



UNIVERSITY OF NAIROBI

**ASSESSING FIELD - LEVEL SORGHUM YIELD VARIABILITY
IN SOUTH SUDAN USING REMOTE SENSING**

BY

JOHN KARONGO

I56/36045/2019

**An MSc Project Submitted for Examination in Partial Fulfillment of the
Requirements for Award of the Degree of Master of Science in Biometry of
the University of Nairobi**

2021

**ASSESSING FIELD - LEVEL SORGHUM YIELD VARIABILITY
IN SOUTH SUDAN USING REMOTE SENSING**

Research Report in Biometry, Number, 2021

John Karongo

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Linear mixed-effects models (LME) include both fixed-effects and random-effects variables. The use of LME is powerful in analyzing repeated measurements, longitudinal data or unbalanced data and the variations between and within-subject observations can be captured by random effects. In this study, remote sensing data were used to understand sorghum yield variability in a context of low input low output extensive farming system such in South Sudan. LME modelling approach helped understanding the sorghum yield variation between the two states of interest in this study and between two different agricultural seasons.

The unbalanced nature, the repeated measures on same statistical units for remotely derived parameters and the longitudinal nature of the data dictated the choice and the appropriateness of Linear Mixed-Effects models (LME) for statistical analysis in this study. The random-effects structures were used to describe the spatio (between states) and temporal (between seasons) specific variations of the sorghum yield during the two agricultural seasons (2018-2019); while the size of cultivated land and the households' size as a proxy of labor were used as fixed-effects variables in addition to remotely derived variables.

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.



20th August 2021

Signature

Date

JOHN KARONGO

Reg No. I56/36045/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

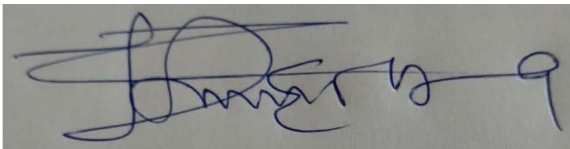


August 20, 2021

Signature

Date

Dr. Nelson Owuor
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke



20th August 2021

Signature

Date

Dr. Vincent Onguso Oeba
Kenya Forestry Research Institute (KEFRI)
Box 20412, 00200.
E-mail: voeba@kefri.org

Dedication

To my wife (Ghislaine Zaina) and Sons (John Karongo and Jordan Karongo).
Thank you for your love and support.

Contents

Abstract	ii
Declaration and Approval	iv
Dedication	vi
List of Figures	ix
List of Tables	x
Acknowledgments	xi
1 INTRODUCTION	1
1.1 RATIONALE.....	2
1.2 RESEARCH OBJECTIVES.....	2
1.2.1 SPECIFIC OBJECTIVES.....	2
1.3 LIMITATIONS OF THE STUDY	3
2 LITERATURE REVIEW	4
3 RESEARCH METHODOLOGY	8
3.1 INTRODUCTION.....	8
3.2 DESCRIPTION OF STUDY AREA.....	8
3.3 STUDY DESIGN.....	8
3.4 DESCRIPTION OF REMOTELY DERIVED FACTORS	9
3.4.1 NDVI.....	9
3.4.2 EVI.....	9
3.4.3 LAI /FPAR ($m^2 * m^{-2}$)	10
3.4.4 Précipitation (mm/day).....	10
3.4.5 Soil moisture ($cm^3 * cm^{-3}$).....	10
3.4.6 Evapo-transpiration.....	11
3.5 DATA COLLECTION	11
3.6 SATELLITE (remotely sensed) DATA ACQUISITION.....	11
3.7 METHODS OF DATA ANALYSIS.....	12
3.7.1 Multiple linear Regression (MLR) Analysis	12
3.7.2 Linear Mixed-Effects Model	15
4 RESULT AND DISCUSSION	19
4.1 CHECKING FOR LINEAR RELATIONSHIPS BETWEEN PARAMETERS	19
4.2 EFFECTS OF REMOTE SENSING DATA AND OTHER COVARIATES ON SORGHUM YIELD IN SE- LECTED SITES.....	19
4.2.1 Descriptive statistics of field data.....	19
4.2.2 Remotely derived data	20
4.2.3 Checking for collinearity between different parameters.....	21

4.2.4	Sorghum harvest analysis during the two agricultural seasons.....	22
4.3	Predicting the effects of remote sensing data on sorghum yield.....	24
	Model (M_6) assumption check with residuals ($Y - NDVI * Cultland$).....	26
4.4	Linear Mixed-Effects Model in predicting sorghum yield variability.....	28
4.4.1	Random-effects structures	29
	Random-effect structure (1) with random intercepts	29
	Random-effect structure (2) with random-intercept including variance-covariance matrix: $\sim (1 State) + (1 Year) +$ $(1 State : Year)$	32
	Comparison of predicted sorghum yield with different random-effects structures	36
	Plotting random slopes for the significant models (S12, S13 and S11b)	39
4.5	Conclusion and Recommendation.....	40
4.5.1	Results discussion and Conclusion	40
4.5.2	Recommendation	41
	REFERENCES.....	42
	ANNEX.....	48
.1	Data exploration	48
.1.1	Covariates description	49
.1.2	Checking for col-linearity between covariates.....	49
.1.3	Some Graphs.....	49
	PART A.....	51
	Linear Multiple Regression Analysis - Modeling.....	51
	Best Model Multiple linear regression	51
	Predicted Sorghum yield vs observed yield.....	52
	PART B.....	57
	LINEAR MIXED EFFECT MODELS	57
	Checking for Best Model Assumption - Residual plots	58
	ANOVA for model S18 and S19	68
	Analysis of variance for the first Random-effects structure.....	68

List of Figures

Figure 1. Table 1: statistic description of the field data	20
Figure 2. Collinearity check between field data and remotely sensed data	21
Figure 3. Sorghum yield plotted against cultivated land for the two agricultural seasons	22
Figure 4. sorghum yield plotted against NDVI	23
Figure 5. Sorghum yield during the two agricultural seasons	23
Figure 6. sorghum yield vs cultivated land	25
Figure 7. Model assumptions check with residuals	26
Figure 8. Predicted Yield plotted against observed yield	27
Figure 9. Sorghum yield plotted against different parameters: (a)NDVI, (b)EVI, (c)Soil-Moisture, (d)Precipitation, (e)FPAR,(f)NDVI vs LAI, (g)LAI, (h) Cultland, (i) HH-Size, (j) Evapotranspiration, (k) LAI vs FPAR.....	28
Figure 10. model S12 assumption check with residuals.....	32
Figure 11. Model S13 assumption check	32
Figure 12. Model S11b diagnostic	35
Figure 13. Comparison of predicted sorghum yield with different random-effects structures.....	38
Figure 14. Plot of random slopes for three significant models S11b, S12 and S13.....	39

List of Tables

Table 1. Average cultivated land and sorghum yield for the two agricultural seasons	19
Table 2. Remotely derived parameters description.....	20
Table 3. Different model fitted for sorghum yield	24
Table 4. Table of coefficient for Model (M_6)	25
Table 5. Different fixed-effect models comparison with random – intercept structure (1) $\sim (1 State) + (1 Year)$	29
Table 6. Anova results for all models with random-intercept structure (1): $\sim (1 State) + (1 Year)$	30
Table 7. Summary of random effects results for model S13	30
Table 8. Summary of fixed effect results for model S13	31
Table 9. Random effects results for model S12.....	31
Table 10. Fixed effects summary results for model S12	31
Table 11. Different-effects models comparison with random-intercept structure (2): $\sim (1 State) + (1 Year) + (1 State : Year)$	33
Table 12. Anova results for all models with random-intercepts structure (2): $\sim (1 State) + (1 Year) + (1 State : Year)$	34
Table 13. Random effects results for model S11b	34
Table 14. Fixed effects results for model S11b.....	35

Acknowledgments

I would like to thank the International Committee of the Red Cross (ICRC) for providing the funding for this MSc studies through iDevelop. The views expressed in this MSc project are those of the author and not necessarily of the ICRC.

I am extremely grateful to Dr Nelson Owuor and Dr Vincent Oeba for providing invaluable support and outstanding mentoring during this research work.

Special thanks are due to all my lecturers and professors for their professionalism and commitment. Thank you so much.

John Karongo

Nairobi, 2021.

1 INTRODUCTION

Remote sensing is defined as the acquisition of information about an object or phenomenon from distance. This involves an instrument, or a sensor mounted on a platform, such as a satellite, an aircraft, an UAV/UGV, or a probe [59]. The sensor measures the electromagnetic radiation that is either reflected or emitted by the target. The type and the usefulness of the information accessible from remote sensing depend on the specific properties and particularities of the instrument and its platform. These properties include: satellite orbitography, UAV/UGV flight/motion plan, field sensor position and orientation, active or passive sensing, detector array and optical lens characteristics [59, 15]. Currently, the climate and satellite data are available within weeks of acquisition and can provide data for operational assessment of crop yields.

Early work by Benedetti et al.[3] showed that National Oceanic and Atmospheric Administration (NOAA) satellite Normalized difference vegetation index (NDVI) data could be used to predict plant photosynthetic capacity and efficiency. In addition, the usefulness and affordability of real-time crop monitoring was made possible using NDVI index. A linear model for estimating wheat yield forecast using NDVI integration during the wheat grain-filling period was developed. Agro-meteorological information was recommended to be added to NDVI for better yield prediction [41].

A number of studies focusing on applying remote sensing products as a proxy indicator for yield estimation has been widely conducted. Such remotely sensed information is essential to explore the relationships of vegetation indices with crop yield. For example, a significant positive correlation was found between maize yield and enhanced vegetation index (EVI), normalized difference vegetation index (NDVI), and wheat yield and EVI in different countries [32, 33, 34]. Another way of using remotely induced vegetation indices to build the link with crop yields is also to explore how these parameters interact with water stress factors, such as surface temperature, soil moisture, rainfall [41], and evapo-transpiration.

Crop yield prediction using remote sensing data have been intensively studied mainly in wheat and maize, but such information is limited in Sorghum and in context of protracted conflict with access restrictions to farms such in South Sudan. The present study proposes a framework for field-level sorghum yield simulation in a country where sorghum cropping is characterized by extensive farming, with low inputs low outputs. The approach that have been used required the collection of remotely sensed data over an adequate time frame and a corresponding record of field crop yields.

In this project a number of remote sensing parameters have been used including Normalized Difference Vegetation Index(NDVI), Enhanced Vegetation Index(EVI), Soil moisture, Precipitation,

Leaf Area index (LAI), evapo-transpiration and Fraction of photo-synthetically Absorbed Radiation(FPAR), to simulate sorghum yield production and understand the yield variations in South Sudan.

1.1 RATIONALE

The difficulty to access ground measurements in South Sudan due the protracted conflict and insecurity, lack of access, inadequate infrastructure and chronic structural problems with inadequate statistics/data service and the challenge to estimate yields over large areas using other monitoring methods such agriculture surveys make remote sensing data a valuable alternative for yield prediction. Widely and freely available remote- sensing indices (MODIS products NDVI, EVI and LAI, FPAR) [9] that simulate above-ground biomass, Soil Moisture and Rainfall data have been combined for the simulation of sorghum crop yield.

For the humanitarian organizations as well as for the government, accurate and timely estimation of sorghum production yielded by small scale farmers who received humanitarian support to rebuild the agriculture sector in South Sudan can be critical as the country deeply depend on humanitarian aid for its agriculture. Modeling represents an opportunity to turn data into insights that will enable decision makers to make strategic planning to meet humanitarian needs and adjust interventions accordingly. Given the unbalanced nature, the repeated measures on same statistical units for remotely derived parameters and the longitudinal nature of the data for this study, Linear Mixed-Effects model (LME) was chosen and appropriate for statistical analysis. The random effects in this study are used to describe the spatio(state) and temporal (inter-annual) specific variations of the sorghum yield during two agricultural seasons(2018-2019). South Sudan was chosen for this study because of the very limited research in remote sensing and the need for a mathematical model that uses remote sensing to simulate sorghum yield to support humanitarian and government efforts in food security and response planning. This study used five(5) years (2016-2020) remote sensing data from MODIS products and the 2018-2019 sorghum yield data from 2 states (Upper Nile and Western Bahr El Gazal states).

1.2 RESEARCH OBJECTIVES

The main objective for this study is to assess field level sorghum yield variability in South Sudan using remote sensing.

1.2.1 SPECIFIC OBJECTIVES

1. To quantify and predict the effects of remote sensing data and other covariates on sorghum yield in the context of selected study sites.
2. To use the linear mixed-effects models to predict sorghum yield variability in the context of low input low output extensive farming system using remotely sensed data.

1.3 LIMITATIONS OF THE STUDY

Despite tremendous efforts developing regression models that relate satellite-derived vegetation indices directly to observed yield data, these models have their limits in the way that they are essentially retrospective and are based empirically on indirect inferences. In addition, because the regression relationship varies largely on a year-to-year basis due to inter-annual variations in climate parameters, water availability (evapo-transpiration, Soil moisture), and farm management practices, the application of these models is limited to the studied regions and periods and is difficult under extreme conditions (e.g., flooding and drought, cloud coverage, etc) beyond historical records.

2 LITERATURE REVIEW

Remote sensing is a technique that aims at acquiring information from space. Fussel et al.[12] defines remote sensing as a set of knowledge and techniques used to determine the physical and biological characteristics of objects or targets through measurements performed from remote locations, without any contact with those objects or targets. Sensors onboard satellites record the radiometric properties of objects observed on the Earth surface in forms of digital images. This technique has the advantage of supplying information over a long period of time and intervals depending on the satellite itself [20] . Using remote sensing is cost-effective compared to traditional data collection from field survey or aerial photography particularly for studying large areas.

Agricultural vegetation develops from planting to harvest as a function of meteorological driving variables such as sunlight, temperature and precipitation. During the crop cycle, the growth is further modified by soil and plant characteristics (genetics) and farming practices. For the last two decades the use of remotely sensed data for crop monitoring and yield modeling has progressed significantly.

Crop yields estimation is an important application of remote sensing [28, 35]. Applications in highly accurate yield estimates and crop disease and water stress detection at sub-pixel level have been operational in Northern America and Europe [8], [54] for last decade. In recent remote sensing studies, the normalized difference vegetation index (NDVI) and Enhanced Vegetation Index (EVI) derived from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery are widely used for crop yield analysis. NDVI, calculated with measurements of reflected light from the red and NIR bands, has long been used as an indirect measure of crop yield, including that of wheat [54], [18]. Siddoway and Aase [1] confirmed the relationship of NDVI and wheat grain yield but noted that the relationship deteriorated rapidly as wheat ripened.

Linear regression models relating NDVI to crop yield have, for example, been developed by Rasmussen [42] and Groten [14] for Burkina Faso and by Maselli et al. [33] for Niger. The same and other investigations showed that yield forecasting can be obtained by the use of NDVI data of specific periods which depend on the eco-climatic conditions of the areas and the types of crop grown.

Idso et al. [18] reported that summing NDVI values from late-season (Feekes 10.5, flowering to grain fill) spectral measurements was useful in predicting the grain yield of wheat. Bartholome [2] reported that accumulated NDVI was a more stable predictor of millet (*Panicum miliaceum* L.) and sorghum [*Sorghum bicolor* (L.)] grain yields than a single spectral measurement. For this reason, this study have used 5 years remotely sensed data to simulate sorghum yield in context of South Sudan.

Ref [42] calculated a sampling-interval weighted average NDVI by integrating multi temporal spectral measurements with time, which improved the grain yield estimates of millet from a single spectral measurement. Ref [51] reported that sensing twice and combining NDVI using a linear model improved correlation with wheat grain yield compared with sensing once.

Studies have shown that the seasonal accumulated NDVI values are correlated well with the reported crop yields in semi-arid regions [14]. Doraiswamy and Cook [11] further demonstrated that accumulating the NDVI values for spring wheat only during the grain-fill period improved the estimates of potential crop yields in North Dakota. Ref [61] used NDVI time series data from moderate resolution imaging spectroradiometer (MODIS) to forecast wheat yield in Kansas and Ukraine and multi-temporal vegetation indexes (VIs) were proposed to improve the yield prediction accuracy [6]. Ref [26, 27] predicted wheat yield with the accumulated VIs such as Σ NDVI(Nir, Green) and Σ RVI(Nir, Red) from jointing to the initial filling stage, and achieved a higher prediction accuracy than with VI in a single stage. Other forecasting yield methods use yield-related agronomic parameters that can be estimated from field and/or remote sensed data. For example, researchers have estimated the absorbed photosynthetically active radiation (PAR) and LAI and used it to predict yield of wheat and maize [40]; [50].

In rain fed agriculture, soil moisture conditions during the crop season are one of the key factors in determining crop yields. Thus, a crop model with a robust and accurate soil-water component was developed (EPIC: Erosion productivity impact Calculator) by [19]. This model has been used widely to simulate spring wheat crop growth and yield, [36], [62].

Several other techniques are being used and that relates VI with final yield at a specific crop growth stage such as at vegetative and reproductive stages during the growing season ([49], [29, 13]). Other techniques associate final yield with historical values of VI such as NDVI obtained during the entire growing season or during a specific period of the growing season such as the vegetative or reproductive stages [23, 34, 57]. These techniques require historical values of NDVI for a specific region and are compared with current values of NDVI to detect NDVI anomalies or deviations from historical values and then after the data are used to estimate yields [21, 17].

Ref [13] used correlations and multiple linear-regression analysis to determine variables to be used to predict final winter wheat grain yield. In their findings both the correlation and regression analyses suggested mid-season NDVI, chlorophyll content, plant height, and total nitrogen uptake to be good predictors of final winter wheat grain yield. Several investigations have shown that NOAA NDVI data accumulated during a rainy season can be related to total rainfall or final primary productivity in the Sahel [33].

In a study about operational maize yield model development and validation based on remote sensing and agro-meteorological in Kenya, [46] found that the land cover NDVI and the actual Evapo-transpiration in his model explain 83% of the maize crop yield variance with a root square mean error (RMSE) of 0.3298 t/ha. Pavlo [31] recent research indicates the suitability of the LAI

and NDVI for the simulation of sweet corn yields. It was determined that LAI is a more suitable index for the crop yield prediction: the R^2 value was 0.92 and 0.94 against 0.85 for the NDVI-based models. It was determined that it is better to use the LAI values obtained at the flowering stage, when R^2 averaged to 0.94, and the NDVI-based models did not depend on the crop stage (the R^2 was 0.85 both for the flowering and ripening stages of the plant development).

