

# Leveraging Lasso Regression for Crop Yield Modelling - Lessons from Work in Burkina Faso

Dr. Sandra Barteit, Heidelberg Institute of Global Health (HIGH), Heidelberg University Hospital, Heidelberg University, Germany



**HEIDELBERG**  
UNIVERSITY  
HOSPITAL



**HEIDELBERG**  
FACULTY OF  
MEDICINE



**HEIDELBERG**  
INSTITUTE OF  
GLOBAL HEALTH

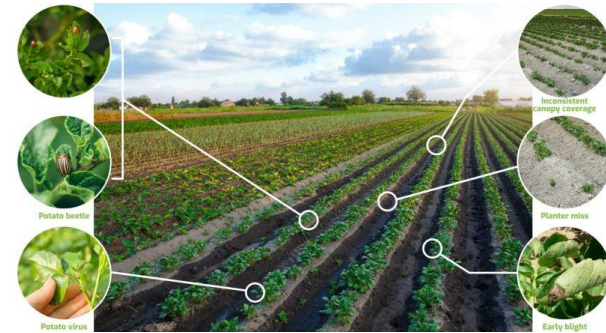
## Learning objectives

- Understand the concept of crop yield modeling and its importance in agriculture and food security.
- Grasp the fundamentals of Lasso regression and its advantages over traditional regression methods in crop yield modeling.
- Learn about the application of Lasso regression in a real-world case study from Burkina Faso.
- Critically analyze the results and implications of the case study.

# What is crop yield modeling?

**Definition:** The process of estimating crop yield using various data sources and statistical methods.

- Crop yield data in agriculture refers to the information on the amount of crop produced per unit area, such as tonnes per hectare
- requires understanding the amount of sunlight and water plants receive, as well as other factors like temperature, humidity, and soil type, which all influence plant growth
- **Importance:**
  - Quantity and quality of crops are crucial for agricultural planning, food security assessment, and resource management - particularly in developing countries where agriculture plays a significant economic role
  - for monitoring progress towards global goals, identifying strengths and weaknesses in farming practices, and making informed decisions about various aspects of farming



# What is crop yield modeling?

## Traditional methods

- **Crop cutting experiments:** Labor-intensive and time-consuming.
- **Statistical surveys:** Limited spatial and temporal resolution.



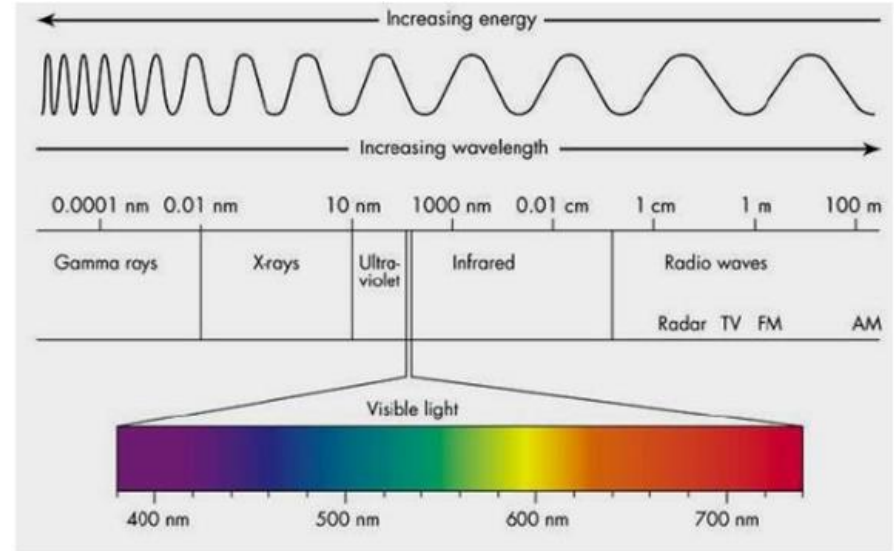
## Modern approaches

- **Remote sensing:** Utilizes satellite data to monitor crop growth and development.
- **Machine learning:** Employs algorithms to analyze large datasets and predict yields.



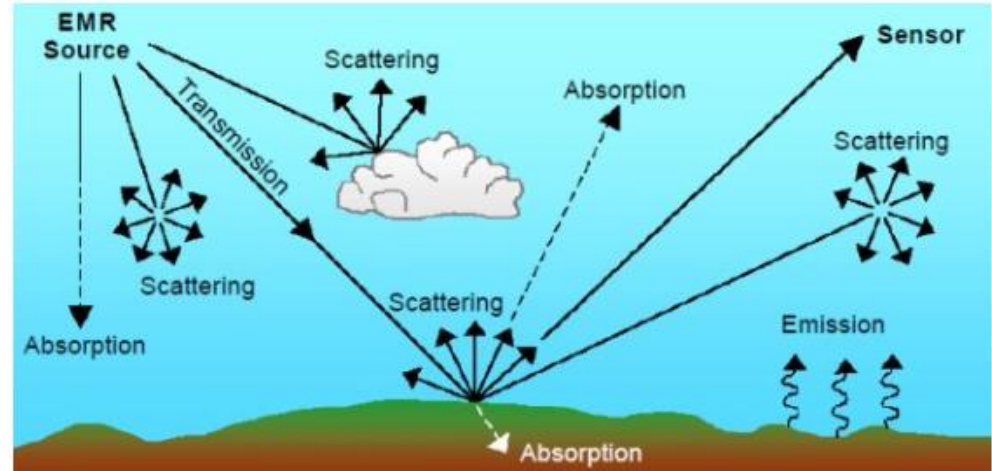
# Remote Sensing: Overview

- Remote sensing involves acquiring information about objects or phenomena **without direct contact**, using techniques like satellite images, aerial photos, radar systems, and lidar
- utilizes electromagnetic radiation (EMR) across various wavelengths, including radio waves, microwaves, infrared, visible light, and ultraviolet light
  - energy of the sun is composed of many kinds of radiation, some of which spans the visible part of the electromagnetic spectrum. Many instruments collect Red, Green, and Blue bands of the spectrum (R,G,B) to create natural color images.
  - meaningful information is also contained in parts of the spectrum outside the range of human vision, including infrared (IR) and ultra-violet (UV).



