Understanding the Drivers of Wheat Yield Variability in Smallholder Systems in India: A Multi-Scale Analysis at the Field and Village scale

by

Jiangpan Bian

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science

(Environment and Sustainability)

in the University of Michigan

Winter 2026

Thesis Committee:

Associate Professor Meha Jain, Chair

Associate Professor Derek Van Berkel, Member

Acknowledgements

I would first like to thank my advisor, Dr. Meha Jain, who provided valuable insights helping me formulate and modify the research questions and methodology. Her patience and feedback helped me overcome a lot of challenges in my research. She consistently pushed me to sharpen my thinking and elevate the quality of my work.

I am also grateful to Professor Derek Van Berkel for his valuable comments and suggestions on my thesis paper, which greatly improved the clarity, depth, and rigor of my writing.

I would like to particularly acknowledge Dr. Victor Hugo Prudente and Dr. Monish Deshpande for their generous support and thoughtful feedback on my research and thesis. Our meetings were always highly productive, helping me refine my ideas and gain meaningful insights. Working with them has been an enriching and enjoyable experience.

In addition, I would like to thank my parents for their support and encouragement throughout my academic journey.

Finally, I would like to thank all the friends I have met at the University of Michigan. The exchange of ideas and perspectives we shared has broadened my horizons and made my life more vibrant and fulfilling. Thank you.

Abstract

Understanding yield variability in smallholder agricultural systems is critical for improving food security under climate change. Existing studies largely rely on field experiments, which often lack spatial representativeness at broader scales. This study investigates the drivers of wheat yield variability in 3 districts, Deoria, Gorakhpur, and Maharajganj, in eastern Uttar Pradesh, India at both the field and village scales.

This study focuses on the 2019 wheat growing season (Early November 2019 to late March 2020). We examine the effects of sowing date, field size, and socio-environmental factors on crop yield. Sowing date and crop yield are estimated using Sentinel-2 imagery. The relationships between yield, sowing date, and field size are analyzed at both the field and village scales, whereas socio-environmental factors are included only at the village scale to investigate the drivers of yield variability across space. The results show that delayed sowing is consistently associated with lower yield, although this relationship is weak at the field level and becomes more evident at the village level, suggesting that spatial aggregation reduces noise in remote sensing-derived estimates. While field size alone has no strong independent effect, its interaction with sowing date is significantly negative, indicating that larger fields are more sensitive to delayed sowing. At the village scale, socio-environmental variables further explain yield variability. Irrigation composition and access, particularly canal irrigation and distance to canals, play important roles, while agricultural labor and asset levels also contribute to observed patterns.

Overall, this study highlights the importance of multi-scale analysis and demonstrates the potential of integrating remote sensing with socio-environmental data to better understand productivity dynamics in smallholder systems.

Keywords: Smallholder agriculture; Wheat yield variability; Sentinel-2; Remote sensing; Sowing date; Multi-scale analysis; Socio-environmental factors; India

1. Introduction

Food security remains one of the most pressing global challenges, and the urgency of ensuring stable agricultural production has intensified due to climate change. Increasing climate variability, including extreme temperature events and irregular precipitation patterns, poses significant threats to crop productivity worldwide. Wheat is a staple crop critical to global food security, serving as a primary source of calories for a substantial share of the global population (Shiferaw et al., 2013). In India, wheat plays a particularly important role in agricultural production, where smallholder farming systems dominate the agricultural landscape (Chatrath, R., et al., 2007). Smallholder agriculture systems are characterized by small, fragmented fields, limited mechanization, and heterogeneous management practices, making them especially vulnerable to climate variability and environmental stressors (Bisht et al., 2020; Williams et al., 2018; Cohn et al., 2017). In addition, farmers often lack access to critical inputs and modern technologies necessary for high-productivity agriculture, further constraining their adaptive capacity and yield potential. Understanding the drivers of yield variability in smallholder systems is one way to improve agricultural productivity and strengthen food security under changing climatic conditions.

Investigating determinants of agricultural production in smallholder regions has several challenges. Previous studies often rely on field experiments or ground-based observations, which are typically limited to small spatial scales and require long-term monitoring (Liu, et al., 2023; Wang et al., 2013). These studies provide detailed insights into crop phenology and its response to management practices. For instance, Liu et al. show that delayed sowing shortens vegetative growth, extends reproductive growth, and reduces pre-winter resources. However, the small and spatially fragmented nature of smallholder farming systems limits the representativeness of in field experiments for capturing agricultural dynamics at regional scales. In addition, many smallholder systems are largely subsistence-oriented, meaning that a considerable proportion of agricultural production is used for household consumption, with only the remainder entering the market (Fan et al., 2013). This production structure further complicates standardized management and the collection of comprehensive agricultural statistics, thereby increasing the uncertainty in regional agricultural assessments.

Remote sensing provides a cost-effective approach for acquiring large-scale land surface information and has been widely applied in agricultural research (Atzberger, 2013; Carfagna et al., 2006). Satellite imagery enables continuous monitoring of vegetation dynamics across large spatial extents, offering new opportunities for studying crop phenology and agricultural production systems. Vegetation indices, such as the Normalized Difference Vegetation Index (NDVI), have been used to detect crop phenology based on a crop's unique temporal signature. However, despite the increasing availability of remote sensing imagery, it has been difficult to map the production or sowing date of smallholder farms, given that the size of an individual field (<2 ha) is typically smaller than the spatial resolution of most accessible satellite

imagery (e.g., 30 m Landsat, 250 m MODIS), which often results in mixed pixels and reduces the accuracy of field-level mapping (Jain et al., 2013).

Recent studies have begun to address these challenges by utilizing remote sensing imagery with higher spatial and temporal resolution, together with machine learning approaches (Duncan et al., 2015; Lungu rt al, 2020; Joshi et al., 2023; Gao et al, 2021; Azzari et al., 2017). For example, Mei et al. (2022) used high-resolution WorldView-3 imagery (0.5 m) combined with a deep learning model to delineate smallholder field boundaries in northeastern India. These advances have greatly enhanced the capacity of remote sensing to study smallholder farming landscapes. Nevertheless, much of the existing research has primarily focused on methodological development, such as improving field boundary detection, crop classification, or yield estimation models (Mei et al, 2022; Ji et al. 2018; Ferencz et al., 2004). Relatively fewer studies have applied these approaches to investigate the drivers of agricultural production in smallholder systems.

Several studies have investigated the factors influencing crop yield, but they mainly rely on field experiments conducted at individual experimental stations (Ran et al., 2018; Berzsenyi et al., 2008). Field-level and village-level analyses remain limited. Furthermore, the effects of the influential factors may vary across spatial scales and regions, making it difficult to generalize findings derived from localized experiments.

This study uses high spatiotemporal resolution Sentinel-2 imagery to estimate sowing date and wheat yield at the pixel level across Deoria, Gorakhpur, and Maharajganj in Uttar Pradesh, India, and then aggregates the results to the field and village levels to investigate the drivers of wheat yield variability. The objectives of this study are as follows:

(i) What is the correlation between sowing date, field size, and wheat yield at the field-level?

(ii) What is the correlation between sowing date, field size, socio-environmental factors, and wheat yield at the village-level?

By identifying the key environmental and management factors that influence yield, researchers and policymakers can design more effective agricultural management strategies and interventions to enhance productivity and resilience in smallholder farming systems.

2. Study area and data

2.1 Study Region

Winter wheat is an important staple crop and is widely cultivated in northern and western India. Major wheat-producing regions are mainly distributed across several states, including Uttar Pradesh, Punjab, and Haryana.

In this study, three districts in eastern Uttar Pradesh, Deoria, Gorakhpur, and Maharajganj, were selected as the study region (Figure 1). These districts are in the eastern part of the Indo-Gangetic Plains, one of the most important agricultural regions in India (Rattan et al., 2021). Agriculture in these districts is dominated by smallholder farming systems. Winter wheat is commonly grown as part of the rice–wheat cropping system, which is the dominant cropping pattern in the region. There is high heterogeneity in sowing date in this region, ranging from mid-November to early January (Liu et al., 2022).

