scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Projections of future agricultural management and crop choice under shared socioeconomic pathways

Sitong Wang[1,2,3], Xiuming Zhang[4], Ouping Deng[1,2,5] & Baojing Gu[1,2,3] ✉

Future agricultural landscapes will likely be shaped by the interplay between socioeconomic developments and natural conditions. However, existing theory-driven, process-based models often rely on idealized assumptions, limiting their capacity to capture real-world complexities fully. To complement these methods through an observational, data-driven approach, we developed a novel global dataset utilizing a statistical fixed-effects model. This paper presents a novel global dataset detailing projections of harvested area allocation for ten major crop groups across 197 countries and regions from 2020 to 2100. The dataset was generated using a statistical fixed-effects model calibrated on historical data. It includes annual projections under six distinct SSP-RCP scenarios (SSP1-2.6, SSP2-4.5, SSP3-7.0, SSP4-3.4, SSP4-6.0, and SSP5-8.5). For each scenario, the dataset provides future trajectories for key national agricultural management inputs—including nitrogen application rates, irrigation extents, and mechanization levels—and the resulting projected cropping shares. This dataset is designed to support assessments of food security, trade policy, and environmental impacts by providing a consistent, data-driven set of future agricultural landscape patterns.

Background & Summary

Agriculture is fundamental to human societies, providing essential resources such as food, fiber, and fuel, and exerting substantial influences on global economies and landscapes[1]. With the global population expected to approach approximately 10 billion by 2050[2], pressures to enhance agricultural productivity and sustainability are intensifying. These challenges are further exacerbated by climate change, resource scarcity, and shifting socioeconomic conditions, all of which significantly affect decisions regarding agricultural land use[3].

A crucial aspect of agricultural dynamics involves the interplay between farmers' land management decisions and their choices regarding crop selection and location[4]. Changes in agricultural management practices, including increased application of irrigation, fertilizers, and machinery, influence crop yields and production costs, thereby reshaping a country's comparative advantage in cultivating particular crops. Consequently, such shifts influence farmers' cropping choices globally[5]. Although existing modeling approaches commonly project future agricultural land use, they frequently lack transparency in explicitly linking broad socioeconomic scenarios, such as the Shared Socioeconomic Pathways (SSPs), to tangible changes in agricultural input intensity, including nitrogen application, irrigation, and mechanization, and subsequently detailing how these management adjustments affect farmers' crop selection based on evolving relative competitiveness. Clearly understanding these causal pathways is essential for accurately evaluating agricultural adaptation strategies and anticipating future landscape compositions.

Many current large-scale land-use projections depend heavily on equilibrium-based, process-driven models such as GLOBIOM and GCAM[6,7]. These frameworks are powerful tools for exploring future scenarios, as they effectively capture broad, theory-driven interactions across agricultural systems. They typically operate under core economic principles such as market equilibrium and rational behavior among participants, which is essential for assessing theoretically optimal adaptation pathways[6,7]. However, there is an acknowledged need for complementary approaches that explore future agricultural landscapes from a different methodological standpoint[8,9].

[1]State Key Laboratory of Soil Pollution Control and Safety, Zhejiang University, Hangzhou, 310058, China. [2]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, 310058, China. [3]Policy Simulation Laboratory, Zhejiang University, Hangzhou, 310058, China. [4]International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361, Laxenburg, Austria. [5]College of Resources, Sichuan Agricultural University, Chengdu, China. ✉e-mail: bjgu@zju.edu.cn

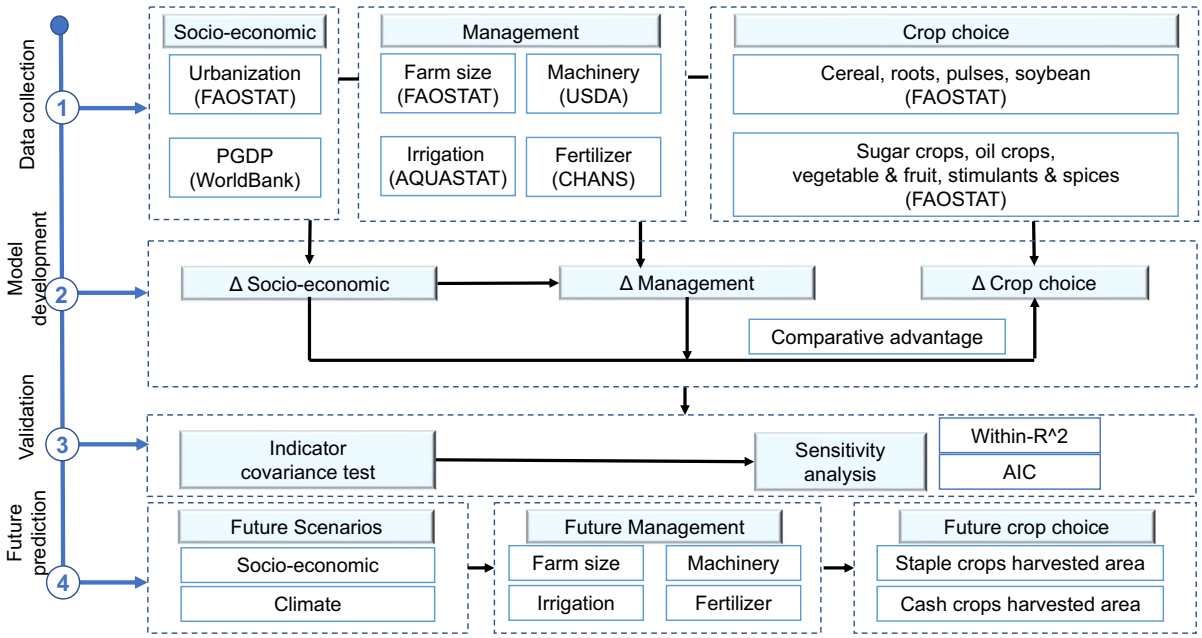

Fig. 1 Flow chart of the methodology of this research.

Statistical models based on historical observations, such as the fixed-effects model employed in this study, offer such a complementary perspective. Instead of being constrained by theoretical assumptions about behavior, these models learn directly from observed, real-world data. By capturing the empirical relationships between macro-level drivers and collective agricultural outcomes, this data-driven approach can reveal patterns shaped by historical inertia, market imperfections, and complex behaviors that may not be fully represented in equilibrium frameworks. Therefore, this study does not aim to replace process-based projections, but rather to enrich the existing landscape of future scenarios. By providing a methodologically distinct, observation-based dataset, we facilitate a more robust understanding of future uncertainties and enable valuable model intercomparison, which is crucial for improving the reliability of projected agricultural outcomes.

To address this need, we have developed a comprehensive dataset containing projections of national cropland allocation among ten major crop groups from 2020 to 2100. The dataset provides these projections under six distinct SSP-RCP scenarios (SSP1-2.6, SSP2-4.5, SSP3-7.0, SSP4-3.4, SSP4-6.0, and SSP5-8.5). The data files are structured to provide two layers of information for each scenario: first, the potential future trajectories for key agricultural management inputs (fertilizer application rates, irrigation extent, and mechanization levels); and second, the resulting national cropland allocations for staple and cash crops that are consistent with these management pathways.