In another study using RapidEye satellite multi-spectral data, Angela et al.[22] estimated LAI and biomass of corn and soybean; the results indicated that the cumulative red-edge simple ratio performed best for estimating LAI and biomass. Ref [25] comparatively, analyzed data from different satellites (Gaofen-1, Huanjing-1, and Landsat-8 multi spectral) data for estimating the leaf area index of winter wheat.

Research result from Ref [7] study estimated LAI of a potato crop with different fertilization levels using Sentinel-2 satellite images and the results demonstrated that the weighted difference vegetation index using high spatial resolution can be used for estimating the LAI. Ref [63] found high correlation (0.75) between LAI and rice yield in China in a study that aimed at predicting grain yield in rice using multi-temporal vegetation indices from Unmanned Aerial Vehicle (UAV)-based multi-spectral and digital imagery.

In a rice trial where different Nitrogen rates, planting patterns and different rice cultivars were involved, Yanyu Wang et al. [58] have used a compact multi-spectral camera mounted on a fixed-wing drone(ebee) to collect data during key growth stages. LME, simple regression (SR), artificial neural networks(ANN) and random forests(RF) models were developed relating growth parameters above ground biomass(AGB) and leaf area index(LAI) to spectral information. Cultivar, growth stage and planting pattern were selected as candidates of random effects for the LME models due to their significant effects on rice growth. The study results revealed that when comparing to other regression models (SR, ANN and RF), the LME model improved the AGB estimation accuracy for all stage groups to varying degrees: the R^2 increased by 0.14–0.35 and the RMSE decreased by 0.88–1.80 t/ha for the whole season, the R^2 in LME increased by 0.07–0.15 and the RMSE decreased by 0.31–0.61 t/ha for pre-heading stages. In addition, the R^2 increased by 0.21–0.53 and the RMSE decreased by 0.72–1.52 t/ha for post-heading stages.

Bolton et al.[5] and Ref [48] mentioned several studies that have focused on other remote indices, including phenological metrics extracted from vegetation indices time series to improve yield estimation. However, these studies typically target the simpler cases of relatively homogeneous landscapes as found in the US and China. In other parts of the world including South Sudan, agricultural landscapes are more heterogeneous with a higher diversity in crop types, crop management practices, and thus field sizes. Ref [30] showed that the correlations between the accumulated FPAR and yields in Europe vary widely depending on the crop type and geographic locations, especially in Northern Europe where observed inter-annual yield variability is too low to be easily detected from remote sensing data.

The issues related to small field sizes is always exacerbated in smallholder landscapes, which are often more likely threatened by food insecurity. The small size of the fields also implies a lack of reliable yield statistics to establish robust empirical relationships. Furthermore, such landscapes are often located in inter-tropical regions where estimating yield from remote sensing observations is complicated by an abundant cloud coverage during the growing season, making the use of optical remote sensing more challenging [52] and [60, 43].

The traditional methods of yield measuring turn to be time-consuming and cannot consider yield variations over a large field or space in addition to insecurity and lack of access in so many parts of South Sudan; therefore they are prone to large errors due to incomplete ground observations, leading to poor crop yield assessment and crop area estimations or predictions. In the light of these limitations, remote sensing methods could be a reliable alternative

3 RESEARCH METHODOLOGY

3.1 INTRODUCTION

This study was based on modelling two years (2018-2019) sorghum grain yield obtained from random surveys carried out in Upper Nile and Western Bar El Gazal states in South Sudan. A combination of Multiple Linear Regression (MLR) and Linear Mixed-Effects (LME) models were performed to assess regression relationship and develop a mathematical model that would fit well the yield and understand the random effects of the states and seasonal variations in rain-fed extensive farming system. In addition to yield data, 5 years satellite-derived data on some key crop parameters were collected remotely and both datasets merged for further analysis.[55, 47]

In particular, satellite-derived vegetation indices, as measures of plant chlorophyll abundance and vegetation radiation absorption, and climatological related data were collected and have proven to be closely related to crop growth in field studies and theoretical models. Accurate and timely estimation of production yielded by the farmers who received humanitarian support to rebuild the agriculture sector in South Sudan can be critical as the country deeply depend on humanitarian aid for its agriculture. This study aims at developing a simple and efficient model-based method to estimate sorghum yield in South Sudan using satellite-derived data from MODerate Resolution Imaging Spectroradiometer (MODIS) products.

3.2 DESCRIPTION OF STUDY AREA

Upper Nile is a state in South Sudan and located at 10030'N32030'E. With Malakal as the head quarter, the state has 13 counties and a total population estimated at 964,353 inhabitants (South Sudan Census, 2008) living in area of 77,823.42km². Western Bahr el Ghazal situated at 07053'N25052'E is another state with Wau as its capital city. It has an area of 93,900km² and is the least populous state(333,431persons) in South Sudan, according to the census conducted in 2008. Western Bahr El Gazal and Upper Nile states are two of 11 states that form the Republic of South Sudan. In these two states, diversified crops are cultivated but Sorghum (*Sorghum bicolor*) is the dominant crop. In general, crop planting is completed by mi-may when the soil moisture is good enough after 2-3 good rains to initiate the germination. Sorghum usually matures after 5 months and harvest happens late November beginning of December.

3.3 STUDY DESIGN

This study relied on the analysis of sorghum yield data randomly collected at household level(self-reported) from farmers who received humanitarian seeds support during 2 agricultural seasons

(2018-2019) and remotely sensed data for 5 years (2016-2020) from MODIS satellite products. Given that in the two states concerned by this study (Upper Nile and Wester Bahr El Gazal states), sorghum is grown under rain fed conditions, the seasonal variability in rainfall patterns could contribute to the variability in crop yields from season to season. This research - in addition to developing a sorghum yield model using remote sensing - it intends to analyse and establish the inter-states and inter-seasonal variations of sorghum in South Sudan.

3.4 DESCRIPTION OF REMOTELY DERIVED FACTORS

3.4.1 NDVI

Vegetation indices (MODIS13A3) from Terra Moderate Resolution Imaging Spectroradiometer) Vegetation Indices (MODIS) Version 6 data are provided monthly at 1 kilometer (km) spatial resolution as a gridded Level 3 product in the sinusoidal projection. The Normalized Difference Vegetation Index (NDVI) from MODIS complements NOAA's Advanced Very High Resolution Radiometer (AVHRR) NDVI products and provides continuity for time series historical applications. The normalized difference vegetation index (NDVI) is derived from the visible and near-infrared (NIR) bands and has been successfully used to monitor vegetation changes at regional scales [54], [10].

$$NDVI = \frac{\rho_{NIR1} - \rho_{red}}{\rho_{NIR1} + \rho_{red}} \quad (1)$$

where ρ_{NIR1} = Near-infrared and ρ_{red} = Red

NDVI data used in this study was downloaded from MODIS13A3 product.

3.4.2 EVI

The enhanced vegetation index (**EVI**) is designed to enhance the vegetation signal with improved sensitivity in high biomass regions and improved vegetation monitoring through a de-coupling of the canopy background signal and a reduction in atmosphere influences. MODIS includes an Enhanced Vegetation Index (EVI) that minimizes canopy background variations and maintains sensitivity over dense vegetation conditions. [9] mentioned that MODIS NDVI and EVI products are computed from surface reflectances corrected for molecular scattering, ozone absorption, and aerosols.

EVI is computed following this equation:

$$EVI = G \times \frac{(NIR - RED)}{(NIR + C1 \times RED - C2 \times Blue + L)} \quad (2)$$

,where NIR/red/blue are atmospherically-corrected and partially atmosphere corrected (Rayleigh and ozone absorption) surface reflectance, L is the canopy background adjustment that addresses

non-linear, differential NIR and red radiant transfer through canopy, C1, C2 are the coefficients of the aerosol resistance term, which uses the blue band to correct for aerosol influences in the red band. In MODIS-EVI equation, the coefficients are L=1, C1 = 6, C2 = 7.5, and G (gain factor) = 2.5.

Then the equation become as follow:

$$EVI = 2.5 \times \frac{(\rho_{NIR1} - \rho_{red})}{(\rho_{NIR1} + 6 \times \rho_{red} - 7.5 \times \rho_{blue} + 1)} \quad (3)$$

The NDVI is chlorophyll sensitive while the EVI is more responsive to canopy structural variations, including leaf area index (LAI), canopy type, plant physiognomy, and canopy architecture. NDVI and EVI, two vegetation indices, complement each other in studies about the global vegetation and improve upon the detection of vegetation changes and extraction of canopy biophysical parameters. [10]. This study used EVI data downloaded from MODIS13A3 product.

3.4.3 LAI /FPAR ($m^2 * m^{-2}$)

The Leaf Area Index (LAI) and the Fraction of Photosynthetically Active Radiation (FPAR) data used in this study are derived from MODIS. An 8-day composite data set with 500 meters (m) pixel size. The algorithm used chooses the “best” pixel available from all the acquisitions of the Terra sensor from within the 8-day period. LAI could be defined as the one-sided green leaf area per unit ground area in broad-leaf canopies and as one-half the total needle surface area per unit ground area in coniferous canopies. FPAR is the fraction of incident photo-synthetically active radiation absorbed by the green elements of a vegetation canopy [37].LAI data were downloaded from MDC15A3H version 6 MODIS product.

3.4.4 Précipitation (mm/day)

Precipitation can be any form of moisture which falls to the earth. This includes rain, snow, hail and sleet. Complex forces are the cause of the water droplets to fall as rainfall. Precipitation can also be defined as any product of the condensation of atmospheric water vapor that falls under gravitational pull from clouds. The main forms of precipitation include drizzling, rain, sleet, snow, ice pellets, etc. Precipitation will occur when part of the atmosphere becomes saturated with water- vapor (with 100% relative humidity), and that the water condenses and precipitates or falls. Rainfall data used in this study are freely available and downloaded from NASA Earth data.

3.4.5 Soil moisture ($cm^3 * cm^{-3}$)

In general, soil moisture refers to the water present in the upper part of the soil and is a variable controlling a wide array of ecological, hydro logical, geo-technical, and meteorological processes. Soil moisture plays also a role of regulator of the partitioning of the incoming solar energy at

land surface level into the outgoing sensible, latent, and surface heat fluxes, mainly through the processes of soil evaporation and plant transpiration. Soil moisture is an important life sustaining entity. The main use of soil moisture is to enable vegetation growth. Water that is stored in the soil has many roles that include the fact that it controls the partitioning of rainfall into runoff and infiltration,[31, 56].

Soil moisture parameter is an important variable in climate system. Accurate prediction and the understanding of the variations of surface temperature, drought, and flood depend critically on knowledge of soil moisture variations, as do impacts of climate change and weather forecasting. Several physical, chemical and biological processes that take place at the land surface are strongly influenced by the amount of water stored within the upper soil layers. Data about soil moisture data in this study come from Soil Moisture Active Passive (SMAP) enhanced L3 Radiometer Global daily 9 km. This enhanced product Level-3 soil moisture gives a composite of daily estimates of global land surface conditions retrieved by (SMAP) radiometer.

3.4.6 Evapo-transpiration

Evapo-transpiration can be defined as the sum of all forms of evaporation plus transpiration, but in the frame of the study it well corresponds to the sum of evaporation from the plant transpiration plus land surface. About evapo-transpiration, literature provides several definitions, here, evapotranspiration will refer to the water lost to the atmosphere from the ground surface, evaporation from the capillary fringe of the groundwater table, and the transpiration of groundwater by plants whose roots tap the capillary fringe of the groundwater table [38].

The evaporation of water from plant leaves is the transpiration aspect of evapo-transpiration. The amount of water that plants transpire varies largely geographically and over time. Several factors determine and influence the transpiration rates, this includes temperature, relative humidity, wind and air movement, soil-moisture availability and type of plant.

3.5 DATA COLLECTION

The data used in this study (Table 1) are about sorghum yield measures that were collected during two-years agricultural seasons 2018 and 2019. Surveys were conducted with small scale farming households in Upper Nile and Western Bahr El Gazal states who received sorghum seeds from humanitarian assistance. 235 household farmers randomly selected were interviewed, and data collected via device magic and then transferred to Excel.

3.6 SATELLITE (remotely sensed) DATA ACQUISITION

Five years (2016-2020) satellite data include 7 standard MODIS products among which vegetation indices: Normalized Difference vegetation Index (NDVI), Enhanced vegetation Index(EVI), the eight day Fraction of Photosynthetically Active Radiation (FPAR) and Leaf Area Index (LAI)

(MOD13A3), and agroclimatic data : Precipitation(mm/day) every month, daily Soil moisture (cm^3/cm^3) and every 8 days evapo-transpiration($kg/m^2/8days$). NDVI and EVI were composited every month of each year, LAI and FPAR every eight days of each month starting from Day-of-Year 1 in each calendar year. Details are available on the MODIS data website.[4]

Remotely sensed data from satellite provide a real-time assessment of the magnitude and variation of crop condition parameters, and this study investigates the use of these parameters as inputs to modelling sorghum yield variation in the context of South Sudan.

3.7 METHODS OF DATA ANALYSIS

3.7.1 Multiple linear Regression (MLR) Analysis

Multiple linear regression (MLR) is a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable. The assumption is that there is no clustering. This technique enables researchers to determine the variation of the model and the relative contribution of each independent variable in the total variance. MLR attempts to establish or to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to measured data. Each value of the independent variable x is associated with a value of the dependent variable y .

The population regression line for k explanatory variables x_1, x_2, \dots, x_k is defined to be:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k \quad (4)$$

, this line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means μ_y and the assumption is that they have the same standard deviation δ . The fitted values b_0, b_1, \dots, b_k estimate the parameters

$$\beta_0, \beta_1, \beta_2, \dots, \beta_k \quad (5)$$

of the population regression line. Since the observed values for y vary about their means μ_y , the multiple regression model includes a term for this variation. The model could be written or expressed as :

$$DATA = FIT + RESIDUAL,$$

where ,
the term FIT represents the expression

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k \quad (6)$$

The "RESIDUAL" term represents the deviations of the observed values y from their means μ_y , which are normally distributed with mean μ and variance σ . The model deviations is written as ε .

The model for multiple linear regression, with n observations, is as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i \quad (7)$$

for $i = 1, 2, \dots, n$. In MLR and particularly in least-squares model, the best-fitting line for the observed data is obtained by minimizing the sum of the squares of the vertical deviations from each data point to the line.

The values fit by the equation :

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i \quad (8)$$

are denoted by \hat{y}_i and the residuals ε_i are equal to $y_i - \hat{y}_i$, the difference between the observed and fitted values. The total of the residuals is equal to zero.

3.7.1.1 Identifying and Controlling for Confounding with Multiple Linear Regression

Multiple regression analysis is a powerful tool and can be used to assess whether confounding exists, and, since it allows us to estimate the association between a given independent variable and the outcome holding all other variables constant; multiple linear regression provides also a way of adjusting for (or accounting for) potentially confounding variables that have been included in the model.

3.7.1.2 Comparison of different Models

The comparisons of different models were done by evaluating their predictive capabilities, which were evaluated by their coefficient of determination (R^2) and the Residual Mean Square Error (RMSE)

3.7.1.3 Coefficient of determination (R^2)

The coefficient of determination (denoted as R^2) is a key output of regression analysis. The (R^2) represents the proportion of the variance in the dependent variable that is predictable from the independent variable. The coefficient of determination informs about the percentage of variation in dependent variable explained by all the independent variables together. The coefficient of determination, R^2 , is a key statistic indicating how well a model including a set of predictors accounts for the variation in the response variable. While it shows the utility of these predictors in fitting the model, it also provides a measure of predictability of the response variable using the

set of predictors. R^2 can also be used to choose the optimal set of predictors when the model size including all predictors, is fixed.

The R^2 is usually presented as the quantity that estimates the percentage of variance of the response variable explained by its (linear) relationship with the explanatory variables. It is computed by means of the ratio:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (9)$$

$$= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (10)$$

where,

ESS, TSS and RSS are respectively the explained, total and residual sum of squares.

When there is an intercept term in the linear model, this coefficient of determination is actually equal to the square of the correlation coefficient between y_i and \hat{y}_i .

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \quad (11)$$

with,

$\bar{\hat{y}}_i$ as the mean predicted Sorghum yield,

y_i observed yield values,

and

\bar{y}_i is the average of observed sorghum yield.

This equation (11) has a great interpretation in that R^2 measures the goodness of fit of the regression model by its ability to predict the response variable and this is measured by the correlation. In Addition, this expression shows that the (unconditional) distribution of the response does not need to be Gaussian to allow for the interpretation of R^2 (**Renaud et Al., 2010**).

3.7.1.3.1 Residual Mean square Error (RMSE)

The Root Mean Square Error (RMSE) is equal to the standard deviation of the residuals (Prediction errors). Residual's measure and provides information on how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. It provide important information about how concentrated the data is around the line of best fit and this is mainly useful for analyzing overall significance of linear regression.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (12)$$

where, x_i and y_i represent the estimated and measured values of Sorghum yield respectively.

3.7.2 Linear Mixed-Effects Model

Longitudinal data are (usually non-uniformly) ordered in time, and unbalanced data are very common. Furthermore, serial measurements of one subject are positively correlated, and between-subject variance is not constant over time due to several biotic and abiotic factors. The linear mixed-effects (LME) model is a suitable model to handle such data [24].

The problem is that the mixed effects model contains two components: a fixed effect (the explanatory variables) and the random effects. There is need to select not only an optimal fixed-effects structure but also an optimal random effects structure. In most cases, the interest is in the fixed effects. But if the random effects are poorly chosen, then this affects the values (biased) and quality of the fixed effects as the random effects work their way into the standard errors of the slopes for the fixed effects. On the other hand, variation in the response variable not modeled in term of fixed effects ends up in the random effects. [64]

The form of the LME model is given by:

$$\mu \sim N(0, G), \varepsilon \sim N(0, R) \quad (13)$$

$$G = \sigma_s^2 H(\phi) \quad (14)$$

$$R = \tau.I \quad (15)$$

In equations (14,15),

G represents the variance-covariance matrix of the random effects.

$H(\cdot)$ is defined by the suitable correlation function,

$f(h, \phi)$, σ_s^2 represents the partial sill,

ϕ represents the range,

and

h represents the lag or distance [39].

R in equation(9) represents a $N \times N$ positive definite variance-covariance matrix of ε .

τ^2 is called the nugget effect.

I represents the identity matrix (diagonal matrix of 1s).

ε is not correlated with the random effect μ , that is, $Cov(\mu, \varepsilon) = 0$.

Simultaneous estimates of correlation parameters (i.e., σ_s , H , ϕ and τ) and fixed effect coefficients (i.e., β) are obtained by Restricted or (residual) Maximum Likelihood estimation (REML) to reflect the loss of degree of freedom due to the estimation of the fixed effects coefficients [39].