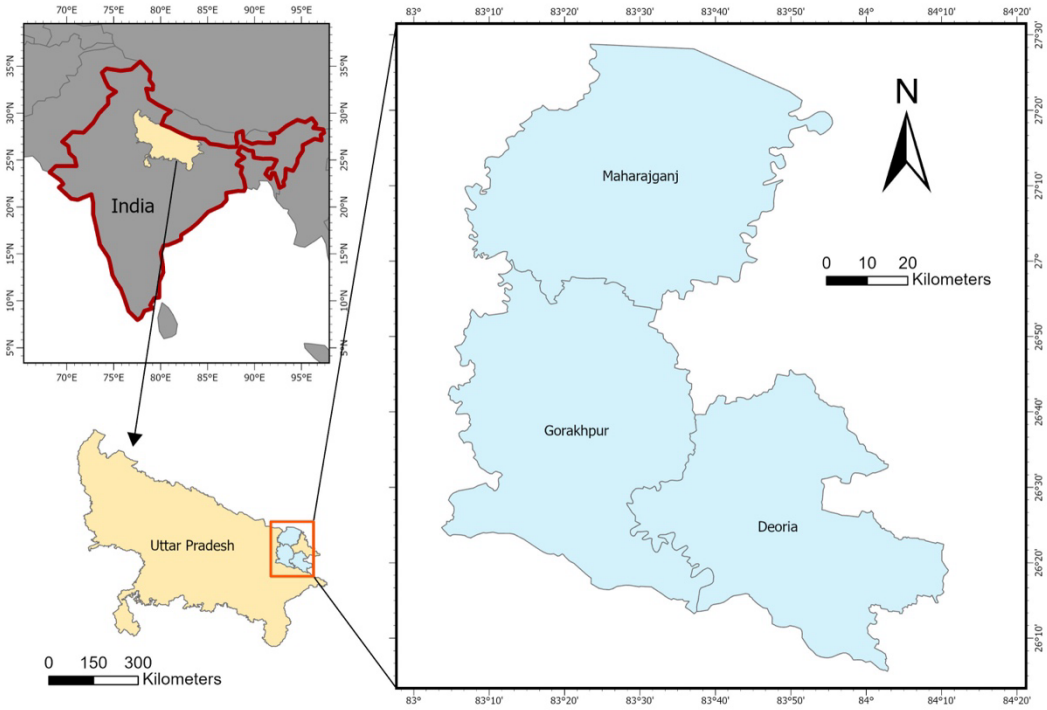


Figure 1. Three districts in eastern Uttar Pradesh, India that represent the study area.

2.2 Remote sensing data

We accessed the surface reflectance product Sentinel-2 Multispectral Instrument Level-2A (COPERNICUS/S2_SR_HARMONIZED) satellite imagery through Google Earth Engine (GEE) (Gorelick et al, 2017) to estimate sowing date and wheat yield. The imagery is at a spatial resolution of 10 m × 10 m, enabling the detection of fine-scale field characteristics and making it suitable for analyzing fragmented fields in smallholder agricultural systems. The administrative boundaries of Deoria, Gorakhpur, and Maharajganj districts were obtained from the FAO Global Administrative Unit Layers dataset (FAO/GAUL/2015/level2) through GEE, and the satellite imagery was clipped to these boundaries to define the study area.

Cloud contamination can substantially affect the availability and reliability of vegetation index observations derived from optical satellite imagery, by obscuring the land surface, introducing noise through cloud shadows and atmospheric scattering, and creating temporal gaps in the time series. All pixels affected by clouds and cloud shadows were masked. We linked the image collection with the Google Cloud Score+ dataset (GOOGLE/CLOUD_SCORE_PLUS/V1/S2_HARMONIZED) and used the cs band as a pixel-level cloud quality indicator. Pixels with cs < 0.4 were masked out, and only pixels with relatively high observation quality were retained for subsequent analysis.

To further constrain the analysis to agricultural areas, a cropland mask derived from the ESRI Global Land Cover 10 m dataset (ESRI Global LULC) was applied, retaining only pixels classified as cropland.

After masking out clouds and non-cropland pixels, we calculated the Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1973) and the Green Chlorophyll Vegetation Index (GCVI) (Gitelson et al., 2003) at the pixel level. NDVI is calculated using the following formula:

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

GCVI is calculated using the following formula:

$$GCVI = \frac{NIR}{GREEN} - 1$$

The Sentinel-2 spectral bands used in this study and their corresponding wavelengths are summarized in Table 1. The red band is strongly absorbed by chlorophyll, the green band is partially reflected by vegetation, and the near-infrared band is highly reflected due to leaf cellular structure, making these bands sensitive to vegetation conditions.

Band	**Name**	**Wavelength (nm)**	**Spatial Resolution**	**Usage in this study**
B3	Green	560	10 m	GCVI calculation
B4	Red	665	10 m	NDVI calculation
B8	Near Infrared (NIR)	842	10 m	NDVI and GCVI calculation

Table 1. Sentinel-2 bands used to calculate NDVI and GCVI

The resulting NDVI time series was used to characterize crop phenological dynamics for each field and to identify the start-of-season (SoS) inflection point, following approaches developed in previous studies (Prudente et al. 2025; Diao, 2020; Liu et al., 2023; Rodigheri et al., 2023; Urban et al., 2018).

2.3 Field boundary data

Field boundary data were processed using methods from Mei et al. (2022). Very-high-resolution WorldView-3 satellite imagery with a spatial resolution of 0.5 m was used as the primary input. Field boundaries were detected using a Mask Region-Based Convolutional Neural Network (Mask R-CNN), a deep learning model designed for object detection and instance segmentation (). The resulting field boundary polygons were used as the basic spatial units for field-level aggregation and analysis in this study.

2.4 Census data

We compiled data from multiple village-level census statistics from different Indian ministries (originally compiled and described in Jain et al. 2021) and merged these data with village-level shapefiles. Specifically, we obtained socio-demographic variables and percent area that was irrigated by different sources from the Houses, Household Amenities, and Assets census (2011) and the India Demographic census (2001). Distance to the nearest permanent canal was calculated using the nearest distance algorithm in QGIS and a shapefile on global canals produced by the Digital Chart of the World (2009). We derived and extracted several different rainfall variables, including monsoon start date, monsoon end date, and total seasonal precipitation using CHIRPS rainfall data (C. Funk, et al, 2015). We also derived and extracted mean winter temperature using the CRU dataset. All weather data were processed and extracted using GEE. For soil type, we used the ISRIC soil grids data (L. Poggio, et al, 2021) and extracted the most common soil type at the village-scale for all villages using QGIS software.

3. Methods

We present the overall workflow and the general methodology in Figure 2. The framework contains three steps. First, pixel-level sowing date and yield were estimated using Sentinel-2 imagery. Second, the pixel-level estimates were aggregated to field polygons using zonal statistics, and the relationships between yield, sowing date, and field size were analyzed at field level. Finally, field-level variables were further aggregated to the village level and integrated with socio-environmental factors to examine the drivers of yield variability in smallholder agricultural systems.