Providing annual estimates for more than 150 countries, this dataset allows for the exploration of a wide range of plausible futures under varying socioeconomic and climate trajectories. Its global scope enables large-scale assessments of agricultural patterns, as well as detailed analyses of national-level land-use changes. By providing an integrated set of socioeconomic conditions, agricultural management inputs, and crop selection outcomes within a coherent framework, this dataset is designed to facilitate critical examinations of food security, sustainable agricultural practices, trade policies, and environmental impacts.

Methods

Figure 1 shows flow chart of the methodology of this research.

Study scale and rationale. All analyses in this study were conducted at the national level. This scale was deliberately chosen for three primary reasons. First, it aligns with our objective to inform national and global-scale assessments of policy, trade, and food security, where the nation-state is the key analytical and decision-making unit. Second, it ensures maximum data consistency and coverage, as reliable, long-term historical panel data for the wide range of socioeconomic and management drivers used in our model are most comprehensively available at the national level from sources like FAOSTAT and the World Bank. Third, it maintains methodological coherence, as our model is designed to capture the relationships between aggregate national-level drivers and the resulting national agricultural landscape patterns. The construction and analysis process for this dataset involved four main stages: data sourcing and processing, statistical analysis, and prediction based on future scenarios.

Data source and processing. The analysis focused on the following ten major crop categories: cereals, starchy roots, pulses, soybeans, oil crops (excluding soybeans), sugar crops, vegetables, fruits, stimulants (e.g., coffee, tea, cocoa), and spices. Harvested area (hectares) for various crops were obtained from FAOSTAT (https://www.fao.org/faostat/en/#data/QCL)[10]. This study utilized annual country-level data spanning the period 1961–2021.

	Ln Actual farm size
Ln Cropland area per agricultural laborer	0.417***
s.e.	−0.051
P values	0.000
Country	Yes
Year	Yes
N	418
Adjust-R2	0.967

Table 1. Correlation between cropland area per agricultural laborer and actual farm size. Note: Each column represents a separate regression model. All equations include country and year fixed effects. The asterisks indicate the statistical significance level based on P values: *p ≤ 0.1, **p ≤ 0.05, ***p ≤ 0.01, which assessed using a two-sided t-test.

Socio-economic. Data for urbanization level (percentage) was obtained from FAOSTAT (https://www.fao.org/faostat/en/#data/OA)[11]. Per capita gross domestic product (PGDP) data were sourced from the World Bank's World Development Indicators (WDI) database (https://datatopics.worldbank.org/world-development-indicators/)[12].

Management. The "farm size" in this paper is defined as the cropland area per unit of agricultural laborer. Due to the limitations of available data, this study employs cropland area per agricultural laborer as a proxy for farm size. This variable was constructed by dividing the total cropland area for each country and year, sourced from FAOSTAT (https://www.fao.org/faostat/en/#data/RL)[13], by the total agricultural labor force data from the USDA's International Agricultural Productivity dataset (https://ers.usda.gov/data-products/international-agricultural-productivity/)[14]. We provide a detailed explanation and validation of this substitution's effectiveness in Table 1.

The data on mechanization ownership were derived from the USDA's International Agricultural Productivity dataset (https://ers.usda.gov/data-products/international-agricultural-productivity/)[14]. In the September 2023 update of this dataset, the USDA provided country-level data on the quantity of total agricultural machinery stock from 1961 to 2021, measured in metric horsepower (CV). Mechanization ownership was calculated by dividing this machinery stock data by the rural population.

The data on irrigated area per capita were derived from the FAO's AQUASTAT database (https://data.apps.fao.org/aquastat/?lang=en)[15], specifically using the actually irrigated area under full control irrigation. This metric was selected because it excludes areas equipped with irrigation infrastructure but not actively irrigated, as well as equipped lowlands area and spate irrigation area. The exclusion of these categories is justified for two reasons. First, incorporating spate irrigation and equipped lowlands would reduce irrigated cropping intensity, thereby failing to accurately reflect the potential benefits of irrigation[16]. Second, since these irrigation methods rely on floodwater, they cannot fully decouple irrigated agriculture from climatic conditions in the same way that full control irrigation does[17]. Irrigated area per capita was then calculated by dividing the actually irrigated area under full control irrigation by the rural population. All historical data were matched and collated by country and year.

The nitrogen fertilizer data utilized in this research were derived from the CHANS model, a nitrogen mass balance framework[18]. CHANS integrates bottom-up nitrogen input and output fluxes from 14 subsystems—including cropland, livestock, grassland, forest, aquaculture, industry, humans, pets, urban green spaces, wastewater treatment, waste disposal, atmosphere, surface water, and groundwater—with top-down reactive nitrogen flux datasets at regional, national, and global scales. This comprehensive approach provides an integrated perspective on nitrogen cycling and fluxes, facilitating a deeper understanding of nitrogen dynamics across multiple scales. The nitrogen budget for cropland used in this study encompasses all crops (including staple crops and cash crops) from 1961 to 2021. The underlying data from the CHANS model are not publicly archived but are available from the corresponding author upon reasonable request.

All historical data were matched and collated by country and year. It is important to note the rationale for using these national-level intensity indicators. In our modeling framework, these metrics serve as powerful proxies for the overall modernization and intensification of a country's entire agricultural system. A nationwide increase in fertilizer availability or machinery stock, for example, reflects a systemic shift in capital investment, technological access, and agricultural policy priorities. These macro-level changes have significant spillover effects and influence resource competition (e.g., for labor and capital) across all agricultural activities, thereby shaping the comparative advantage and profitability of specific crops, even those concentrated in particular subnational regions.

Climate. Climate data from January 1961 to December 2021, including annual average temperature and total annual precipitation, is sourced from the Climatic Research Unit Gridded Time Series (CRU TS v.4.07) (https://doi.org/10.1038/s41597-020-0453-3)[19].

Future scenarios. Data utilized for future scenario analysis encompass projected changes across both socioeconomic and climatic dimensions. Socioeconomic alterations, projected for the period 2020–2100, include variations in urbanization rates and per capita Gross Domestic Product (PGDP). These socioeconomic data are

Variable Category	Description	Data Source (Reference)
Dependent Variables	Natural logarithm of the harvested area proportion of a specific crop category.	FAOSTAT[10] (https://www.fao.org/faostat/en/#data/QCL)
Independent Variables Socioeconomic Factors	Standardized urbanization rate (%) and its squared term.	FAOSTAT[11] (https://www.fao.org/faostat/en/#data/OA)
	Standardized natural logarithm of per capita GDP (constant 2015 USD) and its squared term.	World Bank WDI[12] (https://datatopics.worldbank.org/world-development-indicators/)
	Natural logarithm of farm size (cropland area per agricultural laborer) and its squared term.	FAOSTAT[13] (https://www.fao.org/faostat/en/#data/RL)/USDA[14] (https://ers.usda.gov/data-products/international-agricultural-productivity/)
Farm Structure	Natural logarithm of fertilizer use intensity (kg/ha) and its squared term.	CHANS Model[18] The underlying data from the CHANS model are not publicly archived but are available from the corresponding author upon reasonable request.
Agricultural Inputs	Natural logarithm of machinery use per capita (CV/person) and its squared term.	USDA[14] (https://ers.usda.gov/data-products/international-agricultural-productivity/)
	Natural logarithm of irrigated area per capita (ha/person) and its squared term.	AQUASTAT[15] (https://data.apps.fao.org/aquastat/?lang=en)/USDA (https://ers.usda.gov/data-products/international-agricultural-productivity/)
	Standardized annual average temperature (°C) and its squared term.	CRU TS v.4.07[19] (https://doi.org/10.1038/s41597-020-0453-3)
Climate Factors	Standardized annual total precipitation (mm) and its squared term.	CRU TS v.4.07[19] (https://doi.org/10.1038/s41597-020-0453-3)

Table 2. Description of Variables and Data Sources.

sourced from the International Institute for Applied Systems Analysis (IIASA) Shared Socioeconomic Pathways (SSP) database (https://tntcat.iiasa.ac.at/SspDb/)[20], covering scenarios SSP1, SSP2, SSP3, SSP4, and SSP5. Concurrently, climate change projections address anticipated variations in temperature and precipitation over the same 2020–2100 timeframe. This climate information is derived from the Canadian Earth System Model version 5 (CanESM5) outputs for the Coupled Model Intercomparison Project Phase 6 (CMIP6)[21]. Data were sourced from the Earth System Grid Federation node at IPSL (https://esgf-node.ipsl.upmc.fr/projects/cmip6-ipsl/), incorporating a total of six distinct scenarios: SSP1-2.6, SSP2-4.5, SSP3-7.0, SSP4-3.4, SSP4-6.0, and SSP5-8.5.

