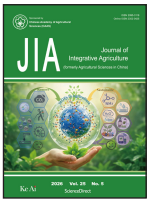
Journal of Integrative Agriculture 2026, 25(5): 2121–2138

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

Uncovering the spatiotemporal evolution and driving mechanisms of soybean planting area in China from 2000 to 2022

Wenbin Liu[1, 2], Shu Li[2], Juan Cao[1#], Jun Xie[3], Jinwei Dong[1], Jichong Han[3], Qinghang Mei[3], Lichang Yin[1], Hongyan Zhang[4], Hong Zhou[1], Fulu Tao[1]

[1] *Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*
[2] *Changjiang Institute of Survey Technical Research, Ministry of Water Resources, Wuhan 430011, China*
[3] *School of National Safety and Emergency Management, Beijing Normal University, Beijing 100088, China*
[4] *School of Computer Sciences, China University of Geosciences, Wuhan 430074, China*

Highlights

- Utilized multi-source data and GEE for nationwide annual soybean mapping.
- Developed ChinaSoyA30m, the first 30-m soybean dataset for China from 2000 to 2022.
- Revealed regional disparities in soybean cultivation trends and centroid movement.
- Identified regional drivers of soybean planting area shifts.

Abstract

Understanding the spatial distribution, temporal dynamics, and driving factors of soybean cultivation is critical for yield estimation, agricultural planning, and national food security. However, high-resolution, long-term, and nationwide datasets of soybean cultivation in China remain scarce. This study developed a 30-m resolution dataset of soybean in China from 2000–2022 using multi-source data (ChinaSoyA30m), and analyzed the spatiotemporal dynamics and driving forces of soybean cultivation. The phenological characteristics of major crops across China were evaluated to generate training samples for supervised classification. Gap statistics, *K*-means clustering, and spectral angle mapping were employed to enhance classification reliability. A supervised classification approach was implemented on Google Earth Engine (GEE) using dense Landsat data to produce annual soybean maps. ChinaSoyA30m demonstrates competitive performance compared to six existed soybean datasets, with strong correlations with provincial, prefectural, and county statistics (R^2=0.95, 0.89, and 0.80), and the F1 scores validated against ground truth data were 70.16, 80.40, and 78.38%. Since 2000, the soybean planting area has exhibited a fluctuating upward trend with distinct regional characteristics. Northern China emerged as the primary production area, characterized by a stable planting centroid and small spatial variation. The primary driver of soybean area dynamics was the "value added of primary industry", while gross power of agricultural machinery was a significant factor in North China, highlighting regional differences in driving mechanisms. This study provides the first long-term, high-resolution soybean planting dataset for China and offers valuable insights into the sustainable development of soybean cultivation.

Keywords: soybean, remote sensing, classification, spatiotemporal dynamics, driving factors

1. Introduction

Soybean (*Glycine max* L.) is a critical cash crop with wide-ranging benefits for human well-being. As a cornerstone of the global food system, it provides an essential and affordable source of protein and animal feed (Singh *et al.* 2008), and serves as a key raw material for biofuel production and chemical industries (Cavalett and Ortega 2010; Zaaboul *et al.* 2022). With growing demand for food and biofuels, soybean cultivation has expanded significantly in recent years, particularly in the United States, Europe,

Received 7 January 2025; Received in revised form 6 March 2025; Accepted 5 June 2025; Available online 17 July 2025
Wenbin Liu, E-mail: liuwenbin@whu.edu.cn; #Correspondence Juan Cao, Mobile: +86-18810183758, E-mail: caojuan@igsnrr.ac.cn

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2025.07.021

and South America, and China (Wang *et al.* 2020; Song *et al.* 2021; Chen *et al.* 2022; Guilpart *et al.* 2022). As the world's largest consumer and importer of soybeans, China is projected to require approximately 133 million tons by 2035 to meet domestic demand (Liu *et al.* 2021). However, insufficient self-sufficiency has made China heavily reliant on imports, primarily from the United States and Brazil (Wu *et al.* 2020; Zhao *et al.* 2021). Recent geopolitical tensions, such as the U.S.–China trade war and the Russia–Ukraine conflict, have underscored the urgency for China to boost domestic soybean production to reduce import dependency (Wu *et al.* 2020; Adjemian *et al.* 2021; Chai *et al.* 2024). Comprehensive spatiotemporal mapping and analysis of soybean cultivation is essential for agricultural management and national food security.

Satellite remote sensing observation technology provides an efficient and economical solution for obtaining soybean planting information. Remote sensing mapping methods for soybean are broadly categorized into unsupervised, sample-free approaches and supervised, sample-driven methods (Mei *et al.* 2024; Xiao G *et al.* 2024). Unsupervised methods rely on spectral and phenological characteristics to distinguish soybeans but are often challenged by the spectral similarity among crops (Chen *et al.* 2023). These methods require detailed knowledge of soybean growth stages and their spectral behavior. For instance, Chen *et al.* (2023) demonstrated that differences in canopy water content and greenness are crucial for soybean identification, leading to the development of the greenness and water content composite index (GWCCI), which combines the normalized difference vegetation index (NDVI) and short-wave infrared (SWIR) bands. By deriving phenological indicators and analyzing crop separability during key growth periods, researchers can establish automated classification rules (Dong *et al.* 2015; Ashourloo *et al.* 2022). However, unsupervised methods are limited by their reliance on phenological timing and data quality (Weiss *et al.* 2020; Zhang *et al.* 2022).

Supervised classification methods, which depend on high-quality training samples, are more commonly used for automated soybean mapping (Mei *et al.* 2024; Zhang *et al.* 2024). The effectiveness of these methods hinges on the representativeness of the training data (Huang *et al.* 2020; Zhang *et al.* 2020). To address the limitations of both approaches, hybrid methods have been developed, combining the strengths of unsupervised and supervised techniques (Wang *et al.* 2019; Belgiu *et al.* 2021). For example, Liu and Zhang (2023) proposed a hybrid approach for rapeseed mapping that integrates multi-source data and automated training sample generation. By fusing spectral and phenological features, this method produces robust training data, enabling reliable supervised classification and providing a promising framework for large-scale, long-term soybean mapping.

China has seen growing efforts to generate high-resolution soybean maps in recent years. You *et al.* (2021) mapped major crops, including soybean, at a 10-m resolution across Northeast China for 2017–2019 using Sentinel-2 imagery. Li H *et al.* (2023) produced a 10-m maize and soybean map for China based on Sentinel-2 imagery from April to October, 2019. Subsequently, Zhang *et al.* (2024) developed a 10-m soybean dataset (ChinaSoybean10) covering ten major soybean-producing provinces from 2019 to 2022, also using Sentinel-2 data. Mei *et al.* (2024) generated a 10-m soybean dataset (ChinaSoyArea10m) for fourteen major soybean-producing provinces from 2017 to 2021. However, the temporal coverage of these datasets is limited to a maximum of five years due to the unavailability of Sentinel-2 data before 2017. This restricts their utility for long-term analyses and comprehensive policy recommendations. For extended temporal studies, some researchers have employed Landsat imagery and Google Earth Engine (GEE) to produce regional soybean datasets. For example, Li *et al.* (2021) generated 30-m soybean maps for Heilongjiang Province from 1984 to 2020, while Xuan *et al.* (2023) mapped soybean areas across Northeast China from 2013 to 2021. Although these long-term datasets are valuable for temporal analysis, their regional scope limits their broader applicability. To date, long-term, high-resolution soybean maps covering all of China are still missing, leaving a significant gap in the ability to monitor soybean cultivation trends and support national agricultural planning.

Hence, this study aims to map soybean planting areas in China at 30-m resolution and comprehensively analyze their spatiotemporal dynamics and driving factors from 2000 to 2022. The objectives were to: (1) automatically generate training samples based on the spectral and phenological characteristics of soybean; (2) refine training samples using gap statistics, *K*-means, and spectral angle mapping to improve classification reliability; (3) map the spatial distribution of soybean cultivation across China using a supervised classification approach on GEE; and (4) analyze the spatiotemporal dynamics and driving factors of soybean cultivation across regions.

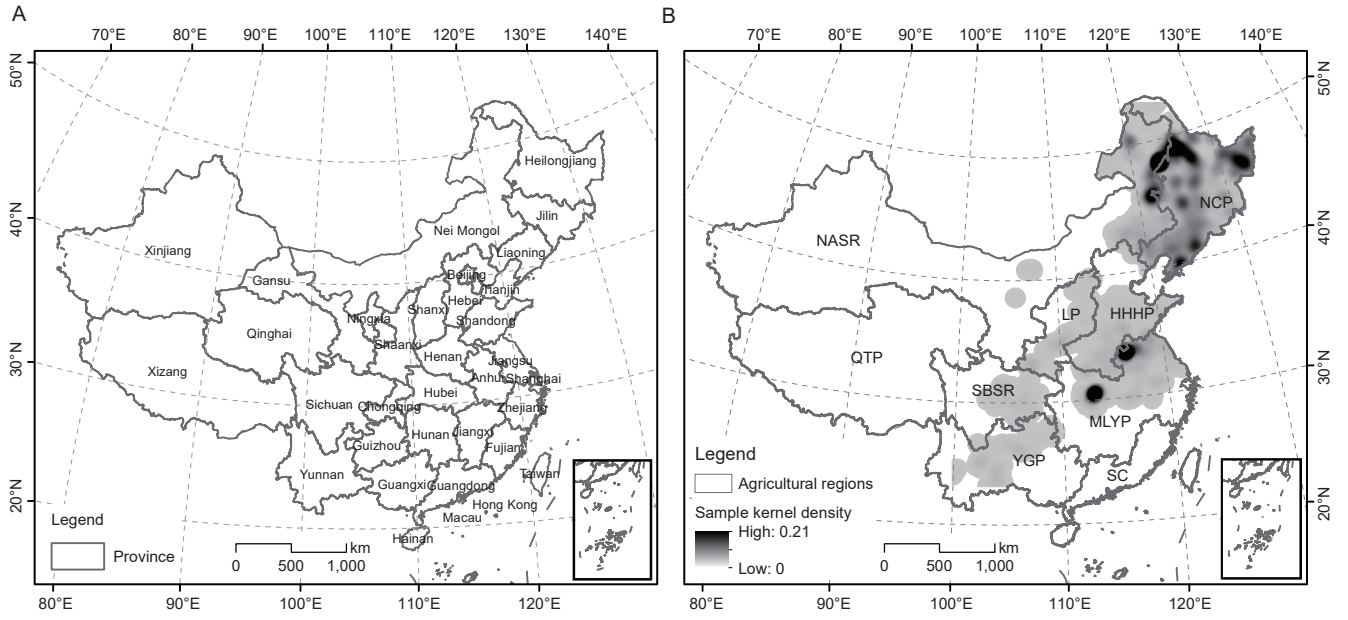
2. Materials and methods

2.1. Study area

The study area encompasses the entirety of China, including 34 provincial-level regions with a total land area of approximately 9.6 million km^2 (Fig. 1-A). The terrain is predominantly mountainous, with mountains, plateaus, and hills accounting for 33, 26, and 10% of the land area. Influenced by diverse topographical and climatic conditions, China is divided into nine major agricultural regions (Fig. 1-B): the Northeast China Plain (NCP), the northern arid and semiarid region (NASR), the Huang-Huai-Hai Plain (HHHP), the Loess Plateau (LP), the Qinghai-Tibet Plateau (QTP), the Middle-lower Yangtze Plain (MLYP), the Sichuan Basin and surrounding regions (SBSR), the Yunnan-Guizhou Plateau (YGP), and the Southern China (SC) (Pan *et al.* 2023; Liang *et al.* 2024).