# Remote Sensing: Overview

- **Energy of the sun is absorbed or scattered** by the atmosphere before it reaches earth.
- In Remote Sensing analysis we aim to learn about objects on Earth through **studying the radiation reflected and/or emitted by them.**

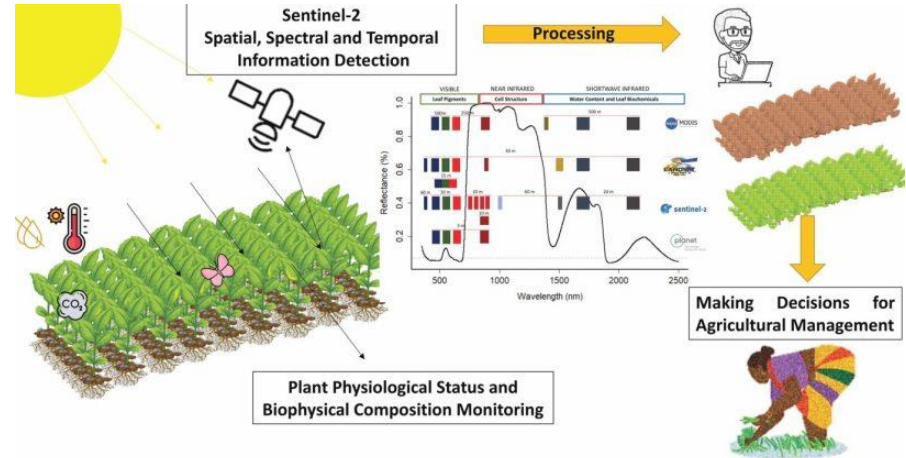


Reflected and emitted radiation

# Remote Sensing: Overview

## Remote sensing data provides insights into:

- crop growth conditions over time
- estimates crop yields
- helps determine optimal harvest times
- measures land-use changes
- detect soil moisture and salinity levels
- assess pest infestations, and monitor environmental pollution levels



## Remote sensing for agriculture monitoring: Sentinel-2 features and precision agriculture



# Remote Sensing

## Spatial Resolution

- **Landsat-7 (30m):** Offers a good balance between spatial resolution and coverage area, making it suitable for large-scale land cover mapping, monitoring deforestation, and studying regional agricultural practices.
- **Sentinel-2 (10m):** Provides high spatial resolution for detailed analysis of land use, urban areas, and environmental monitoring. Its multispectral capabilities capture information on vegetation health, water quality, and coastal zones.
- **Worldview (50cm):** Delivers very high spatial resolution, enabling the identification of individual objects and fine-scale features like vehicles and trees. This makes it useful for urban planning, infrastructure monitoring, disaster response, and precision agriculture.



Fig. 1.4 Landsat-7 (30 m)

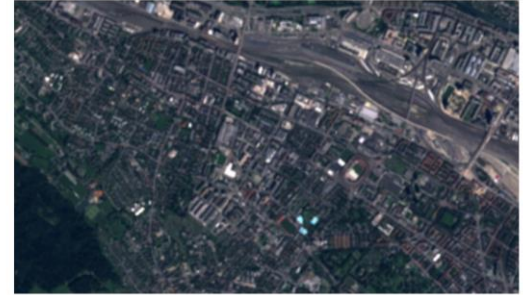


Fig. 1.5 Sentinel-2 (10 m)



Fig. 1.6 Worldview (50 cm)

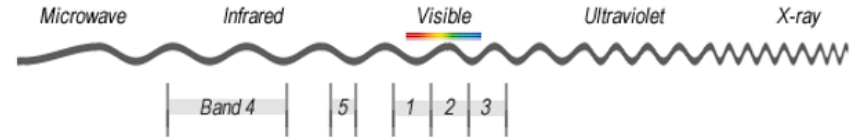


# Remote Sensing

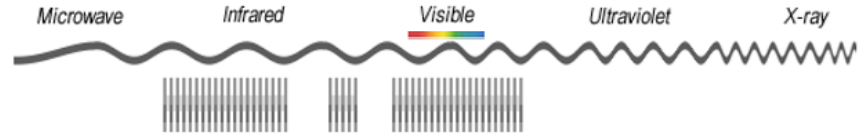
## Spectral Resolution

### Spectral resolution:

- how many and how narrow the color bands a sensor can capture
  - higher spectral resolution means the sensor can detect more specific colors because it has more, narrower bands
- 
- **Multispectral imagery** usually has 3 to 10 color bands.
  - **Hyperspectral imagery** has hundreds or even thousands of much narrower bands, providing more detailed color information.
  - **Panchromatic imagery** uses just one wide band to capture a broad range of wavelengths, essentially in black and white.



Multispectral imagery



Hyperspectral imagery

# Remote Sensing

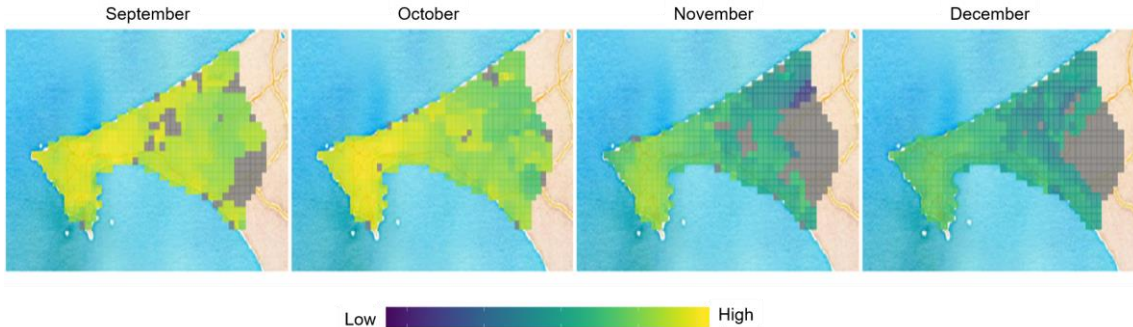
## Temporal Resolution

- temporal resolution is the frequency at which a sensor captures images of the same area on Earth's surface ("revisit time")
- satellite with a larger swath width (the width of the area it can capture from left to right during one pass) generally has a higher temporal resolution → satellite can cover more of the Earth's surface in a single pass, allowing it to revisit the same area more frequently

---

**Summaries at different temporal resolutions by city: Monthly Summary Maps by year/product**

Example: MODIS Land Surface Temperature (LST) day time 2016



# Remote Sensing

## Example Satellite Data - Sentinel 2

Part of ESA's Copernicus Programme: Sentinel-2 is a satellite mission developed by the European Space Agency, dedicated to land monitoring.