All data sources are listed in Table 2.

Statistical analysis. To investigate the relationships between socio-economic, management, and crop harvested area proportions, we employed a data-driven, statistical approach. Our methodological choice is rooted in providing a complementary perspective to existing process-based, equilibrium models. Whereas process-based models simplify reality through theoretical and behavioral idealizations (e.g., perfect rationality), our statistical approach represents a different form of simplification based on parsimonious variable selection. We focus on a key set of observable drivers for which consistent, long-term global panel data is available. The strength of this approach is its ability to learn complex, real-world relationships directly from historical observations, generating projections grounded in empirical evidence and historical inertia.

Given the nature of our dataset, which tracks numerous countries over several decades, a key analytical challenge is to control for the vast unobserved heterogeneity between countries. Factors such as deep-seated institutional settings, cultural preferences, and fundamental agro-ecological endowments are critical drivers of agricultural patterns, yet they are largely time-invariant and difficult to measure.

To address this, we selected a panel data fixed-effects (FE) model as our primary analytical framework. The choice of the FE specification over alternative statistical models, such as a pooled OLS or a random-effects (RE) model, is motivated by its superior ability to mitigate omitted variable bias. By design, the FE estimator controls for all time-invariant heterogeneity by analyzing the variation within each country over time. This aligns perfectly with our research objective: to isolate the dynamic impact of changes in time-varying drivers on crop allocation decisions.

A two-stage regression strategy within a fixed-effects framework was implemented:

Initially, we estimated a fixed effects model to analyze the impact of urbanization (incorporating both linear and quadratic terms) on farm size and the input of management measures. This analysis controlled for per capita GDP (PGDP) and climate factors (including temperature and precipitation). Furthermore, country-specific and year-specific fixed effects were included to account for unobserved, time-invariant country characteristics and common time trends. The model specification is as follows:

$$\ln M_{it} = \alpha_0 + \alpha_1 \times \ln U_{it} + \alpha_2 \times (\ln U_{it})^2 + \sum_n \theta_n q_{nit} + \varphi_i + \varepsilon_t + \mu_{it}$$

In this model, the subscripts i and t represent the country and year, respectively. M_{it} is the management indicators, which includes farm size, fertilizer use intensity, machinery use per capita and irrigation per capita. $ln\ U$ is the logarithm of urbanization level. The control variables q_n include PGDP, temperature and precipitation. α_0 is the constant term, while φ_i, ε_t and μ_{it} are error items. α_1, α_2 and θ are the coefficients to be estimated.

Subsequently, we estimated another fixed effects model. This second model aimed to quantify the impact of urbanization (again considering both linear and quadratic terms), alongside farm size and management measure inputs specific to crop categories, on the proportion of planted area allocated to various crops. A key feature of this stage is that we estimate a separate and distinct coefficient for the effect of each national-level management indicator on each of the ten specific crop categories. This "crop-category level" estimation allows our model to capture the heterogeneous impacts of a general, nationwide increase in agricultural modernization on the

relative competitiveness of different crops. The set of control variables was specified identically to the preceding model. The model specification is as follows:

$$\ln HAP_{it} = \alpha'_0 + \alpha'_1 \times \ln U_{it} + \alpha'_2 \times (\ln U_{it})^2 + \alpha'_3 \times \ln M_{it} + \alpha'_4 \times (\ln M_{it})^2 + \sum_n \theta'_n q_{nit} + \varphi'_i + \varepsilon'_t + \mu'_{it}$$

In this model, the subscripts i and t represent the country and year, respectively. α'_0 is the constant term, while φ'_i, ε'_t and μ'_{it} are error items. α'_1, α'_2 and θ' are the coefficients to be estimated.

A key feature of this two-stage regression strategy is its ability to mitigate potential endogeneity arising from the simultaneity between management inputs (M) and crop harvested area proportions (HAP). A farmer's decision to invest in a management input (e.g., irrigation) is often jointly determined with the decision to plant a specific crop. To address this, our framework establishes a causal hierarchy. In Stage 1, we first model management inputs as a function of broader, more exogenous socioeconomic and climatic drivers. In Stage 2, we then use the predicted values of the management inputs from the first stage, rather than the observed values, as explanatory variables for crop area proportions. This approach, analogous to an instrumental variable (IV) strategy, uses the component of management variation that is driven by exogenous macro-level factors to identify its effect on crop choice, thereby breaking the simultaneity loop and reducing potential bias in the estimated coefficients. The inclusion of country-specific fixed effects in both stages further controls for time-invariant unobserved confounders.

A key consideration in our modeling, which employs a log-transformed dependent variable, was the treatment of zero harvested area shares. Many country-crop combinations in the raw data exhibit structural zeros, where a crop is never cultivated. To avoid the biases associated with arbitrarily modifying these zero values, our analysis focused on active cultivation instances. For each crop-category-specific model, the estimation sample was restricted to countries that reported non-zero harvested area for that crop in at least one year during the historical period. Consequently, the model is specified to explain allocation changes among active producers. For prediction, countries excluded from a model's estimation sample retain a zero share for that crop throughout the projection period.

All statistical analyses were performed using Stata version 17.0. The model validation process is detailed in the Technical Validation section.

Future scenario prediction. The estimated coefficients from our historical regressions were used to project future outcomes for the period 2020–2100. Our projection methodology is built upon a two-stage sequential simulation framework, where future management indicators are projected first, and these projections then serve as inputs for the subsequent projection of harvested area proportions. Within each stage, we employ a differencing approach to isolate the net impact of future SSP-RCP scenarios.

Stage 1: Projecting Future Management Indicators. The first stage focuses on projecting the four key management indicators (fertilizer use, mechanization, irrigation, and farm size). For each indicator, the process is as follows:

Estimating Scenario Impacts (Delta): Using the coefficients from the first-stage model (Eq. 169), we generate two sets of predictions for each future year under a given SSP-RCP scenario: one for a "reference" future (without the SSP-RCP shock, i.e., base = = 0) and one for a "scenario" future (with the shock, base = = 1). The difference between these two predictions yields the net scenario impact (Δ), which represents the pure effect of the changes in all relevant drivers (urbanization, PGDP, and climate).

Generating Final Projections: This calculated impact (Δ) is then added to a stable historical baseline (the 2017–2021 average) to produce the final time-series projection for that management indicator under the specific scenario.