2.2. Data source

We utilized multi-source data such as Landsat-5/7/8/9 imagery, land cover products, existed soybean datasets, statistical data, ground truth data, and various crop layers to support soybean mapping and spatiotemporal analysis, as detailed in Appendix A.



审图号 : GS 京 (2026) 0655 号

Fig. 1 Overview of the study area. A, provincial boundaries. B, the nine agricultural regions and field survey locations. NCP, Northeast China Plain; NASR, northern arid and semiarid region; HHHP, Huang-Huai-Hai Plain; LP, Loess Plateau; QTP, Qinghai-Tibet Plateau; MLYP, Middle-lower Yangtze Plain; SBSR, Sichuan Basin and surrounding regions; YGP, Yunnan-Guizhou Plateau; SC, southern China. Validation sample density is represented using a kernel density surface calculated with a 1-km search radius, where darker shades indicate higher sample concentration.

Landsat series imagery Landsat satellite imagery (2000–2022) served as the primary dataset for soybean mapping. The sensors utilized include the thematic mapper (TM), enhanced thematic mapper plus (ETM+), operational land imager (OLI), and the most recent operational land imager 2 (OLI-2) on Landsat-9. All available Landsat surface reflectance data were processed on the GEE platform to enable long-term, national-scale soybean mapping. A harmonic fusion approach was employed to integrate Landsat product series and address spectral discrepancies among TM/ETM+ and OLI sensors (Roy *et al.* 2016; Savage *et al.* 2018; Vogeler *et al.* 2018). Four sensing scenarios were created: Landsat-5 and -7 (2000–2011), Landsat-7 (2012), Landsat-7 and -8 (2013–2021), and Landsat-7, -8, and -9 (2022). Furthermore, the Landsat quality assessment (QA) band was used to mask clouds, cloud shadows, and snow, ensuring the reliability and accuracy of the observations.

Land cover dataset The 30-m annual land cover product of China (CLCD) served as a reference dataset for land cover classification. Produced using training samples from the China Land Use/Cover Dataset, satellite time-series data, and Google Earth imagery, the CLCD was generated through the Random Forest (RF) algorithm on GEE (Yang and Huang 2021). This dataset, spanning 1985–2022, provides valuable insights into land-use trends and patterns across China. The cropland layer from the CLCD was extracted annually to exclude non-cropland pixels from the soybean mapping analysis, ensuring spatial accuracy and reliability.

Existed soybean dataset Six publicly available soybean datasets were used as reference materials. The 30-m dataset from Li *et al.* (2021) cover Heilongjiang Province from 1984 to 2020. The 30-m maps from Xuan *et al.* (2023) cover the NCP from 2013 to 2021. You *et al.* (2021) provide 10-m maps for Northeast China from 2017 to 2019. The 10-m ChinaSoyArea10m dataset by Mei *et al.* (2024) includes 14 provinces from 2017 to 2021. The 10-m map from Li H *et al.* (2023) covers the entire country for the year 2019. Zhang *et al.* (2024) released the 10-m ChinaSoybean10 dataset, spans 10 provinces from 2019 to 2022. These datasets differ in their spatial extent, ranging from single provinces to nationwide coverage.

Statistical data Statistics were collected from Chinese statistical authorities to assess the reliability of the ChinaSoyA30m dataset and uncover the driving mechanisms of soybean planting dynamics. Specifically, we collected 31 provincial, 284 prefecture, and 2,384 county-level soybean area from 2000 to 2022 to compare the consistency of the remotely sensed soybean area with the statistical area. In addition, eight prefectural-level economic and social factors were collected to support the driving force analysis, such as “primary industry employees”, ”value added of primary industry“, “pesticide usage”, gross power of agricultural machinery”, ”agricultural fertilizer usage, “agricultural plastic film usage”, ”total area of crops planted”, and “effective irrigated area”.

Field samples A total of 21,592 ground truth samples were utilized for validation (Fig. 1-B and Table 1). The dataset incorporated 17,870 crop samples collected by You *et al.* (2021) for Northeast China from 2017 to 2019, derived from field surveys, Sentinel-2 imagery, and Google Earth (You and Dong 2020; You *et al.* 2021). Additional validation samples were drawn from major soybean-producing regions

(14 provinces) in 2019, as reported by Mei *et al.* (2024) were added for ground reference. Finally, field survey data collected in Henan, Anhui, and Hubei provinces were included to ensure robust validation of the ChinaSoyA30m dataset.

Other crop layers Five publicly available crop datasets were employed to extract crop-specific phenological features from Landsat imagery. These include CARM30, a long-term 30-m rapeseed map for China from 2000–2022 (Liu *et al.* 2024); 10-m rice maps from 2017–2022 (Shen *et al.* 2023); 30-m wheat maps from 2016–2018 (Dong *et al.* 2020); and a 10-m crop map for 2019 covering both soybean and maize, provided a dual crop layer reference (Li H *et al.* 2023).

2.3. Methodology

The workflow of this study is depicted in Fig. 2. Initially, the time-series phenological characteristics of major crops were analyzed across China's agricultural regions. Based on these analyses, soybean phenological samples were automatically generated using the GWCCI. To enhance the quality and representativeness of the samples, noise and errors were filtered through gap statistics and *K*-means clustering, while spectral angle measurements were applied to remove outliers from primary clusters. The purified samples were then combined with Landsat series imagery on GEE to map soybean distribution. The ChinaSoyA30m dataset was validated against existed soybean maps, agricultural statistics, and field survey data. The spatiotemporal dynamics and driving factors of soybean cultivation were finally systematically examined.

Table 1 The ground truth dataset used in this study

Year	Type		Total
	Soybean	Others	
2017	1,464	3,337	4,801
2018	1,924	5,985	7,909
2019	1,584	7,298	8,882

Generating initial training samples Training samples are essential for machine learning classifiers, yet their acquisition is often constrained by the limited availability of field survey data (Zhang *et al.* 2022; Gao *et al.* 2023). Recent advancements in automated classification methodologies, which integrate unsupervised and supervised algorithms, have been applied to crop mapping, including rapeseed (Liu *et al.* 2024), wheat (Yang *et al.* 2021), and paddy rice (Zhang *et al.* 2023). These approaches significantly reduce the economic and temporal costs of sample collection and mitigate the limitations posed by traditional field surveys (Wang *et al.* 2019; Belgiu *et al.* 2021). In this study, we employed the GWCCI-assisted method to automatically generate initial soybean training samples. The GWCCI is a spectral index proposed by Chen *et al.* (2023) for identifying soybean crops based on phenological characteristics, which has proven effective in key soybean-producing countries, including China, the United States, Brazil, and Argentina. The formula for the GWCCI is as follows:

$$GWCCI=NDVI_t\times\rho SWIR_t \quad (1)$$

$$NDVI=\frac{\rho NIR-\rho Red}{\rho NIR+\rho Red} \quad (2)$$

where ρNIR, ρRed, and ρSWIR represent the surface

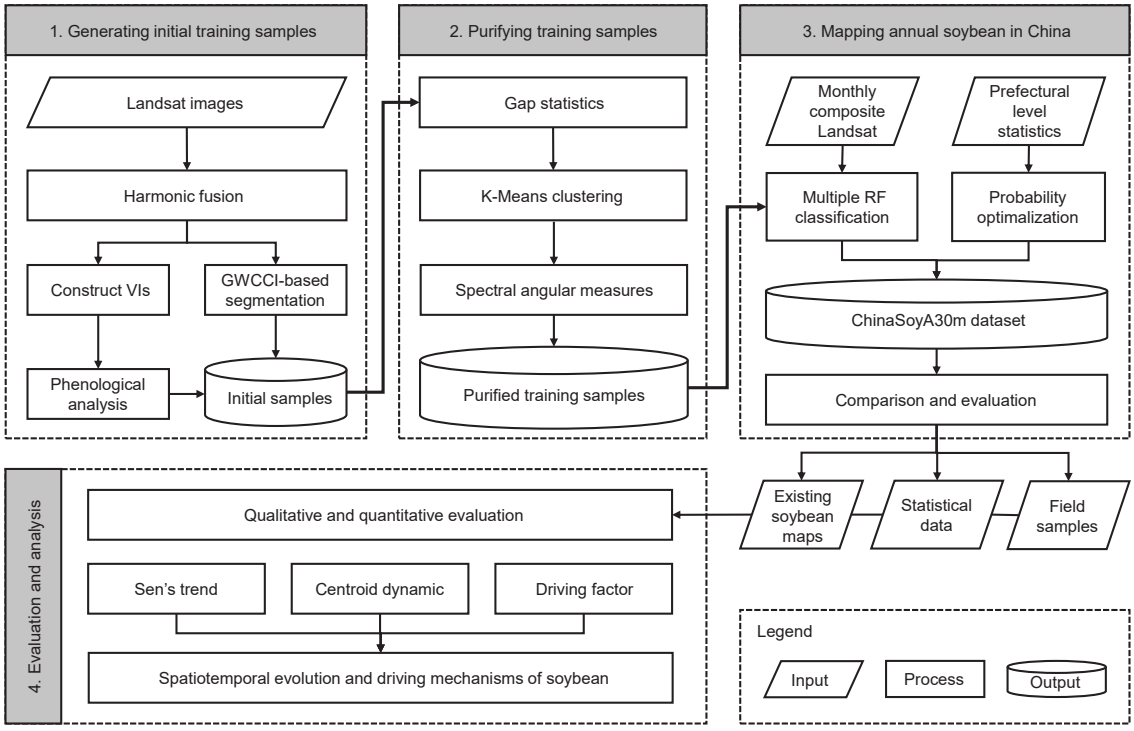


Fig. 2 The workflow of this study. VIs, vegetation indices; GWCCI, greenness and water content composite index; RF, random forest.

reflectance of the near-infrared, red, and short-wave infrared band, t denotes the peak growing season of soybean.

To evaluate the effectiveness of the GWCCI, we analyzed the time-series phenological profiles of major crops across nine agricultural regions in China. The NDVI and GWCCI for each Landsat surface reflectance image from 2018 to 2019 were calculated. Subsequently, 1,000 random samples were generated for each agricultural region based on existed crop layers (Appendix B). The resulting time-series phenological curves are presented in Fig. 3.