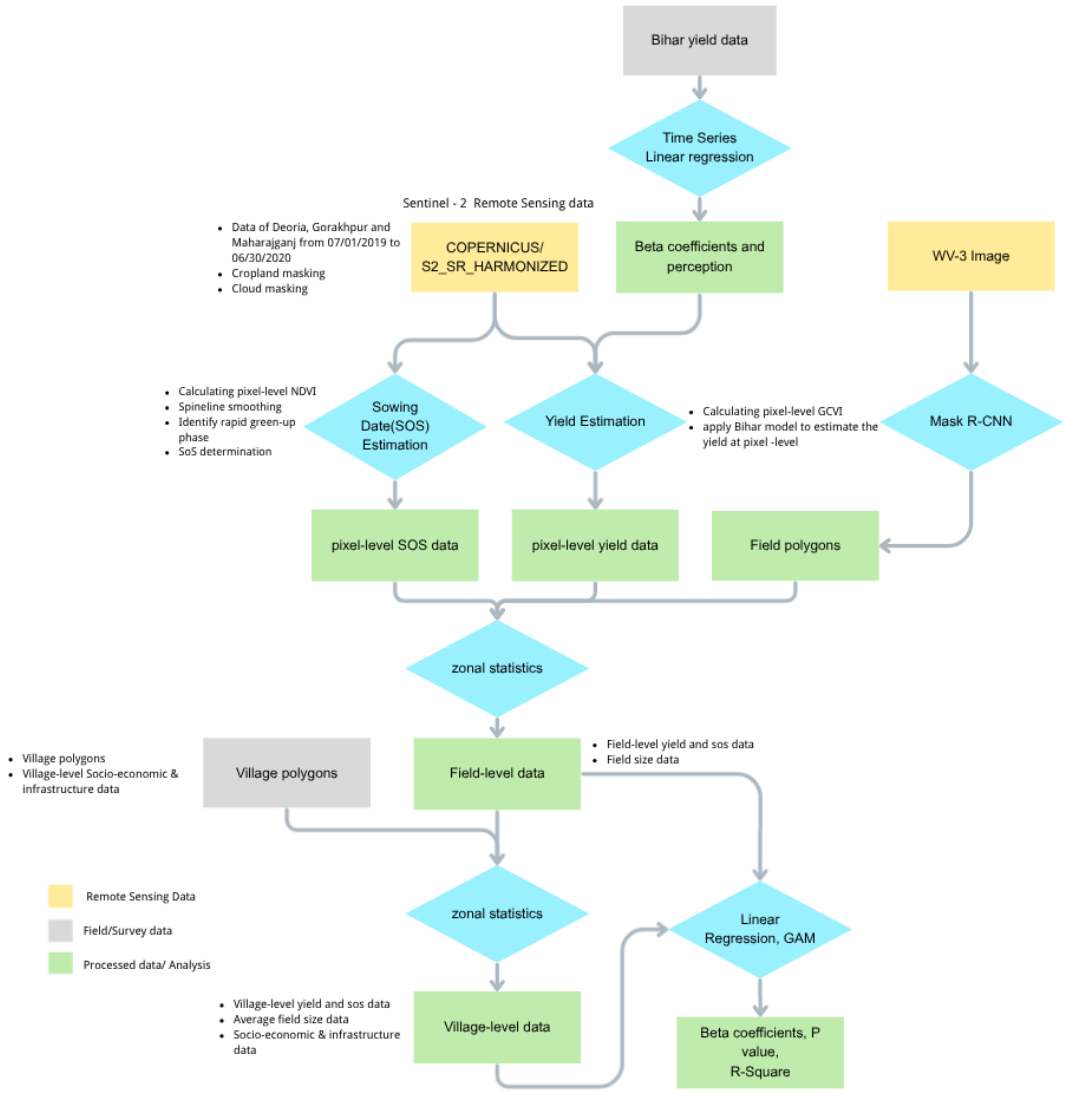


Figure 2. Workflow for data production and analyses

3.1 Sowing date estimation

3.1.1 Pixel-level sowing date estimation

The sowing date of winter wheat was estimated based on the start-of-season (SoS) derived from the NDVI time series.

The approach proposed by Prudente et al. (2025) was adapted to operate at the pixel level for each cropland pixel. The NDVI time series for each cropland pixel was then smoothed using a spline-based method to reduce noise while preserving key phenological signals. Then, the first derivative of the smoothed NDVI time series was

computed, and the earliest positive inflection point within a predefined search window was selected as the SoS. The search window was defined based on regional agronomic knowledge, starting from the earliest possible sowing date (November 1) and ending at the date of maximum NDVI.

Since SoS corresponds to crop emergence rather than the actual sowing date, a 10-day correction factor was applied to estimate the sowing date (SoS − 10 days), reflecting the typical time interval between sowing and crop emergence reported in previous studies. Previous studies have demonstrated that this approach can provide reliable estimates of crop sowing dates in smallholder agricultural systems, where field-level management data are often unavailable. This approach allows sowing date to be estimated without ground calibration data and is suitable for large-scale monitoring in smallholder agricultural systems.

Sowing date estimation was conducted independently for each Sentinel-2 image tile to improve computational efficiency. The resulting tile-level outputs were subsequently mosaicked to produce a continuous sowing date map for the entire study area (Figure 3). In regions with overlapping tiles, the first valid observation was retained to assign pixel values. Finally, the estimated sowing dates were expressed as Day of Year (DOY) for subsequent analysis.

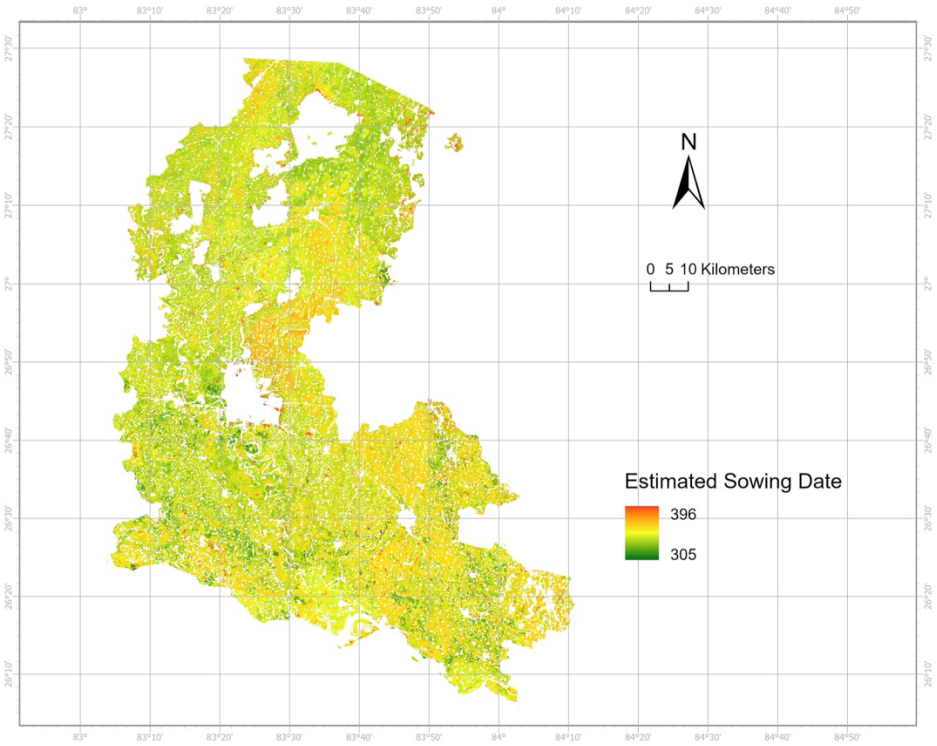


Figure 3. Map of estimated sowing dates across the study region.

3.1.2 Field-level sowing date estimation

Field-level sowing dates were derived by aggregating pixel-level estimates to field polygons. Pixel-level sowing dates, obtained from the NDVI time-series analysis, were summarized within each field boundary using zonal statistics. Zonal statistics were calculated by averaging all pixels whose centroids fall within each field polygon, thereby ensuring that only pixels primarily associated with a given field were included in the aggregation.

This aggregation reduces pixel-level noise and provides a more stable estimate of sowing timing at the management unit scale. The resulting field-level sowing date was then used in regression analyses to examine its relationship with wheat yield and its interaction with field size.

3.1.3 Village-level sowing date estimation

Village-level sowing dates were calculated by aggregating field-level sowing dates to the village scale. To better reflect the contribution of fields of different sizes, an area-weighted average was applied:

$$\bar{X} = \frac{\sum_{i=1}^{n} A_i * X_i}{\sum_{i=1}^{n} A_i}$$

$\bar{X}$ represents the aggregated variable (e.g., sowing date or yield) for the village, X_i is the value for field *i,* A_i is the area of field *i, n is* the is the total number of fields within village.