Stage 2: Projecting Future Harvested Area Proportions. The second stage utilizes the outputs from the first. The process is repeated for each of the ten crop categories:

Estimating Scenario Impacts (Delta): Using the coefficients from the second-stage model (Eq. 183), we again calculate the net scenario impact (Δ) on the harvested area proportion. Crucially, this calculation uses the projected management indicators from Stage 1 as inputs, alongside the projected changes in all other drivers (urbanization, PGDP, and climate).

Generating Final Projections: This impact (Δ) is then added to the crop's stable historical baseline (2017–2021 average) to produce the final projection. Finally, a normalization correction is applied across all crop proportions within a given country and year to ensure their sum equals 1.

This two-stage structure is a critical feature of our methodology. It explicitly models the causal pathway where broader socioeconomic and climatic changes first influence on-farm management decisions (Stage 1), and these evolving management practices, in turn, influence farmers' crop allocation choices (Stage 2). This approach helps to mitigate potential endogeneity issues and provides a more mechanistically plausible set of projections. For all projections, 90% confidence intervals were also calculated to represent the statistical uncertainty of the model's parameter estimates.

Data Records

This dataset provides projections of changes in harvested area proportions for various crops across 197 countries globally, spanning the period from 2020 to 2100 in five-year increments under five Shared Socioeconomic Pathway scenarios (SSP1 to SSP5), allowing users to explicitly assess scenario uncertainty in their own analyses. The dataset is publicly available through the figshare repository (https://doi.org/10.6084/

File Name in Repository	Corresponding Dependent Variable
regression_outputs-farm_size.xlsx	Farm size
regression_outputs-machinery.xlsx	Machinery
regression_outputs-irrigation.xlsx	Irrigation
regression_outputs-fertilizer.xlsx	Fertilizer
regression_outputs-cereal_share.xlsx	Cereal harvested area proportion
regression_outputs-roots_share.xlsx	Roots harvested area proportion
regression_outputs-pulses_share.xlsx	Pulses harvested area proportion
regression_outputs-soybean_share.xlsx	Soybean harvested area proportion
regression_outputs-sugarcrops_share.xlsx	Sugar crops harvested area proportion
regression_outputs-oilcrops_share.xlsx	Oil crops (ex. soy) harvested area proportion
regression_outputs-vegetables_share.xlsx	Vegetables harvested area proportion
regression_outputs-fruit_share.xlsx	Fruit harvested area proportion
regression_outputs-stimulants_share.xlsx	Stimulants harvested area proportion
regression_outputs-spices_share.xlsx	Spices harvested area proportion

Table 3. Index of detailed regression output files available in the data repository. This table provides a guide to the individual Excel files containing the full statistical outputs of the nested regression models discussed in the Technical Validation section.

Column Name	Description
areacode	Unique numerical identifier for each country/region (FAO code).
country	Country name.
year	The projection year (2020, 2025,…, 2100).
scenario	The SSP-RCP scenario identifier.
crlb_ssp	Projected central estimate for farm size (ha).
crlb_lb/crlb_ub	Lower and upper bounds of the 90% confidence interval for the farm size projection.
ferNpL_ssp	Projected central estimate for fertilizer application rate (kg ha^{-1} yr^{-1}).
ferNpL_lbn/ferNpL_ub	Lower and upper bounds of the 90% CI for the fertilizer projection.
malb_ssp	Projected central estimate for machinery stock (CV cap^{-1}).
malb_lb/malb_ub	Lower and upper bounds of the 90% CI for the machinery projection.
irlb_ssp	Projected central estimate for irrigation extent (ha cap^{-1}).
irlb_lb/irlb_ub	Lower and upper bounds of the 90% CI for the irrigation projection.
Rh_[Crop]_ssp	Projected central estimate for the harvested area proportion of a specific crop (e.g., Rh_Cereal_ssp).
Rh_[Crop]_lb/Rh_[Crop]_ub	Lower and upper bounds of the 90% CI for the specific crop share projection.

Table 4. Description of columns in the projection data files (predict_Rh_*.csv).

m9.figshare.28838930)[22]. The repository contains three categories of files: (1) the main projection data in comma-separated value (CSV) format, (2) tables with detailed regression results in Excel format (A complete index of the regression output files is provided in Table 3), and (3) the Stata code file used to generate the projections.

The main projection data are provided in six separate CSV files, one for each SSP-RCP scenario. File Naming Convention: The files are named predict_Rh_[scenario].csv, where [scenario] corresponds to the specific SSP-RCP combination (e.g., ssp126 for SSP1-2.6, ssp245 for SSP2-4.5, etc.). File Content: Each file contains the projected annual values for 197 countries and regions for the period 2020–2100 in five-year increments. The columns in each file are described in Table 4. Figures 2–5 show the model outputs.

The repository includes 14 Excel files containing the detailed regression outputs that form the basis of our projections. The files are descriptively named (e.g., regression_outputs-cereal_share.xlsx) to indicate the dependent variable of the model within. or a detailed list and description of each file, please refer to Table 3.

The Stata do-file code-projection.do is provided in the repository. This file contains the full set of commands used to run the two-stage fixed-effects regressions and generate all future projections, ensuring full reproducibility of the dataset.

Technical Validation

To assess the robustness and suitability of the chosen model specification for projecting national cropland allocation, we performed a comparative analysis of different model configurations using historical data. The core objective was to validate that the comprehensive model incorporating all identified driver categories, provides a statistically superior fit and robust predictive power, thereby justifying its use for generating the future projections presented in this dataset. This process involved three key components: (1) a comparative analysis of