Crop growth patterns were analyzed using time-series phenological curves. To accurately capture the phenological dynamics of soybean, we defined a broad soybean growing period for each agricultural region based on the time-series NDVI, field surveys, and crop calendar (https://ipad.fas.usda.gov/rssiws/al/crop_calendar/che.aspx). Regions such as the NCP, NASR, and QTP predominantly cultivate single-season crops with growth cycles of about 210 days. In contrast, the HHHP, LP, and MLYP feature double-peak pattern in NDVI curves, characteristic of double-cropping systems where soybeans are rotated with winter crops, resulting in a growth cycle of around 150 days. The SBSR, YGP, and SC exhibit bi-annual or tri-annual cropping systems, producing less regular phenological pattern (Huang *et al.* 2023; Li X *et al.* 2023). A comparison of the GWCCI and NDVI curves revealed that soybeans typically exhibit a growth cycle of approximately 180 days, from May to October, with the GWCCI effectively distinguishing soybeans

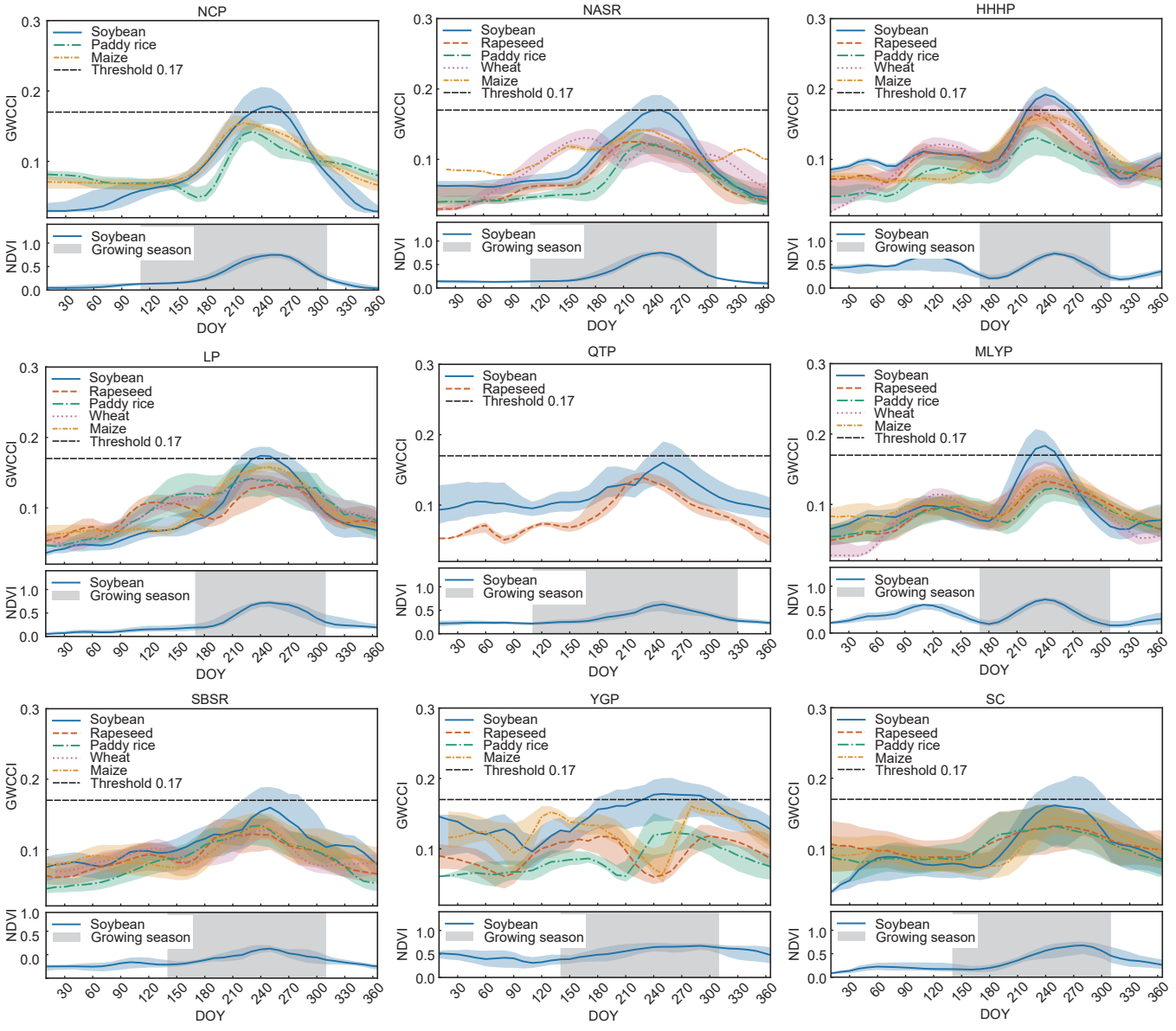


Fig. 3 Time-series phenological profiles of major crops in nine agricultural regions (Julian day). NCP, Northeast China Plain; NASR, northern arid and semiarid region; HHHP, Huang-Huai-Hai Plain; LP, Loess Plateau; QTP, Qinghai-Tibet Plateau; MLYP, Middle-lower Yangtze Plain; SBSR, Sichuan Basin and surrounding regions; YGP, Yunnan-Guizhou Plateau; SC, southern China. DOY, day of year. For each crop, the median curve of greenness and water content composite index (GWCCI) or normalized difference vegetation index (NDVI) is shown as a solid line in a distinct color, with shaded bands representing the interquartile range (25th to 75th percentile). The soybean growing season is indicated by a gray-filled area. Wheat data are from 2018, while data for all other crops are from 2019.

from other crops.

Based on the above phenological analysis, we adopted the threshold segmentation method proposed by Chen *et al.* (2023) to extract initial soybean training samples. Pixels with GWCCI values exceeding the predefined threshold were classified as soybean, as shown in eq. (3):

$$\text{Sample}_{\text{initial}}=\begin{cases}1, \text{ if GWCCI}_{\text{TW}}\geq\text{threshold}\\0, \text{ otherwise}\end{cases} \quad (3)$$

where $\text{Sample}_{\text{initial}}$ is the initial training sample, the threshold was set to 0.17 in reference to the suggestion of Chen *et al.* (2023), and TW is the peak growing season window of soybean. According to the phenological analysis, the peak growth times of soybean for NCP, NASR, HHHP, LP, MLYP, and SBSR were around DOY 240; for QTP, around DOY 210; and for YGP and SC, around DOY 270. To capture the peak growth signal of soybeans, TW was defined as two months before and after the peak growth period for soybeans in each region.

Purifying training samples To refine the initial training samples generated by the phenology-derived approach, we employed the gap statistic and *K*-means clustering methods. We clustered the initial samples using *K*-means, an unsupervised clustering method that calculates Euclidean distances to assign samples to clusters, thus creating *k* clusters (Jain 2010). This method has been applied in land cover mapping and is valuable for supervised classification without ground truth samples (Wang *et al.* 2019; Belgiu *et al.* 2021). To determine the optimal number of clusters (*k*), we used the gap statistic method proposed by Tibshirani *et al.* (2001) to compare the clustering of the original data with a null reference dataset to determine the optimal *k*, as shown in eqs. (4) and (5):

$$Gap(k)=\frac{1}{B}\sum_{b=1}^{B}\log(W_k^b)-\log(W_k) \quad (4)$$

$$W_k=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|^2 \quad (5)$$

where k is the number of clusters, B is the number of reference datasets, W_k represents the within-cluster sum of squares for clusters k, and W_k^b is the within-cluster sum of squares for the b-th reference dataset. In the formula for W_k, C_i is the i-th cluster, x represents a data point in cluster C_i, μ_i is the centroid (mean) of cluster C_i, and $\left\|x-\mu_i\right\|^2$ is the squared Euclidean distance between x and μ_i.

The smallest k where the gap value no longer significantly increases was chosen as the optimal number of clusters (eqs. (6) and (7)):

$$Gap(k)\geq Gap(\text{k}+1)-S_{k+1} \quad (6)$$

$$S_k=\sqrt{1+\frac{1}{B}}\times\text{sd}(\log(W_k^b)) \quad (7)$$

where $\text{sd}(\log(W_k^b))$ is the standard deviation of $\log(W_k^b)$, S_k represents the standard error of the gap statistic for clusters k.

We further applied the spectral angle mapping (SAM) method to measure the spectral angle between unknown and reference pixels, enabling the exclusion of pixels with large deviations from the reference class (Dennison *et al.* 2004; Mafanya *et al.* 2022). This method helps eliminate non-soybean pixels by comparing the spectral angle θ with a predefined threshold (Liu *et al.* 2024), as shown in eq. (8):

$$\theta=\cos^{-1}\left(\frac{\sum_{i=1}^{n}t_i r_i}{\sqrt{\sum_{i=1}^{n}t_i^2}\sqrt{\sum_{i=1}^{n}r_i^2}}\right) \quad (8)$$

where t is the spectral vector of the unknown pixel, r is the spectral vector of the reference class, and n is the length of the vector. θ ranges from 0 to π, and a smaller angle means a higher similarity between t and r. In this study, r was defined as the mean of the soybean clusters obtained from the *K*-means and gap statistic methods, and the threshold value of θ was defined as the maximum spectral angle between r and the mean spectrum perturbed by three times standard deviation of the soybean clusters. Pixels with a θ greater than the threshold are excluded from the soybean category.

Mapping soybean extent Cloud contamination and missing values in satellite time-series images can significantly hinder accurate crop mapping (Li *et al.* 2022). The multi-RF fusion mapping method, based on ensemble learning, addresses this issue by training multiple classifiers on distinct time-series images and integrating their classification probability maps to produce seamless results (Zhang *et al.* 2022, 2023). We applied this method to generate an average soybean probability map by combining classification probabilities from multiple time-series datasets, as shown in eq. (9):

$$P_i=\frac{\sum_{j=1}^{m}p_{i,j}}{m} \quad (9)$$

where i represents a pixel, j represents a temporal phase, m represents the length of the time series over the entire soybean growing season, $p_{i,j}$ represents the probability of pixel i being classified as soybean in temporal phase j, and P_i represents the probability that pixel i is eventually classified as soybean.

The multi-RF fusion mapping approach was implemented on the GEE platform. Monthly median-composited Landsat images and purified soybean and non-soybean samples (with a 1:1 ratio) were used for training. The RF classifier was configured with 100 decision trees (ntree) and a number of predictors (mtry) equal to the square root of the total input features. A 10-fold cross-validation method was employed to reduce biases and validate model performance. The models exhibited high accuracy, confirming their reliability for soybean mapping (Appendix C).

While remote sensing techniques provide objective insights into the spatial distribution and dynamics of crops, uncertainties in training sample quality, image variability, and model performance can introduce mapping errors (Verburg *et al.* 2011). To address these issues, we employed the iterative area gap spatial allocation method proposed by Hu *et al.* (2021). This method compares statistical data with remotely sensed crop areas to adjust the probability map (Hu *et al.* 2021; Yang *et al.* 2024). Specifically, it defines the area gap (Agap) between the remote sensing-derived soybean area and the statistical data, using an area adjustment

parameter (Ap) for each prefectural-level city, as described in eqs. (10) and (11):

$$Agap=\left|(RArea-SArea)/SArea\right| \quad (10)$$

$$Ap=SArea/RArea \quad (11)$$

where *RArea* denotes the remotely sensed soybean area and *SArea* denotes the soybean area derived from agricultural statistics. In this study, when *Agap*<0.1 and *RArea*<1,000 ha, the soybean probability map was retained without further adjustment. Otherwise, the probability map is multiplied by the adjustment parameter *Ap* in iterations to dynamically adjust to minimize *Agap*. After adjusting the map, threshold segmentation was applied to finalize the soybean classification result.