In this study, the Linear Mixed-Effects Model (LME) was used to capture the effects of fixed parameters (cultivated land and households' size) and spatial (in each state) and temporal (inter annual

variations) random effects in different models using remotely sensed data. Cultivated land as well household size as a proxy of labor were set as fixed-effects parameters given the interest of the study and as it is considered as a key factor in low input low output extensive farming system such in the South Sudan.

The random effects had different choices, the interest was on understanding the random effects generated by the spatial variations at State level(differences between the two states: Upper Nile and Wester Bahr El Gazal) and inter-annual variations for the two agricultural seasons (2018–2019). Moreover, with these random effects, more group combinations with remotely sensed factors (EVI, LAI) and precipitations or Soil-moisture were utilized to improve the robustness of Yield/NDVI/Cultivated land models in this study. Optimal random effects were determined by comparing the REML values.

3.7.1.1 Model Selection in Mixed Effects Modelling

As it could be done in linear regression, there are two main options for model selection. The first option is based on selection tools like the Akaike Information Criteria (AIC), or the Bayesian Information Criteria (BIC). The AIC and BIC both contain in their equations two terms that measure the fit and the complexity of the model. The likelihood value including Maximum Likelihood (ML) and the Restricted Maximum Likelihood (RML) are also used in defining the measure of the model fit.

From Ref [64] the AIC is defined as twice the difference between the value of the likelihood L (measure of fit) and the number of parameters (penalty for model complexity) in the model. For the Bayesian Information Criterion(BIC), the number of observations is also taken into account, which means that more significant increases in the likelihood are required for larger data sets to confirm a model as better. In below formula, p is the number of parameters in the model (θ), L can be either the maximum Likelihood (ML) or the Restricted maximum Likelihood (REML), and for ML, we have $n^* = n$ but for REML,
 $n^* = n - p$

$$AIC = -2 \times L(\theta) + 2 \times p \quad (16)$$

$$BIC = -2 \times L(\theta) + 2 \times p \times \ln(n^*) \quad (17)$$

This means that an AIC based on REML can not be comparable with an AIC obtained by ML. This is equally valid for the BIC.

3.7.1.2 The Restricted Maximum Likelihood

Restricted Maximum Likelihood (REML), a statistical methodology that is a particular form of LME, does not base estimates on a maximum likelihood fit of all the information. In statistics, the REML uses a likelihood function that is derived from a transformed set of data so that parameters that could bring noise have no effect. From its formula, the estimator for the variance obtained by maximum likelihood is biased by a factor $(n-2)/n$. In case the model contains p explanatory variables, then the bias is $(n-p)/n$. The reason that the maximum likelihood estimator is biased is because it ignores the fact that the intercept and slope are estimated as well (as opposed to being known for certain). So, there is need for a mechanism that gives better ML estimators, and this is what REML does.

REML works as follows:

The linear regression model:

$$Y_i = \alpha + \beta \times X_i + \varepsilon_i$$

can be written as:

$$Y_i = X_i \times \beta + \varepsilon_i.$$

This is based on simple matrix notation using $X_i = (1X_i)$,

and

the first element of β is the intercept,

and

the second element is the original β .

The normality assumption implies that,

$$Y_i = N(X_i \times \beta, \sigma^2) \tag{18}$$

The problem with the ML estimator is that we have to estimate the intercept and the slope, which are in β in Equation (3.14). Obviously, the problem is solved if there is no β . The REML avoids having any β in Equation (5.18). It does this by finding a special matrix A of dimension $n \times (n-1)$, and special means orthogonal to (or independent of) X' , multiplies $Y = (Y_1, \dots, Y_n)$ with this matrix and continues with ML estimation. Orthogonal means that if A and X are multiplied, the result is 0.

Hence, we get

$$A' \times Y = A' \times X \times \beta + A' \times \varepsilon = 0 + A' \times \varepsilon$$

. The distribution for $A' \times Y$ is now given by:

$$A' \times Y \sim N(0, \sigma^2 \times A' \times A) \tag{19}$$

which no longer depends on β .

Applying ML on $A' \times Y$ gives an unbiased estimator for σ^2 .

3.7.1.2 Hypothesis testing

The second approach to find the optimal model is via hypothesis testing. There are two options here: the t-statistic and the F-statistic.

4 RESULT AND DISCUSSION

4.1 CHECKING FOR LINEAR RELATIONSHIPS BETWEEN PARAMETERS

In part A, data were analysed using the Multiple Linear Regression (MLR) procedure in R (R Language [53, 16]) to test the effect of cultivated land and vegetation indices remotely derived measurements NDVI, EVI, LAI, FPAR ON sorghum yield. In addition, MLR analysis was used to model final yield using NDVI and Cultivated land measurements and household size as predictors. [45] The coefficient of determination (R^2) and Root Mean Square Error (RMSE) were used as the criteria to determine if remotely sensed vegetation indexes could be used as linear predictors of sorghum yield in context of farmers self-reported data and extensive low input low output farming system.

In Part B a Linear Mixed-Effects Model (LME) regression analysis was used to model final yield using cultivated land, household size, vegetation indices (NDVI, EVI, LAI, FPAR) and combining with agrometeorological measurements (Precipitation, soil moisture, evapotranspiration) as fixed-effects parameters **between-states and between-seasons variations were captured using States and Year as random-effects** parameters. The coefficient of determination (R^2), Aka lke information criterion (AIC), Bayesian Information Criterion (BIC), Loglikelihood were used as the criteria to determine the best model which provide the best combination of agro (cultivated land) and remotely sensed predictors.

4.2 EFFECTS OF REMOTE SENSING DATA AND OTHER COVARIATES ON SORGHUM YIELD IN SELECTED SITES

4.2.1 Descriptive statistics of field data

The average cultivated land during the two agricultural seasons was 0.674 ± 0.076 and the self-declared sorghum yield average was 453.75 ± 49.21 . The maximum yield obtained was obtained in Upper State and was at $2,345 \text{ kg.ha}^{-1}$ with a maximum cultivation of 3 ha.

	Cultivated land (ha)		Sorghum yield kg.ha^{-1}	
	2018	2019	2018	2019
Upper Nile	0.673 ± 0.075	0.817 ± 0.12	453.75 ± 0.075	565.23 ± 78.68
Western B Gazal	0.736 ± 0.095	0.673 ± 0.075	506.22 ± 60.83	453.73 ± 49.21

Table 1. Average cultivated land and sorghum yield for the two agricultural seasons

Parameter	House_hold Size	Cultivated land (ha)	Sorghum Harvest (Kg/ha)
Mean	9.419	0.674	453.755
Standard Error	0.276	0.038	24.980
Median	9	0.5	380
Standard Deviation	4.217	0.590	382.942
Kurtosis	2.902	5.134	5.171
Skewness	1.147	2.329	2.038
Minimum	1	0.2	17.5
Maximum	29	3	2345
Count	234	235	235
Confidence Level(95%)	0.543	0.076	49.215

Figure 1. Table 1: statistic description of the field data.

4.2.2 Remotely derived data

	NDVI	EVI	Precipitation	FPAR	LAI	Evapotranspir	SoilMoisture
median	0.462	0.243	2.072	0.480	1.500	1.655e+01	0.256
mean	0.475	0.278	2.519	7.715	8.844	9.273e+02	0.270
SE.mean	0.0046	0.003	0.054	0.922	0.918	1.268e+02	0.003
CI.mean.0.95	0.009	0.006	0.107	1.808	1.800	2.48e+02	0.007
var	0.045	0.024	6.089	1784.894	1769.764	2.897e+07	0.011
std.dev	0.212	0.154	2.467	42.248	42.068	5.383e+03	0.103
coef.var	0.447	0.554	0.979	5.476	4.757	5.804e+00	0.379

Table 2. Remotely derived parameters description

4.2.3 Checking for collinearity between different parameters

Data indicate a positive high correlation between sorghum harvest and cultivated land (0.884). In addition, strong positive correlation is found between NDVI and EVI (0.966) and between precipitation and EVI(0.702). NDVI was also highly correlated to precipitation (0.682). NDVI was also highly correlated to precipitation (0.682)

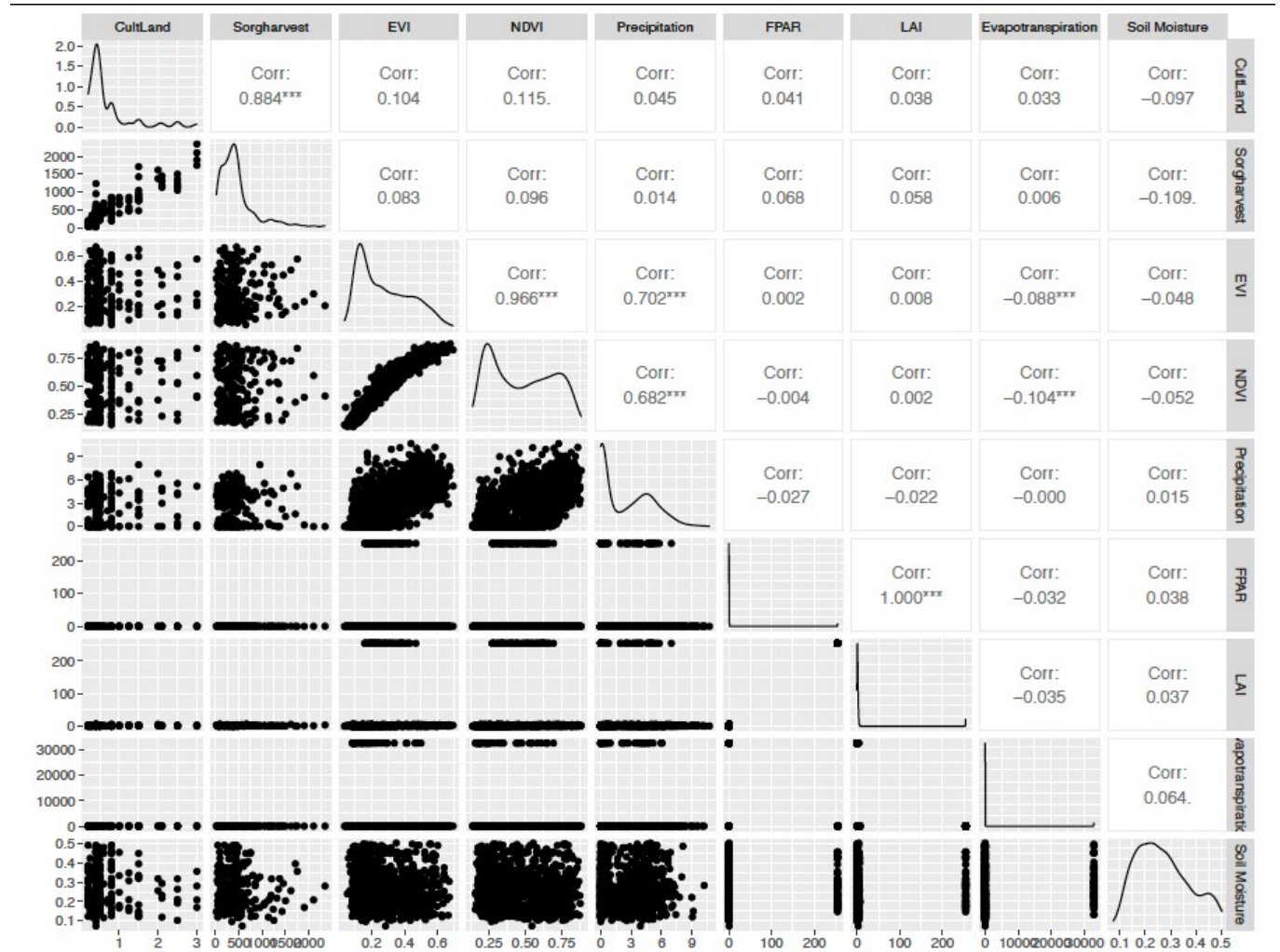


Figure 2. Collinearity check between field data and remotely sensed data

4.2.4 Sorghum harvest analysis during the two agricultural seasons

Higher yields were observed in Upper Nile state compare to Western Bahr El Gazal. In addition, the agriculture year 2019 was better with the maximum yield compared to year 2018 for both states.

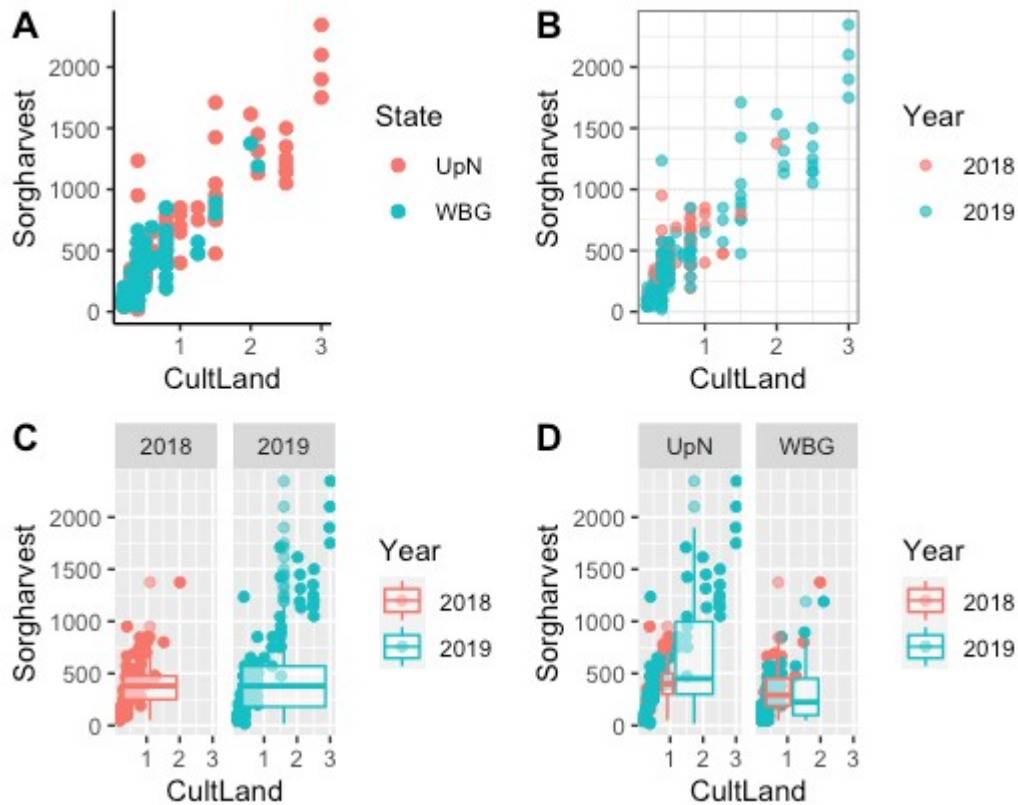


Figure 3. Sorghum yield plotted against cultivated land for the two agricultural seasons

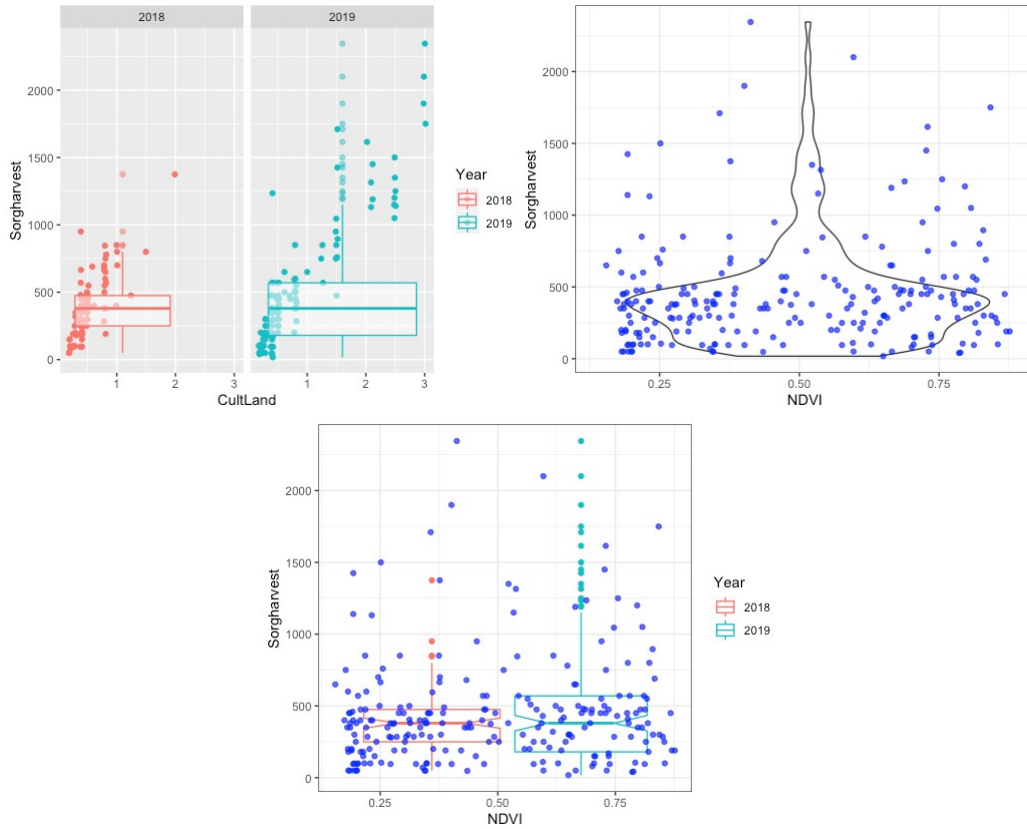


Figure 4. sorghum yield plotted against NDVI

The beeswarm plot in Figure 5 illustrates well that the majority of sorghum data fall between quartile 1 and quartile 3 and majority of people produced less than $500kg.ha^{-1}$ below the average sorghum yield in Sub-Saharan countries which is around $800kg.ha^{-1}$ (FAOSTAT, 2018)

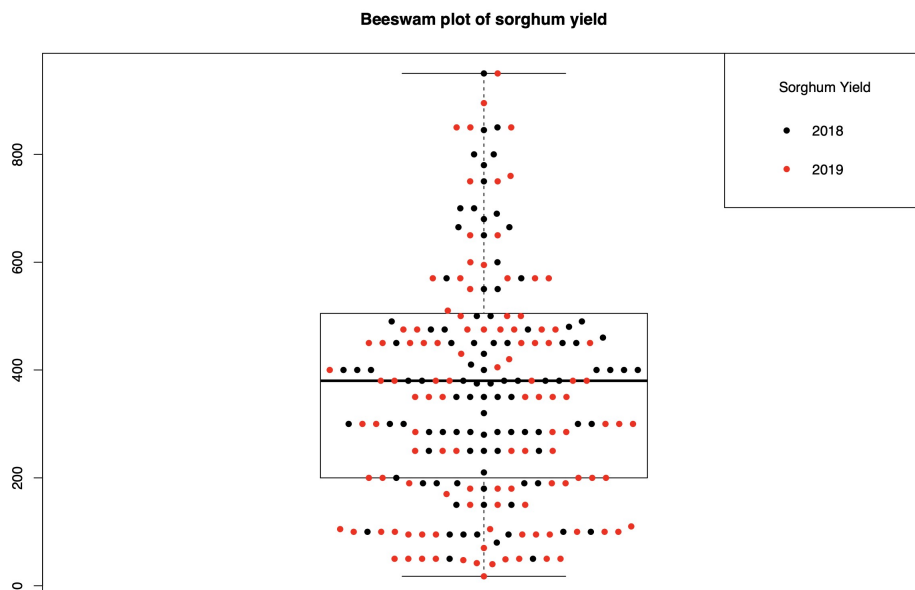


Figure 5. Sorghum yield during the two agricultural seasons

4.3 Predicting the effects of remote sensing data on sorghum yield

The combination of NDVI and Cultivated land provided a better estimate of sorghum yield than the NDVI or EVI on their own. There was a strong linear regression relationship between sorghum yield and cultivated land ($R^2 = 0.781, p = 2.2e - 16$). In addition, the association between cultivated land and NDVI improved the model ($R^2 = 0.786$). All models that included cultivated land and NDVI as ones of the predictors were significant ($R^2 > 0.78$ and $RMSE < 180.6kg.ha^{-1}, p = 2.2.e - 16$). Soil moisture was negatively related to sorghum yield.