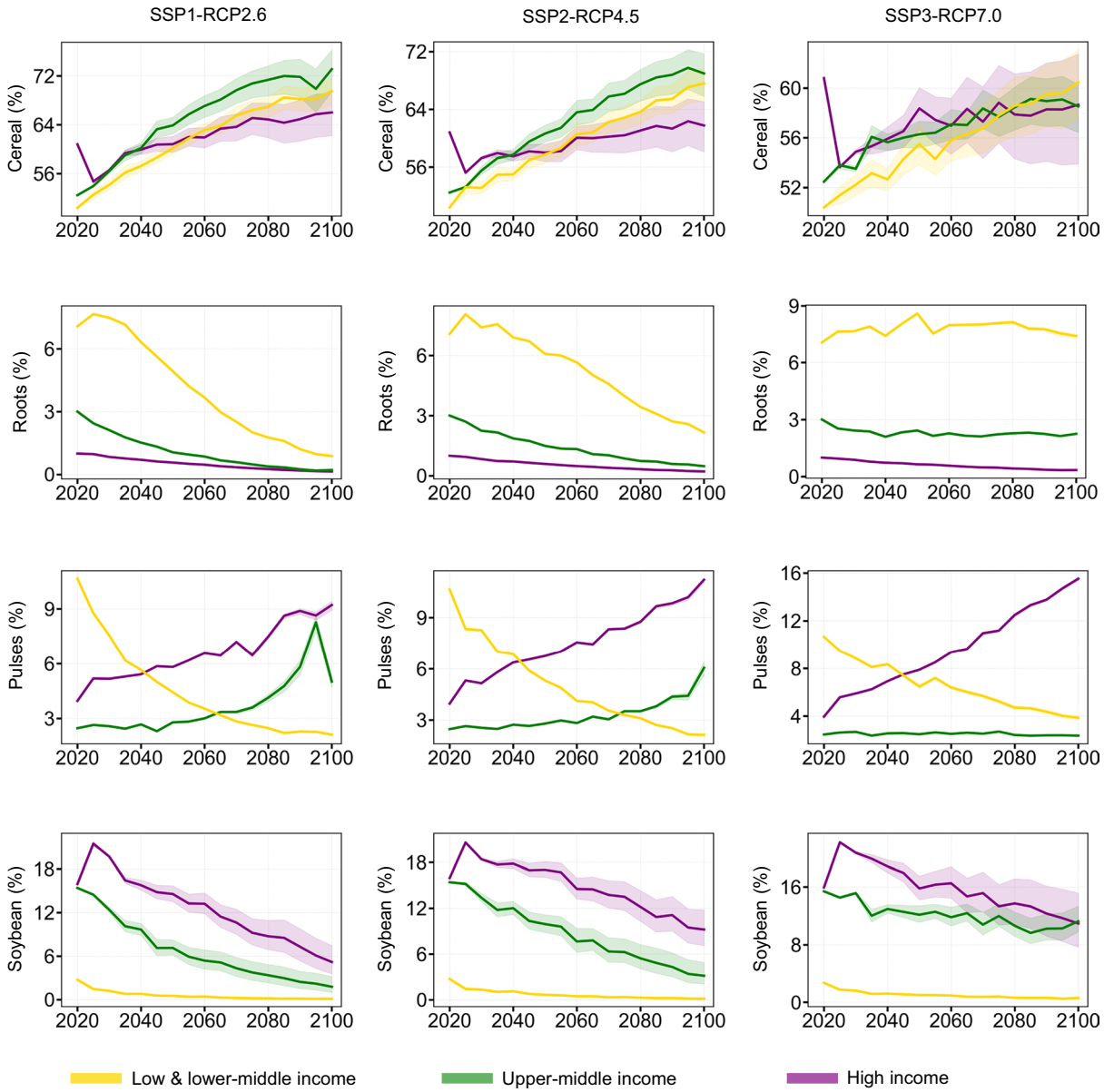

Fig. 2 Prediction of staple crop harvested area proportion from 2020 to 2100 under SSP 1-RCP2.6, SSP2-RCP4.5 and SSP3-RCP7.0 in different income groups. (**a–c**) Cereal; (**d–f**), Roots; (**g,h**), Pulses; (**j–l**), Soybean. Shaded areas around the lines represent the 90% confidence intervals of the projections, indicating model parameter uncertainty.

different model configurations using historical data, (2) an out-of-sample predictive performance test, and (3) a robustness check to account for the influence of time-invariant eco-environmental factors.

Model specification and goodness-of-Fit. We systematically evaluated the contribution of four distinct categories of predictor variables, reflecting the key drivers outlined in the Background & Summary: Socioeconomic factors variables representing broader development context, including urbanization rate and per capita GDP; Farm Structure variables capturing characteristics of the agricultural landscape, i.e. farm size; Agricultural inputs variables representing the intensity of management practices, such as fertilizer application, machinery availability, and irrigation infrastructure; Climate Factors variables reflecting key climatic conditions, specifically temperature and precipitation.

To determine the optimal specification for the management practice models, we also constructed a series of nested models. The full statistical outputs for this model comparison analysis are provided in a series of individual Excel files in the project's data repository, which are indexed and described in Table 3. We began with the 'full model' (Model A), which incorporated both linear and quadratic terms for our two primary socioeconomic drivers: the urbanization rate and the natural logarithm of per capita GDP (ln PGDP). Subsequently, we estimated three more parsimonious specifications: Model B excluded the quadratic term for urbanization; Model C

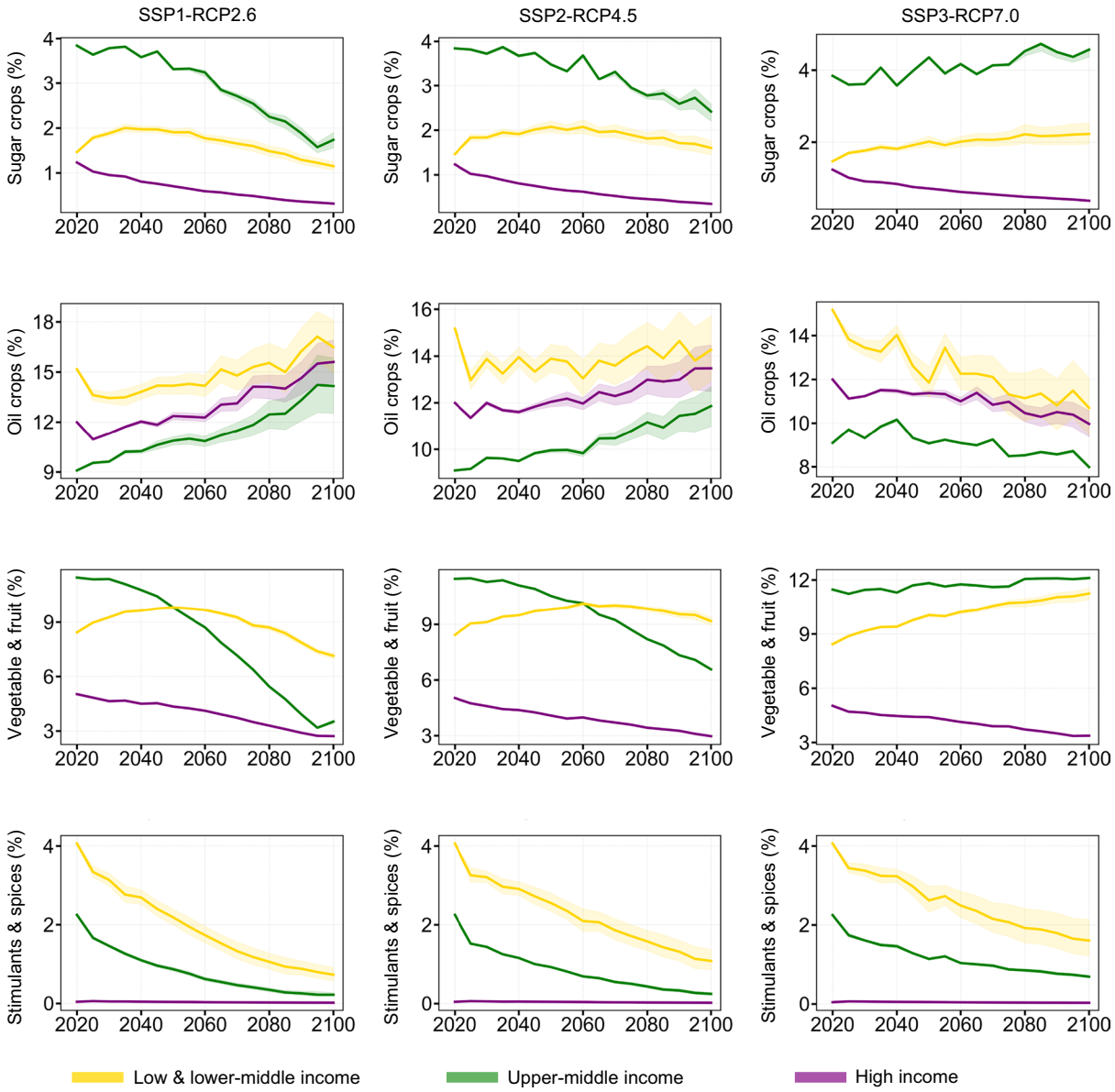

Fig. 3 Prediction of cash crop harvested area proportion from 2020 to 2100 under SSP 1-RCP2.6, SSP2-RCP4.5 and SSP3-RCP7.0 in different income groups. (**a**–**c**) Oil crops; (**d**–**f**) Sugar crops; (**g**–**i**) Vegetables and Fruits; (**j**–**l**) Stimulants and Spices. Shaded areas around the lines represent the 90% confidence intervals of the projections, indicating model parameter uncertainty.

excluded both the linear and quadratic terms for ln PGDP; and Model D excluded only the quadratic term for ln PGDP.