Evaluation and analysis We assessed the soybean mapping results using regression analysis and confusion matrices. The consistency and errors of seven soybean datasets were evaluated at provincial, prefectural, and county levels using official statistical data. Pixel-level accuracy was validated against field survey data. Statistical evaluation metrics included the determination coefficient (R^2) and root mean square error (RMSE), while pixel-level verification used producer accuracy (PA), user accuracy (UA), overall accuracy (OA), and F1 score to quantify performance.

Subsequently, we analyzed the temporal dynamics of soybean planting, identified patterns of planting center shifts, and explored driving mechanisms using Sen's trend analysis (Gocic and Trajkovic 2013), centroid dynamic analysis (Luo *et al.* 2020), and geographic detectors (Xu *et al.* 2024), respectively. The formulas are shown in eqs. (12)–(14).

$$S=Median\left(\frac{X_j-X_i}{j-i}\right),\ 1\le i<j\le n \quad (12)$$

where X_j and X_i are soybean area in year j and i, and n is the total number of years. A positive S indicates an upward trend, while a negative S indicates a downward trend.

$$x_c=\frac{\sum_{i=1}^{n}x_iA_i}{\sum_{i=1}^{n}A_i},\ y_c=\frac{\sum_{i=1}^{n}y_iA_i}{\sum_{i=1}^{n}A_i} \quad (13)$$

where x_c and y_c represent the longitude and latitude of the centroid, x_i and y_i are the coordinates of the ith region, and A_i is the soybean planting area in the corresponding region.

$$q=1-\frac{\sum_{h=1}^{L}N_h\sigma_h^2}{N\sigma^2} \quad (14)$$

where N is the total number of samples, σ^2 is the variance of the soybean planting area, N_h and σ_h^2 represent the sample size and variance within the h-th sub-region, and L is the number of sub-regions. A higher q-value indicates a stronger explanatory power of the factor for the spatial distribution.

3. Results

3.1. Comparison with existed soybean datasets

Figs. 4 and 5 compare the spatial distribution of seven soybean datasets. The reference year of 2019 was selected to align with the coverage periods of the available datasets, ensuring a consistent temporal framework for comparison. As the maps from Li *et al.* (2021) (Fig. 4-B, 30-m), Xuan *et al.* (2023) (Fig. 4-C; 30-m), and You *et al.* (2021) (10-m) (Fig. 4-D; 10-m) primarily focused on Northeast China, this region was selected for detailed analysis. Landsat-8 images from August 2019, corresponding to the podding stage of soybean, served as the reference. The ChinaSoyA30m dataset (Fig. 4-A; 30-m) demonstrated high spatial consistency with satellite imagery. However, the map from Li *et al.* (2021) overclassified soybean areas and introduced noise, particularly at Site 3. The map from Xuan *et al.* (2023) showed overall agreement but misclassified some non-soybean areas (e.g., Sites 3 and 4). The map from You *et al.* (2021) captured finer detail (e.g., Sites 2 and 4) but misidentified soybeans at Sites 1 and 3.

Further comparisons in four typical regions (Fig. 5) highlighted similar trends. At Site 1, a non-soybean area, the ChinaSoyArea10m map from Mei *et al.* (2024) (Fig. 5-B; 10-m) overestimated soybean pixels. At Site 2, both the ChinaSoyA30m map (Fig. 5-A; 30-m) and the map from Li H *et al.* (2023) (Fig. 5-C; 10-m) accurately mapped soybeans, while ChinaSoyArea10m (Fig. 5-B) and the ChinaSoybean10 map from Zhang *et al.* (2024) (Fig. 5-D; 10-m) underperformed. At Site 3, all datasets except ChinaSoybean10 identified soybean fields correctly. At Site 4, only ChinaSoyA30m accurately mapped the central soybean field. Despite its 30-m resolution, ChinaSoyA30m consistently aligned well with satellite imagery, often outperforming or matching higher-resolution datasets in accuracy and reliability.

3.2. Comparison with agricultural statistical data

Regression analyses of the seven datasets at provincial, prefecture, and county levels demonstrate the strong performance of the ChinaSoyA30m dataset (Appendices D–N). At the provincial level, it achieved an R^2 exceeding 0.9. At the prefecture level, R^2 values ranged from 0.63 to 0.975, with RMSEs between 24,491 and 71,441 ha. The lowest agreement occurred in 2001, with R^2 values of 0.63 nationally and 0.598 in Northeast China, while the highest agreement was in 2022, with R^2 values of 0.975 nationally and 0.985 in the Northeast. At the county level, the dataset maintained robust performance, with R^2 values between 0.721 to 0.889 nationwide and 0.69 to 0.913 in the Northeast. These results confirm the robustness of the ChinaSoyA30m dataset across multiple spatial scales.

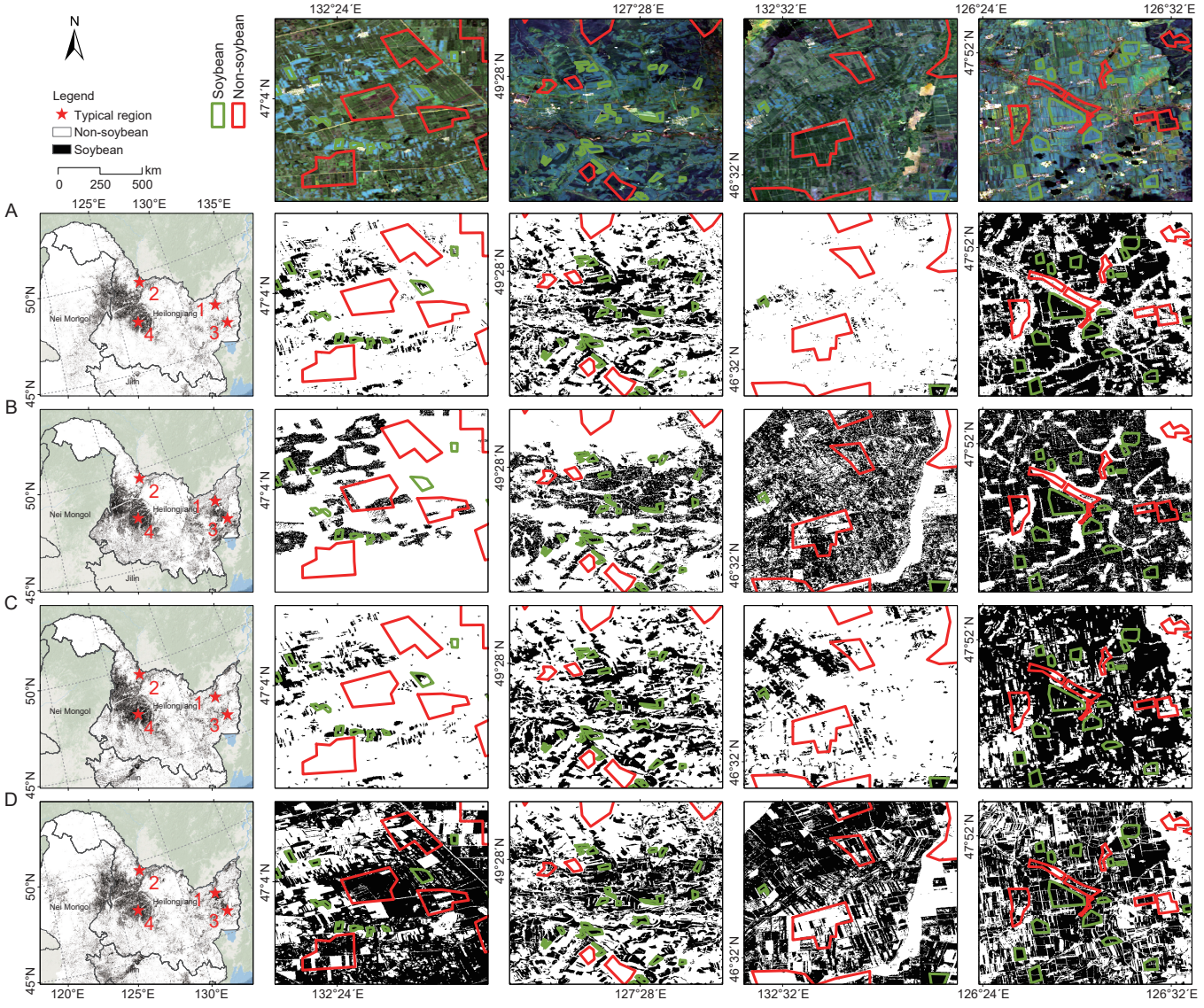
A benchmark comparison with six other soybean datasets (Fig. 6) further underscores the strengths of ChinaSoyA30m. The *x*-axis represents the accuracy, and the *y*-axis represents the different datasets. The datasets used for validation vary across provincial, prefectural, and county levels due to differences in spatial coverage and temporal availability. Each point in the figure represents a single validation result for a specific year. The median values for each dataset are shown as solid black lines, and the interquartile range is represented by dashed black lines. At the provincial level, ChinaSoyA30m achieved the second-highest R^2, following ChinaSoybean10, and recorded the lowest RMSE. At the prefecture and county levels, it ranked second in R^2 and RMSE. Although other datasets outperformed at certain scales, their limited temporal coverage restricts broader

application.