It was observed that the addition of other remote sensed predictors did not improve the model $M_6 < - Sorgharvest \sim NDVI *Cultland$.

	Model equation	R^2	$RMSEkg.ha^{-1}$	p value
M_1	$Y = 165.07 * NDVI + 373.68$	0.00923	382	0.142
M_2	$Y = 195.2 * EVI + 398.93$	0.00687	382.4	0.205
M_3	$Y = 574 * Cultland + 66.86$	0.781	179.2	2.2e-16
M_4	$Y = 154 * NDVI - 385 * Soil - Moisture + 485.21$	0.02007	380.7	0.0952
M_5	$Y = 10.36LAI + 149.11 * NDVI + 360.01$	0.01031	382.6	0.3004
M_6	$Y = 116.2 * NDVI + 679.4 * Cultland + 194.5 * NDVI : Cultland + 6.24$	0.786	178.3	2.2e-16**
M_7	$Y = 77.269.37 * NDVI + 574.9 * Cultland + 0.61 * HHSize$	0.782	180.3	2.2e-16
M_8	$Y = 105.55 + 35.23 * NDVI + 77.36 * SoilMoisture + 573.36 * Cultland + 7.26 * Precipitation + 0.79 * Evapotranspiration$	0.784	179.9	2.2e-166
M_9	$Y = 100.06 + 573.5 * Cultland + 0.77 * HHsize + 25.55 * NDVI + 9.01 * LAI + 7.79 * Precipitation + 75.08 * Soilmoisture + 0.79 * Evapotranspiration$	0.785	180.6	2.2e-16
Y= Sorghum yield ** best model selected based on high R^2 and small RMSE				

Table 3. Different model fitted for sorghum yield

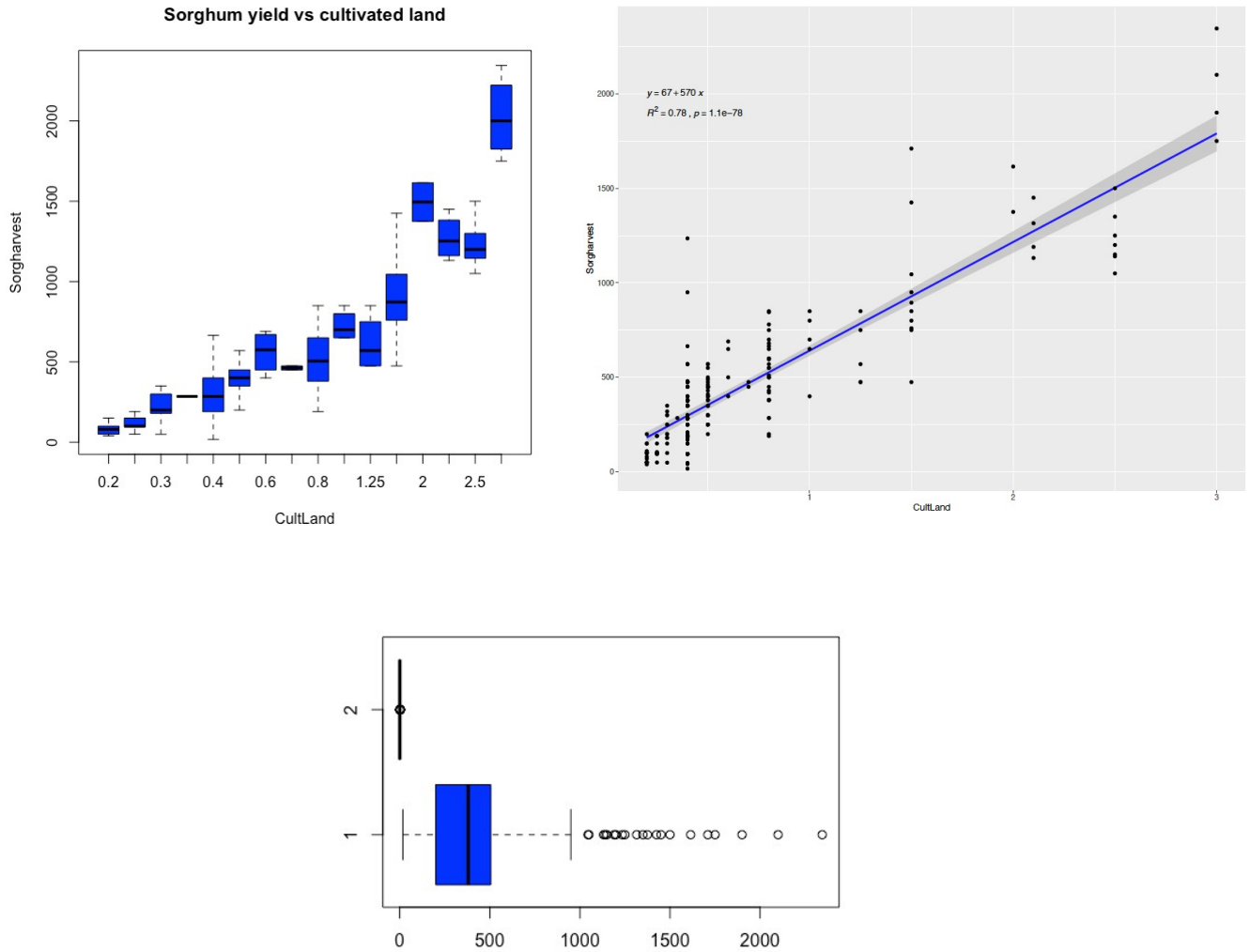


Figure 6. sorghum yield vs cultivated land

Model(M6) $Ssm6 <- lm(Sorgharves \sim NDVI * CultLand)$

Coefficients:				
	Estimate	Std. Error	t value	$Pr(> t)$
(Intercept)	6.242	43.064	0.145	0.885
NDVI	116.204	79.968	1.453	0.148
CultLand	679.361	53.671	12.658	$< 2e - 16$ ***
NDVI:CultLand	-194.495	92.713	-2.098	0.037 *

Table 4. Table of coefficient for Model (M6)

Model (M_6) assumption check with residuals ($Y = NDVI * Cultland$)

$$Y = 116.2 * NDVI + 679.4 * Cultland + 194.5 * NDVI : Cultland + 6.24$$

The model M_6 was better fit with a higher $R^2 = 0.786$ and a lower $RMSE = 178.3 kg.ha^{-1}$. The normality assumption does not fully hold for the residuals. The qq-plot of the residuals versus a normal distribution shows some skewness from the normality assumption. This is due to values in the qq-plot that deviate from the qq-line at the upper end of the graph and could be attributed to the outliers in data. (Figure 4.7)

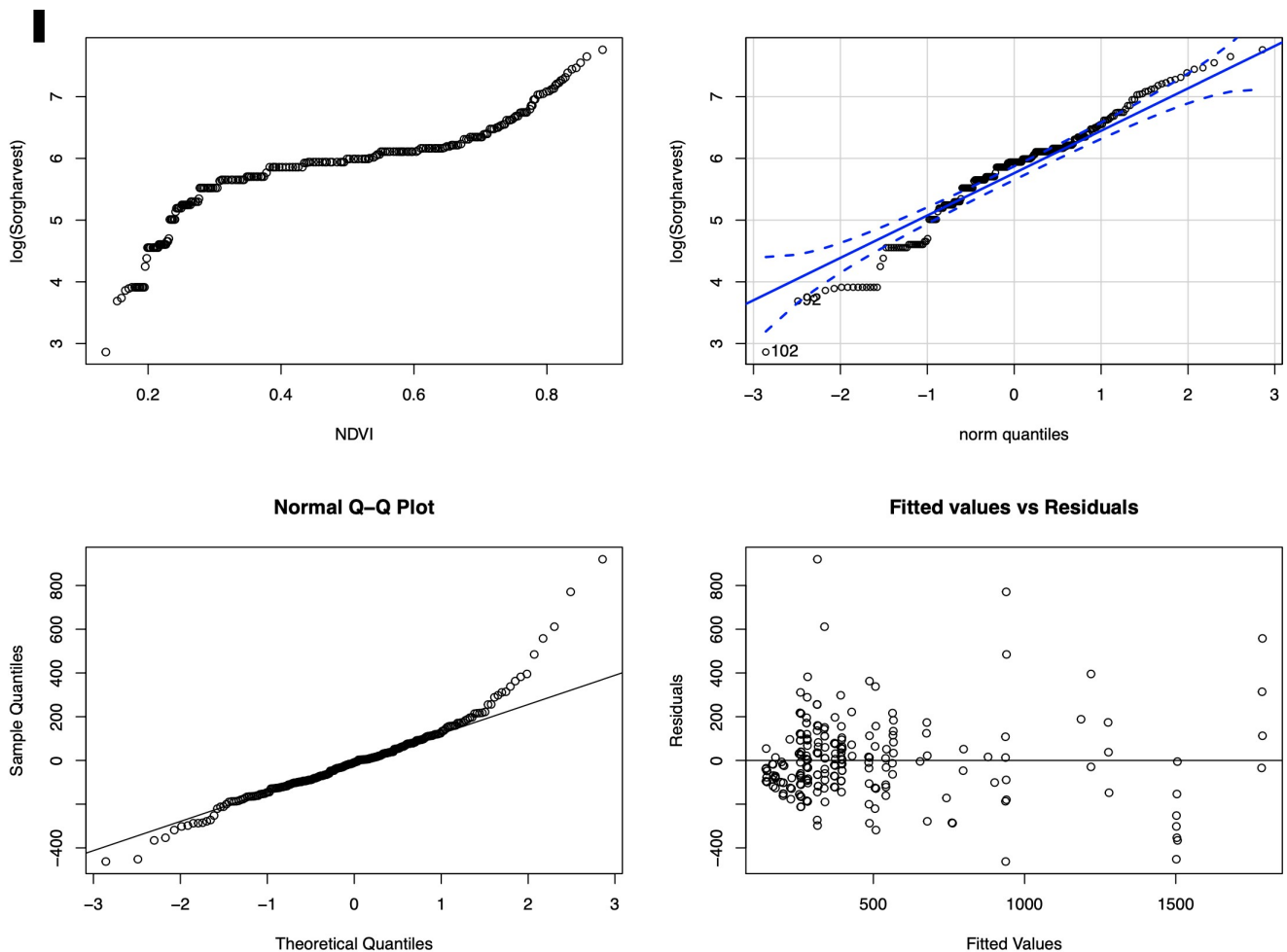


Figure 7. Model assumptions check with residuals

Observed sorghum yield vs predicted. ($R^2 = 0.786$)

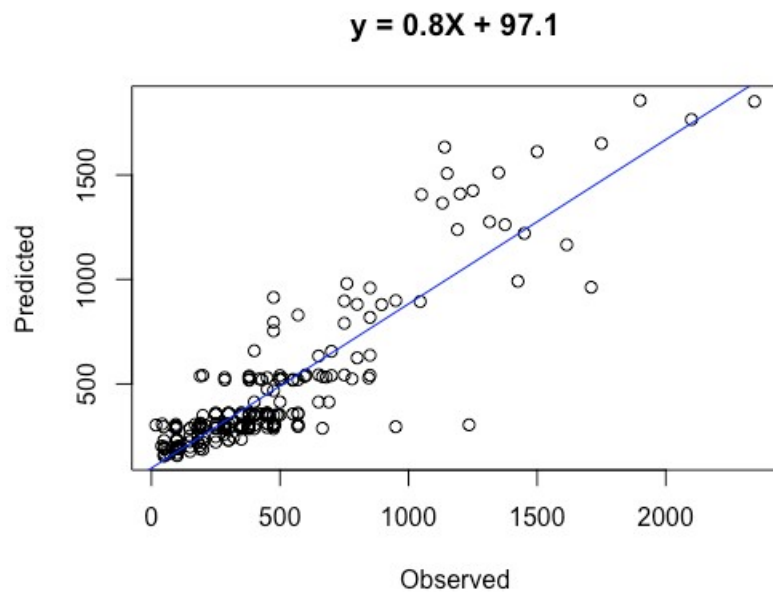


Figure 8. Predicted Yield plotted against observed yield

Linear regression relationships between all remote sensed variables were tested and none of them alone was significant as yield predictor of self-reported sorghum yield.

Compare to other models, cultivated land was the best predictor with a significantly higher accuracy of the estimation of sorghum yield. There was a strong linear regression between cultivated land and sorghum yield ($R^2 = 0.78$ and $p = 2.2e - 16$). The LAI was highly related to NDVI ($p = 2.9e - 05$) and FPAR also significantly related to LAI ($p = 8.6e - 100$)

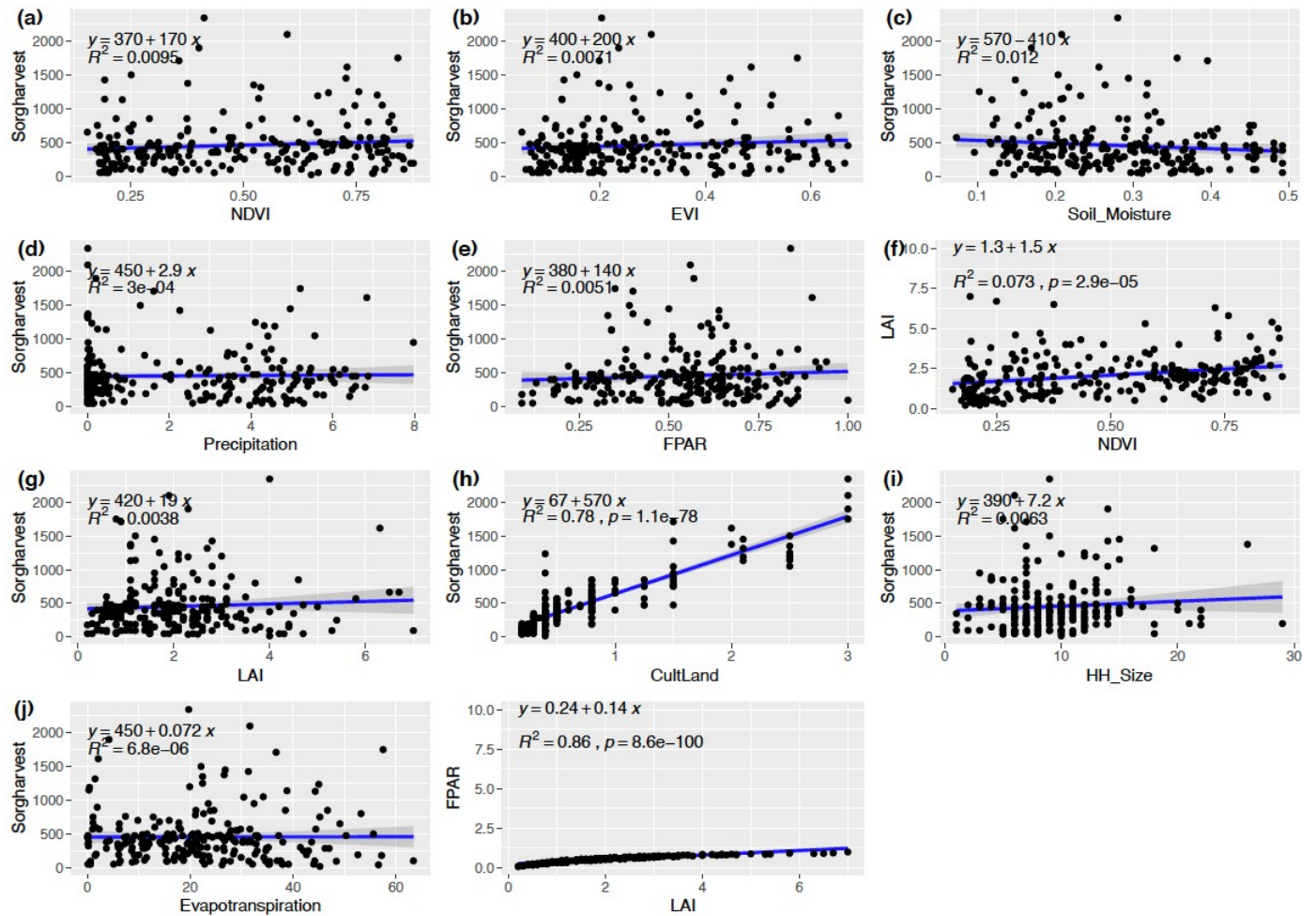


Figure 9. Sorghum yield plotted against different parameters: (a)NDVI, (b)EVI, (c)Soil-Moisture, (d)Precipitation, (e)FPAR,(f)NDVI vs LAI, (g)LAI, (h) Cultland, (i) HH-Size, (j) Evapotranspiration, (k) LAI vs FPAR

4.4 Linear Mixed-Effects Model in predicting sorghum yield variability

The LME models were executed using lme4 package in R software.
 library(lme4) # for linear Mixed-effects models using lmer function.
 library(merTools) # confidence interval of predictors in LME.
 library(MuMIn) # to have R^2
 library(lmerTest)# provide p-value for fixed-effects predictors

Several models were constructed starting from the full model with all seven remotely sensed predictors as fixed-effect models and the agricultural season (year) and State were used as random-effects variables. It was important to identify the appropriate model which explains the maximum variations while keeping the cultivated land as the main sorghum predictor variable in the logic of an extensive farming system.