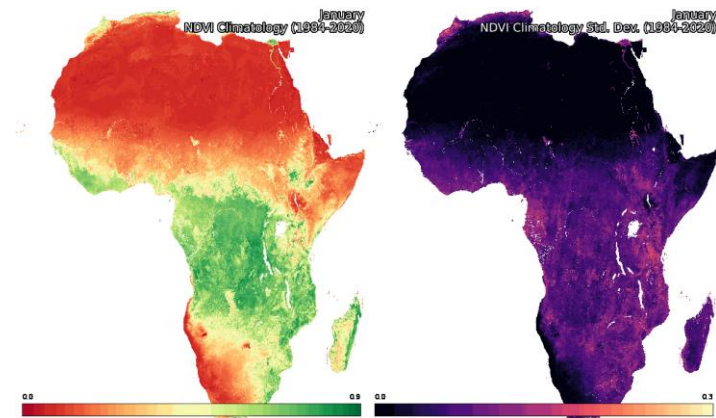
- Spatial resolution of **10 meters**
- **Temporal Resolution:** Comprising two satellites, Sentinel-2A and Sentinel-2B, the mission revisits the same spot on Earth every 5 days.
- **Global Coverage:** Monitors Earth's land surfaces, large islands, and inland and coastal waters.
- **Vegetation Indices:** Used to calculate various vegetation indices such as normalized difference vegetation index (NDVI) for assessing plant growth and vegetation health.
- **Data Accessibility:** Sentinel-2 data is freely available and widely used by researchers and policymakers.



# Vegetation Indices

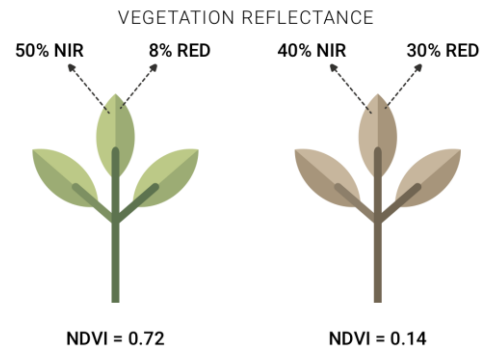
## Normalized Difference Vegetation Index (NDVI)

- What is NDVI?
  - Measures vegetation health using satellite imagery.
  - What It Measures:
    - NDVI measures the difference between near-infrared light (which vegetation strongly reflects) and red light (which vegetation absorbs).
    - Healthy vegetation absorbs most of the visible light that hits it and reflects a large portion of the near-infrared light.
    - Unhealthy or sparse vegetation reflects more visible light and less near-infrared light.
- Value Range
  - -1 to +1
    - Negative: Non-vegetated (e.g., water)
    - Near 0: Bare soil/minimal vegetation
    - Positive: Healthy vegetation (higher is healthier)
- Color Mapping
  - Brown = Low NDVI (sparse vegetation)
  - Green = High NDVI (dense vegetation)



HEALTHY

STRESS



$$NDVI = \frac{NIR - RED}{NIR + RED}$$

# Vegetation Indices

## NDVI, NDRE, NDWI

### NDVI (Normalized Difference Vegetation Index):

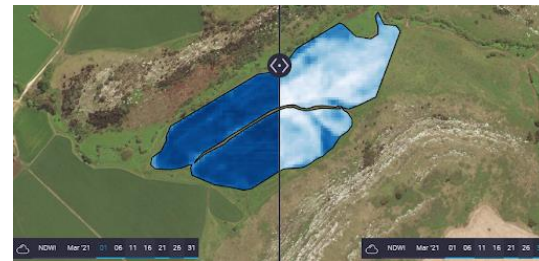
- NDVI is an index used to measure vegetation health and density by comparing the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs).
- It ranges from -1 to +1, where higher values indicate healthier, denser vegetation.

### NDRE (Normalized Difference Red Edge Index):

- NDRE is similar to NDVI but uses the red edge band instead of the red band, making it useful for assessing plant health and chlorophyll content, especially in more mature or stressed vegetation.
- It provides better sensitivity to variations in vegetation health, particularly in crops (ranges from -1 to +1)

### NDWI (Normalized Difference Water Index):

- NDWI is an index used to detect water content in vegetation and soil by comparing near-infrared and shortwave infrared reflectance.
- It helps in monitoring water stress in plants and identifying bodies of water or moist areas (ranges from -1 to +1)



## Case Study: Nouna HDSS, Burkina Faso

- **Context:** Sub-Saharan Africa: Faces significant challenges in food security due to climate change and population growth.
  - **Research question:** Can we develop accurate crop yield models for major food crops in Burkina Faso using satellite data and LASSO regression?
-

# Nouna Health and Demographic Surveillance System

## Study Area

- **Location:** Nouna Health and Demographic Surveillance System (HDSS) area, Burkina Faso.
- **Characteristics:** Semi-arid climate, subsistence agriculture.
- **Crops:** Millet, sorghum, maize, and rice.
- **Population:** Predominantly rural, dependent on agriculture.
- **Importance:** Region vulnerable to climate change with limited data availability