3.3. Pixel-wise validation using field samples

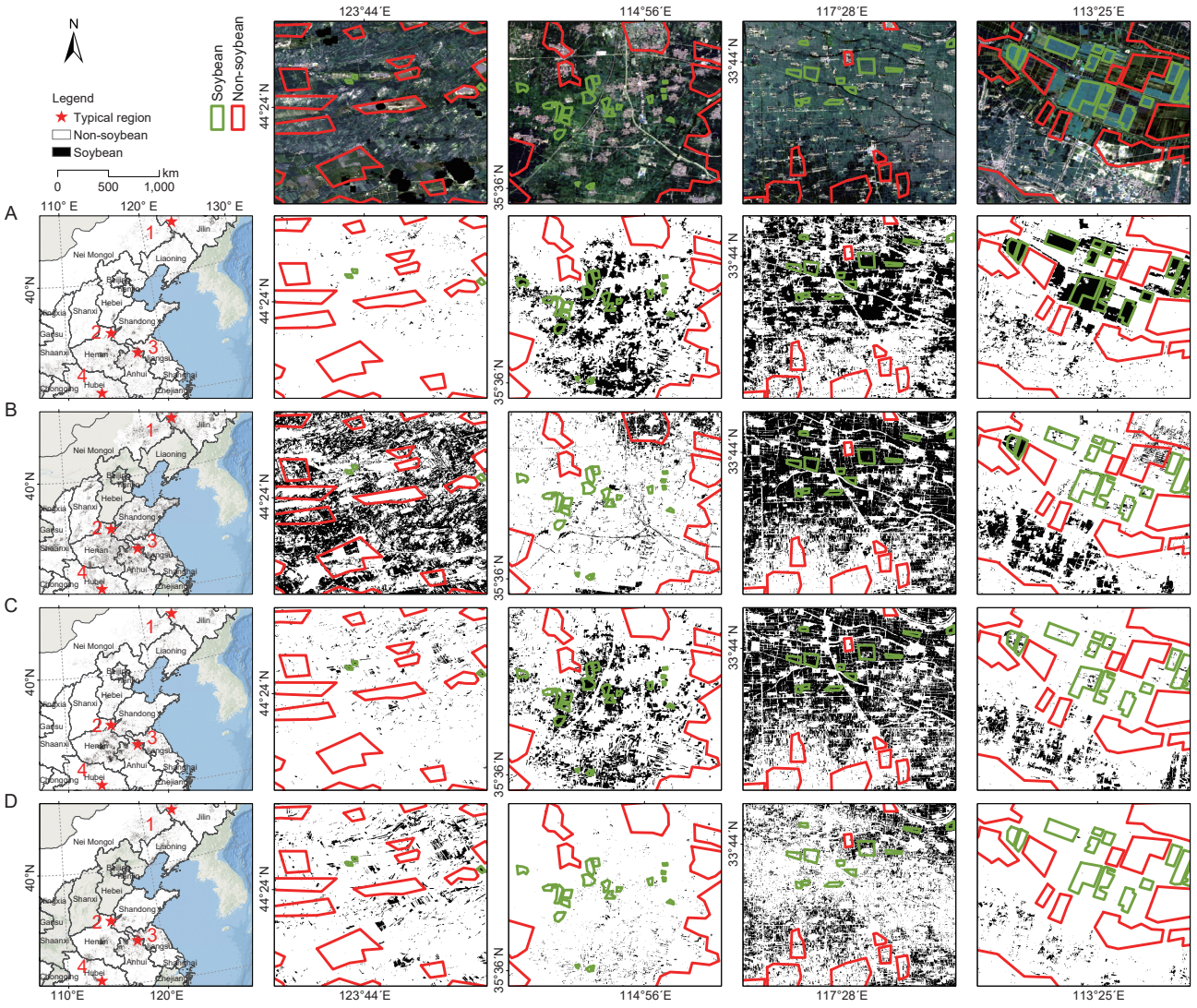
Pixel-wise validation of seven soybean datasets using field data is summarized in Table 2. In 2017, the map from You *et al.* (2021) achieved the highest F1 score (83.41%), while ChinaSoyA30m ranked third (70.16%). The map from Li *et al.* (2021) had the lowest accuracy (F1=59.87%). In 2018, ChinaSoyA30m outperformed others with an F1 score of 80.4%, followed by the map from You *et al.* (2021) (77.84%), whereas the map from Li *et al.* (2021) again ranked the lowest (F1=37.47%). In 2019, the map from You *et al.* (2021) maintained the highest accuracy (F1=85.76%), with ChinaSoyA30m ranking second (F1=78.38%). ChinaSoyA30m leveraged the largest validation sample set due to its nationwide coverage, ensuring comprehensive evaluation. While the map from You *et al.* (2021) showed higher accuracy in specific years, ChinaSoyA30m consistently outperformed other 30-m resolution maps and demonstrated competitive pixel-level accuracy overall.

Appendix O presents the confusion matrices of seven soybean maps across nine agricultural regions. Due to limited validation samples, accuracy assessment in 2017 was restricted to the NCP and NASR. The map from You *et al.* (2021) achieved the highest accuracy in both regions, with an F1 score of 93.18 and 63.57%. ChinaSoyA30m ranked second with a F1 score of 50.18% in NASR. In 2018, validation samples were extended to HHHP and MLYP. In the NCP, the maps from You *et al.* (2021) and Xuan *et al.* (2023) ranked first and second, with F1 scores of 90.80 and 86.96%. ChinaSoyA30m performed the best in NASR, HHHP, and MLYP, with F1 scores of 72.51, 87.34, and



审图号：GS 京 (2026) 0655 号

Fig. 4 Inner comparison of four soybean maps based on 2019 data and Landsat-8 August composites. A, 30-m ChinaSoyA30m map in this study. B, 30-m map from Li *et al.* (2021). C, 30-m map from Xuan *et al.* (2023). D, 10-m map from You *et al.* (2021). The second to fifth columns show enlarged views of four representative regions, with areas of approximately 215, 360, 370, and 210 km^2.



审图号：GS 京 (2026) 0655 号

Fig. 5 Inner comparison of four soybean maps based on 2019 data and Landsat-8 August composites. A, 30-m ChinaSoyA30m map in this study. B, 10-m ChinaSoyArea10m map from Mei *et al.* (2024). C, 10-m map from Li H *et al.* (2023). D, 10-m ChinaSoybean10 map from Zhang *et al.* (2024). The second to fifth columns show enlarged views of four representative regions, with areas of approximately 280, 150, 230, and 160 km^2.

90.49%. In 2019, validation expanded to LP, SBSR, and YGP. In the NCP, the maps from Li H *et al.* (2023) and Xuan *et al.* (2023) ranked first and second, with F1 scores of 94.30 and 93.66%, while ChinaSoyA30m scored 91.17%. In the NASR, the map from Li H *et al.* (2023) performed the best, with an F1 score of 81.37%, followed by ChinaSoyA10m at 74.69%. As soybean-positive samples were absent in HHHP, LP, MLYP, and YGP, F1 scores were unavailable. In these regions, ChinaSoyA30m achieved the highest OA, while ChinaSoybean10 generally ranked second. Overall, while ChinaSoyA30m showed no significant advantage, it performed well across multiple regions.

3.4. Spatiotemporal dynamics of soybean cultivation

Fig. 7 depicts the spatial distribution of soybean cultivation across China, with solid black curves representing soybean area distributions along latitudes and longitudes and gray bars indicating area fluctuations from 2000 to 2022. Soybean cultivation was primarily concentrated in Northeast China, particularly in Heilongjiang Province (123°–127°E, 45°–50°N), followed by the North China Plain (114°–118°E, 32°–34°N). In southern China, soybean cultivation was more scattered and less intensive.

Fig. 8 shows annual changes in soybean planting area across nine agricultural regions from 2000 to 2022. The NCP consistently had the largest soybean planting area, while the QTP recorded the smallest. Nationally, soybean cultivation exhibited an overall increasing trend with fluctuations, particularly before 2012. Key turning points in 2008 and 2013, identified *via* a five-year moving average, highlight the impact of government policies. The temporary grain reserve purchase policy introduced in 2008 was associated with an increase in soybean prices, which rose from approximately 3,700 CNY t^{-1} in 2008 to about 4,600 CNY t^{-1} by 2013, initially

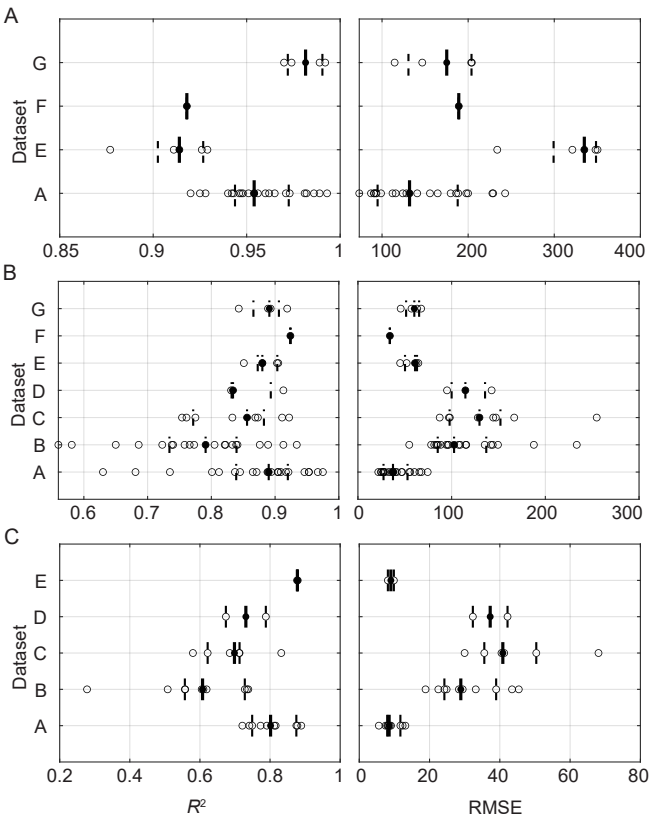


Fig. 6 R^2 and RMSE calculated for the seven soybean maps using statistical data. A–C correspond to the provincial, prefectural, and county levels, respectively. Datasets A to G represent: 30-m ChinaSoyA30m map in this study, the 30-m map from Li *et al.* (2021), the 30-m map from Xuan *et al.* (2023), the 10-m map from You *et al.* (2021), the 10-m ChinaSoyArea10m map from Mei *et al.* (2024), the 10-m map from Li H *et al.* (2023), and the 10-m ChinaSoybean10 map from Zhang *et al.* (2024). Black dots and dashed lines indicate the medians and interquartile ranges, reflecting the distribution of the accuracy metrics.

boosting production enthusiasm (Liu *et al.* 2019; Wang and Wei 2021). However, the policy failed to sustain long-term output growth and was replaced in 2014 by soybean target price subsidies, which significantly incentivized farmers and expanded planting (Huang and Liu 2024).

County-level temporal changes were analyzed using Mann-Kendall and Sen's slope tests (Fig. 9). From 2000 to 2008, 287 counties showed increased soybean cultivation, with significant growth in Heilongjiang, while 54 counties experienced declines. Between 2008 and 2013, soybean areas increased in 46 counties but declined in 34 counties, notably in Heilongjiang, due to rising production costs and reduced profitability, prompting shifts to corn (Xin *et al.* 2020; Hu *et al.* 2022). From 2013 to 2022, cultivation expanded in 387 counties, mainly in Northeast China and the HHHP, while declines were recorded in 193 counties, mostly in Southwest and South China. Over the entire period (2000–2022), 1,183 counties showed increases, while 233 saw declines, particularly in Heilongjiang (e.g., Harbin, Suihua), Jilin (e.g., Changchun, Tonghua), and Liaoning (e.g., Fushun, Benxi). Declines were also evident in less prominent soybean-growing regions, such as Fujian, Guangdong, and Guangxi.