4.4.1 Random-effects structures

Analyzing sorghum yield variability in the two states during the two agricultural seasons and considering all combinations of random-effects given the assumption about spatial and temporal variability, two (2) different structures were developed to capture the random-effects of the State (spatial random-effects) and the Year (temporal random-effects variations in each agricultural season).

Random-effect structure (1) with random intercepts

for factors State and Year: $\sim (1|State) + (1|Year)$.

Model summary	R^2	RMSE	AIC	BIC	Loglik
S11<-lmer(Sorgharvest~CultLand+NDVI+(1 Year) + (1 State))	0.784	175.45	3100.0	3120.8	-1544.0
S12<-lmer(Sorgharvest~CultLand+LAI + (1 Year) + (1 State))	0.785	174.98	3099.1	3119.8	-1543.5*
S13<-lmer(Sorgharvest~CultLand+HH-Size+NDVI+LAI+Precipitation+Soil-Moisture+Evapotranspiration+ (1 Year) + (1 State))	0.787	173.88	3106.6	3144.6	-1542.3*
S14<-lmer(Sorgharvest~CultLand + NDVI + LAI + EVI + Soil-Moisture + Precipitation + Evapotranspiration + (1 State) + (1 Year))	0.787	174.18	3107.0	3145.0	-1542.5
S18a<-lmer(Sorgharvest~CultLand + NDVI + Precipitation + (1 State) + (1 Year))	0.785	175.07	3101.3	3125.5	-1543.7
S19a<-lmer(Sorgharvest~CultLand+ NDVI + Precipitation + LAI + EVI + (1 Year) + (1 State))	0.786	174.4	3103.8	3134.9	-1542.9
S22<-lmer(Sorgharvest~NDVI + LAI + Soil-Moisture + Precipitation + Evapotranspiration + (1 Year) + (1 State))	0.084	359.5	3444.7	3475.0	-1713.4
*Model is significant.					

Table 5. Different fixed-effect models comparison with random – intercept structure (1) $\sim (1|State) + (1|Year)$

All models with cultivated land as fixed-effect predictor had a higher coefficient of determination ($R^2 > 0.784$) as opposed to the model with only remotely derived predictors which shows a very high residual mean square error of $359.5kg.ha^{-1}$. For this random-structure, the model S13 is the best fit for it has high $R^2 = 0.787$ and smaller Akaike Information criterion (AIC= 3106) and relatively small $RMSE = 173.88kg.ha^{-1}$ [55]

The analysis of variance for all mixed-effects models in above table is summarized in the below table:

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	$Pr(> Chisq)$
S11	6	3100.1	3120.8	-1544.0	3088.1			
S12	6	3099.1	3119.8	-1543.5	3087.1	1.0164	0	$< 2e - 16 ***$
S18a	7	3101.3	3125.5	-1543.7	3087.3	0.0000	1	1.0000
S19a	9	3103.8	3134.9	-1542.9	3085.8	1.5092	2	0.4702
S22	9	3444.7	3475.8	-1713.4	3426.7	0.0000	0	1.0000
S13	11	3106.6	3144.7	-1542.3	3084.6	342.0951	2	$< 2e - 16 ***$
S14	11	3107.0	3145.0	-1542.5	3085.0	0.0000	0	1.0000

Table 6. Anova results for all models with random-intercept structure (1): $\sim (1|State) + (1|Year)$

The Anova analysis reveals that models S12 and S13 are statistically significant ($p = 2e - 16$) for this first random-intercept structure. From these outputs, it is clear the models S12 and S13 are better than all the other and their difference are highly significant.

Random effects for model S13			
Groups	Name	Variance	Std. Dev
State	Intercept	897.2	29.95
Year	Intercept	419.5	20.48
Residual		30529.5	174.71

Table 7. Summary of random effects results for model S13

Model S13 predict a yield variability of $897.3kg.ha^{-1}$ at state level as compare to the annual variations which is predicted at $419.5kg.ha^{-1}$. There was a much lower yield variability predicted by model S12 at seasonal level ($241kg.ha^{-1}$) as compared to S13.

Fixed effects of Model S13					
	Estimate	Std. Error	df	t value	$Pr(> t)$
(Intercept)	98.87	63.44	43.79	1.55	0.126
CultLand	568.46.	20.40	227.25	27.85	$< 2e - 17 ***$
HH-Size	-1.98	2.82	228.70	-0.70	0.484
NDVI	29.05	68.76	232.45	0.42	0.63
LAI	10.34	9.61	230.85	1.07	0.283
Precipitation	-0.80	6.84	232.53	-1.18	0.238
Soil-Moisture	-36.63	116.07	224.78	-0.31	0.753
Evapotranspiration	-0.73	0.86.	231.18	-0.84	0.397

Table 8. Summary of fixed effect results for model S13

Random effects for model S12			
Groups	Name	Variance	Std. Dev
State	Intercept	241.0	15.53
Year	Intercept	934.6	30.58
Residual		30891.3	175.76

Table 9. Random effects results for model S12

Model S12 predict much higher variability at state level ($934.6\text{kg}\cdot\text{ha}^{-1}$) as compare to the seasonal variation ($241\text{kg}\cdot\text{ha}^{-1}$)

Fixed effects: Model S12					
	Estimate	Std. Error	df	t value	$Pr(> t)$
(Intercept)	53.66	35.069	5.530	1.530	0.181
CultLand	564.87	20.25	222.37	27.889	$< 2e - 16 ***$
LAI	9.34	9.15	231.08	1.021	0.308

Table 10. Fixed effects summary results for model S12

Model diagnostic (S12) – check for assumption

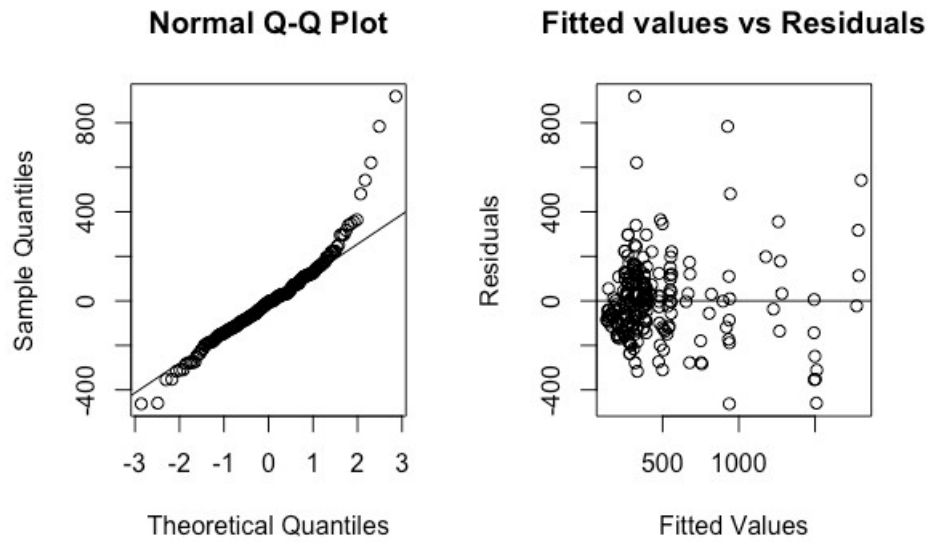


Figure 10. model S12 assumption check with residuals

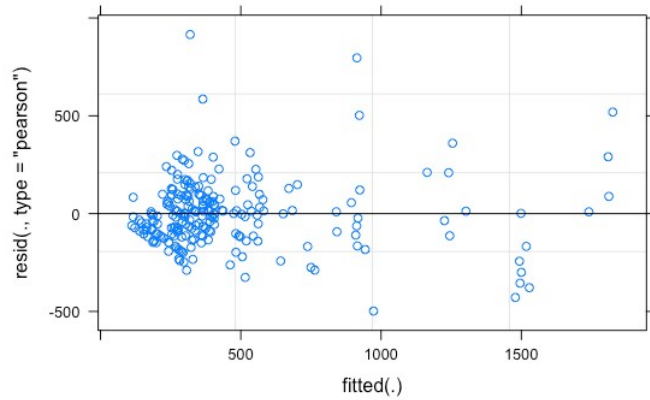


Figure 11. Model S13 assumption check

Random-effect structure (2) with random-intercept including variance-covariance matrix: $\sim (1|State) + (1|Year) + (1|State : Year)$

Model summary	R^2	RMSE	AIC	BIC	Loglik
S11a<-lmer(Sorgharvest~CultLand + NDVI +(1 State) + (1 Year) + (1 State : Year))	0.784	175.45	3102.1	3126.3	-1544.0
S11b<-lmer(Sorgharvest~CultLand*NDVI +(1 State) + (1 Year) + (1 State : Year))	0.788	173.85	3099.8	3127.5	-1541.9*
S12a<-lmer(Sorgharvest~CultLand + LAI + (1 State) + (1 Year) + (1 State : Year))	0.785	174.98	3101.1	3125.2	-1543.5
S13a<-lmer(Sorgharvest~CultLand + HH-Size + NDVI + LAI + Precipitation + Soil-Moisture + Evapotranspiration + (1 Year) + (1 State) + (1 State : Year))	0.787	173.88	3108.6	3150.1	-1542.3
S14<-lmer(Sorgharvest~CultLand + NDVI + LAI + EVI + Soil-Moisture + Precipitation + Evapotranspiration +(1 State) + (1 Year) + (1 State : Year))	0.787	174.18	3109.0	3150.5	-1542.5
S18<-lmer(Sorgharvest~CultLand + NDVI + Precipitation + (1 State) + (1 Year) + (1 State : Year))	0.785	175.07	3103.3	3131.0	-1543.7
S19<-lmer(Sorgharvest~CultLand + NDVI + Precipitation + LAI + (1 Year) + (1 State) + (1 State : Year))	0.786	174.42	3103.8	3134.9	-1542.9
S19b<-lmer(Sorgharvest~CultLand + Precipitation + LAI + EVI + (1 State) + (1 Year) + (1 State : Year))	0.786	174.45	3103.9	3135.0	-1542.9
S20<-lmer(Sorgharvest~CultLand + NDVI + LAI + EVI + FPAR + Soil-Moisture + Precipitation + Evapotranspiration + (1 State) + (1 Year) + (1 State : Year))	0.787	174.14	3110	3155.6	-1542.3
* Good model based on high R2, low RMSE and low AIC					

Table 11. Different-effects models comparison with random-intercept structure (2):
 $\sim (1|State) + (1|Year) + (1|State : Year)$

The analyse of variance of all crossed models in below table suggests Model S11b is a fairly good model with a higher R^2 (0.788) and a smaller AIC (3099.8).

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	$Pr(> Chisq)$
S11a	7	3102.1	3126.3	-1544.0	3088.1			
S12a	7	3101.1	3125.2	-1543.5	3087.1	1.0164	0	
S11b	8	3099.8	3127.4	-1541.9	3083.8	3.2464	1	0.07158.
S18	8	3103.3	3131.0	-1543.7	3087.3	0.0000	0	
S19	9	3103.8	3134.9	-1542.9	3085.8	1.4946	1	0.22151
S19b	9	3103.9	3134.9	-152.9	3085.9	0.0000	0	
S14a	11	3107.1	3145.1	-1542.6	3085.1	0.7397	2	0.69082
S13a	12	3108.7	3150.1	-1542.3	3084.7	0.4680	1	0.49390
S20	13	3110.6	3155.6	-1542.3	3084.6	0.0054	1	0.94152

Table 12. Anova results for all models with random-intercepts structure (2):
 $\sim (1|State) + (1|Year) + (1|State : Year)$

Tables below summarize the significant model S11b

Random effects for model S11b			
Groups	Name	Variance	Std. Dev
State	Intercept	238.6	15.45
Year	Intercept	811.1	28.48
Residual		30489.0	174

Table 13. Random effects results for model S11b

Comparing Model S11b, S12 and S13, the Model S11b predicts a much lower yield variability ($811.1kg.ha^{-1}$) at state level as opposed to model S12 ($934kg.ha^{-1}$) and model S13($897.2kg.ha^{-1}$). The lower annual(seasonal) random effects was predicted by model S11b($283.6kg.ha^{-1}$) followed by model S12($241kg.ha^{-1}$) and then model S13 ($419.5kg.ha^{-1}$).

Fixed effects: Model S11b					
	Estimate	Std. Error	df	t value	$Pr(> t)$
(Intercept)	11.76	48.22	21.98	0.244	0.8095
CultLand	668.37	53.06	233.67	12.597	$< 2e - 16^{***}$
NDVI	115.64	78.53	231.90	1.473	0.1422
CultLand:NDVI	-188.60	90.88	230.77	-2.075	0.0391*

Table 14. Fixed effects results for model S11b

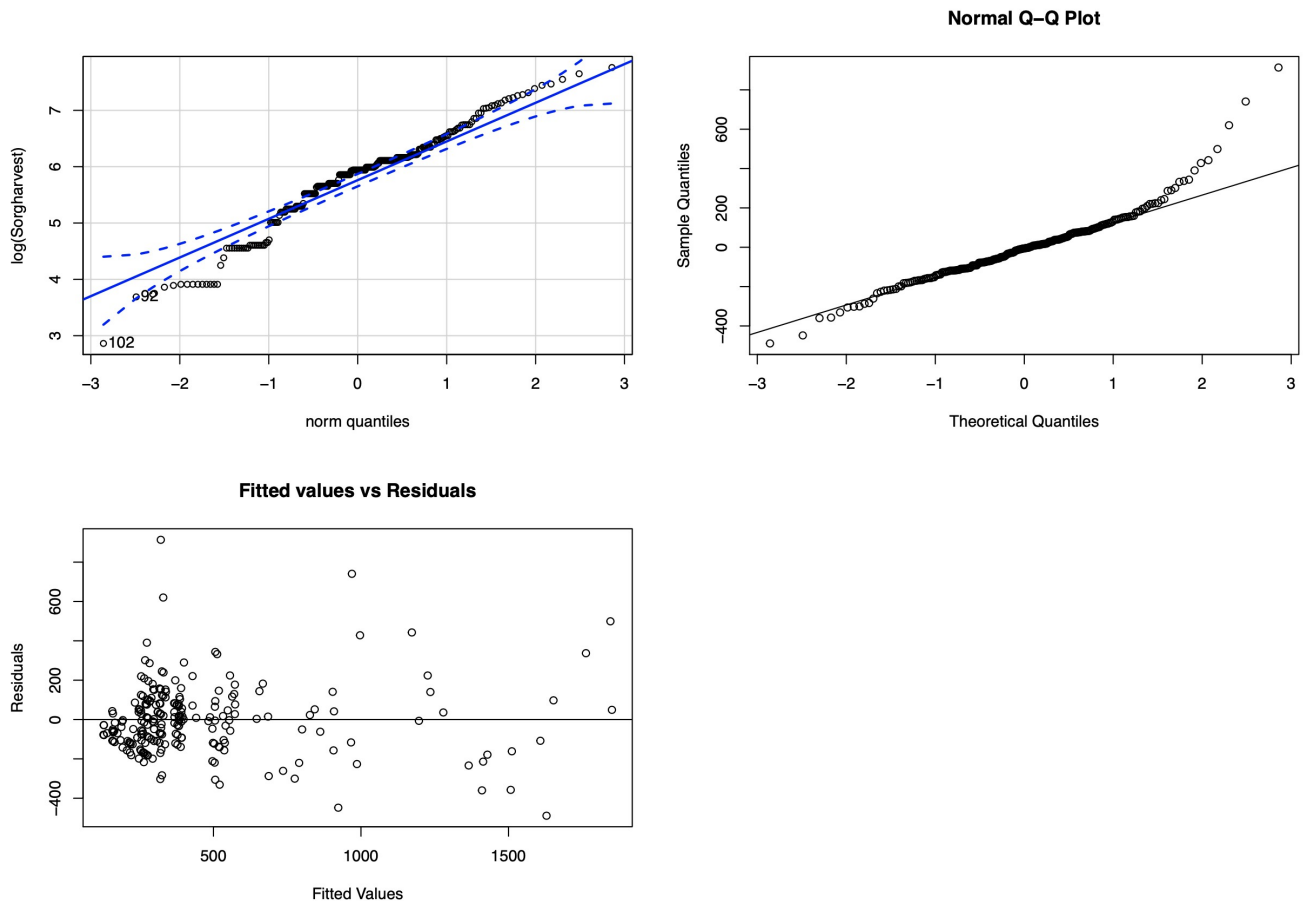
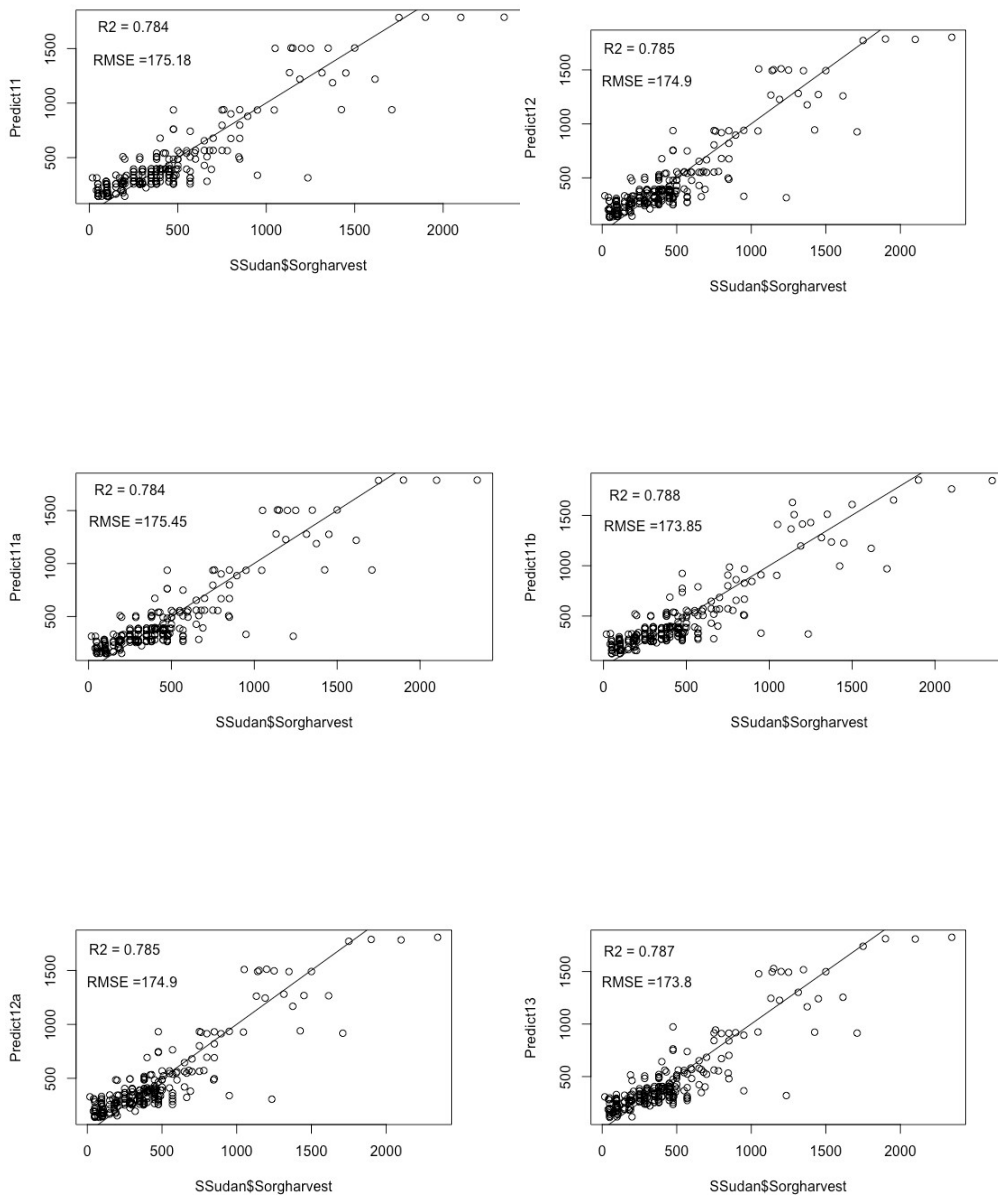
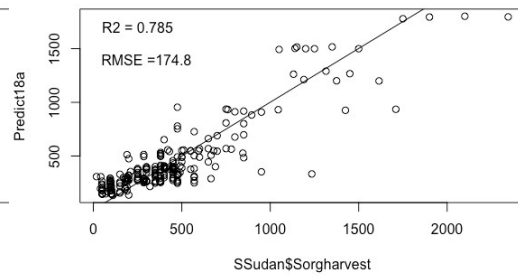
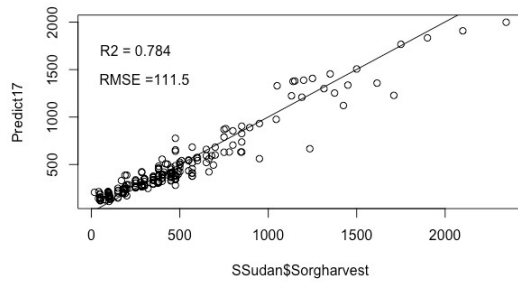
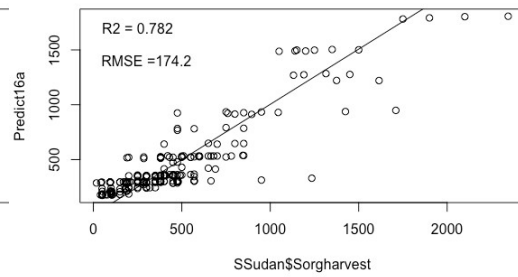
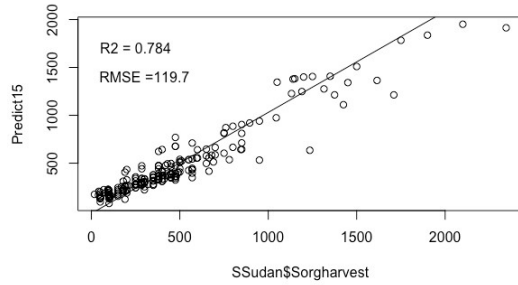
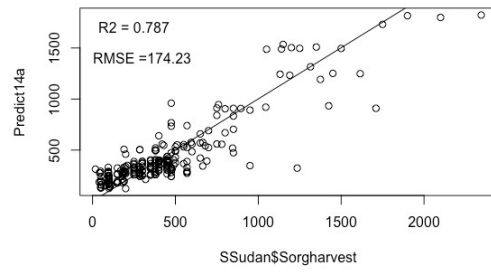
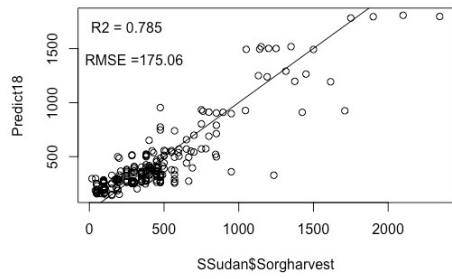
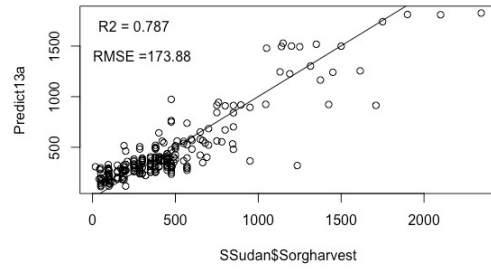
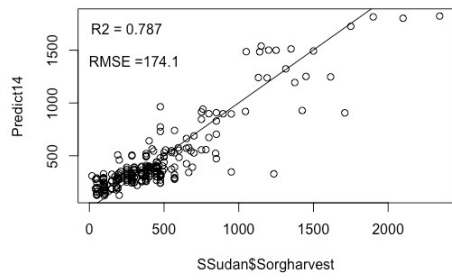