Burkina Faso



Kossi province

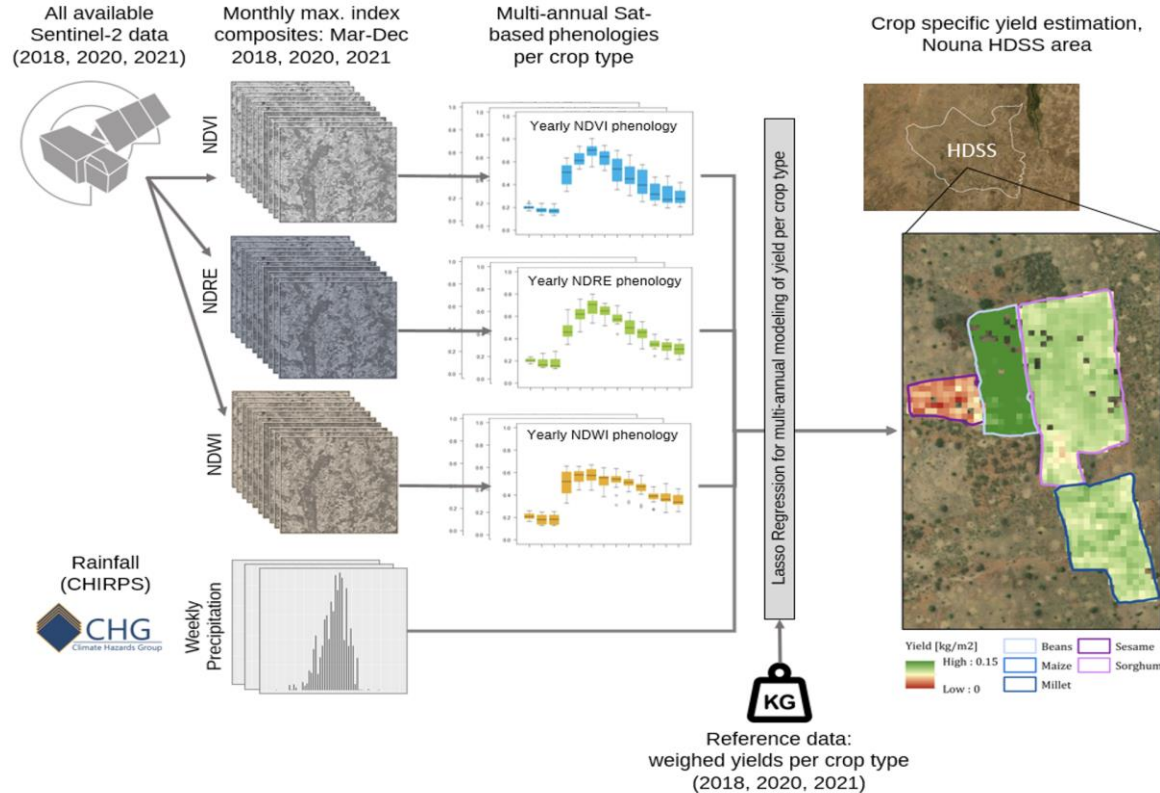


Nouna HDSS area





# Crop Yield Modelling: Nouna HDSS, Burkina Faso



# Data and Methodology

- **Data sources:**

- Sentinel-2 satellite imagery: Provides high-resolution vegetation indices (NDVI, NDRE, NDWI).
- CHIRPS rainfall data: Offers weekly accumulated precipitation estimates.
- In-situ yield measurements: Collected over three years (2018, 2020, 2021) for maize, millet, sorghum, beans, and sesame.

- **Methodology:**

- Preprocessing of satellite and rainfall data.
  - LASSO regression model development and validation using cross-validation.
  - Yield prediction for all sampled fields.
  - Comparison with national yield statistics for plausibility check.
-

# Rainfall Data

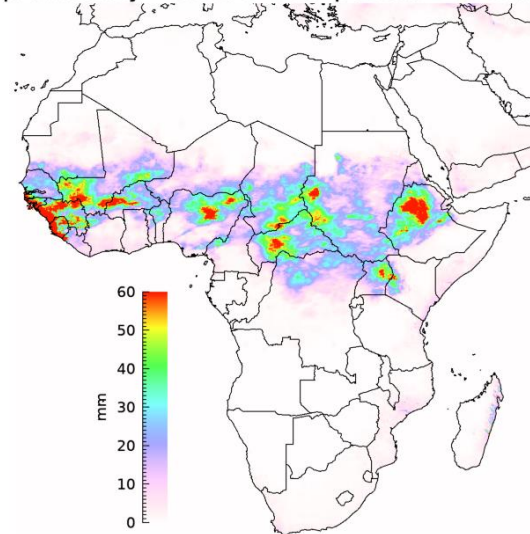
- **CHIRPS Overview:**

- CHIRPS is a global dataset providing over 35 years of rainfall data, covering areas between 50°S and 50°N, using satellite imagery and ground station data.
- It offers high-resolution gridded data (0.05°) for analyzing trends and monitoring seasonal droughts, from 1981 to the present.

- **Data Processing:**

- For each calendar week of the crop growing season, daily rainfall data were accumulated including all preceding weeks.
- Used as input variables for the model from March onwards (except November and December as harvest months).
- Total of 35 variables (Fig. 3).

preliminary CHIRPS v2.0 pentad 2024.07.5

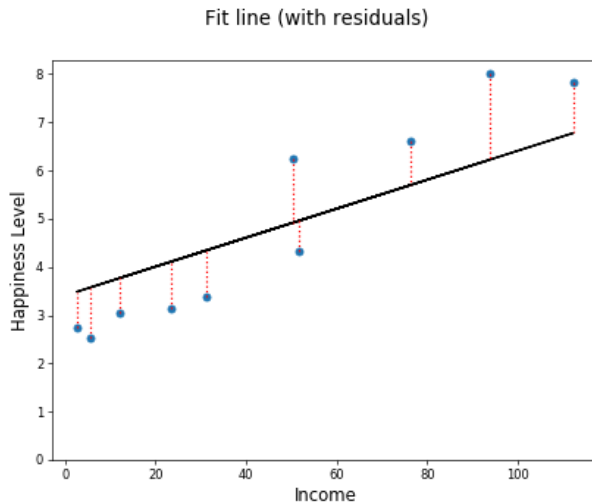


# Lasso Regression

## Review Linear Regression

### Linear Regression:

- **Fundamental Concept:** Linear regression is a basic but essential technique in machine learning, used to find the best-fit line through a set of data points.
- **Residuals:** The differences between the actual data points and the predicted values from the line are called residuals.
- **Sum of Squared Residuals (SSR):** The residuals are squared and summed to calculate the SSR, which measures the accuracy of the model.
- **Loss Function:** Linear regression aims to minimize the SSR, which serves as the loss function, by adjusting the line to better fit the data points.
- Predicts a target variable based on a linear relationship with input features.
- Includes all features in the model without any regularization.
- Can overfit, especially with many features.
- Does not automatically select important features.