3.5. Relocation of soybean planting centroid

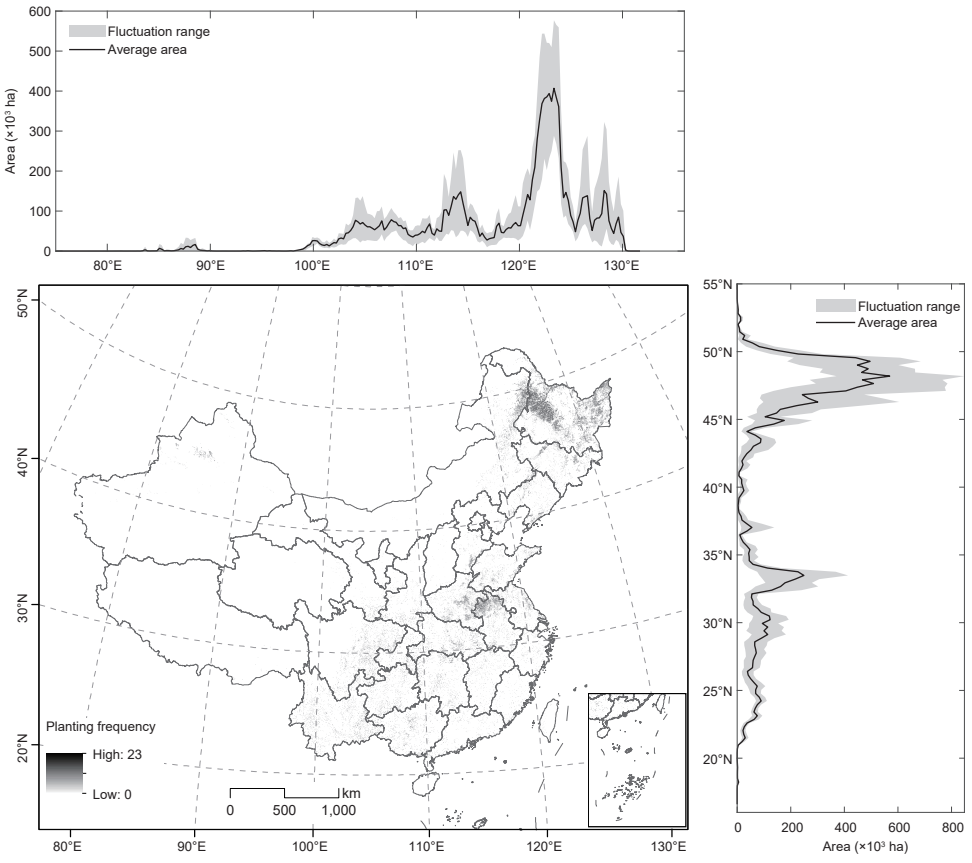
Fig. 10 shows the annual centroids and movement trajectories of soybean cultivation in China from 2000 to 2022. Circular markers represent yearly centroid positions, transitioning in color from white (2000) to blue (2022). Solid black lines connect these points, illustrating temporal shifts in planting patterns.

The spatial dynamics of soybean centroids varied across the nine regions, reflecting differences in agricultural practices, land availability, policy interventions, and environmental constraints. In the NCP, centroids remained concentrated in northern and central Heilongjiang Province with minimal displacement, indicating the stability of well-

Table 2 Accuracy of ChinaSoyA30m dataset *vs.* six existed soybean maps validated against ground truth data

Year	Metric[1)] (%)	Li *et al.* (2021) maps (30-m)	Xuan *et al.* (2023) maps (30-m)	You *et al.* (2021) maps (10-m)	Mei *et al.* (2024) ChinaSoyArea10m (10-m)	Li H *et al.* (2023) map (10-m)	Zhang *et al.* (2024) ChinaSoybean10 (10-m)	ChinaSoyA30m (30-m) in this study
2017	*n*	4,631	4,759	4,792	4,730	–	–	4,797
	UA	52.39	<u>68.44</u>	**89.14**	65.10	–	–	67.76
	PA	69.83	**83.29**	<u>78.38</u>	69.82	–	–	72.73
	OA	77.84	<u>86.08</u>	**89.17**	80.49	–	–	82.41
	F1	59.87	<u>75.14</u>	**83.41**	67.37	–	–	70.16
2018	*n*	5,991	6,357	6,366	7,660	–	–	7,895
	UA	25.45	40.56	**90.61**	77.18	–	–	<u>80.93</u>
	PA	70.97	**86.79**	68.23	73.72	–	–	<u>79.89</u>
	OA	80.43	85.65	<u>88.78</u>	87.39	–	–	**90.39**
	F1	37.47	55.28	<u>77.84</u>	75.41	–	–	**80.40**
2019	*n*	4,431	6,533	6,568	7,032	8,882	8,781	8,882
	UA	44.08	60.33	**89.51**	87.71	<u>86.05</u>	64.20	81.00
	PA	67.39	**90.84**	<u>82.32</u>	59.86	63.66	63.36	75.92
	OA	74.09	89.53	**93.32**	84.04	88.75	86.85	<u>92.01</u>
	F1	53.30	72.51	**85.76**	71.16	73.18	63.78	<u>78.38</u>

[1)] UA, user accuracy; PA, producer accuracy; OA, overall accuracy; F1, F1 score.
Data in bold are the highest performance, data underlined are the second highest performance; –, no data.



审图号 : GS 京 (2026) 0655 号

Fig. 7 Spatial distribution of soybean in China on the longitudinal and latitudinal gradients. The planting frequency of soybean in China from 2000 to 2022 is shown in black. The line graphs indicate the distribution of soybeans across different latitudinal and longitudinal gradients. The solid black curve indicates the average area of soybean from 2000 to 2022, while the gray bar indicates the fluctuation range of soybean area.

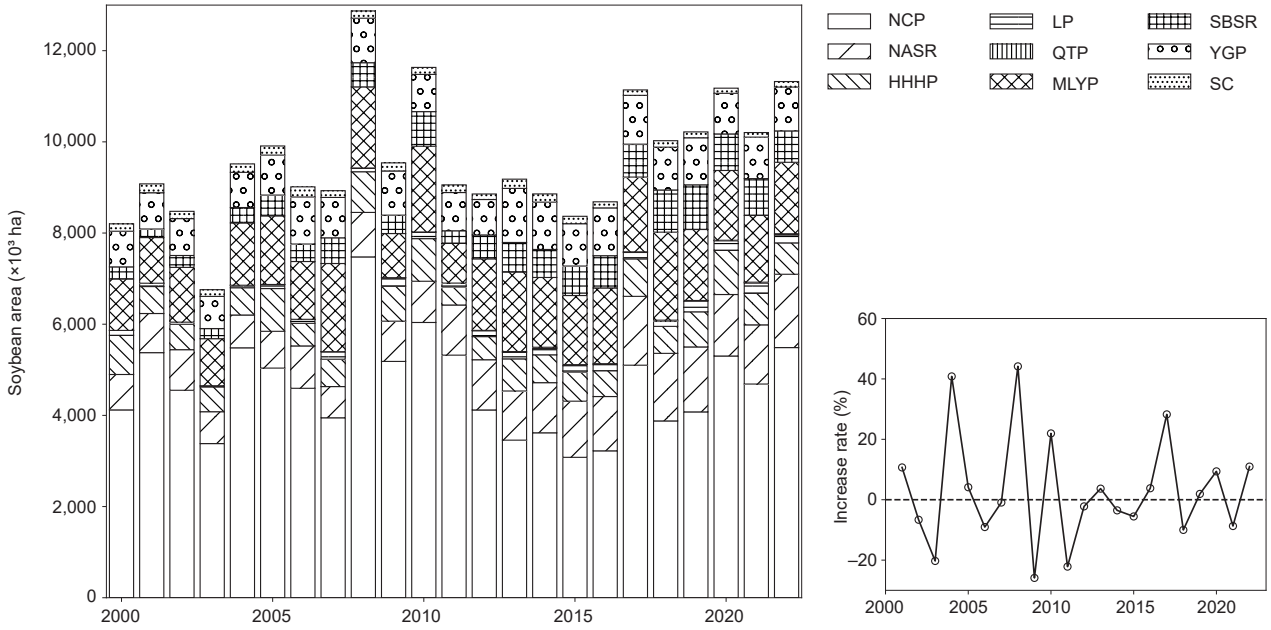
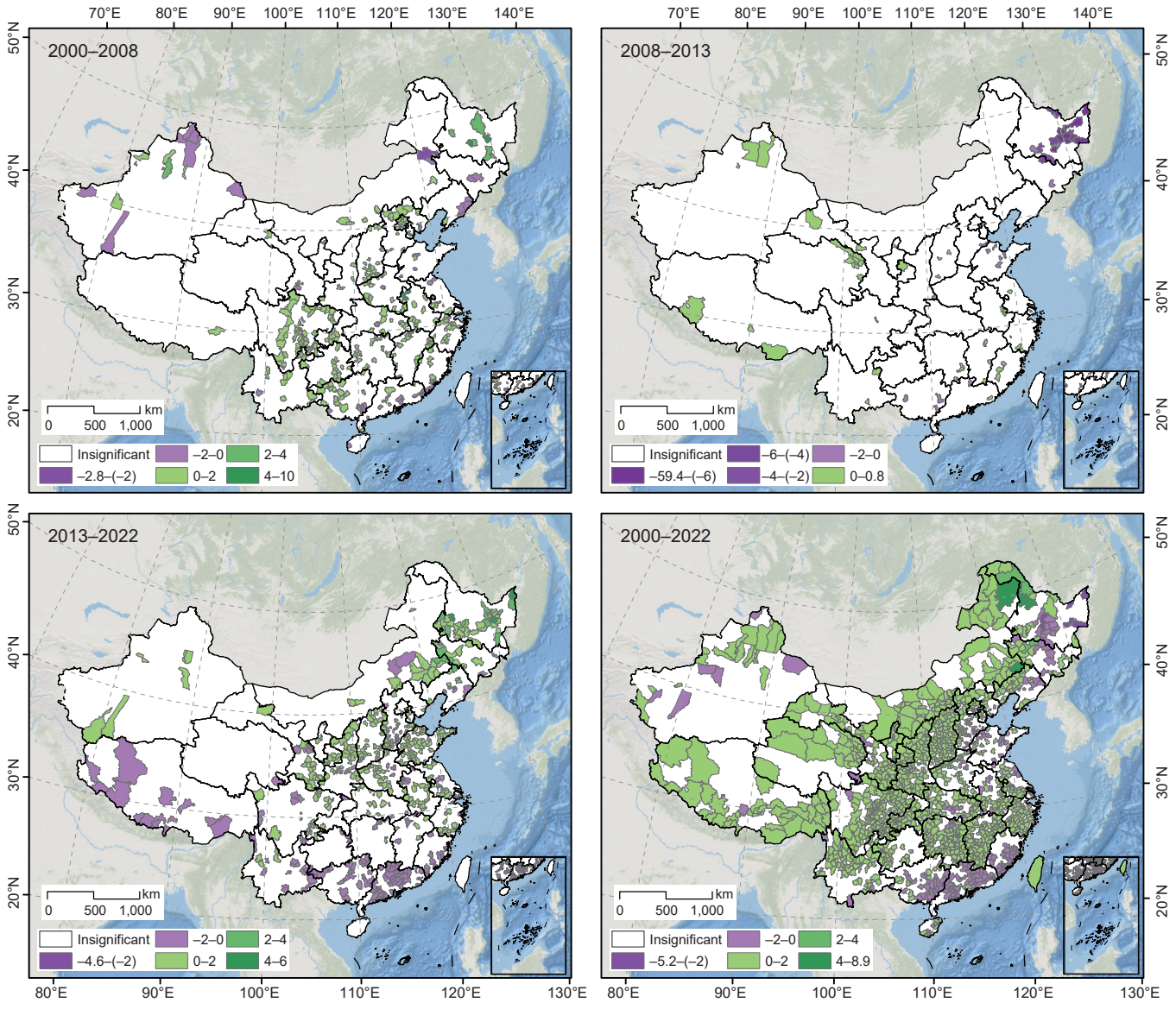