Figure 12. Model S11b diagnostic

Comparison of predicted sorghum yield with different random-effects structures

Below is a series of figures showing the scatter diagrams sorghum yield values versus predicted values from different type of LME models, which were constructed with different random effects structures. From the overall performance, the model S11b achieved the highest accuracy using the cultivated land and all the remotely sensed predictor as fixed-effects predictors. This model yielded more accurate estimation results ($R^2 = 0.788$) and ($RMSE = 173.85kg.ha^{-1}$) follow by model S12 that used cultivated land and LAI as fixed effects predictors($R^2 = 0.785$) and ($RMSE = 174.9kg.ha^{-1}$) these two models used State and Year in the random-effects structures. Model S13 had a coefficient of determination ($R^2 = 0.787$) and ($RMSE = 173.80kg.ha^{-1}$).





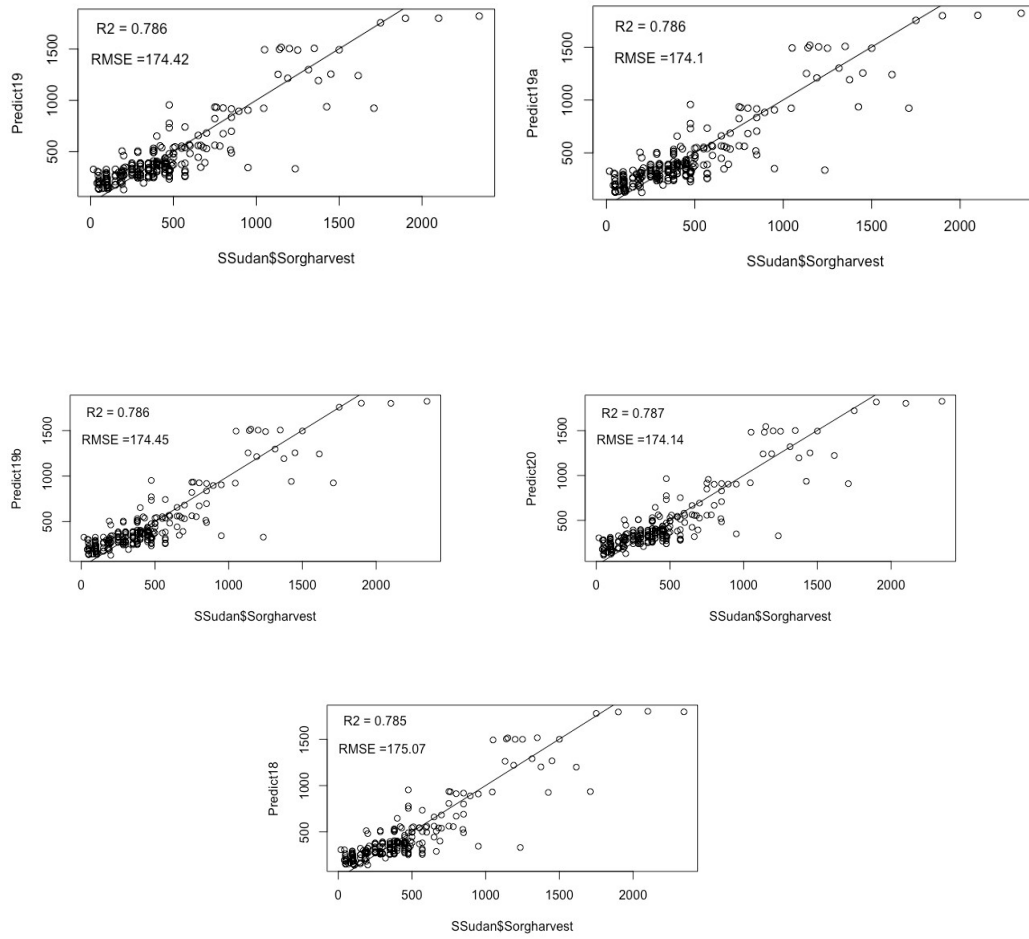


Figure 13. Comparison of predicted sorghum yield with different random-effects structures.

Plotting random slopes for the significant models (S12, S13 and S11b)

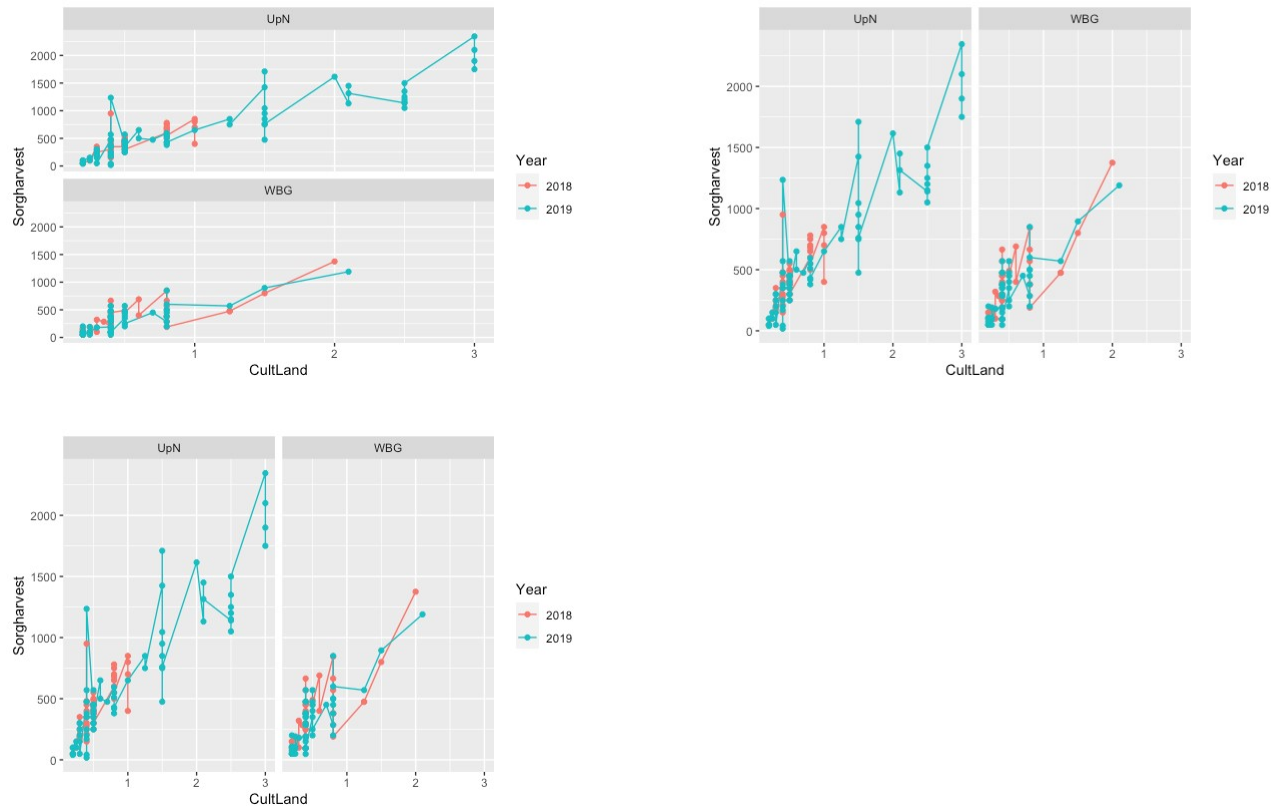


Figure 14. Plot of random slopes for three significant models S11b, S12 and S13.

It is noticeable that significant variation in sorghum yield is predicted for Upper Nile state as opposed to Western Bahr El gazal and that for the two years the two lines are clearly different suggesting seasonal variability as well.

4.5 Conclusion and Recommendation

4.5.1 Results discussion and Conclusion

For the humanitarian organizations working in South Sudan as well as for the government, understanding the crop yield variability between different states and between agricultural seasons is extremely important for humanitarian planning purposes as well as for government to better meet farmers needs. Remote sensing data have been widely used to model crop yield simulation and forecasting. In this study remote sensing data were used to compare sorghum yield variability in two states (Upper Nile and Western Bahr El gazal) during two agricultural seasons (2018-2019). Using linear mixed-effects model approach provided a great tool to understanding sorghum yield variations between the two producing states.

Given the unbalanced nature, the repeated measures on same statistical units for remotely derived parameters and the longitudinal nature of the data for this study, Linear Mixed-Effects model (LME) was chosen and appropriate for statistical analysis. The random-effects in this study were used to describe the spatio (state) and temporal (inter-annual) specific variations of the sorghum yield during the two agricultural seasons (2018-2019).

The LME models with more explanatory remotely derived variables was poorer than the optimum models (*S12, S11b*), which included only a few predictors. *S11b* : $R^2 = 0.788$, $RMSE = 173.85kg.ha^{-1}$, *S12* : $R^2 = 0.785$, $RMSE = 174.98kg.ha^{-1}$ as compare to the full model *S20* . Furthermore, the 9 constructed and independent models (*S11a, S12a, S11b, S18, S19, S19b, S14a, S13a, S20*) with the second random-effects structures in table (12) had an average $R^2 = 0.786$ with average $RMSE$ of $174.49kg.ha^{-1}$ while for the 7 models build with second random-effects structure (table6) the average coefficient of determination was much lower ($R^2 = 0.685$) and with a higher average $RMSE=201.06kg.ha^{-1}$.

Comparing Model *S11b, S12* and *S13*, the Model *S13* predicted a much higher sorghum yield variability ($897.2kg.ha^{-1}$) at state level as opposed to model *S12* ($241.0kg.ha^{-1}$) and model *S11b* ($238.6kg.ha^{-1}$). In addition, higher annual (seasonal) random effects was predicted by model *S12*($934.6kg.ha^{-1}$) and *S11b*($811.1kg.ha^{-1}$) and then model *S13* ($419.5kg.ha^{-1}$); therefore model *S12* and *S13* would be better choices to understanding the sorghum yield variability in self-reported data and in the context of South Sudan.

Cultivated land had a significant effect on sorghum yield variability in all the models that were considered, for instance model *M4* had a p -value: $0.095(p>0.05)$ without cultivated land as a predictor and model *M6* had a p -value of $2.2e - 16(p < 0.05)$ when cultivated land was included. Combining NDVI or EVI with cultivated land had a significant effect on model *M6* (Interaction NDVI: Cult land, $p - value = 0.037(p < 0.005)$).

In addition, in LME, the models which did not include cultivated land as fixed effect parameter were not significant. This means remote sensing parameters taken alone as fixed predictors were not significant in the context of this study.

In conclusion, this study results show that Linear Mixed-Effects (LME) models not only offer advantages in understanding the appropriateness of sorghum yield modeling using remote sensing data; they also provide a good insight in yield variations between the two states as well as between the two agricultural seasons. For application of LME models in context of this study, using fewer remotely derived factors is promising. In the context of South Sudan modeling crop yield by using remote sensing data seems to be a viable option to assess yield variability.

4.5.2 Recommendation

This study assessed the potential of using LME models to study sorghum yield variability in South Sudan using farmers self-reported data and remote sensing. For greater precision, and given the need for the country in term of cereal yield prediction and forecasting, we would recommend following:

- There would be need to validate optimum models from this study with future sorghum yield in South Sudan or relevant statistical yield when they are available.
- More precise on-farm field research in cereal crop modeling using remote sensing data would be needed for cereal yield modeling and forecasting:
 - Explore the possibility of crop simulation combining with mathematical modeling for South Sudan context.
 - Consider using UAV (drones) in field research when security allows for better understanding of crop yield parameters mixed with remote sensing data and crop yield prediction in South Sudan'
- Model the effects of insecurity and conflict on cereal yield as result of poorly tended sorghum crops

REFERENCES

- [1] JK Aase and FH Siddoway. Assessing winter wheat dry matter production via spectral reflectance measurements. *Remote Sensing of Environment*, 11:267–277, 1981.
- [2] Etienne Bartholome. Radiometric measurements and crop yield forecasting some observations over millet and sorghum experimental plots in mali. *International journal of remote sensing*, 9(10-11):1539–1552, 1988.
- [3] Roberto Benedetti and Paolo Rossini. On the use of ndvi profiles as a tool for agricultural statistics: the case study of wheat yield estimate and forecast in emilia romagna. *Remote Sensing of Environment*, 45(3):311–326, 1993.
- [4] Awetahegn Niguse Beyene, Hongwei Zeng, Bingfang Wu, Liang Zhu, Tesfay Gebretsadkan Gebremicael, Miao Zhang, and Temesgen Bezabh. Coupling remote sensing and crop growth model to estimate national wheat yield in ethiopia. *Big Earth Data*, pages 1–18, 2021.
- [5] Douglas K Bolton and Mark A Friedl. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013.
- [6] G Cai, Y Xue, Y Hu, Y Wang, J Guo, Y Luo, C Wu, S Zhong, and S Qi. Soil moisture retrieval from modis data in northern china plain using thermal inertia model. *International Journal of Remote Sensing*, 28(16):3567–3581, 2007.
- [7] Jan GPW Clevers, Lammert Kooistra, and Marnix MM Van den Brande. Using sentinel-2 data for retrieving lai and leaf and canopy chlorophyll content of a potato crop. *Remote Sensing*, 9(5):405, 2017.
- [8] Arthur P Cracknell. *Advanced very high resolution radiometer AVHRR*. CRC Press, 1997.
- [9] K Didan. Mod13a3 modis/terra vegetation indices monthly l3 global 1km sin grid v006. *NASA EOSDIS Land Processes DAAC*, 10, 2015.
- [10] K Didan. Mod13a3 modis/terra vegetation indices monthly l3 global 1km sin grid v006 modis/terra vegetation indices 16-day l3 global 1km sin grid v061 [data set]. *NASA EOSDIS Land Processes DAAC*, 10, 2021.
- [11] Paul C Doraiswamy and Paul W Cook. Spring wheat yield assessment using noaa avhrr data. *Canadian Journal of Remote Sensing*, 21(1):43–51, 1995.
- [12] Jay Fussell, Donald RUNDQUIST, and JA Harrington. On defining remote sensing. *Photogrammetric Engineering and Remote Sensing*, 52(9):1507–1511, 1986.