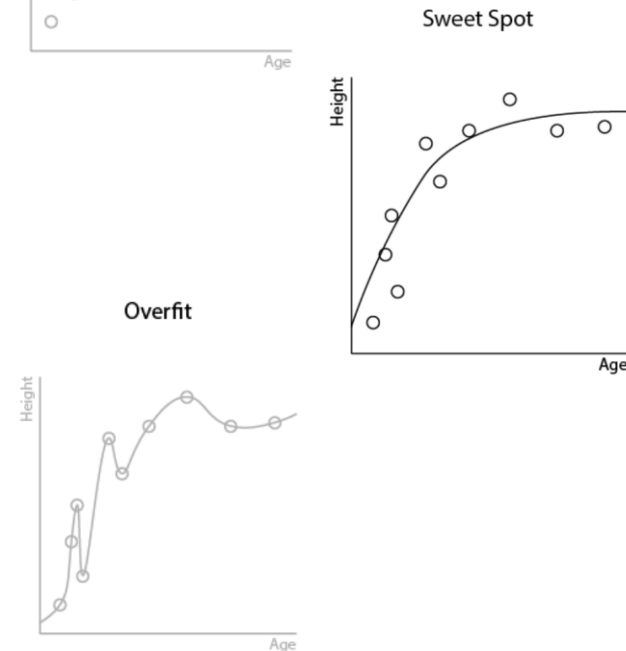
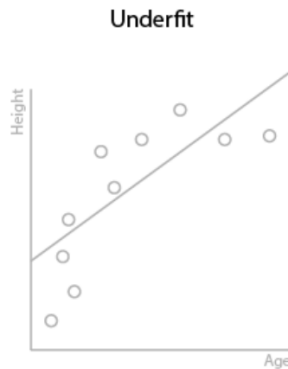
$$SSR = \sum (y - \hat{y})^2$$

$$L = \sum (y - \hat{y})^2$$

# Lasso Regression

## Review Linear Regression

- **Overfitting Issue:** Overfitting occurs when a model performs well on training data but poorly on test data, indicating it has learned too much from the training data, including noise.
- **Identifying Overfitting:** It can be identified by comparing the model's performance on training versus test data, with a significant performance drop indicating overfitting.
- **Preventing Overfitting:** Using fewer, more relevant features can help prevent overfitting by reducing the model's complexity.
- **Balancing Complexity:** The goal is to find a balance between model complexity and simplicity, avoiding both overfitting and underfitting to accurately describe the data relationships.

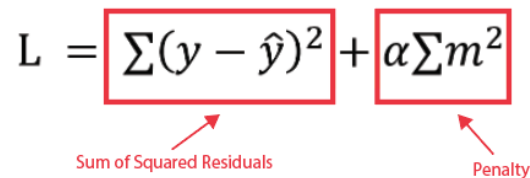


# Lasso Regression

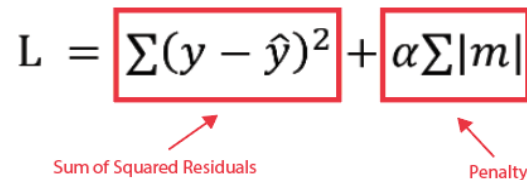
## Using Regularization

- **Regularization Overview:** Regularization adds a penalty to the loss function in linear regression to prevent overfitting, using either Ridge (L2) or Lasso (L1) techniques.
- **Ridge (L2) Regularization:** This method adds a penalty proportional to the sum of the squared coefficients, which shrinks the coefficients but doesn't necessarily eliminate them.
- **Lasso (L1) Regularization:** This technique adds a penalty based on the sum of the absolute values of the coefficients, which can shrink some coefficients to zero, effectively eliminating less important features (feature selection).
- **Effect on Model:** Both methods simplify the model by reducing the magnitude of coefficients, with Lasso also capable of removing irrelevant features entirely, thus controlling model complexity and reducing overfitting.

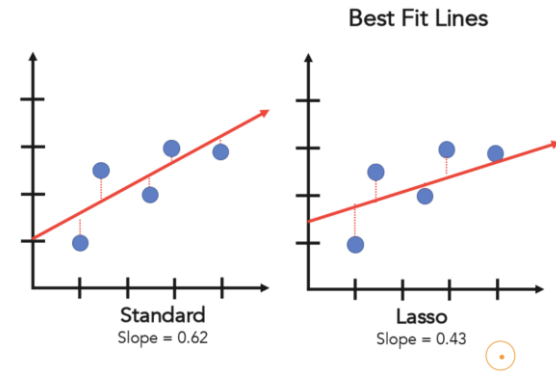
$$L = \boxed{\sum (y - \hat{y})^2} + \boxed{\alpha \sum m^2}$$



$$L = \boxed{\sum (y - \hat{y})^2} + \boxed{\alpha \sum |m|}$$



# Lasso Regression



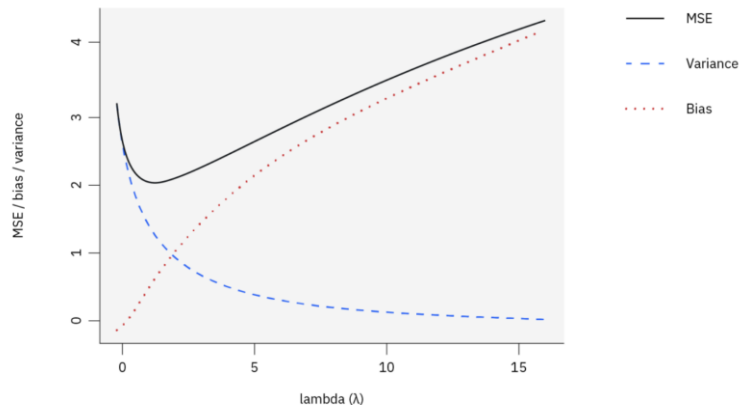
- Lasso Regression:
  - Lasso stands for Least Absolute Shrinkage and Selection Operator
  - Adds L1 regularization to shrink some feature coefficients to zero.
  - Automatically selects important features by excluding irrelevant ones.
  - Helps prevent overfitting by simplifying the model.
  - Focuses on the most significant features, making the model more interpretable.



# Lasso Regression

## Bias-Variance-Tradeoff

- **Bias-Variance Tradeoff Definition:** The tradeoff describes how bias measures the average difference between predicted and true values, while variance measures the variability of predictions across different data sets.
- **Impact on Accuracy:** High bias leads to less accurate predictions on training data, while high variance leads to less accurate predictions on new, unseen data.
- **Role of Regularization:** Techniques like Lasso regression help balance the tradeoff by adjusting the hyperparameter lambda ( $\lambda$ ). Increasing  $\lambda$  reduces variance but increases bias, resulting in a simpler model.
- **Lambda's Effect:** A higher  $\lambda$  leads to fewer model parameters and potentially more bias, while a lower  $\lambda$  increases the model's complexity and variance. When  $\lambda$  is zero, the model becomes standard linear regression without regularization.