We constructed a series of nested models. The full statistical outputs for this model comparison analysis are provided in a series of individual Excel files in the project's data repository, which are indexed and described in Table 3. First, models including only one category of predictors were estimated (Models 1–4). Subsequently, models combining the core 'Agricultural Inputs' category with each of the other categories were tested (Models 5–7). Further combinations were explored (Models 8–9), culminating in the 'full model' (Model 10) that incorporates all four categories of predictors simultaneously.

A critical aspect of model comparison using information criteria is ensuring that all models are estimated on the exact same set of observations. Due to listwise deletion of missing values inherent in regression analysis, simply running models with different variable sets often results in different estimation samples. To address this, we first identified the estimation sample used by the full model (which contains all predictors and thus typically has the most missing-data constraints). We then explicitly restricted the estimation of all simpler models to this identical sample. This rigorous approach ensures that the log-likelihoods and resulting information criteria are directly comparable across all model specifications. We used the Akaike Information Criterion (AIC) as the primary metric for model selection[23]. AIC provides a measure of relative model quality by balancing

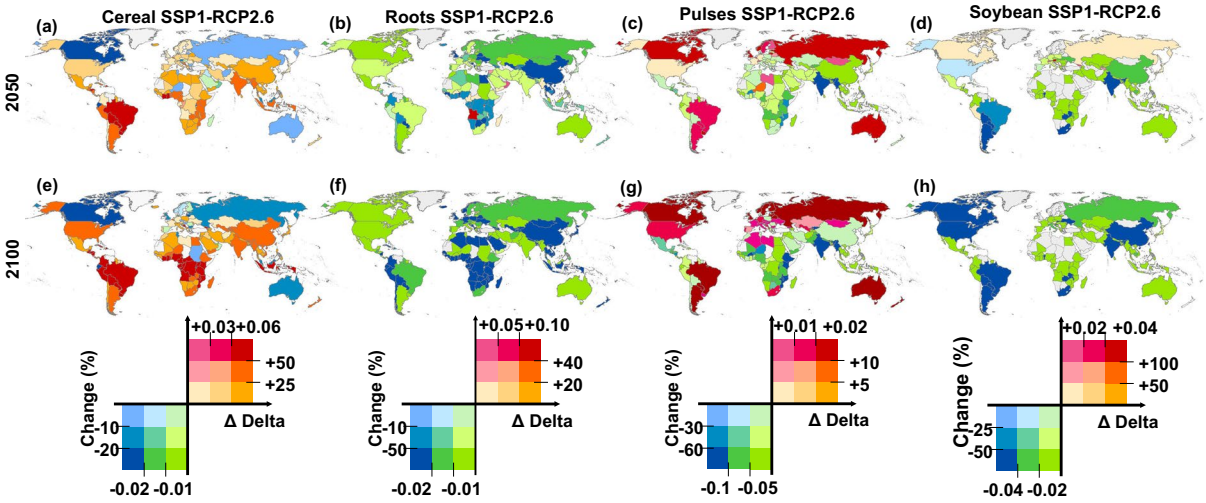

Fig. 4 Predicting changes of staple crop harvested area proportion in 2050 and 2100 under SSP 1-RCP2.6. (**a–c**) Cereal; (**d–f**) Roots; (**g,h**) Pulses; (**j-i**) Soybean.

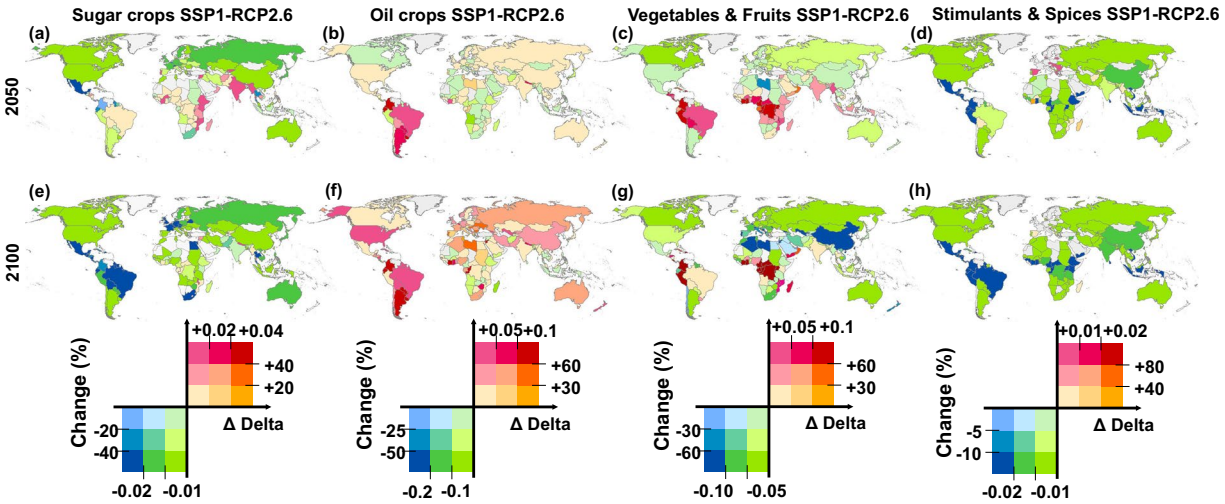

Fig. 5 Predicting changes of cash crop harvested area proportion in 2050 and 2100 under SSP 1-RCP2.6. (**a–c**) Oil crops; (**d–f**) Sugar crops; (**g–i**) Vegetables and Fruits; (**j–l**) Stimulants and Spices.

goodness-of-fit (log-likelihood) against model complexity (number of parameters), penalizing models with more parameters. A lower AIC value indicates a preferred model among the candidate set.

Across the vast majority of crop categories, the 'full model' (Model 10)—which incorporates Agricultural Inputs, Farm Structure, Socioeconomic Factors, and Climate Factors simultaneously—consistently yielded the lowest Akaike Information Criterion (AIC) value. In a few isolated cases, a slightly more parsimonious model exhibited a marginally lower AIC. However, for the sake of methodological consistency and to ensure a unified theoretical framework across all crop projections, we selected the full Model 10 as the final specification for all categories. This decision is justified as Model 10 was overwhelmingly the best-performing specification and provides the most comprehensive explanatory structure. The finding that a model incorporating all four interconnected driver categories generally provides the best statistical fit underscores the importance of a holistic approach when modeling cropland allocation dynamics.