- [13] Kefyalew Girma, KL Martin, RH Anderson, DB Arnall, KD Brixey, MA Casillas, B Chung, BC Dobey, SK Kamenidou, SK Kariuki, et al. Mid-season prediction of wheat-grain yield potential using plant, soil, and sensor measurements. *Journal of Plant Nutrition*, 29(5):873–897, 2006.
- [14] SME Groten. Ndvi—crop monitoring and early yield assessment of burkina faso. *Remote Sensing*, 14(8):1495–1515, 1993.
- [15] Noemi Guindin-Garcia, Anatoly A Gitelson, Timothy J Arkebauer, John Shanahan, and Albert Weiss. An evaluation of modis 8-and 16-day composite products for monitoring maize green leaf area index. *Agricultural and Forest Meteorology*, 161:15–25, 2012.
- [16] Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):346–363, 2008.
- [17] Jingfeng Huang, Xiuzhen Wang, Xinxing Li, Hanqin Tian, and Zhuokun Pan. Remotely sensed rice yield prediction using multi-temporal ndvi data derived from noaa’s-avhrr. *PloS one*, 8(8):e70816, 2013.
- [18] SB Idso, RD Jackson, PJ Pinter Jr, RJ Reginato, and JL Hatfield. Normalizing the stress-degree-day parameter for environmental variability. *Agricultural meteorology*, 24:45–55, 1981.
- [19] CA Jones, CV Cole, AN Sharpley, and JR Williams. A simplified soil and plant phosphorus model: I. documentation. *Soil Science Society of America Journal*, 48(4):800–805, 1984.
- [20] George Joseph. *Fundamentals of remote sensing*. Universities Press, 2005.
- [21] Jude H Kastens, Terry L Kastens, Dietrich LA Kastens, Kevin P Price, Edward A Martinko, and Re-Yang Lee. Image masking for crop yield forecasting using avhrr ndvi time series imagery. *Remote Sensing of Environment*, 99(3):341–356, 2005.
- [22] Angela Kross, Heather McNairn, David Lapen, Mark Sunohara, and Catherine Champagne. Assessment of rapideye vegetation indices for estimation of leaf area index and biomass in corn and soybean crops. *International Journal of Applied Earth Observation and Geoinformation*, 34:235–248, 2015.
- [23] MP Labus, GA Nielsen, RL Lawrence, R Engel, and DS Long. Wheat yield estimates using multi-temporal ndvi satellite imagery. *International Journal of Remote Sensing*, 23(20):4169–4180, 2002.
- [24] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.
- [25] Guangxin Li, Chao Wang, Meichen Feng, Wude Yang, Fangzhou Li, and Ruiyun Feng. Hyperspectral prediction of leaf area index of winter wheat in irrigated and rainfed fields. *PloS one*, 12(8):e0183338, 2017.

- [26] Yan Li, Qingguo Zhou, Jian Zhou, Gaofeng Zhang, Chong Chen, and Jing Wang. Assimilating remote sensing information into a coupled hydrology-crop growth model to estimate regional maize yield in arid regions. *Ecological modelling*, 291:15–27, 2014.
- [27] Zhenhai Li, Jihua Wang, Xingang Xu, Chunjiang Zhao, Xiuliang Jin, Guijun Yang, and Haikuan Feng. Assimilation of two variables derived from hyperspectral data into the dssat-ceres model for grain yield and quality estimation. *Remote Sensing*, 7(9):12400–12418, 2015.
- [28] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.
- [29] David B Lobell, J Ivan Ortiz-Monasterio, C Lee Addams, and Gregory P Asner. Soil, climate, and management impacts on regional wheat productivity in mexico from remote sensing. *Agricultural and Forest Meteorology*, 114(1-2):31–43, 2002.
- [30] Raúl López-Lozano, Gregory Duveiller, Lorenzo Seguini, Michele Meroni, Sara García-Condado, Josh Hooker, Olivier Leo, and Bettina Baruth. Towards regional grain yield forecasting with 1 km-resolution eo biophysical products: Strengths and limitations at pan-european level. *Agricultural and Forest Meteorology*, 206:12–32, 2015.
- [31] Pavlo Lykhovyd. Sweet corn yield simulation using normalized difference vegetation index and leaf area index. *Journal of Ecological Engineering*, 21(3), 2020.
- [32] Michael L Mann and James M Warner. Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach. *Field crops research*, 201:60–74, 2017.
- [33] F Maselli, C Conese, L Petkov, and MA Gilabert. Environmental monitoring and crop forecasting in the sahel through the use of noaa ndvi data. a case study: Niger 1986–89. *International Journal of Remote Sensing*, 14(18):3471–3487, 1993.
- [34] Manasah S Mkhabela, Milton S Mkhabela, and Nkosazana N Mashinini. Early maize yield forecasting in the four agro-ecological regions of swaziland using ndvi data derived from noaa’s-avhrr. *Agricultural and Forest Meteorology*, 129(1-2):1–9, 2005.
- [35] M Susan Moran, Yoshio Inoue, and EM Barnes. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote sensing of Environment*, 61(3):319–346, 1997.
- [36] AP Moulin and HJ Beckie. Evaluation of the ceres and epic models for predicting spring wheat grain yield over time. *Canadian Journal of Plant Science*, 73(3):713–719, 1993.
- [37] Ranga Myneni, Yuri Knyazikhin, and Taejin Park. Mod15a2h modis/terra leaf area index/fpar 8-day l4 global 500 m sin grid v006. *NASA EOSDIS Land Processes DAAC*, 2015.
- [38] Samuel Ortega-Farias, Suat Irmak, and RH Cuenca. Special issue on evapotranspiration measurement and modeling, 2009.

- [39] Alessio Pollice and Massimo Bilancia. Kriging with mixed effects models. *Statistica*, 62(3):405–429, 2002.
- [40] Sanatan Pradhan, KK Bandyopadhyay, Rabi Narayan Sahoo, Vinay Kumar Sehgal, Ravender Singh, Vinod Kumar Gupta, and DK Joshi. Predicting wheat grain and biomass yield using canopy reflectance of booting stage. *Journal of the Indian Society of Remote Sensing*, 42(4):711–718, 2014.
- [41] Anup K Prasad, Lim Chai, Ramesh P Singh, and Menas Kafatos. Crop yield estimation model for iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33, 2006.
- [42] Michael S Rasmussen. Assessment of millet yields and production in northern burkina faso using integrated ndvi from the avhrr. *International Journal of Remote Sensing*, 13(18):3431–3442, 1992.
- [43] William R Raun, John B Solie, Gordon V Johnson, Marvin L Stone, Erna V Lukina, Wade E Thomason, and James S Schepers. In-season prediction of potential grain yield in winter wheat using canopy reflectance. *Agronomy Journal*, 93(1):131–138, 2001.
- [44] Felix Rembold, Clement Atzberger, Igor Savin, and Oscar Rojas. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing*, 5(4):1704–1733, 2013.
- [45] Olivier Renaud and Maria-Pia Victoria-Feser. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140(7):1852–1862, 2010.
- [46] O Rojas. Operational maize yield model development and validation based on remote sensing and agro-meteorological data in kenya. *International Journal of Remote Sensing*, 28(17):3775–3793, 2007.
- [47] Steven W Running, Ramakrishna R Nemani, Faith Ann Heinsch, Maosheng Zhao, Matt Reeves, and Hirofumi Hashimoto. A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, 54(6):547–560, 2004.
- [48] Toshihiro Sakamoto, Anatoly A Gitelson, and Timothy J Arkebauer. Near real-time prediction of us corn yields based on time-series modis data. *Remote Sensing of Environment*, 147:219–231, 2014.
- [49] John F Shanahan, James S Schepers, Dennis D Francis, Gary E Varvel, Wallace W Wilhelm, James M Tringe, Mike R Schlemmer, and David J Major. Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal*, 93(3):583–589, 2001.
- [50] Adam M Sibley, Patricio Grassini, Nancy E Thomas, Kenneth G Cassman, and David B Lobell. Testing remote sensing approaches for assessing yield variability among maize fields. *Agronomy Journal*, 106(1):24–32, 2014.

- [51] RCG Smith, J Adams, DJ Stephens, and PT Hick. Forecasting wheat yield in a mediterranean-type environment from the noaa satellite. *Australian Journal of Agricultural Research*, 46(1):113–125, 1995.
- [52] K Sowmya, CM John, and NK Shrivasthava. Urban flood vulnerability zoning of cochin city, southwest coast of india, using remote sensing and gis. *Natural Hazards*, 75(2):1271–1286, 2015.
- [53] R Core Team. (2015). *R: A language and environment for statistical computing*, 2013.
- [54] Compton J Tucker. Remote sensing of leaf water content in the near infrared. *Remote sensing of Environment*, 10(1):23–32, 1980.
- [55] Florin Vaida and Suzette Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.
- [56] Raisa Vozhehova, Mykola Maliarchuk, Iryna Biliaieva, Pavlo Lykhovyd, Anastasiia Maliarchuk, and Anatoliy Tomnytskyi. Spring row crops productivity prediction using normalized difference vegetation index. *Journal of Ecological Engineering*, 21(6):176–182, 2020.
- [57] Lenny Wall, Denis Larocque, and Pierre-Majorique Léger. The early explanatory power of ndvi in crop yield modelling. *International Journal of Remote Sensing*, 29(8):2211–2225, 2008.
- [58] Yanyu Wang, Ke Zhang, Chunlan Tang, Qiang Cao, Yongchao Tian, Yan Zhu, Weixing Cao, and Xiaojun Liu. Estimation of rice growth parameters based on linear mixed-effect model using multispectral images from fixed-wing unmanned aerial vehicles. *Remote sensing*, 11(11):1371, 2019.
- [59] Marie Weiss, Frédéric Jacob, and G Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236:111402, 2020.
- [60] Alyssa K Whitcraft, Inbal Becker-Reshef, and Christopher O Justice. A framework for defining spatially explicit earth observation requirements for a global agricultural monitoring initiative (geoglam). *Remote Sensing*, 7(2):1461–1481, 2015.
- [61] Alyssa K Whitcraft, Eric F Vermote, Inbal Becker-Reshef, and Christopher O Justice. Cloud cover throughout the agricultural growing season: Impacts on passive optical earth observations. *Remote sensing of Environment*, 156:438–447, 2015.
- [62] Yaojie Yue, Jian Li, Xinyue Ye, Zhiqiang Wang, A-Xing Zhu, and Jing-ai Wang. An epic model-based vulnerability assessment of wheat subject to drought. *Natural Hazards*, 78(3):1629–1652, 2015.
- [63] X Zhou, HB Zheng, XQ Xu, JY He, XK Ge, X Yao, T Cheng, Y Zhu, WX Cao, and YC Tian. Predicting grain yield in rice using multi-temporal vegetation indices from uav-based multispectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:246–255, 2017.

- [64] Alain Zuur, Elena N Ieno, Neil Walker, Anatoly A Saveliev, and Graham M Smith. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009.

ANNEX

```

library(ggplot2)
library(ggpmisc)
library(ggpubr)
library(gridExtra)
library(ggeffects)
library(effects)
library(psych)
library(pastecs)
library(beeswarm)
library(sjPlot)

#Attaching the dataset
attach(FinalDataset2021JK)
str(FinalDataset2021JK)
head(FinalDataset2021JK)
FinalDataset2021JK$Year<-as.factor(Year)
str(FinalDataset2021JK)

SSudan<-na.omit(FinalDataset2021JK)
str(SSudan)
describe(FinalDataset2021JK$Sorgharvest)

Fielddata<-cbind(HH_Size, CultLand, Sorgharvest)
stat.desc(Fielddata, basic = F)

Remotedata<-cbind(FinalDataset2021JK$EVI,FinalDataset2021JK$NDVI, FinalD
FinalDataset2021JK$LAI, FinalDataset2021JK$Evapotranspiration,
FinalDataset2021JK$Soil_Moisture)

stat.desc(Remotedata, basic = F)
stat.desc(Evapotranspiration, basic = F)

```

.1 Data exploration

.1.1 Covariates description

```
describe(FinalDataset2021JK$CultLand)
describe(FinalDataset2021JK$Sorgharvest)
Stats<-tapply(Sorgharvest, CultLand, mean)
Stats
describe(NDVI)
describe(EVI)
describe(LAI)
describe(HH_Size)
stat.desc(SSudan$Sorgharvest, basic = F)
```

.1.2 Checking for col-linearity between covariates

```
pairs.panels(FinalDataset2021JK)
ggpairs(FinalDataset2021JK[, 4:12])
ggpairs(FinalDataset2021JK[, 4:9])
```

.1.3 Some Graphs

```
par(mfrow=c(2,2))
```

```
plot(CultLand, Sorgharvest)
```

```
qplot(CultLand, Sorgharvest, data = na.omit(FinalDataset2021JK), colour = Year)
```

```
boxplot(Sorgharvest ~ CultLand, main = "Boxplot production vs cultivated land", col="blue")
```

```
boxplot(Sorgharvest, CultLand, main = "Boxplot production vs cultivated land", col="blue")
```

```
boxplot(Sorgharvest ~ CultLand, data = FinalDataset2021JK, main = "Sorghum yield vs cultivated land", # do not duplicate outliers: outline = FALSE, col = "blue")
```

1.3.1 Box plot and Bee swarm

```
boxplot(Sorgharvest, outline = FALSE,
main = "Beeswam plot of sorghum yield")
```

```
beeswarm(Sorgharvest,
data = as.data.frame(FinalDataset2021JK),
add = TRUE, pwc = as.numeric(Year), pch = 16)
```

```
legend("topright", title = "Sorghum Yield", legend = levels(Year),
pch = 16, col = 1:2)
```

```
boxplot(Sorgharvest, CultLand, col = c("blue", "green"),
horizontal = TRUE)
```

1.3.2 ggplots

```
g1<-ggplot(SSudan, aes(x = CultLand, y = Sorgharvest, colour = State))
+
geom_point(size = 2) +
theme_classic() +
theme(legend.position = "right")
g1
```

```
g2<-ggplot(data = na.omit(FinalDataset2021JK),
aes(x= CultLand, y=Sorgharvest, col = Year))+
geom_point(alpha = 0.7) + theme_bw()
g2
```

```
g3<-ggplot(SSudan, aes(x=CultLand, y=Sorgharvest, col=Year))+
geom_jitter() +
geom_boxplot(alpha=0.5) + facet_wrap(~Year)
g3
```

```
g4<-ggplot(SSudan, aes(x=CultLand, y=Sorgharvest, col=Year)) +
geom_jitter() +
geom_boxplot(alpha=0.5) + facet_wrap(~State)
g4
```

```
ggarrange(g1, g2, g3, g4,
```

```
labels = c("A", "B", "C", "D"),
ncol = 2, nrow = 2)
```

```
ggplot(data=na.omit(FinalDataset2021JK),
aes(x=NDVI, y=Sorgharvest, col=Year))+
geom_boxplot(notch = TRUE) +
geom_point(alpha = 0.7, col="blue") +
theme_bw()
```

```
ggplot(data=FinalDataset2021JK, aes(x=NDVI, y=Sorgharvest))+
geom_violin(notch = FALSE) +
geom_point(alpha = 0.7, col="blue")+theme_bw()
```

```
*****
```

PART A

Linear Multiple Regression Analysis - Modeling

```
*****
```

```
SSm1<-lm(Sorgharvest~NDVI)
summary(SSm1)
```

```
SSm2<-lm(Sorgharvest~EVI)
summary(SSm2)
```

```
SSm3<-lm(Sorgharvest~CultLand) # Significant model
summary(SSm3)
```

```
SSm4<-lm(Sorgharvest~NDVI+Soil_Moisture)
summary(SSm4)
summary(anova(SSm4))
```

```
SSm5<-lm(Sorgharvest~LAI+NDVI)
summary(SSm5)
```

Best Model Multiple linear regression

```
SSm6<-lm(Sorgharvest~NDVI*CultLand)
summary(SSm6) # good model Cultivated land and
```



```
NDVI effect statistically significant
AnovaSSm6<-anova(SSm6)
AnovaSSm6
```

```
par(mfrow=c(2,2))
plot(SSm6)
par(mfrow=c(1,1))
```

Predicted Sorghum yield vs observed yield.

```
Predicted<-predict(SSm6); length(PP)
Predicted
describe(Predicted) #The model SSm6 is able to predict up 79% of
observed sorghum yield
describe(Sorgharvest)
```

```
SimulateSorg_yield<-simulate(SSm6)
describe(SimulateSorg_yield)
```

```
length(na.omit(FinalDataset2021JK$Sorgharvest))
length(SSudan$Sorgharvest)
```

```
Observed<-na.omit(FinalDataset2021JK$Sorgharvest); length(Observed)
```

equation of the line :

```
regpp<-lm(Predicted~Observed)
summary(regpp)
eq = paste0("y = ", round(coeff[2],1), "X + ",
round(coeff[1],1), p-value)
eq
with(FinalDataset2021JK, plot(Observed,Predicted, main = eq))
abline(regpp, col= "blue")
```

```
coeff=coefficients(regpp)
coeff
```

Adding the equation

```

eq = paste0("y = ", round(coeff[2],1), "X + ", round(coeff[1],1))
eq
with(FinalDataset2021JK,
plot(Observed,Predicted, main = "Sorghum predicted yield vs observed"))
abline(regpp, col= "blue")

coeff=coefficients(regpp)
coeff

```

Checking for model assumptions

```

par(mfrow=c(2,2))
qqplot(NDVI, log(Sorgharvest))
qqPlot(log(Sorgharvest), distribution = "norm")

qqnorm(residuals(SSm6)); qqline(residuals(SSm6))
plot(fitted(SSm6), residuals(SSm6),
abline(h=0), xlab="Fitted Values",
ylab="Residuals", main = "Fitted values vs Residuals")
par(mfrow=c(1,1))

SSm7<-lm(Sorgharvest~EVI*CultLand)
summary(SSm7)