# Lasso Regression

## Introduction to LR in the study

- **Purpose:** The study utilized Lasso regression to develop multi-annual yield models for key food crops (maize, millet, sorghum, beans, and sesame) in rural Burkina Faso.
  - **Objective:** Accurately estimate crop yields at the household field level, reducing reliance on extensive in-situ data collection. The study targeted fields with an average size of less than 2 hectares.
  - **Approach:** Leveraging a unique three-year dataset (2018, 2020, 2021) of in-situ measurements, Lasso regression was employed for its ability to handle high-dimensional data and perform automatic feature selection.
-

# Lasso Regression

## Methodology and Application

- **Data Sources: Satellite**
  - **Data:** Monthly NDVI, NDRE, and NDWI composites derived from Sentinel-2 imagery at a 10 m spatial resolution.
  - **Rainfall Data:** Weekly accumulated rainfall data from the CHIRPS dataset, used to capture the impact of rainfall on crop growth.
- **Lasso Implementation:**
  - **Model Choice:** Lasso regression was chosen over stepwise multiple linear regression to avoid overfitting and accurately identify significant predictors.
- **Cross-Validation:** A 5-fold cross-validation approach was used, repeated 1000 times, to determine the optimal  $\lambda$  (regularization parameter). This ensured the selection of a robust model with minimized prediction error.
- **Model Development:** The best  $\lambda$  value was selected based on the ratio of  $R^2$  and RMSE, resulting in crop-dependent models.

# Lasso Regression

## Results and Benefits

- **Model Accuracy:**
  - **Maize:** The three-year model achieved an  $R^2$  of 0.62 and an nRMSE of 12.70%.
  - **Millet:** The model showed an  $R^2$  of 0.32 and an nRMSE of 16.28%.
  - **Sorghum:** The  $R^2$  was 0.30 with an nRMSE of 13.06%.
  - **Beans:** An  $R^2$  of 0.54 and an nRMSE of 13.49% were achieved.
  - **Sesame:** The model had an  $R^2$  of 0.59 and an nRMSE of 14.39%.
- **Outcomes:** The models effectively captured inter-annual yield variability and demonstrated good fit with national yield statistics. This capability is crucial for monitoring and predicting food security, especially in response to climate variability and its impact on small-scale farmers.
- **Impact:** The successful application of Lasso regression in this study highlights its potential for generating accurate, field-level yield estimates, providing valuable insights for policymakers and stakeholders in food security and nutrition. The use of multi-annual data further enhances the model's robustness, making it a reliable tool for future applications.

# Model Outcomes

- **Model Fit:**

- Maize and Millet had generally good  $R^2$  values, indicating strong models
- Sorghum and Beans had moderate fits, while Sesame showed variable performance

- **Prediction Error:**

- Maize had consistently low RMSE and nRMSE, indicating accurate predictions
- Millet and Sorghum showed more variability in error rates

- **Feature Use:**

- Number of predictors varied, with Millet using the most in 2020 (24), indicating differing model complexities across crops and years

the table on the single page.

Crop type	Statistic parameters	2018	2020	2021	3-year model
<b>Maize</b>	<i>N</i>	32	28	23	83
	<i>No. of predictors</i>	10	9	–	13
	$R^2$	0.78	0.52	–	0.62
	<i>Adj. <math>R^2</math></i>	0.68	0.28	–	0.55
	<i>RMSE (kg/m<sup>2</sup>)</i>	0.056	0.033	–	0.065
	<i>Range (kg/m<sup>2</sup>)</i>	0.456	0.256	–	0.512
<b>Millet</b>	<i>nRMSE (%)</i>	12.28	12.89	–	12.70
	<i>N</i>	44	30	29	103
	<i>No. of predictors</i>	9	24	21	7
	$R^2$	0.46	0.95	0.64	0.32
	<i>Adj. <math>R^2</math></i>	0.32	0.71	–0.44	0.27
	<i>RMSE (kg/m<sup>2</sup>)</i>	0.053	0.013	0.029	0.056
<b>Sorghum</b>	<i>Range (kg/m<sup>2</sup>)</i>	0.328	0.224	0.260	0.344
	<i>nRMSE (%)</i>	16.16	5.80	11.15	16.28
	<i>N</i>	57	35	28	120
	<i>No. of predictors</i>	6	12	3	11
	$R^2$	0.41	0.56	0.23	0.30
	<i>Adj. <math>R^2</math></i>	0.34	0.32	0.13	0.23
<b>Beans</b>	<i>RMSE (kg/m<sup>2</sup>)</i>	0.050	0.029	0.032	0.047
	<i>Range (kg/m<sup>2</sup>)</i>	0.348	0.172	0.168	0.360
	<i>nRMSE (%)</i>	14.37	16.86	19.05	13.06
	<i>N</i>	31	0	12	43
	<i>No. of predictors</i>	9	–	–	10
	$R^2$	0.59	–	–	0.54
<b>Sesame</b>	<i>Adj. <math>R^2</math></i>	0.41	–	–	0.40
	<i>RMSE (kg/m<sup>2</sup>)</i>	0.033	–	–	0.034
	<i>Range (kg/m<sup>2</sup>)</i>	0.252	–	–	0.252
	<i>nRMSE (%)</i>	13.10	–	–	13.49
	<i>N</i>	0	28	29	57
	<i>No. of predictors</i>	–	22	13	17
	$R^2$	–	0.84	0.65	0.59
	<i>Adj. <math>R^2</math></i>	–	0.14	0.35	0.41
	<i>RMSE (kg/m<sup>2</sup>)</i>	–	0.007	0.023	0.019
	<i>Range (kg/m<sup>2</sup>)</i>	–	0.080	0.132	0.132
	<i>nRMSE (%)</i>	–	8.75	17.42	14.39