This weighting scheme ensures that larger fields, which contribute more to overall production, have a proportionally greater influence on the aggregated sowing date. The resulting village-level sowing date was used in subsequent analyses to assess its relationship with yield and its interaction with other socio-environmental factors.

3.2 Wheat yield estimation

3.2.1 Field-level yield estimation

Previous studies have demonstrated that the GCVI is strongly correlated with crop biomass and yield, making it a reliable predictor for yield estimation using remote sensing data (Li et al., 2024). To estimate wheat yield at the field level, we applied a linear regression model based on GCVI time series derived from Sentinel-2 imagery.

The linear regression model requires valid observations within each time window. Persistent cloud contamination can lead to missing GCVI values during certain periods of the growing season. To reduce the impact of missing observations, we analyzed cloud coverage across the 2019 wheat growing season in the three districts of Uttar Pradesh. Time windows with relatively low cloud coverage were identified

and used to extract GCVI features for the yield estimation model.

GCVI values were calculated from Sentinel-2 imagery and aggregated across several key time windows during the wheat growing season. These temporal GCVI features capture the crop growth dynamics and were used as predictors in a linear regression model to estimate wheat yield.

Due to the lack of field-level ground truth yield data for Uttar Pradesh in 2019, the yield estimation model was trained using yield observations from the neighboring state of Bihar in 2022, where field survey data were available. Bihar shares similar agro-ecological conditions, cropping systems, and wheat management practices with eastern Uttar Pradesh, making it a suitable proxy for model training.

Specifically, field-level yield observations from Bihar were linked with corresponding GCVI time-series features derived from Sentinel-2 imagery, and a linear regression model was trained to establish the relationship between vegetation dynamics and crop yield:

$$Yield = \beta_0 + \sum_{i=1}^{8} \beta_i * GCVI_i + \epsilon$$

$GCVI_i$ denotes the GCVI value for at time window i, β_i is the coefficient corresponding to each time window, and ϵ is the error term.

The performance of the linear regression model was evaluated using the coefficient of determination (R^2), mean squared error (MSE), and root mean squared error (RMSE). The model achieved an R^2 of 0.3492, indicating that approximately 34.9% of the variability in wheat yield can be explained by the GCVI-based predictors. The MSE and RMSE were 0.4777 and 0.6912, respectively, reflecting a moderate level of prediction error.

The resulting beta coefficients and intercept are shown in Table 2. They were then applied to the study area to estimate pixel-level wheat yield based on GCVI time-series data. The resulting spatial distribution of estimated yield is presented in Figure 4.

Variable	Time window	Coefficient (β)
β_0 (Intercept)	-	3.7167070756688876
β_1	11/01/2019 – 11/16/2019	-0.0756
β_2	11/16/2019 – 12/01/2019	-0.3559
β_3	12/01/2019 – 12/16/2019	0.0952
β_4	01/16/2020 – 02/01/2020	-0.2197
β_5	02/01/2020 – 02/16/2020	0.2409
β_6	03/01/2020 – 03/16/2020	-0.0793
β_7	03/16/2020 – 04/01/2020	-0.0038
β_8	04/01/2020 – 04/15/2020	0.3407

Table 2. The time periods for GCVI mosaics used to develop the yield prediction

algorithm.

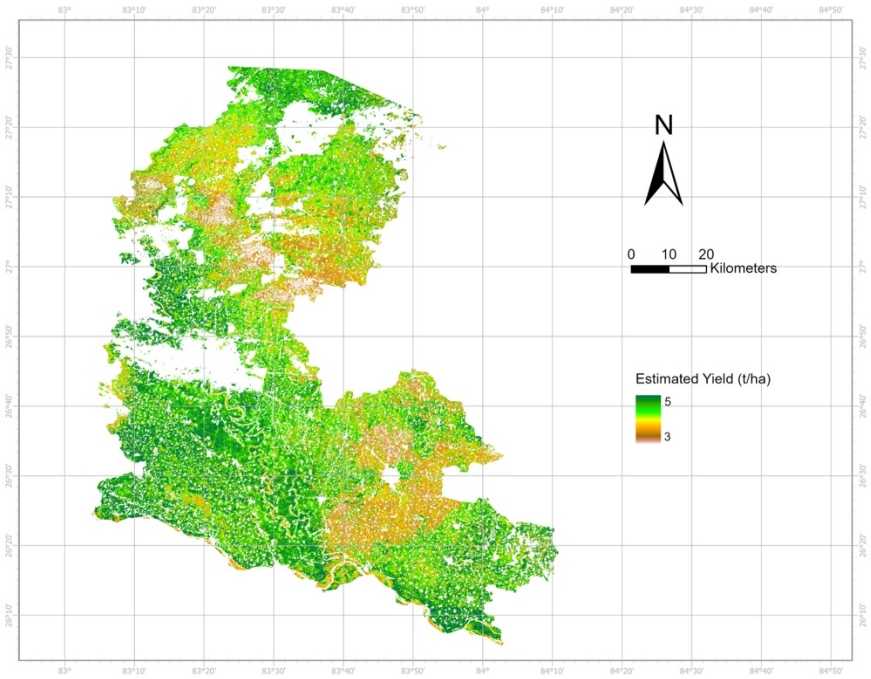


Figure 4. Map of estimated yield across the study region.

3.2.2 Village-level yield estimation

Village-level wheat yield was estimated by aggregating field-level yield values using an area-weighted average, consistent with the approach applied to sowing date (see Section 3.1.2). This ensures that larger fields contribute proportionally more to the aggregated estimate.

3.3 Stratified random sampling of field polygons

To ensure representative sampling across different field sizes and spatial regions, a stratified random sampling approach was applied to the field polygon dataset. The study area was divided into eight spatial zones (Z1–Z8) to facilitate efficient data processing and to capture spatial heterogeneity across the region.

Field polygons were then stratified into three groups based on field area: small fields (< 500 m²), medium fields (500–1000 m²), and large fields (> 1000 m²). This classification was designed to explicitly account for the fragmented nature of smallholder agricultural systems and to ensure balanced representation of different field sizes in the analysis.

Within each spatial zone, 1,000 field polygons were randomly sampled from each size group, resulting in a total of 24,000 sampled fields across all zones (Figure 6). This stratified design ensures that both spatial variability and field size heterogeneity are

adequately captured, while also maintaining computational efficiency for large-scale analysis.