```

Good model cultivated land and EVI good predictors of Sorghum Yield

```

SSm8<-lm(Sorgharvest~NDVI+CultLand+HH_Size)
summary(SSm8) # This is a significant model.

SSm9<-lm(Sorgharvest~NDVI+CultLand*Soil_Moisture)
summary(SSm9) # This is another significant model

SSm10<-lm(Sorgharvest~NDVI+Soil_Moisture+
CultLand+Precipitation+Evapotranspiration)
summary(SSm10)

```

Full model not significance.

```

SSm11<-lm(Sorgharvest~CultLand+HH_Size+NDVI+LAI+

```

```
Precipitation+Soil_Moisture+Evapotranspiration)
summary(SSm11)
```

Anova for significant models SSm3, SSm6, SSm7, SSm8 & SSm9

```
AnovaSSm3679<-anova(SSm3, SSm6, SSm7, SSm9,SSm10)
AnovaSSm3679
summary(AnovaSSm367910)
```

Plot simple models *Plot1*

```
qplot(NDVI, Sorgharvest, data = na.omit(FinalDataset2021JK),
      geom=c("point", "smooth"), method="lm", formula=y~x,
      main="Regression of Sorghum Production and NDVI",
      ylab="Sorghum production in Kg", xlab="NDVI", se = TRUE)
```

Plot2

```
qplot(EVI, Sorgharvest, data = na.omit(FinalDataset2021JK),
      geom=c("point", "smooth"), method="lm", formula=y~x,
      main="Regression of Sorghum Production and EVI",
      ylab="Sorghum production in Kg", xlab="EVI", se =TRUE)
```

Plot3

```
qplot(Sorgharvest, Soil_Moisture, data = na.omit(FinalDataset2021JK),
      geom=c("point", "smooth"), method="lm", formula=y~x,
      main="Regression of Sorghum Production and soil moisture",
      ylab="Sorghum production in Kg", xlab="Soil moisture", se = TRUE)
```

Model Production predicted by NDVI

```
S1a<-ggplot(na.omit(FinalDataset2021JK),
            aes(x = NDVI, y = Sorgharvest)) + geom_point()
S1a
print(S1)
S1 + stat_smooth(method = "lm", formula = y ~ x, size = 1)
S1 + stat_smooth(method = "loess", formula = y ~ x, size = 1)
S1 + stat_smooth(method = "gam", formula = y ~ s(x), size = 1)
```

Sorghum yield plotted against all study predictors. Regression plot with equation on the graph

```
S1 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = NDVI, y = Sorgharvest)) +
geom_smooth(method = "lm", se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800, aes(label = paste(..rr.label..)))+
stat_regline_equation(label.y = 2000)
S1
```

```
S2 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = EVI, y = Sorgharvest)) +
geom_smooth(method = "lm", se = TRUE,
color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800, aes(label = paste(..rr.label..)))+
stat_regline_equation(label.y = 2000)
S2
```

Use of EVI as predictor

```
S3 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = Soil_Moisture, y = Sorgharvest)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label..., sep = "~`, `~")))+
stat_regline_equation(label.y = 2000)
S3
```

```
S4 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = Precipitation, y = Sorgharvest)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label..., sep = "~`, `~")))+
stat_regline_equation(label.y = 2000)
```

S4

```
S5 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = FPAR, y = Sorgharvest)) +
geom_smooth(method = "lm", se=TRUE,
color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label.., sep = "~`, `~")))+
stat_regline_equation(label.y = 2000)
S5
```

```
S6 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = NDVI, y = LAI)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 8,
aes(label = paste(..rr.label.., ..p.label.., sep = "~`, `~")))+
stat_regline_equation(label.y = 10)
S6
```

```
S6a <- ggplot(na.omit(FinalDataset2021JK),
aes(x = LAI, y = FPAR)) +
geom_smooth(method = "lm", se=TRUE,
color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 8,
aes(label = paste(..rr.label.., ..p.label.., sep = "~`, `~")))+
stat_regline_equation(label.y = 10)
S6a
```

Good model linear relationship Rainfall and Sorghum harvest

```
S7 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = LAI, y = Sorgharvest)) +
geom_smooth(method = "lm", se=TRUE,
color="blue", formula = y ~ x) +geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label.., sep = "~`, `~")))+
stat_regline_equation(label.y = 2000)
S7
```

```
S8 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = CultLand, y = Sorgharvest)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label.., ..p.label.., sep = "~`,`~")))+
stat_regline_equation(label.y = 2000)
S8
```

```
S9 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = HH_Size, y = Sorgharvest)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label.., sep = "~`,`~")))+
stat_regline_equation(label.y = 2000)
S9
```

```
S10 <- ggplot(na.omit(FinalDataset2021JK),
aes(x = Evapotranspiration, y = Sorgharvest)) +
geom_smooth(method = "lm",
se=TRUE, color="blue", formula = y ~ x) +
geom_point()+
stat_cor(label.y = 1800,
aes(label = paste(..rr.label.., sep = "~`,`~")))+
stat_regline_equation(label.y = 2000)
S10
```

```
ggarrange(S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S6a,
labels = c("(a)", "(b)", "(c)", "(d)", "(e)", "(f)", "(g)",
"(h)", "(i)", "(j)"), ("k"), ncol = 3, nrow = 4)
```

PART B

LINEAR MIXED EFFECT MODELS

```

library(lme4) # for linear Mixed-effects models using lmer function
library(merTools) # confidence interval of predictors in LME
library(MuMIn) # to have Rsquared
library(lmerTest) #provide p-value for fixed-effects predictors
library(sjPlot)
library(sjmisc)
library(emmeans)
ggeffect(S11b, "NDVI")

plot_models(S11b)

p<-ggpredict(S11b)
plot(p)

#

FinalDataset2021JK$State<-as.factor(FinalDataset2021JK$State)
str(FinalDataset2021JK)
S11<-lmer(Sorgharvest~CultLand + NDVI + (1|Year) + (1|State) ,
data = na.omit(FinalDataset2021JK), REML = FALSE)
S11 # Good LME model
plot(S11)
summary(S11)
RMSE.merMod(S11) # provides RMSE for the LME model .
RMSE=175.18
r.squaredLR(S11) # Rsquared = 0.784
plot(SSudan$Sorgharvest, Predict11)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =175.18", "R2 = 0.784"))
abline(Regression12a)

```

Checking for Best Model Assumption - Residual plots

Checking normality assumption with qqplot

```

par(mfrow=c(2,2))
qqplot(NDVI, log(Sorgharvest))
qqPlot(log(Sorgharvest), distribution = "norm")
qqnorm(residuals(S11)); qqline(residuals(S11))
plot(fitted(S11), residuals(S11), abline(h=0),
xlab="Fitted Values", ylab="Residuals",

```

```
main = "Fitted values vs Residuals")
par(mfrow=c(1,1))

S11a<-lmer(Sorgharvest~CultLand + NDVI +
(1|State) + (1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S11a
summary(S11a)
RMSE.merMod(S11a)
r.squaredLR(S11a)
Predict11a<-predict(S11a)
Regression11a<-lm(SSudan$Sorgharvest~Predict11a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict11a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =175.45", "R2 = 0.784"))
abline(Regression11a)
```

```
S11b<-lmer(Sorgharvest~CultLand*NDVI +
(1|State) + (1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S11b
summary(S11b)
RMSE.merMod(S11b)
r.squaredLR(S11b)
Predict11b<-predict(S11b)
Regression11b<-lm(SSudan$Sorgharvest~Predict11b)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict11b)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =173.85", "R2 = 0.788"))
abline(Regression11b)
```

Check for model assumption

```
par(mfrow=c(2,2))
```



```

qqPlot(log(Sorgharvest), distribution = "norm")

qqnorm(residuals(S11b)); qqline(residuals(S11b))
plot(fitted(S11b), residuals(S11b), abline(h=0),
xlab="Fitted Values", ylab="Residuals",
main = "Fitted values vs Residuals")
par(mfrow=c(1,1))
plot(S11b)

S12<-lmer(Sorgharvest~CultLand + LAI + (1|Year) +(1|State),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S12
simulate(S12)
a<-summary(SSm6)
RMSE.merMod(S12)
r.squaredLR(S12)
Predict12<-predict(S12)
Regression12<-lm(SSudan$Sorgharvest~Predict12)

```

plot lmer

```

plot(SSudan$Sorgharvest, Predict12)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.9", "R2 = 0.785"))
abline(Regression12)

```

Model diagnostics

```

par(mfrow=c(1,2))
qqPlot(log(Sorgharvest), distribution = "norm")

qqnorm(residuals(S12)); qqline(residuals(S12))
plot(fitted(S12), residuals(S12), abline(h=0),
xlab="Fitted Values", ylab="Residuals",
main = "Fitted values vs Residuals")
par(mfrow=c(1,1))

plot(S12)
mixed.mod.visual(S12, rand.intercept)

```

```
emmeans(S12,list(pairwise~State), adjust ="Tukey")
```

```
S12a<-lmer(Sorgharvest~CultLand + LAI +
  (1|State) +(1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S12a
summary(S12a)
RMSE.merMod(S12a)
r.squaredLR(S12a)
Predict12a<-predict(S12a)
Regression12a<-lm(SSudan$Sorgharvest~Predict12a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict12a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.98", "R2 = 0.785"))
abline(Regression12a)
```

```
ranef(S12a)
```

```
S13<-lmer(SSudan$Sorgharvest~CultLand +
HH_Size+ NDVI+ LAI +Precipitation +
Soil_Moisture + Evapotranspiration + (1|State) + (1|Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S13
summary(S13)
RMSE.merMod(S13)
r.squaredLR(S13)
Predict13<-predict(S13)
Regression13<-lm(SSudan$Sorgharvest~Predict13)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict13)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =173.8", "R2 = 0.787"))
abline(Regression13)
```

```
confint.merMod(S13, level = 0.95)
# compute the confidence interval for the intercepts
```

Model diagnostics

```
par(mfrow=c(2,2))
qqPlot(log(Sorgharvest), distribution = "norm")
```

```
qqnorm(residuals(S13)); qqline(residuals(S13))
plot(fitted(S13), residuals(S13), abline(h=0),
xlab="Fitted Values", ylab="Residuals",
main = "Fitted values vs Residuals")
plot(S13)
par(mfrow=c(1,1))
```

```
S13a<-lmer(Sorgharvest~ CultLand + HH_Size +
NDVI + LAI + Precipitation + Soil_Moisture +
Evapotranspiration + (1|Year)+(1|State) +(1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S13a
summary(S13a)
RMSE.merMod(S13a)
r.squaredLR(S13a)
Predict13a<-predict(S13a)
Regression13a<-lm(SSudan$Sorgharvest~Predict13a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict13a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =173.88", "R2 = 0.787"))
abline(Regression12)
```

```
confint.merMod(S13a, level = 0.95)
# compute the profile confidence intervals for predictors
```

```
S14<-lmer(Sorgharvest~CultLand + NDVI + LAI +
EVI + Soil_Moisture + Precipitation +
Evapotranspiration + (1|State) +(1|Year)+(1|State:Year),
```

```

data = na.omit(FinalDataset2021JK), REML = FALSE)
S14
summary(S14)
RMSE.merMod(S14)
r.squaredLR(S14)
Predict14<-predict(S14)
Regression14<-lm(SSudan$Sorgharvest~Predict14)

```

plot lmer

```

plot(SSudan$Sorgharvest, Predict14)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.1", "R2 = 0.787"))
abline(Regression14)

```

```

S14a<-lmer(Sorgharvest~CultLand + NDVI +
LAI + Soil_Moisture + Precipitation +
Evapotranspiration + (1|State) + (1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S14a #when combine NDVI fixed effect is much greater in this model
and Yearly influence of remotely sensed data is higher in State(UpN)
summary(S14a)
RMSE.merMod(S14a)
r.squaredLR(S14a)
Predict14a<-predict(S14a)
Regression14a<-lm(SSudan$Sorgharvest~Predict14a)

```

plot lmer

```

plot(SSudan$Sorgharvest, Predict14a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.23", "R2 = 0.787"))
abline(Regression12)

```

```

AnovaS13111214<-anova(S13,S11, S11a, S12, S12a, S13, S13a, S14, S14a)
AnovaS13111214 # Model S12 and S12a with LAI index is significant

```

```
S15<-lmer(Sorgharvest~CultLand + (1|Soil_Moisture) +
(1|Evapotranspiration),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S15
summary(S15)
```

```
RMSE.merMod(S15)
r.squaredLR(S15)
Predict15<-predict(S15)
Regression15<-lm(SSudan$Sorgharvest~Predict15)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict15)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =119.7", "R2 = 0.784"))
abline(Regression15)
```

```
S16<-lmer(Sorgharvest~CultLand + NDVI + EVI + Precipitation+
(1|Soil_Moisture) + (1|Evapotranspiration),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S16
summary(S16)
RMSE.merMod(S16)
r.squaredLR(S16)
PredictS16<-predict(S16)
Regression16<-lm(SSudan$Sorgharvest~PredictS16)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict14a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =117.6", "R2 = 0.784"))
abline(Regression16)
```

```
S16a<-lmer(Sorgharvest~CultLand + NDVI + EVI + (1|Soil_Moisture)+
(0 +Soil_Moisture|Evapotranspiration), d
ata = na.omit(FinalDataset2021JK), REML = FALSE)
summary(S16a)
```

```
RMSE.merMod(S16a)
r.squaredLR(S16a)
Predict16a<-predict(S16a)
Regression16a<-lm(SSudan$Sorgharvest~Predict16a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict16a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.2", "R2 = 0.782"))
abline(Regression16a)
```

```
S17<-lmer(Sorgharvest~CultLand + Precipitation + (1|NDVI) + (1|State),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S17
```

```
RMSE.merMod(S17)
r.squaredLR(S17)
Predict17<-predict(S17)
Regression17<-lm(SSudan$Sorgharvest~Predict17)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict17)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =111.5", "R2 = 0.784"))
abline(Regression12)
```

```
S18<-lmer(Sorgharvest~CultLand + NDVI + Precipitation +
(1|State)+(1|Year) +(1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
```

```
S18
```

```
summary(S18)
RMSE.merMod(S18)
r.squaredLR(S18)
Predict18<-predict(S18)
Regression18<-lm(SSudan$Sorgharvest~Predict18)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict18)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =175.07", "R2 = 0.785"))
abline(Regression18)
```

```
S18a<-lmer(Sorgharvest~CultLand + NDVI + Precipitation +
(1|State)+ (1|Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S18a
summary(S18a)
RMSE.merMod(S18a)
r.squaredLR(S18a)
Predict18a<-predict(S18a)
Regression18a<-lm(SSudan$Sorgharvest~Predict18a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict18a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.8", "R2 = 0.785"))
abline(Regression18a)
```

```
S19<-lmer(Sorgharvest~CultLand + NDVI + Precipitation + LAI +
(1|Year) +(1|State)+ (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S19
summary(S19)
RMSE.merMod(S19)
r.squaredLR(S19)
```

```
Predict19<-predict(S19)
Regression19<-lm(SSudan$Sorgharvest~Predict19)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict19)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.42", "R2 = 0.786"))
abline(Regression19)
```

```
S19a<-lmer(Sorgharvest~CultLand+ NDVI + Precipitation + LAI + EVI +
(1|State) + (1|Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S19a
summary(S19a)
RMSE.merMod(S19a)
r.squaredLR(S19a)
```

```
Predict19a<-predict(S19a)
Regression19a<-lm(SSudan$Sorgharvest~Predict19a)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict19a)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.1", "R2 = 0.786"))
abline(Regression19a)
```

```
S19b<-lmer(Sorgharvest~CultLand + Precipitation + LAI + EVI + (1|State) +
(1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S19b
summary(S19b)
RMSE.merMod(S19b)
r.squaredLR(S19b)
```

```
Predict19b<-predict(S19b)
Regression19b<-lm(SSudan$Sorgharvest~Predict19b)
```

plot lmer

```
plot(SSudan$Sorgharvest, Predict19b)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.45", "R2 = 0.786"))
abline(Regression19b)
```

```
S20<-lmer(Sorgharvest~CultLand + NDVI + LAI + EVI + FPAR +
```



```

Soil_Moisture + Precipitation + Evapotranspiration +
(1|State) + (1|Year) + (1|State:Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S20
summary(S20)
RMSE.merMod(S20)
r.squaredLR(S20)

Predict20<-predict(S20)
Regression20<-lm(SSudan$Sorgharvest~Predict20)

```

plot lmer

```

plot(SSudan$Sorgharvest, Predict20)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =174.14", "R2 = 0.787"))
abline(Regression20)

```

```

S22<-lmer(Sorgharvest~NDVI + LAI + Soil_Moisture +
Precipitation + Evapotranspiration + (1|State) + (1|Year),
data = na.omit(FinalDataset2021JK), REML = FALSE)
S22
summary(S22)
RMSE.merMod(S22)
r.squaredLR(S22)

Predict22<-predict(S22)
Regression22<-lm(SSudan$Sorgharvest~Predict22)

```

plot lmer

```

plot(SSudan$Sorgharvest, Predict22)
text(c(x=300, 250), y=c(1400, 1700),
labels = c("RMSE =155.7", "R2 = 0.757"))
abline(Regression22)

```

ANOVA for model S18 and S19

Analysis of variance for the first Random-effects structure

Anova for table 6 in the Thesis document

```
AnovaS1112131418a19a22<-anova(S11,S12,S13, S14, S18a, S19a, S22)
AnovaS1112131418a19a22
```

Anova for table 10 in the thesis document

```
AnovaS11a11b12a13a14a181919b20<-
anova(S11a, S11b, S12a, S13a, S14a, S18, S19, S19b, S20)
AnovaS11a11b12a13a14a181919b20
```

Anova for the 3 significant models

```
Anova121311b<-anova(S12,S13,S11b)
Anova121311b # this analyse show S12 is the best model given
smaller AIC and loglik
```

Plotting random slope variation with lmer

```
*****
ggplot(S11b, aes(x=CultLand, y=Sorgharvest, color=Year))+
geom_line()+
geom_point(data=SSudan, aes(x=CultLand, y=Sorgharvest))+
facet_wrap(~State, nrow=3)

ggplot(S12, aes(x=CultLand, y=Sorgharvest, color=Year))+
geom_line()+
geom_point(data=SSudan, aes(x=CultLand, y=Sorgharvest))+
facet_wrap(~State)

ggplot(S13, aes(x=CultLand, y=Sorgharvest, color=Year))+
geom_line()+
geom_point(data=SSudan, aes(x=CultLand, y=Sorgharvest))+
facet_wrap(~State)
```