidence, Likert scale)
- Asks how confident they are about continuing to farm next season if they are forced to do expenditure cut (confidence, Likert scale)

Self-regulation – Planning, control, barrier, remembering, commitment

Irrigation

- Do you have a specific plan on how and when to use the irrigation system? (planning, Likert scale)
- Do you pay attention to how much water you use for the irrigation? (control, Likert scale)
- How often do you forget to irrigate your crops? (remembering, Likert scale)

Expenditure cuts

- Do you have a specific plan on how to reduce expenditure? (planning, Likert scale)
- Do you pay close attention to how much savings, loan, costs, and revenue you have? (control, Likert scale)

Other Questions

- Have you attended an agricultural promotion/education from the government?
- Have you experienced drought previously?
  - Provide pictures of drought to streamline the definition across different farmers
  - Elicit experiential vs conjectural responses (have and have not experienced)
    - Did/will you seek help from government, did they provide help? Financial or otherwise
    - Did/will you sell livestock to cover for losses
- What symptoms would cause you to provide extra irrigation? Create a list of symptoms and score the answers based on the severity of the symptoms mentioned.

Appendix I

Implementation of Two Buckets System in the Soil Column

Fig. I.1 gives an overview of the preliminary two buckets system tested. This method have only been tested manually for their responses from various inputs but have not been validated. The two bucket system is thought to be more accurate in representing the vertical movement of water and thus a more accurate dynamics of the overall fluxes in the system. Further discretization of the soil column into smaller grid cells should be done with a precaution since the time step of the model (daily) can be too large for the fluxes and boundary conditions.

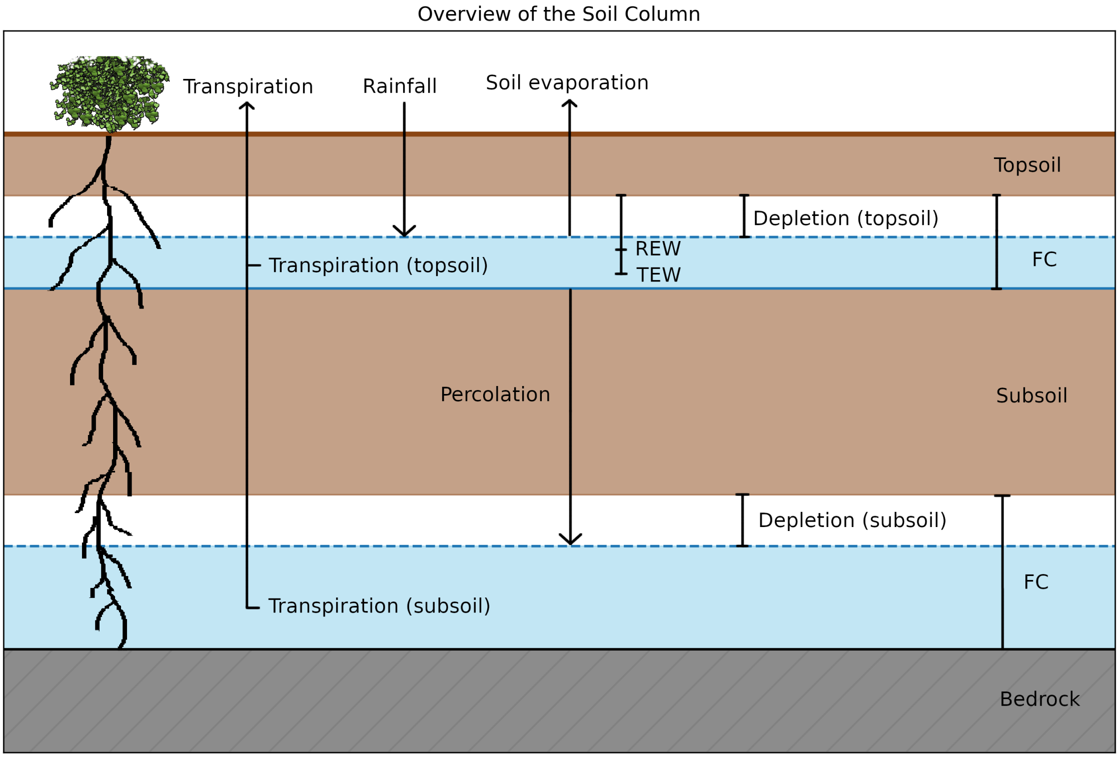


Figure I.1: Implementation of the two bucket system in the soil column