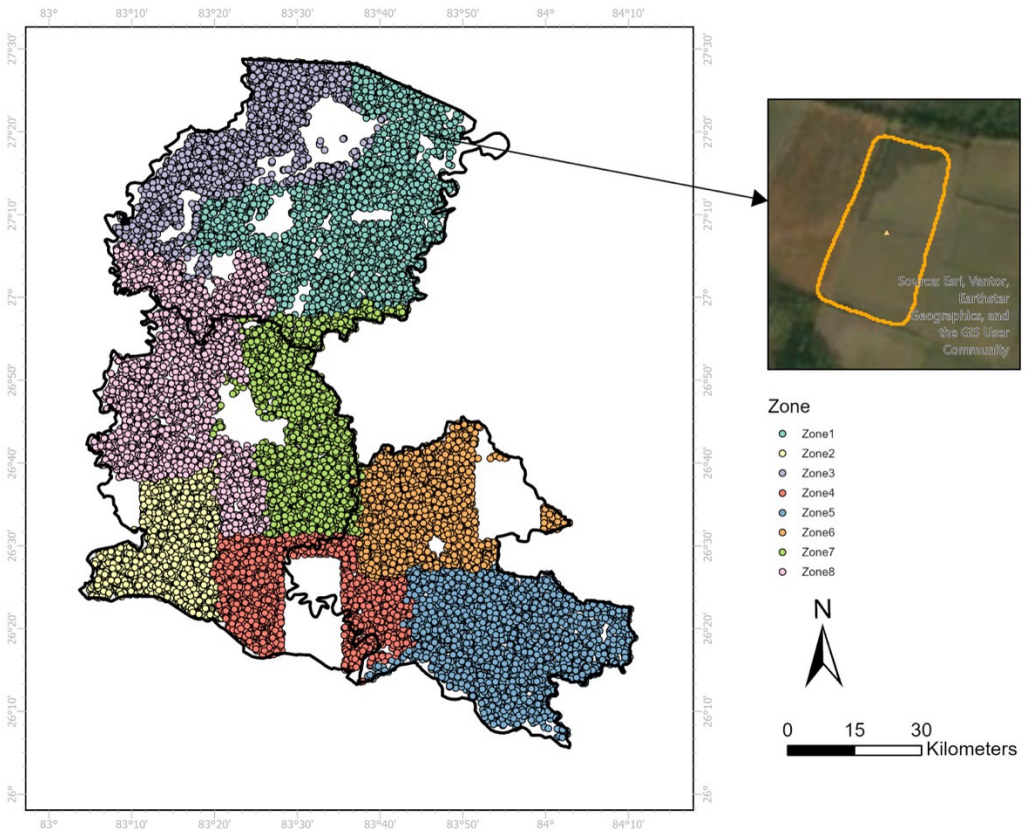


Figure 6. Zones used for stratified sampling of small, medium, and large fields.

3.4 Data analysis

Data analysis was conducted at both the field level and village level to examine the drivers of wheat yield variability across scales.

At the field level, the area variable was log-transformed prior to analysis due to its highly right-skewed distribution, with a large proportion of small fields and a few relatively large fields. Applying a logarithmic transformation helps reduce the influence of extreme values and improves the stability and interpretability of the regression model. Specifically, three regression models were fitted: (1) yield as a function of sowing date, (2) yield as a function of log (field area), and (3) yield as a function of sowing date, log (field area), and their interaction:

$$Yield = \beta_0 + \beta_1 * Sowing\ date + \beta_2 * \log(area) + \beta_3 * [Sowing\ date * \log(area)]$$

At the village level, we replicated the same set of analyses conducted at the field level, including: (1) modeling yield as a function of sowing date, (2) modeling yield

as a function of field area, and (3) modeling yield as a function of sowing date, field area, and their interaction:

$Yield = + * Sowing\ date + * area + * (Sowing\ date * area)$

A key difference is that the logarithmic transformation of field area was not applied at the village level, as the distribution of mean field area across villages was approximately normal rather than highly skewed.

In addition, linear regression models were also used to evaluate the combined effects of biophysical and socio-environmental factors on wheat yield. These variables include irrigation characteristics, infrastructure accessibility, soil type, and socio-economic indicators, which are shown in Table 3.

Category	Variable	Description
Irrigation	irrper	percent of agricultural land that is irrigated
	canal_ip	% of irrigation from canals
	well_ip	percent of irrigation from groundwater
Socio-economic	assetind	"Asset index (measure of how wealthy households are on average, higher numbers are more wealthy villages)"
	aglabper	% of agricultural labor population
Environmental	soilwb	Soil type (world bank classification)
	permcanal_dist	Distance to permanent canal
Fixed effects	DISTRICT	District-level fixed effects

Table 3. Socio-environmental factors included in village-level analysis

We used the following linear regression model to examine village-level wheat yield variability:

$Yield \sim Average\ sowing\ date \times Average\ field\ area + socio_environmental\ covariates.$

To further investigate how irrigation interacts with crop management, additional models were fitted including interaction terms between sowing date and irrigation variables:

$Yield \sim Average\ sowing\ date \times irrigation\ rate + socio_environmental\ covariates.$

$Yield \sim Average\ sowing\ date \times canal\ irrigation\ rate + socio_environmental\ covariates.$

$$Yield \sim Average\ sowing\ date\ \times well\ irrigation\ rate\ +\ socio_environmental\ covariates.$$

All regression analyses were performed using ordinary least squares (OLS), and model performance was evaluated using standard metrics including R^2 and statistical significance of coefficients.

4. Results

4.1 Field-level Relationships Between Sowing Date, Field Size, and Yield

First, the relationship between yield and sowing date was analyzed (Figure 7). The results indicate a weak but consistent negative relationship, suggesting that delayed sowing tends to reduce yield. This finding is consistent with previous studies in the Indo-Gangetic Plains, which show that late sowing of wheat leads to significant yield declines due to a shortened growing period and increased exposure to terminal heat stress (Mkondiwa et al., 2022; McDonald et al., 2022). However, the explanatory power of the model is limited ($R^2 \approx 0.065$), suggesting that sowing date alone cannot fully explain yield variability, and that other biophysical and management-related factors may also play important roles.

Second, the relationship between yield and field size was evaluated (Figure 7). Field size alone showed no statistically significant effect on yield ($R^2 \approx 0$), indicating that field size does not directly determine productivity in smallholder systems. This is consistent with previous studies highlighting that the farm size–productivity relationship is highly sensitive to measurement error and may vary depending on data accuracy and contextual factors (Carletto et al., 2013).

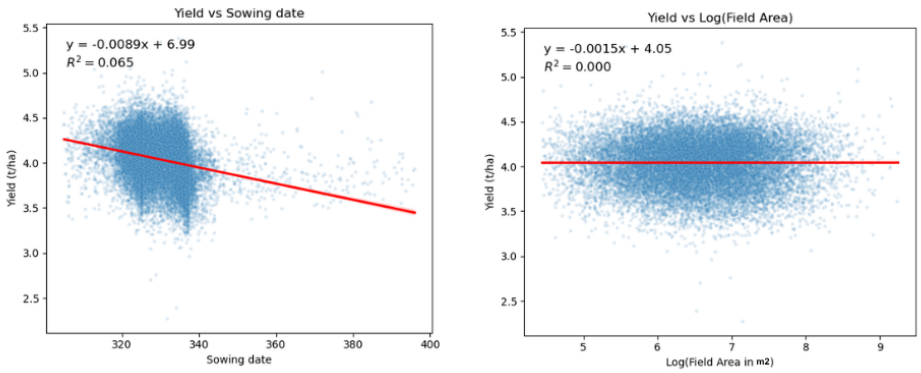


Figure 7. Relationships between wheat yield, sowing date, and field size at the field level

To further explore potential interactions, a model including sowing date, log-transformed field area, and their interaction was fitted.

In this model, sowing date alone was not statistically significant (p-value ≈ 0.834), while field size exhibited a positive effect on yield. Importantly, the interaction between sowing date and field size was strongly negative and statistically significant,

indicating that the negative effect of delayed sowing is minimal in small fields but becomes increasingly pronounced as field size increases.

	coefficient	p-value
β0	3.8784	0
β1	0.0004	0.834
β2	0.4718	0
β3	-0.0014	0

Table 4. Coefficients for the field-level linear regression model

4.2 Village-level Relationships Between Sowing Date, Field Size, and Yield

First, the relationship between average yield and average sowing date was examined. The results show a negative association, indicating that villages with later sowing dates tend to have lower yields (Figure 8).

Second, the relationship between yield and average field size was analyzed. Field size alone exhibits limited explanatory power, indicating that it is not a strong independent predictor of yield at the village level (Figure 8).

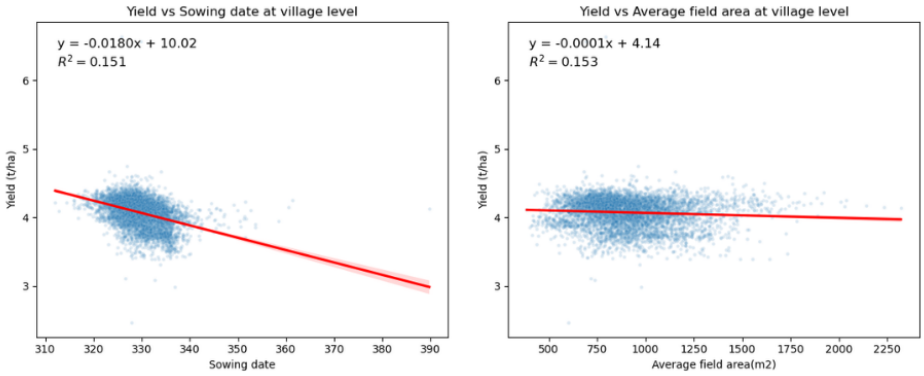


Figure 8. Relationships between wheat yield, sowing date, and field size at the village level

To further explore potential interactions, a model including average sowing date, average field size, and their interaction term was fitted.