# Lasso Regression

## Why was multi-year modeling more robust?

- **Overfitting Reduction:**
  - Single-year models, trained on limited data, tend to select too many variables, leading to overfitting
  - Multi-year models, with a wider range of data, avoid this by focusing on the most consistently important predictors
- **Inter-annual Variability:**
  - Crop yields vary from year to year due to changing weather patterns, pests, and other factors
  - Multi-year models capture this variability better than single-year models, leading to more accurate predictions across different years
- **Wider Range of Yields:**
  - The three-year dataset used in the study encompassed a broader range of yield values than any single year
  - This wider range helps the model generalize better and make more accurate predictions for new data
- **Robustness:**
  - While multi-year models might have slightly lower R-squared values than the best single-year models, they are more robust against overfitting and thus more reliable for predicting yields in future years

# Nouna Health and Demographic Surveillance System

## Field Data

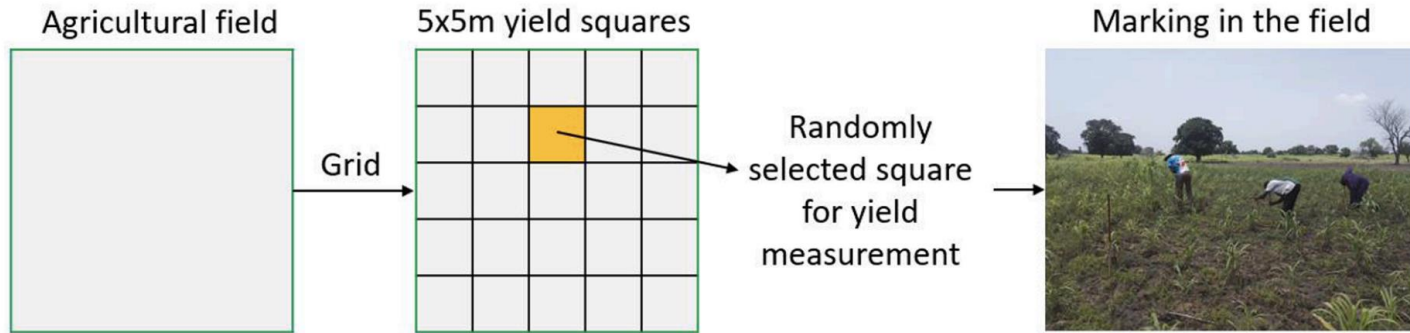
- **Collection:** Ground surveys, GPS measurements.
- **Metrics:** Crop type, planting dates, harvest dates, yield measurements.
- **Purpose:** Ground truth for calibrating and validating satellite-based models.
- **Challenges:** Labor-intensive, limited spatial coverage.
- **Importance:** Essential for model accuracy and validation.





# Nouna Health and Demographic Surveillance System

## Field Data



**Fig. 2.** Schematic representation of in-situ yield measurements. Each field, that was selected for in-situ measurements, was divided into 5x5m squares using a grid. The yield square, for which the harvest was measured, was randomly selected and then marked and protected in the field by circumferences. Picture copyright by Isabel Mank.