Following the selection of the full model based on AIC, we performed additional diagnostic checks to assess potential multicollinearity among the predictor variables within the estimation sample. Pearson correlation coefficients were calculated, results revealing a high positive correlation (approx. 0.80) was observed between the standardized urbanization rate and the standardized logarithm of per capita GDP, suggesting a strong linear association between these two socioeconomic indicators. To further evaluate the potential impact of these relationships on model stability, we calculated Variance Inflation Factors (VIFs). According to the updated VIF analysis results, multicollinearity appears to be low to moderate. The Mean VIF based on the variables included in this specific VIF calculation was 3.43. These results suggest that despite the notable pairwise correlation between urbanization and GDP per capita, the potential impact of multicollinearity on the model's stability

	Urbanization	Urbanization^2	Ln PGDP	Ln PGDP^2
Urbanization	1.0000			
Urbanization^2	−0.7260	1.0000		
Ln PGDP	0.7956	−0.5344	1.0000	
Ln PGDP^2	−0.6075	0.6678	−0.7312	1.0000

Table 5. Pearson correlation coefficients between urbanization rate and ln PGDP.

Variable	VIF
Urbanization	2.92
Urbanization^2	2.88
Ln PGDP	1.41
Ln PGDP^2	3.05
Mean VIF	3.43

Table 6. Variance Inflation Factor (VIF) for Independent Variables. Notes: Mean VIF below the threshold of 5, suggesting no severe multicollinearity.

and standard errors is minimal according to standard VIF diagnostics[24]. This strengthens the confidence in the reliability of the parameter estimates from the full model (Tables 5, 6).

The comparative model analysis, based on the AIC criterion and conducted on a consistent sample, provides strong statistical support for the chosen full model specification. The finding that incorporating variables representing agricultural inputs, farm structure, socioeconomic context, and climate factors together leads to the best statistical performance underscores the importance of considering these interconnected drivers simultaneously when modeling cropland allocation dynamics. This validation exercise increases confidence in the capacity of the chosen model structure to capture the key historical relationships, forming a robust basis for the forward-looking projections under different SSP-RCP scenarios presented in this dataset. While AIC selects the best model among the candidates considered, and the model captures statistical associations rather than definitive causal pathways. This validation exercise confirms that the comprehensive model specification provides a statistically superior fit to the historical data, which strengthens the confidence in the reliability of the projections generated from this framework.

Out-of-sample predictive validation. To further validate the robustness of the chosen model structure, we conducted a rigorous out-of-sample validation exercise. For each crop category, the full model was trained using historical data only from the period 1961–2000. The estimated coefficients were then used to generate predictions for the subsequent 2001–2021 period, representing data the model had not previously seen. The model's predictive accuracy was evaluated using the Out-of-Sample R-squared (OOS-R^2), which measures the proportion of the variation in the unseen data that is explained by the model's predictions.

The results (Table 7) demonstrate the model's strong predictive power for future outcomes based on past relationships. The OOS-R^2 was substantial across most major crop categories. For instance, the model explained 56.7% of the out-of-sample variation for Sugar Crops, 54.9% for Fruit, 45.6% for Vegetables, and 43.0% for Cereals. This strong out-of-sample performance provides compelling evidence that the relationships between socioeconomic drivers and crop choices captured by our model are not mere statistical artifacts of the training period but are robust and persistent over time, strengthening the justification for its use in forward-looking projections.

Robustness Check: accounting for time-invariant eco-environmental factors. A potential concern regarding our statistical approach is the exclusion of explicit eco-environmental variables, such as soil characteristics and topography. Our model addresses this challenge through the inherent properties of the fixed-effects specification, a choice reinforced by the temporal nature of different drivers and practical data constraints.

The inclusion of country-specific fixed effects (α_c) is a core feature designed to capture the combined influence of all relatively stable, country-specific characteristics, including the underlying eco-environmental endowments. We acknowledge that these environmental variables are not perfectly static over long time horizons. However, they are typically "slow-moving" variables (e.g., soil formation), in contrast to the "fast-moving" socioeconomic drivers (e.g., per capita GDP) at the core of our study. Furthermore, the scarcity of consistent, long-term annual panel data for these environmental factors at a global scale makes their inclusion as time-varying predictors impractical. The fixed-effects approach is therefore a methodologically sound and practical choice, as it controls for the baseline influence of these slow-moving or hard-to-measure variables.

To empirically validate that our fixed effects have effectively proxied for these critical baseline factors, we conducted a comprehensive quantitative robustness check. We compiled a cross-sectional dataset of four key country-level eco-environmental variables representing average conditions within agricultural lands: mean soil organic carbon density (SOC, in hg/m^3) from SoilGrids database (https://files.isric.org/soilgrids/latest/data/soc/)[25], mean bulk density (in cg/cm^3) from SoilGrids (https://files.isric.org/soilgrids/latest/

Crop Category	Out-of-Sample R^2	RMSE	MAE
Sugar Crops	0.567	1.266	0.027
Fruit	0.549	1.023	−0.238
Vegetables	0.456	0.888	−0.13
Oil crops	0.44	1.163	0.235
Cereal	0.43	0.863	0.023
Spices	0.418	1.718	0.01
Stimulants	0.324	2.421	0.064
Pulses	0.291	1.219	0.132
SoyBean	0.257	2.425	0.556
Roots	0.19	1.544	−0.292

Table 7. Out-of-Sample Predictive Performance of the Final Models. Note: The models for each crop category were trained on historical data from 1961–2000. The performance metrics were then calculated based on the models' predictions for the out-of-sample period of 2001–2021. Out-of-Sample R-squared (OOS-R^2) measures the proportion of variation in the unseen data explained by the model. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) measure the average magnitude of the prediction errors in the original logarithmic scale of the dependent variable.

	mean_soc	mean_bulk_density	mean_ph	mean_tri
Cereal	−0.217	0.365	0.328	−0.181
Fruit	0.167	−0.265	−0.184	0.239
Oilcrops	−0.201	0.208	−0.200	−0.294
Pulses	−0.083	−0.017	−0.169	−0.078
Roots	0.336	−0.318	−0.257	0.164
SoyBean	0.223	−0.272	−0.358	−0.006
SugarCrops	0.182	−0.382	−0.313	−0.124
Vegetables	0.203	−0.018	0.183	0.217

Table 8. Pearson correlation coefficients between country fixed effects and eco-environmental variables. Note: The table displays the Pearson correlation coefficients between the country-specific fixed effects (α_c), estimated from the final regression models for each of the eight crop categories, and the four country-level eco-environmental variables. All environmental variables represent the mean values within cropland areas, derived from zonal statistics. These variables include mean soil organic carbon density (mean_soc), mean bulk density (mean_bulk_density), mean soil pH (mean_ph), and the mean Terrain Ruggedness Index (mean_tri). The fixed effects capture time-invariant heterogeneity across countries.

data_aggregated/1000m/bdod/)[26], mean soil pH from SoilGrids database (https://files.isric.org/soilgrids/latest/data_aggregated/1000m/phh2o/)[27], and the mean Terrain Ruggedness Index (TRI) (https://diegopuga.org/data/rugged/)[28].

We then analyzed the relationship between the country fixed effects estimated from our final models and these four variables. The results, summarized in Tables 8, 9, are highly significant. We found that the four eco-environmental variables jointly explain a statistically significant and agronomically meaningful portion of the cross-country variation in the fixed effects. The explanatory power (Adjusted R-squared) is notably high for major commodity and high-value crops, reaching 18.7% for Sugar Crops, 18.5% for Oilcrops (excluding soybeans), and 14.4% for Cereals.

We found that the four eco-environmental variables jointly explain a statistically significant and agronomically meaningful portion of the cross-country variation in the fixed effects across nearly all crop categories. The explanatory power (Adjusted R-squared) is notably high for major commodity and high-value crops, reaching 18.7% for Sugar Crops, 18.5% for Oilcrops (excluding soybeans), 14.4% for Cereals, and 12.6% for Vegetables.