In this model, average sowing date shows a significant negative effect on yield (p-value ≈ 0), in contrast to the field-level analysis where its effect was weak and not statistically significant. Meanwhile, average field size exhibits a positive effect on yield. Importantly, the interaction between sowing date and field size is negative and statistically significant, indicating that the adverse effect of delayed sowing is amplified in villages with larger average field sizes.

	coefficient	p-value
β0	7.2283	0
β1	-0.0095	0
β2	0.0029	0.001
β3	-0.000008857	0.001

Table 5. Coefficients for the village-level linear regression model

These findings suggest that the negative effect of delayed sowing is amplified in villages with larger average field sizes. Compared to the field-level analysis, the relationships observed at the village level are stronger and more consistent, indicating that spatial aggregation may reduce noise and reveal underlying patterns in yield variability.

4.3 Effects of Socio-environmental Factors on Yield at village level

To better explain the variation in wheat yield at the village level, socio-environmental variables were incorporated into the regression model alongside average sowing date and average field size.

The results show that several socio-environmental factors have significant effects on yield (Table 6). Among irrigation-related variables, the overall irrigation percentage does not exhibit a statistically significant effect. However, the composition of irrigation sources plays an important role. Canal irrigation is associated with a significant effect on yield, while well irrigation shows a negative effect. In addition, greater distance to canals is negatively associated with yield, suggesting that limited access to irrigation infrastructure may constrain agricultural productivity.

Socio-economic factors also contribute to yield variation across space. The asset index shows a negative relationship with yield, which may reflect livelihood diversification rather than direct agricultural investment in wealthier households. In contrast, the proportion of agricultural labor has a positive effect on yield, indicating that greater labor availability may enhance farm management and productivity.

Regional differences are also observed, with Gorakhpur showing higher yield levels compared to Deoria and Maharajganj, highlighting spatial heterogeneity in agricultural performance across the study area.

Variable	Coef.	Std. Err.	Significance
Intercept	5.844	0.795	***
Soil type (ref = base)			
Soil (27)	0.055	0.159	
Soil (48)	0.136	0.161	
Soil (56)	-0.047	0.159	
Soil (79)	0.061	0.159	

Soil (116)	0.016	0.159	
District (ref = Deoria)			
Gorakhpur	0.238	0.008	***
Maharajganj	-0.064	0.008	***
Mean sowing date	-0.0056	0.002	**
Mean field area	0.0030	0.001	***
Sowing date × Area	-9.13e-06	2.46e-06	***
Irrigation (%)	0.010	0.012	
Canal irrigation (%)	-0.059	0.010	***
Well irrigation (%)	-0.019	0.005	***
Distance to canal	-0.420	0.125	***
Asset index	-0.0003	0.00007	***
Agricultural labor (%)	0.081	0.043	*
Observations	5056		
R^2	0.431		

Table 6. Village-level regression results for wheat yield with socio-environmental controls

In models that include interaction terms between sowing date and irrigation variables, the results show that neither the overall irrigation percentage nor well irrigation has a significant interaction with sowing date (Tables 7 and 8).

While canal irrigation is negatively associated with yield in the baseline model (Table 6), the interaction model (Table 9) reveals a different pattern: canal irrigation exhibits a positive main effect, but a negative interaction with sowing date.

This indicates that while canal irrigation generally enhances yield, it also amplifies the negative impact of delayed sowing. In other words, in villages with higher reliance on canal irrigation, the yield penalty associated with late sowing becomes more pronounced. This pattern suggests that canal-based irrigation systems, although beneficial for overall production, may be less flexible in accommodating suboptimal sowing timing compared to other irrigation sources.

Overall, these findings suggest that yield variability at the village level is jointly shaped by both biophysical conditions and socio-economic factors, as well as their interactions with crop management practices. Compared to models using sowing date

and field size alone, the inclusion of socio-environmental variables substantially improves the explanatory power, highlighting the importance of considering broader system-level dynamics in smallholder agricultural systems.

Variable	Coef.	Std. Err.	Significance
Intercept	7.055	0.815	***
Mean sowing date	-0.0093	0.002	***
Irrigation (%)	1.819	0.882	**
Sowing date × Irrigation	-0.0055	0.003	**
Mean field area	-9.96e-06	1.08e-05	
Canal irrigation (%)	-0.058	0.010	***
Well irrigation (%)	-0.0195	0.005	***
Observations	5056		
R^2	0.430		

Table 7. Village-level regression results for wheat yield including interaction between sowing date and irrigation

Variable	Coef.	Std. Err.	Significance
Intercept	8.616	0.297	***
Mean sowing date	-0.0140	0.001	***
Well irrigation (%)	0.0566	0.353	
Sowing date × Well irrigation	-0.0002	0.001	
Mean field area	-1.01e-05	1.08e-05	
Irrigation (%)	0.0093	0.012	
Canal irrigation (%)	-0.0587	0.010	***
Observations	5056		
R^2	0.430		

Table 8. Village-level regression results for wheat yield including interaction between sowing date and well irrigation

Variable	Coef.	Std. Err.	Significance
Intercept	8.378	0.246	***
Mean sowing date	-0.0133	0.001	***
Canal irrigation (%)	2.708	0.664	***
Sowing date × Canal irrigation	-0.0084	0.002	***
Mean field area	-9.64e-06	1.08e-05	
Irrigation (%)	0.0107	0.012	
Well irrigation (%)	-0.0197	0.005	***
Observations	5056		
R^2	0.432		
Adj. R^2	0.430		

Table 9. Village-level regression results for wheat yield including interaction between sowing date and canal irrigation

(The tables report key variables of interest, while soil and other control variables are omitted for clarity. Full results are available in Appendix)

5. Discussion

This study finds that sowing date has a negative effect on wheat yield, with delayed sowing associated with reduced productivity. This result is consistent with previous studies showing that late sowing shortens the growing period and increases exposure to terminal heat stress during critical growth stages.

However, the effect of sowing date differs across spatial scales. At the field level, sowing date alone shows a weak and statistically insignificant effect when accounting for its interaction with field size. In contrast, at the village level, sowing date emerges as a significant predictor of yield. This scale-dependent pattern suggests that aggregation reduces noise in pixel- or field-level estimates and reveals more robust relationships between phenology and yield.

Field size alone does not show a strong independent effect on yield at either scale, but the interaction between sowing date and field size is consistently negative and significant. This may indicate that the negative impact of delayed sowing is amplified in larger fields. One possible explanation is that larger fields are more difficult to manage precisely, limiting farmers' ability to optimize sowing timing across the entire field. In contrast, smaller fields may allow for more flexible and timely management, reducing the yield penalty associated with delayed sowing.

At the village level, irrigation patterns help explain the observed relationship between sowing date and yield. A higher share of canal irrigation may indicate underlying water constraints, such as insufficient rainfall or limited access to alternative water sources. In such contexts, agricultural production becomes more dependent on canal water delivery. However, canal irrigation systems are often less flexible in timing compared to decentralized sources like wells, as water distribution is typically scheduled and managed at a larger scale. This limited flexibility means that farmers may have less control over precisely when they can irrigate their fields, which can delay optimal sowing. As a result, in villages with higher reliance on canal irrigation, the negative effects of delayed sowing on yield may be more pronounced. In contrast, farmers with access to more flexible irrigation sources can better align sowing with optimal conditions, thereby reducing the yield penalty associated with late planting.