# Nouna Health and Demographic Surveillance System

## Challenges Data Collection



# Nouna Health and Demographic Surveillance System

## Challenges Data Collection

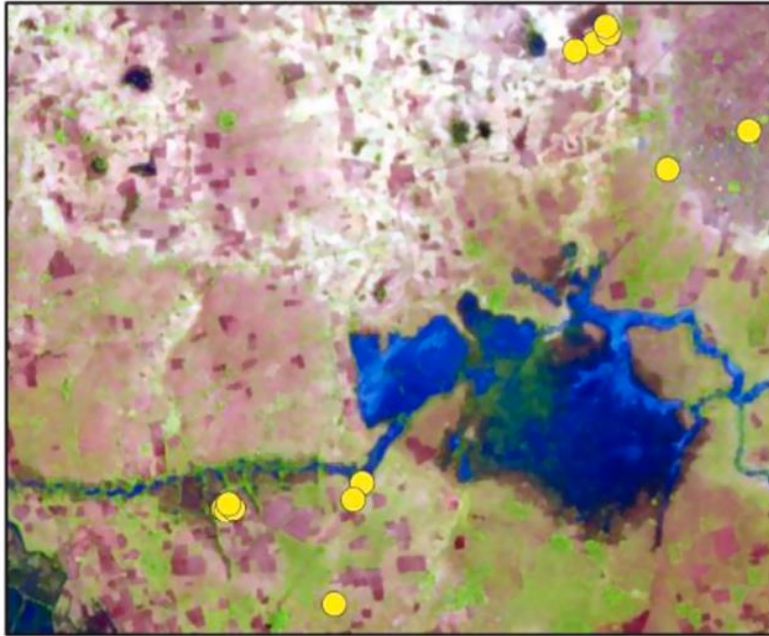
Flooding



# Nouna Health and Demographic Surveillance System

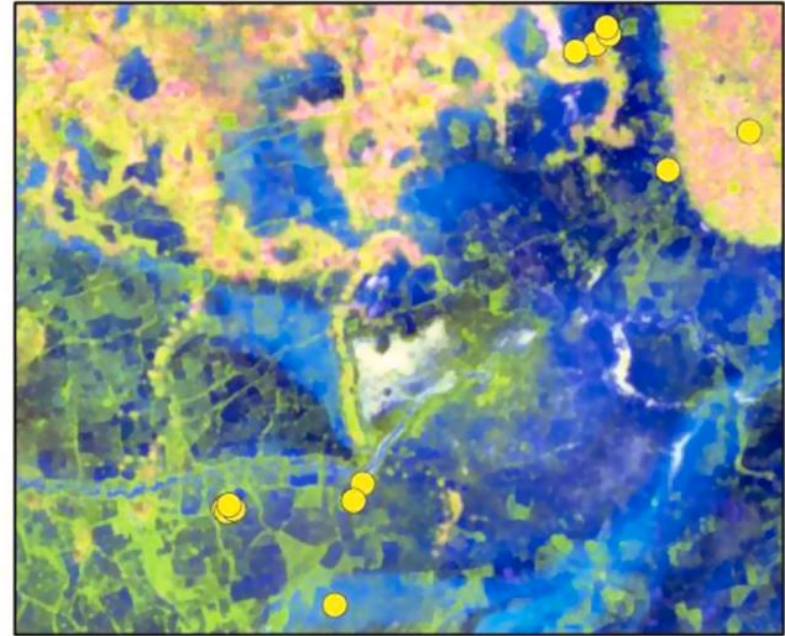
## Challenges Data Collection

Sentinel-2: 01/07/2021



0 0,5 1 2  
Kilometers

Sentinel-2: 20/08/2021





# Crop Yield

## Data Collection Outcomes

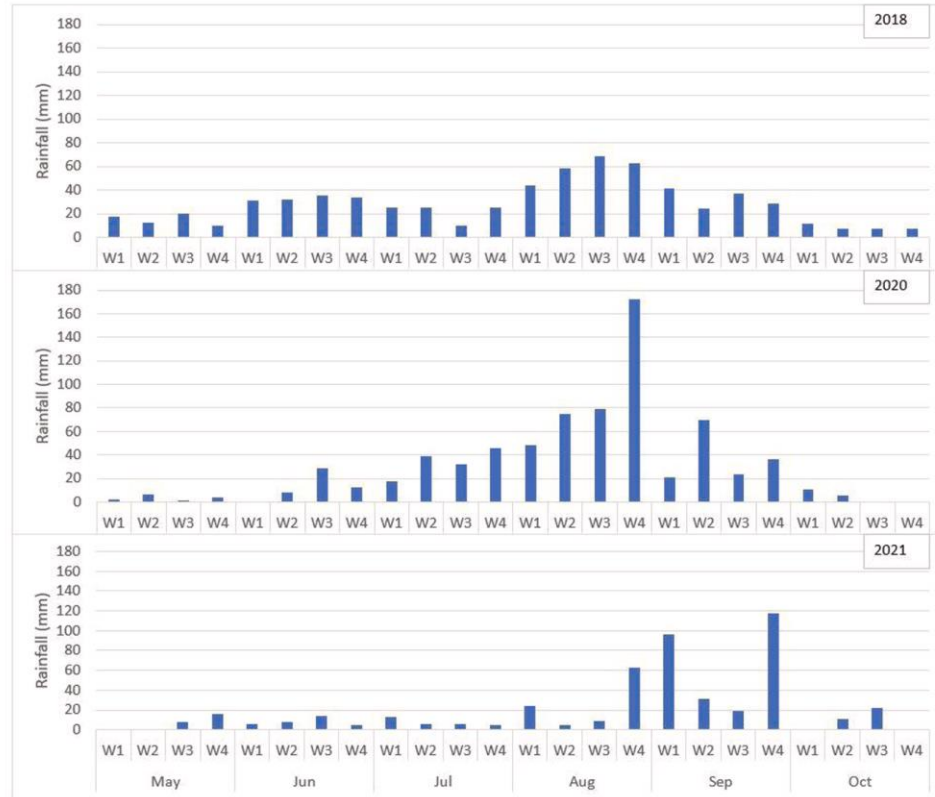
TABLE 1

Number of sampled yield squares and field boundaries per crop type in the years 2018, 2020, and 2021.

	2018		2020		2021		Total	
	Yield plots	Boun-daries	Yield plots	Boun-daries	Yield plots	Boun-daries	Yield plots	Boundaries
Maize	33	44	29	35	23	136	85	215
Millet	45	44	30	30	29	163	104	237
Sorghum	57	61	35	36	28	175	120	272
Beans	31	52	0	0	12	84	43	136
Sesame	0	0	30	33	29	134	59	167
Total	166	201	124	134	121	692	411	1027

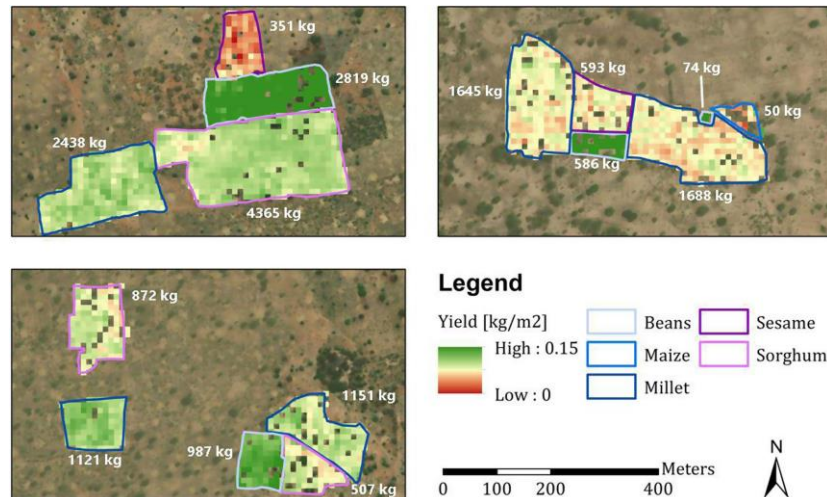
# Rainfall

## Data Collection Outcomes



# Key Findings

- Three-year models outperformed single-year models in terms of robustness and capturing inter-annual variability.
- More training data did not always lead to better model performance, highlighting the importance of data quality and representativeness.
- The models successfully predicted yield estimates at the household field level, providing valuable insights for agricultural interventions.





# Implications and Future Directions

- **Implications for agriculture and food security:**

- Improved yield monitoring and forecasting for better decision-making.
- Targeted interventions to address yield gaps and improve productivity.
- Enhanced food security assessments and early warning systems for droughts and other climate-related risks.

- **Future research directions:**

- Incorporate additional data sources, such as soil moisture and temperature, to improve model accuracy.
  - Explore the use of more advanced machine learning techniques, such as deep learning, for yield prediction.
  - Expand the study to other regions and crops to assess the generalizability of the approach.
-

# Group Activity

## Instructions:

4 groups - after 20 minutes of discussion, each group will share their key points with the class

## Questions:

1. What are the potential benefits and limitations of using satellite-based crop yield models in developing countries?
  2. How might the implementation of satellite-based crop yield models affect the social and economic aspects of rural communities?
  3. In what ways can satellite-based crop yield models contribute to enhancing food security in Sub-Saharan Africa? What limitations should be considered?
  4. What strategies can be employed to introduce and utilize crop yield models effectively, empowering farmers and improving their livelihoods? How can we overcome potential barriers to the adoption of crop yield models among small-scale farmers?
-

## Key Takeaways

- Crop yield modeling is essential for agricultural planning and food security.
  - LASSO regression is a powerful tool for crop yield modeling, offering advantages over traditional methods.
  - The case study from Burkina Faso demonstrates the potential of satellite data and LASSO regression for yield prediction in developing countries.
  - Continued research and innovation are needed to improve crop yield modeling and address the challenges of food security in a changing climate.
-