Considering that the country fixed effect is a complex composite that also includes other unobservable factors (e.g., institutions, culture), the ability to explain up to nearly one-fifth of its variation with just four physical variables is a powerful validation of our model. This quantitative evidence demonstrates that a substantial portion of the heterogeneity in agricultural potential driven by stable soil and topographic conditions has been successfully captured, strengthening our model's ability to isolate the impacts of the time-varying socioeconomic drivers.

Usage notes and limitations. While these validation steps confirm the statistical robustness of our dataset, users should be aware of several key limitations when applying it to their own research. The primary limitation is the national-scale aggregation of the data. Our projections represent national averages and, consequently, do not capture the significant subnational heterogeneity in socioeconomic conditions or eco-environmental constraints.

Crop Category	Adjusted R-squared
Cereal	0.14
Fruit	0.07
Oil crops	0.18
Pulses	0.01
Roots	0.11
Soybean	0.10
Sugar crops	0.19
Vegetables	0.13

Table 9. Joint explanatory power of eco-environmental variables on country fixed effects. This table presents the results of a series of multiple linear regression analyses. For each of the eight crop categories, the estimated country-specific fixed effect (α_c) was regressed on the four country-level eco-environmental variables (mean_soc, mean_bulk_density, mean_ph, and mean_tri). The Adjusted R-squared value is reported, quantifying the proportion of the total cross-country variation in the fixed effects that can be jointly explained by these four key soil and topographic factors. This analysis serves as a quantitative robustness check to validate that the fixed-effects specification has effectively captured heterogeneity driven by stable environmental endowments.

As such, the results may not fully reflect the dynamics within specific, agriculturally significant regions of a country. Future research could build upon our framework by developing methods to downscale these national projections where sufficient subnational data is available. A second limitation is that our statistical model, by design, projects future patterns based on historically observed relationships and does not endogenously model potential future structural breaks or novel policy interventions not seen in the historical record. The dataset is therefore best interpreted as a robust, empirically-grounded projection of future crop choices under the specific socioeconomic and climatic trajectories defined by the SSP-RCP scenarios.

Data availability

The dataset is publicly available through the figshare repository (https://doi.org/10.6084/m9.figshare.28838930)[22].

Code availability

Code for this study are available within figshare files.

Received: 24 April 2025; Accepted: 6 October 2025;
Published online: 18 November 2025

References

1. Kastner, T. *et al*. Global agricultural trade and land system sustainability: Implications for ecosystem carbon storage, biodiversity, and human nutrition. *One Earth* **4**, 1425–1443 (2021).
2. Population Division of the Department of Economic and Social Affairs of the United Nations. 2024 Revision of World Population Prospects. https://population.un.org/wpp/ (2024).
3. Li, M. *et al*. Sustainable management of agricultural water and land resources under changing climate and socio-economic conditions: A multi-dimensional optimization approach. *Agric. Water Manage.* **259**, 107235 (2022).
4. Zhu, Y., Zhang, Y., Ma, L., Yu, L. & Wu, L. Simulating the dynamics of cultivated land use in the farming regions of China: A social-economic-ecological system perspective. *J. Clean. Prod.* **478**, 143907 (2024).
5. Varma, V., Mosedale, J. R., Alvarez, J. A. G. & Bebber, D. P. Socio-economic factors constrain climate change adaptation in a tropical export crop. *Nat. Food* **6**, 343–352 (2025).
6. Schneider, U. A. *et al*. Impacts of population growth, economic development, and technical change on global food production and consumption. *Agric. Syst.* **104**, 204–215 (2011).
7. Joint Global Change Research Institute. GCAM v7.1 Documentation: Global Change Analysis Model. *Pacific Northwest National Laboratory*. https://jgcri.github.io/gcam-doc/ (2023).
8. Eini, M. R., Salmani, H. & Piniewski, M. Comparison of process-based and statistical approaches for simulation and projections of rainfed crop yields. *Agric. Water Manage.* **277**, 108107 (2023).
9. Lobell, D. B. & Asseng, S. Comparing estimates of climate change impacts from process-based and statistical crop models. *Environ. Res. Lett.* **12**, 15001 (2017).
10. FAOSTAT. Harvested Area; https://www.fao.org/faostat/en/#data/QCL (2025).
11. FAOSTAT. Urbanizatrion, population and land use data; https://www.fao.org/faostat/en/#data/OA (2025).
12. World Development Indicators (World Bank); https://datatopics.worldbank.org/world-development-indicators/ (2023).
13. FAOSTAT. Land use data; https://www.fao.org/faostat/en/#data/RL (2023).
14. International Agricultural Productivity Database (USDA); https://ers.usda.gov/data-products/international-agricultural-productivity/ (2023).
15. AQUASTAT (FAO), FAOs Global Information System on Water and Agriculture; https://data.apps.fao.org/aquastat/?lang=en (2024).
16. AQUASTAT (FAO), Methodology thematic discussion of irrigation and drainage; https://www.fao.org/aquastat/en/overview/methodology/irrig-drainage/ (2024).
17. AQUASTAT (FAO), Irrigated crop calendars; https://www.fao.org/aquastat/en/data-analysis/irrig-water-use/irrigated-crop-calendars (2024).
18. Gu, B., Ju, X., Chang, J., Ge, Y. & Vitousek, P. M. Integrated reactive nitrogen budgets and future trends in China. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8792–8797 (2015).
19. Harris, I., Osborn, T.J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data* **7**, https://doi.org/10.1038/s41597-020-0453-3 (2020).

20. International Institute for Applied Systems Analysis (IIASA). Shared Socioeconomic Pathways Database; https://tntcat.iiasa.ac.at/SspDb/ (2025).
21. Canadian Centre for Climate Modelling and Analysis. CMIP6 Monthly data for surface air temperature (tas) and precipitation (pr). https://esgf-node.ipsl.upmc.fr/projects/cmip6-ipsl/ (2024).
22. Wang, S., Zhang, X., Deng, O. & Gu, B. Projection of future crop choice under shared socioeconomic pathways. *figshare* https://doi.org/10.6084/m9.figshare.28838930 (2025).
23. Portet, S. A primer on model selection using the Akaike Information Criterion. *Infect. Dis. Model.* **5**, 111–128 (2020).
24. Craney, T. A. & Surles, J. G. Model-Dependent Variance Inflation Factor Cutoff Values. *Qual. Eng.* **14**, 391–403 (2002).
25. ISRIC. International Soil Reference and Information Centre, SoilGrids250m 2.0 - Soil organic carbon; https://files.isric.org/soilgrids/latest/data/soc/ (2022).
26. ISRIC. International Soil Reference and Information Centre, SoilGrids250m 2.0 - Bulk density; https://files.isric.org/soilgrids/latest/data_aggregated/1000m/bdod/ (2022).
27. ISRIC. International Soil Reference and Information Centre, SoilGrids250m 2.0 - Soil pH in H2O; https://files.isric.org/soilgrids/latest/data_aggregated/1000m/phh2o/ (2022).
28. Nunn, N. & Puga, D. Grid-cell-level data on terrain ruggedness. https://diegopuga.org/data/rugged/ (2012).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (42261144001 and 42325707).

Author contributions

B.G. designed the research. S.W. conducted the research and performed the analysis. X.Z. provided the nitrogen supply data. S.W. wrote the first draft. And all authors contributed to the discussion and revision of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.