This study provides a multi-scale perspective on yield variability in smallholder agricultural systems by integrating field-level and village-level analyses. By combining remote sensing–derived indicators (e.g., sowing date and yield) with socio-environmental variables, it captures both biophysical and human dimensions of agricultural production. The comparison across spatial scales is a key strength, as it reveals scale-dependent relationships that would be overlooked in single-scale analyses. In particular, the finding that aggregation at the village level helps uncover clearer relationships between sowing date and yield highlights the value of linking fine-resolution data with broader contextual factors.

Several limitations should be acknowledged. First, the yield estimates are derived from remote sensing models trained on proxy data from another region, which may introduce uncertainty and reduce the accuracy of field-level relationships. Second, cloud contamination and temporal gaps in satellite observations may affect the reliability of phenological metrics such as sowing date. Third, the analysis relies on aggregated village-level variables, which may mask within-village heterogeneity in management practices and environmental conditions. In addition, some socio-environmental variables, such as irrigation characteristics, are relatively coarse and do not fully capture differences in water availability, timing, or management flexibility. Finally, the observational nature of the study limits the ability to establish causal relationships between sowing date, irrigation, and yield.

Future research could improve the analysis by using more accurate and locally calibrated models for yield estimation. In this study, yield was derived from remote sensing–based prediction, which introduces uncertainty, especially at the field level. Applying more robust models, such as machine learning approaches, nonlinear models, or models trained with local ground-truth yield data, could reduce estimation error and better capture the complex relationships between crop growth, management practices, and environmental conditions. More precise yield estimates would also help clarify whether the weak relationships observed at the field level reflect true variability in smallholder systems or are partly driven by model uncertainty. Methodologically, incorporating nonlinear models or process-based approaches may better capture complex interactions between climate, management, and crop growth.

Finally, extending the analysis across multiple years would allow for a more robust assessment of how climate variability influences the relationships identified in this study.

6. Conclusion

This study examines the drivers of wheat yield variability in smallholder agricultural systems in eastern Uttar Pradesh using Sentinel-2 data at both field and village scales. The results highlight that delayed sowing is consistently associated with lower yield, although this effect is only clearly detectable at the village level, suggesting that spatial aggregation helps reduce noise in remote sensing estimates.

While field size alone shows no strong independent effect, its interaction with sowing date is significantly negative, indicating that larger fields are more vulnerable to delayed sowing. At the village level, socio-environmental factors further explain yield variability. Irrigation composition and access, particularly canal irrigation and distance to canals, play important roles, while agricultural labor and asset levels also contribute to observed patterns.

This study demonstrates the value of a multi-scale approach in revealing relationships that are not evident at finer scales. However, limitations remain, including uncertainty in yield estimation and the use of observational data.

Future work should focus on improving yield estimation with more advanced models and incorporating additional data sources and temporal analyses. Overall, this study provides insights into the complex interactions shaping productivity in smallholder systems and contributes to understanding pathways toward improved agricultural resilience.

References

Atzberger, C., 2013. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. Remote Sensing 5, 949–981. https://doi.org/10.3390/rs5020949

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014

Berzsenyi, Z., Dang, Q., 2008. Effect of various crop production factors on the yield and yield stability of maize in a long-term experiment. Cereal Research Communications 36, 167–176. https://doi.org/10.1556/CRC.36.2008.1.17

Bisht, I.S., Rana, J.C., Pal Ahlawat, S., 2020. The Future of Smallholder Farming in India: Some Sustainability Considerations. Sustainability 12, 3751. https://doi.org/10.3390/su12093751

Carfagna, E., Gallego, F.J., 2006. Using Remote Sensing for Agricultural Statistics. International Statistical Review 73, 389–404. https://doi.org/10.1111/j.1751-5823.2005.tb00155.x

Carletto, C., Savastano, S., Zezza, A., 2013. Fact or artifact: The impact of measurement errors on the farm size–productivity relationship. Journal of Development Economics 103, 254–261. https://doi.org/10.1016/j.jdeveco.2013.03.004

Chatrath, R., Mishra, B., Ortiz Ferrara, G., Singh, S.K., Joshi, A.K., 2007. Challenges to wheat production in South Asia. Euphytica 157, 447–456. https://doi.org/10.1007/s10681-007-9515-2

Diao, C., 2020. Remote sensing phenological monitoring framework to characterize corn and soybean physiological growing stages. Remote Sensing of Environment 248, 111960. https://doi.org/10.1016/j.rse.2020.111960

Duncan, J.M.A., Dash, J., Atkinson, P.M., 2015. The potential of satellite-observed crop phenology to enhance yield gap assessments in smallholder landscapes. Front. Environ. Sci. 3. https://doi.org/10.3389/fenvs.2015.00056

Fan, Shenggen and Brzeska, Joanna and Keyzer, Michiel and Halsema, Alex, 2013. From subsistence to profit: Transforming smallholder farms. Intl Food Policy Res Inst, Washinton, DC.

Ferencz, Cs., Bognár, P., Lichtenberger, J., Hamar, D., Tarcsai†, Gy., Timár, G., Molnár, G., Pásztor, Sz., Steinbach, P., Székely, B., Ferencz, O.E., Ferencz-Árkos, I., 2004. Crop yield estimation by satellite remote sensing. International Journal of Remote Sensing 25, 4113–4149. https://doi.org/10.1080/01431160410001698870

Gao, F., Zhang, X., 2021. Mapping Crop Phenology in Near Real-Time Using

Satellite Remote Sensing: Challenges and Opportunities. J Remote Sens 2021, 2021/8379391. https://doi.org/10.34133/2021/8379391

Gitelson, A.A., Gritz †, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of Plant Physiology 160, 271–282. https://doi.org/10.1078/0176-1617-00887

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guo, Z., Chamberlin, J., You, L., 2023. Smallholder maize yield estimation using satellite data and machine learning in Ethiopia. Crop and Environment 2, 165–174. https://doi.org/10.1016/j.crope.2023.07.002

Jain, M., Mondal, P., DeFries, R.S., Small, C., Galford, G.L., 2013. Mapping cropping intensity of smallholder farms: A comparison of methods using multiple sensors. Remote Sensing of Environment 134, 210–223. https://doi.org/10.1016/j.rse.2013.02.029

Quoc, T.T.P., Linh, T.T., Minh, T.N.T., 2020. Comparing U-Net Convolutional Network with Mask R-CNN in Agricultural Area Segmentation on Satellite Images, in: 2020 7th NAFOSTED Conference on Information and Computer Science (NICS). Presented at the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), IEEE, Ho Chi Minh City, Vietnam, pp. 124–129. https://doi.org/10.1109/NICS51282.2020.9335856

Jain, M., Singh, B., Srivastava, A.A.K., Malik, R.K., McDonald, A.J., Lobell, D.B., 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. Environ. Res. Lett. 12, 094011. https://doi.org/10.1088/1748-9326/aa8228

Jain, M., Srivastava, A., Balwinder-Singh, Joon, R., McDonald, A., Royal, K., Lisaius, M., Lobell, D., 2016. Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data. Remote Sensing 8, 860. https://doi.org/10.3390/rs8100860

Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y., 2018. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. Remote Sensing 10, 75. https://doi.org/10.3390/rs10010075

Jin, Z., Azzari, G., Burke, M., Aston, S., Lobell, D., 2017. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sensing 9, 931. https://doi.org/10.3390/rs9090931

Joshi, A., Pradhan, B., Gite, S., Chakraborty, S., 2023. Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. Remote Sensing 15, 2014. https://doi.org/10.3390/rs15082014

Li, S., Huang, J., Xiao, G., Huang, H., Sun, Z., Li, X., 2024. Improved Winter Wheat Yield Estimation by Combining Remote Sensing Data, Machine Learning, and Phenological Metrics. Remote Sensing 16, 3217. https://doi.org/10.3390/rs16173217