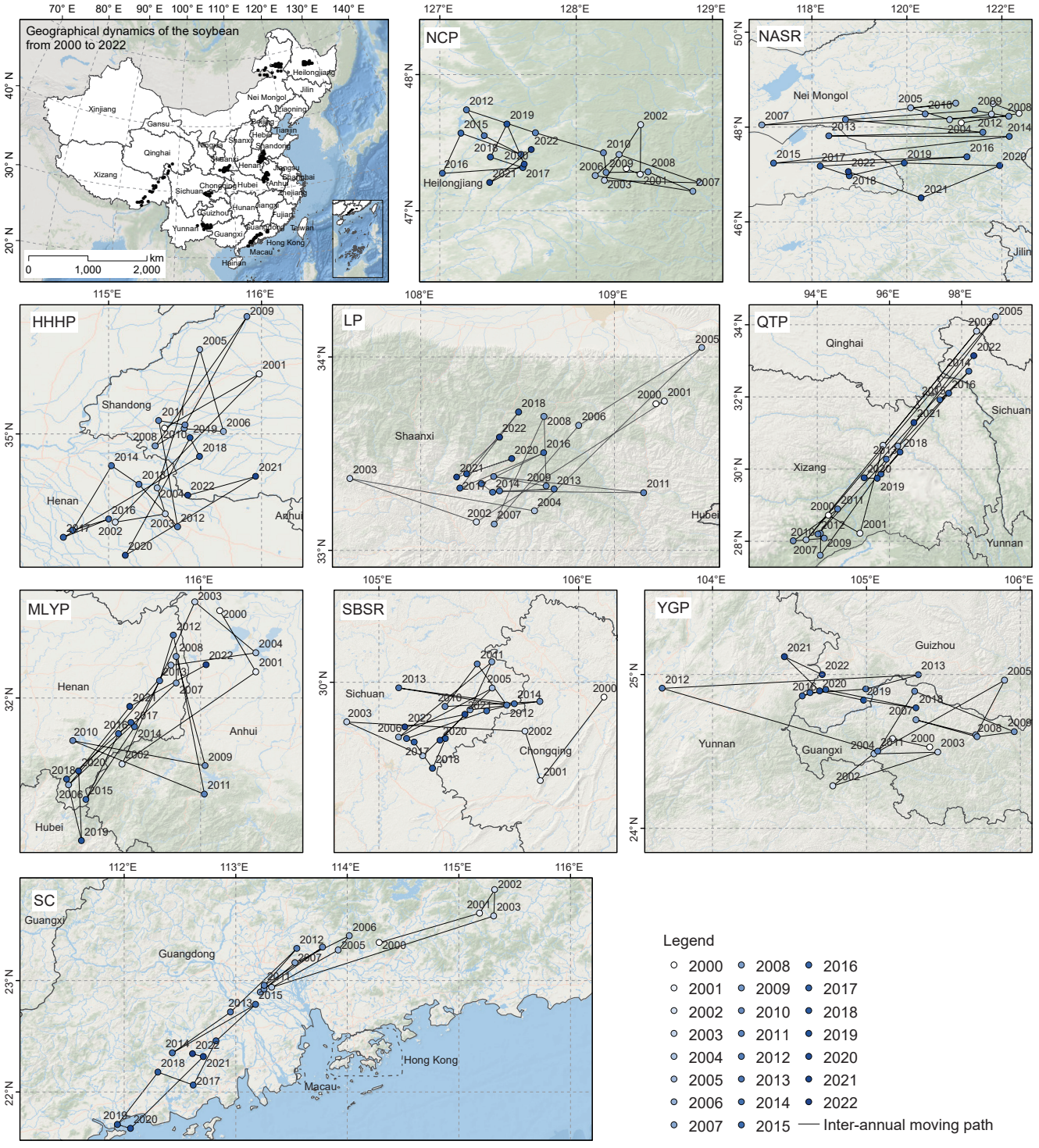
Fig. 8 Temporal dynamics of soybean areas in China from 2000 to 2022. Stacked bar charts represent the dynamics of soybean area in the nine agricultural regions, while line charts represent the dynamics of annual increase rate of soybean across China. NCP, Northeast China Plain; NASR, northern arid and semiarid region; HHHP, Huang-Huai-Hai Plain; LP, the Loess Plateau; QTP, Qinghai-Tibet Plateau; MLYP, Middle-lower Yangtze Plain; SBSR, Sichuan Basin and surrounding regions; YGP, Yunnan-Guizhou Plateau; SC, southern China.



审图号：GS 京 (2026) 0655 号

Fig. 9 Sen's trend of soybean area at county level in China at different stages. Counties in purple and green indicate a decreasing and increasing trend in soybean area, while counties that do not pass the Mann-Kendall test are shown in white (P>0.05).

established cultivation zones. Similarly, in the SBSR, centroid positions exhibited little movement, constrained by topographic limitations and stable cropland allocation, where the dominance of smallholder farming and intensive land use likely restricted large-scale shifts. In the NASR, centroids initially located in eastern Nei Mongol gradually shifted southward, likely driven by changes in cropping systems and climate variability, as rising temperatures and shifting precipitation patterns have altered the agroecological suitability of soybean cultivation (Yin *et al.* 2016; Xiao X *et al.* 2024). A similar southward shift in the MLYP, particularly near the Anhui-Hubei border, suggests agricultural restructuring and increased adoption of rice–soybean rotations, influenced by economic incentives and cropping diversification (Wang *et al.* 2018; Qiu *et al.* 2023). In the HHHP, centroid positions oscillated along a north–south axis across Henan and Shandong provinces, reflecting the influence of alternating crop rotations among soybeans, maize, and wheat, driven by market demand and policy incentives (Zhang *et al.* 2019). In the LP, centroids in the Guanzhong Plain, Shaanxi Province, shifted slightly westward, possibly due to policies promoting crop diversification and adjustments in water resource allocation. In the QTP, highly dispersed centroid movements reflected the fragmented distribution of arable land in river valleys, where marginal suitability for large-scale soybean production likely contributed to planting inconsistencies. In the YGP, centroids near the Yunnan-Guizhou border shifted westward, potentially in response to land consolidation and climate adaptation strategies. In the SC region, centroids in Guangdong Province migrated along a northeast–southwest axis, reflecting pressures from urbanization, land



审图号 : GS 京 (2026) 0655 号

Fig. 10 Geographical dynamics of soybean planting area centroids from 2000 to 2022. NCP, Northeast China Plain; NASR, northern arid and semiarid region; HHHP, Huang-Huai-Hai Plain; LP, Loess Plateau; QTP, Qinghai-Tibet Plateau; MLYP, Middle-lower Yangtze Plain; SBSR, Sichuan Basin and surrounding regions; YGP, Yunnan-Guizhou Plateau; SC, southern China. The centroids for different years are represented by different colored dots, while the solid black lines indicate the inter-annual moving paths of the centroids. The inset (upper left) summarizes centroid locations by provincial administrative units, whereas the remaining panels depict centroid shifts across major agricultural regions.

conversion, and declining agricultural land availability due to economic development.

Overall, centroid movements were more pronounced in southern regions, where urban expansion and land policy adjustments played a greater role, whereas northern regions remained dominant due to well-established agricultural infrastructure and favorable climatic conditions. These shifts underscore the role of policy-driven transformations and climate

adaptation strategies in shaping planting patterns, warranting further investigation through the integration of socioeconomic datasets, climate models, and field-based surveys.

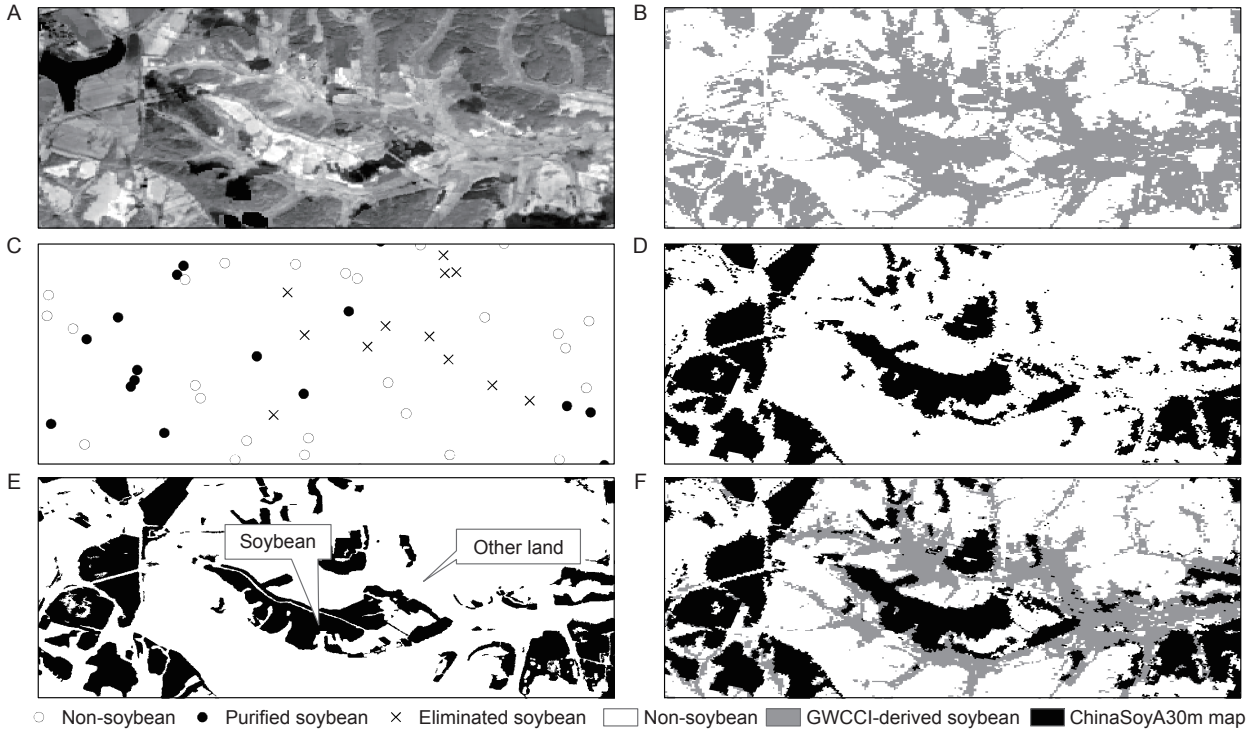
4. Discussion

4.1. Advantages of our soybean mapping strategy

Accurate soybean cultivation information is crucial for farmers, sustainable agricultural planning, and international food trade. Various remote sensing methods have been developed to map the spatial distribution of soybeans in China, resulting in several regional and national datasets (Li *et al.* 2021; You *et al.* 2021; Li H *et al.* 2023; Xuan *et al.* 2023; Mei *et al.* 2024; Zhang *et al.* 2024). Our methodology offers several advantages over existing efforts, primarily in terms of automation and dataset competitiveness.