Liu, J., He, Q., Zhou, G., Song, Y., Guan, Y., Xiao, X., Sun, W., Shi, Y., Zhou, K., Zhou, S., Wu, Y., Ma, S., Wang, R., 2023. Effects of Sowing Date Variation on Winter Wheat Yield: Conclusions for Suitable Sowing Dates for High and Stable Yield. Agronomy 13, 991. https://doi.org/10.3390/agronomy13040991

Liu, Y., Rao, P., Zhou, W., Singh, B., Srivastava, A.K., Poonia, S.P., Van Berkel, D., Jain, M., 2022. Using Sentinel-1, Sentinel-2, and Planet satellite data to map field-level tillage practices in smallholder systems. PLoS ONE 17, e0277425. https://doi.org/10.1371/journal.pone.0277425

Lungu, O.N., Chabala, L.M., Shepande, C., 2020. Satellite-Based Crop Monitoring and Yield Estimation—A Review. JAS 13, 180. https://doi.org/10.5539/jas.v13n1p180

Madhukar, A., Kumar, V., Dashora, K., 2020. Spatial and Temporal Trends in the Yields of Three Major Crops: Wheat, Rice and Maize in India. Int. J. Plant Prod. 14, 187–207. https://doi.org/10.1007/s42106-019-00078-0

Mandapati, R., Gumma, M.K., Metuku, D.R., Bellam, P.K., Panjala, P., Maitra, S., Maila, N., 2023. Crop Yield Assessment Using Field-Based Data and Crop Models at the Village Level: A Case Study on a Homogeneous Rice Area in Telangana, India. AgriEngineering 5, 1909–1924. https://doi.org/10.3390/agriengineering5040117

McDonald, A.J., Balwinder-Singh, Keil, A., Srivastava, A., Craufurd, P., Kishore, A., Kumar, V., Paudel, G., Singh, S., Singh, A.K., Sohane, R.K., Malik, R.K., 2022. Time management governs climate resilience and productivity in the coupled rice–wheat cropping systems of eastern India. Nat Food 3, 542–551. https://doi.org/10.1038/s43016-022-00549-0

Mehak Rattan, G.S. Sidhu, Sandeep Kumar Singh, 2021. HISTORY OF LAND USE IN THE INDO GANGETIC PLAINS, INDIA AND ITS IMPACT ON POPULATION: A REVIEW. Plant Archives 21, 532–537. https://doi.org/10.51470/PLANTARCHIVES.2021.v21.S1.082

Mei, W., Wang, H., Fouhey, D., Zhou, W., Hinks, I., Gray, J.M., Van Berkel, D., Jain, M., 2022. Using Deep Learning and Very-High-Resolution Imagery to Map Smallholder Field Boundaries. Remote Sensing 14, 3046. https://doi.org/10.3390/rs14133046

Mkondiwa, M., Kishore, A., Sherpa, S., Urfels, A., Pinjarla, B., Kumar, V., Peramaiyan, P., McDonald, A., 2026. Time is money: Spatially explicit system analysis for rice-wheat cropping systems of Eastern Indo-Gangetic Plains, India. Agricultural Systems 233, 104648. https://doi.org/10.1016/j.agsy.2026.104648

Prudente, V.H.R., Garcia-Medina, M., Krishna, V., Euler, M., Bhattarai, N., Lerner, A.M., McDonald, A.J., Sherpa, S., Rajan, H., Urfels, A., Carneiro De Santana, C.T.,

Jain, M., 2025. Mapping grain crop sowing date in smallholder systems using optical imagery. Remote Sensing Applications: Society and Environment 39, 101660. https://doi.org/10.1016/j.rsase.2025.101660

Ran, Y., Chen, H., Ruan, D., Liu, H., Wang, S., Tang, X., Wu, W., 2018. Identification of factors affecting rice yield gap in southwest China: An experimental study. PLoS ONE 13, e0206479. https://doi.org/10.1371/journal.pone.0206479

Rouse, J. W.; Haas, R. H.; Schell, J. A.; Deering, D. W., 1973. Monitoring Vegetation Systems in the Great Okains with ERTS., in: 3rd ERTS Symposium. Presented at the Third Earth Resources Technology Satellite-1 Symposium (ERTS-1), NASA, Washington DC, pp. 309–317.

Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., Muricho, G., 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Sec. 5, 291–317. https://doi.org/10.1007/s12571-013-0263-y

Wang, X., Otto, I.M., Yu, L., 2013. How physical and social factors affect village-level irrigation: An institutional analysis of water governance in northern China. Agricultural Water Management 119, 10–18. https://doi.org/10.1016/j.agwat.2012.12.007

Williams, P.A., Crespo, O., Abu, M., Simpson, N.P., 2018. A systematic review of how vulnerability of smallholder agricultural systems to changing climate is assessed in Africa. Environ. Res. Lett. 13, 103004. https://doi.org/10.1088/1748-9326/aae026

Jain, M., Fishman, R., Mondal, P., Galford, G.L., Bhattarai, N., Naeem, S., Lall, U., Balwinder-Singh, DeFries, R.S., 2021. Groundwater depletion will reduce cropping intensity in India. Sci. Adv. 7, eabd2849. https://doi.org/10.1126/sciadv.abd2849

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci Data 2, 150066. https://doi.org/10.1038/sdata.2015.66

Appendix 1

Variable	Coef.	Std. Err.	Significance
Intercept	7.055	0.815	***
Soil type (ref = base)			
Soil (27)	0.054	0.159	
Soil (48)	0.124	0.161	
Soil (56)	-0.049	0.159	
Soil (79)	0.058	0.159	
Soil (116)	0.010	0.159	
District (ref = Deoria)			
Gorakhpur	0.237	0.008	***
Maharajganj	-0.066	0.008	***
Mean sowing date	-0.0093	0.002	***
Irrigation (%)	1.819	0.882	**
Sowing date × Irrigation	-0.0055	0.003	**
Mean field area	-9.96e-06	1.08e-05	
Canal irrigation (%)	-0.058	0.010	***
Well irrigation (%)	-0.0195	0.005	***
Asset index	-0.0003	0.00007	***
Distance to canal	-0.449	0.125	***
Agricultural labor (%)	0.0812	0.043	*
Observations	5056		
R^2	0.430		

Appendix 2

Variable	Coef.	Std. Err.	Significance
Intercept	8.378	0.246	***
Soil type (ref = base)			
Soil (27)	0.050	0.159	
Soil (48)	0.135	0.160	
Soil (56)	-0.050	0.158	
Soil (79)	0.058	0.159	
Soil (116)	0.005	0.159	
District (ref = Deoria)			
Gorakhpur	0.235	0.008	***
Maharajganj	-0.068	0.008	***
Mean sowing date	-0.0133	0.001	***
Canal irrigation (%)	2.708	0.664	***
Sowing date × Canal irrigation	-0.0084	0.002	***
Mean field area	-9.64e-06	1.08e-05	
Irrigation (%)	0.0107	0.012	
Well irrigation (%)	-0.0197	0.005	***
Asset index	-0.0003	0.00007	***
Distance to canal	-0.475	0.125	***
Agricultural labor (%)	0.0826	0.043	*
Observations	5056		
R^2	0.432		
Adj. R^2	0.430		

Appendix 3

Variable	Coef.	Std. Err.	Significance
Intercept	8.616	0.297	***
Soil type (ref = base)			
Soil (27)	0.053	0.159	
Soil (48)	0.128	0.161	
Soil (56)	-0.050	0.159	
Soil (79)	0.057	0.159	
Soil (116)	0.009	0.159	
District (ref = Deoria)			
Gorakhpur	0.237	0.008	***
Maharajganj	-0.066	0.008	***
Mean sowing date	-0.0140	0.001	***
Well irrigation (%)	0.0566	0.353	
Sowing date × Well irrigation	-0.0002	0.001	
Mean field area	-1.01e-05	1.08e-05	
Irrigation (%)	0.0093	0.012	
Canal irrigation (%)	-0.0587	0.010	***
Asset index	-0.0003	0.00007	***
Distance to canal	-0.445	0.125	***
Agricultural labor (%)	0.0801	0.043	*
Observations	5056		
R^2	0.430		