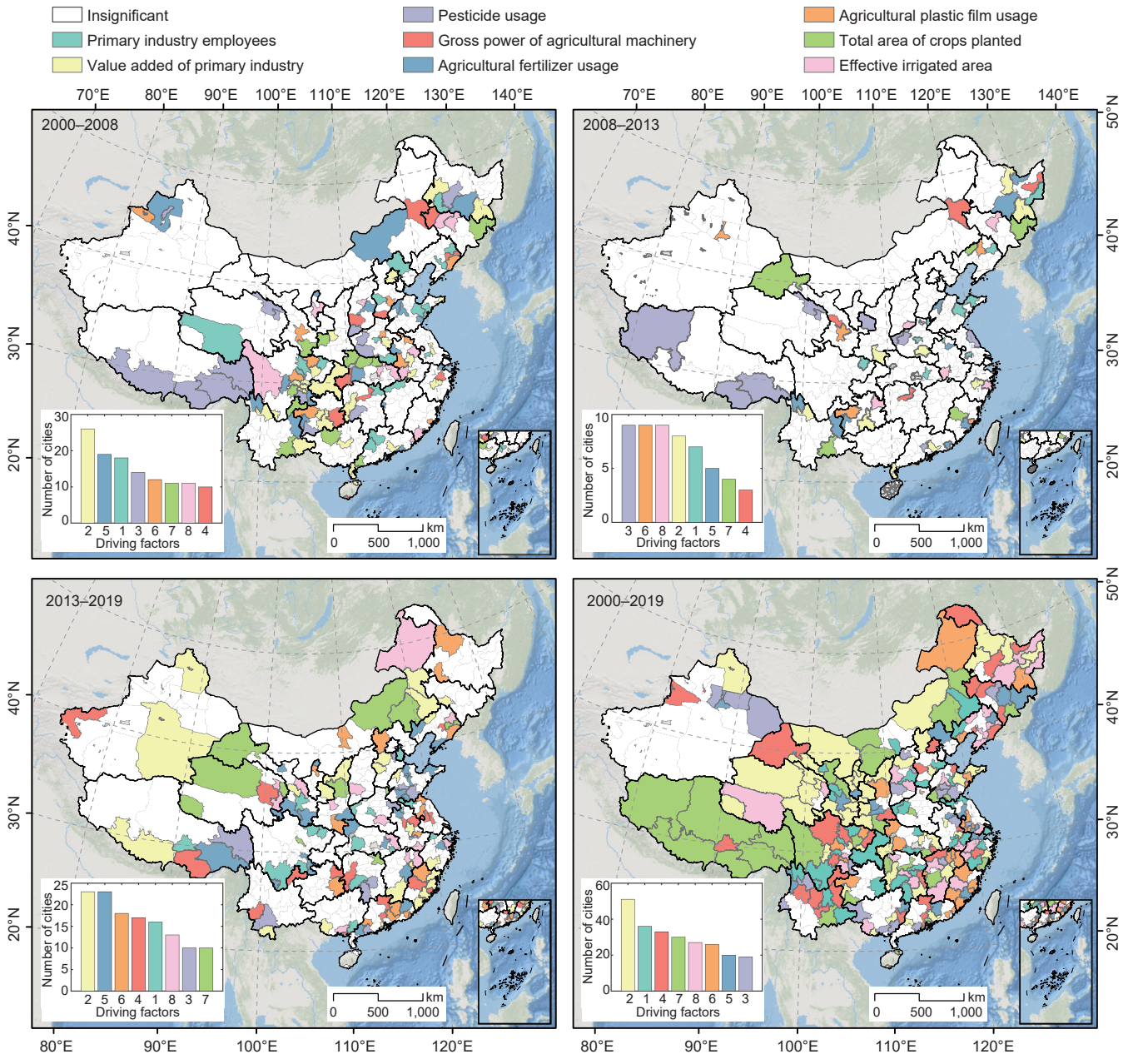
First, our approach enables efficient soybean mapping through an automated classification strategy. Machine learning has become a leading method in crop mapping due to its effectiveness and robustness (Weiss *et al.* 2020). However, the availability of sufficient historical training samples is often a limitation for these methods (Maxwell *et al.* 2018). To address this, we generated and refined annual training samples using a phenology-based approach, enabling the creation of accurate annual soybean maps using a machine learning classifier. The mapping process, as illustrated in Fig. 11, begins by calculating GWCCI bands (Fig. 11-A) from Landsat imagery during key phenological periods of soybean growth. These bands are then classified using threshold segmentation (Fig. 11-B) to differentiate between soybeans and non-soybean areas, generating initial training samples (Fig. 11-C). Since the initial classification may include errors, we refined the samples by applying gap statistics, *K*-means, and the SAM method to eliminate noise. In Fig. 11-C, black dots represent pure soybean samples, crossed dots denote rejected noise, and white dots correspond to non-soybean samples. Finally, we used a multiple RF fusion approach to map soybean distribution across time-series Landsat images (Fig. 11-D). The true distribution of soybean is shown in Fig. 11-E (with reference to field samples). The overlay comparison demonstrates strong spatial agreement between our results and the GWCCI-derived soybean map, with our method showing superior accuracy in certain regions (Fig. 11-F).

A second key strength of our approach is its comparative advantage over other publicly available soybean datasets. With a 30-m resolution, our map captures spatial detail similar to the maps from Li *et al.* (2021) and Xuan *et al.* (2023), although it is less detailed than the Sentinel-2-based datasets. In terms of time span, our 23-year map ranks second only to the dataset from Li *et al.* (2021), which spans 37 years but is limited to Heilongjiang Province. Notably, the ChinaSoyA30m dataset provides a nationwide coverage of all 34 provincial regions in China, offering broader spatial extent than most existed maps. In contrast, while the map from Li H *et al.* (2023) covers the entire country, it is restricted to a single year. Overall, despite the lower resolution compared to Sentinel-2-derived maps, the extended time span and nationwide coverage of the ChinaSoyA30m dataset represent



审图号：GS 京 (2026) 0655 号

Fig. 11 Schematic representation of our proposed soybean mapping approach. A, the greenness and water content composite index (GWCCI) layer. B, soybean obtained by GWCCI. C, initial training samples and purified samples. D, ChinaSoyA30m map. E, true distribution of soybean. F, overlay of ChinaSoyA30m map with GWCCI-derived soybean.



审图号：GS 京 (2026) 0655 号

Fig. 12 Driving forces of soybean planting area changes at the prefecture level in China at different stages. Different colors represent different driving factors, while cities that do not pass the significance test are shown in white (P>0.05). Inset bar charts indicate the number of cities dominated by each driving factor, where x-axis values 1–8 represent: primary industry employees, value added of primary industry, pesticide usage, gross power of agricultural machinery, agricultural fertilizer usage, agricultural plastic film usage, total area of crops planted, and effective irrigated area.

its primary strengths, making it particularly valuable for long-term analysis of soybean cropping patterns.

In summary, our methodology combines phenological analysis and machine learning to map soybean distribution using Landsat data. By generating unsupervised training samples based on phenological cycles, we reduce the reliance on costly and time-consuming field surveys, thus addressing the challenge of limited historical training data (Belgiu *et al.* 2021; Zhang *et al.* 2022). Validation against agricultural statistics and field data confirms the reliability of our maps, demonstrating significant spatial and temporal advantages over the six other soybean datasets.

4.2. Drivers of soybean planting area shifts

Fig. 12 illustrates the factors driving changes in soybean planting areas at the prefecture level. Different colors represent specific driving factors, while white indicates cities where no factor was statistically significant (P>0.05). In the lower left corner of each subfigure, the number of cities predominantly influenced by different driving factors during the

given time period is displayed in descending order, with the *x*-axis representing the numerical identifiers of these factors. From 2000 to 2008, soybean area changes in 121 cities were significantly influenced by factors such as the "value added of primary industry" (26 cities) and "agricultural fertilizer usage" (19 cities). Between 2008 and 2013, dominant drivers were observed in only 54 cities, with "pesticide usage", "agricultural plastic film usage", and "effective irrigated area" emerging as key factors. From 2013 to 2019, the number of cities affected by significant drivers rose to 130, with a wider range of influences. The "value added of primary industry" and "agricultural fertilizer usage" remained prominent, while the impact of "gross power of agricultural machinery" expanded to 17 cities, reflecting the growing role of mechanization in soybean production. Over the entire period (2000–2019), significant drivers were identified in 242 cities, with the "value added of primary industry", "primary industry employees", and "gross power of agricultural machinery" consistently dominant.

Regional variations in driving mechanisms were evident. The "value added of primary industry" had the broadest influence, underscoring the economic impact on planting adjustments. In northern China, "gross power of agricultural machinery" was a critical driver, reflecting the significance of mechanization in intensive agriculture (Gong 2018; Wu and Ding 2021). In regions such as Xizang and Nei Mongol, the expansion of arable land and irrigation areas played a larger role, emphasizing the importance of land reclamation and water management (Cao *et al.* 2021). The increasing share of cities without significant driving factors, particularly in the NCP, suggests heightened vulnerability of soybean cultivation to external uncertainties, such as global market fluctuations and policy changes (Adjemian *et al.* 2021; Barboza Martignone *et al.* 2024).

4.3. Uncertainty and improvement of the soybean maps

Uncertainty in soybean mapping arises mainly from the quality of Landsat imagery, with cloud contamination posing a significant challenge for crop monitoring (Whitcraft *et al.* 2015). From 2000 to 2022, Landsat sensor constellations with varying revisit times enabled long-term annual soybean mapping. The mapping accuracy was evaluated using the ratio of cloud-free images during the soybean growing season. Reliability was the lowest in 2012, when only Landsat-7 data were available, but improved significantly in 2022 due to multi-sensor integration. Cloud cover, especially in southern China, reduces mapping accuracy and creates spatial discontinuities. However, multi-sensor data integration effectively mitigates this issue (Whitcraft *et al.* 2015; Zhang *et al.* 2022; Liu and Zhang 2023). Despite challenges in rainy southwestern regions, soybean maps for major production areas in Northeast China showed high reliability. Future efforts will focus on developing global soybean datasets at 10-m resolution using Sentinel-2 imagery to enhance accuracy and reduce uncertainty.

5. Conclusion

This study presents a fully automated framework for generating high-resolution, long-term soybean maps across China, culminating in the ChinaSoyA30m dataset spanning 2000–2022. By integrating phenology-based sample generation with a multi-temporal RF ensemble and leveraging dense Landsat time series, the proposed method enables reliable, large-scale crop mapping with minimal reliance on manual labeling. This approach addresses key limitations in existing datasets, particularly regarding spatial coverage, temporal consistency, and automation. We also reveal significant spatiotemporal dynamics in soybean cultivation across China, shaped by agricultural policy shifts, mechanization, and regional land-use patterns. While northern production zones remain stable, southern regions show greater spatial fluidity, reflecting broader structural changes in land management and climate adaptation. Statistical analyses further highlight economic and infrastructural drivers underlying planting area variations. Overall, this work contributes a scalable, validated mapping strategy and an open-access dataset that supports long-term agricultural monitoring and policy evaluation. Future research will focus on multi-sensor data fusion and fine-resolution mapping to further improve the accuracy in cloudy regions and enhance understanding of crop-environment interactions under global change.

Acknowledgements

This study was supported by the China Postdoctoral Science Foundation (2023M743450 and GZC20232614).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT 4.0 to assist in improving the clarity and readability of the language. All content generated was carefully reviewed and revised by the authors as necessary, and the authors take full responsibility for the final content of the article.

Data availability

The ChinaSoyA30m dataset and associated reliability layers for China from 2000 to 2022 are available at Zenodo Open Data Repository (https://doi.org/10.5281/zenodo.13806551).

Appendices associated with this paper are available at https://doi.org/10.1016/j.jia.2025.07.021

References

Adjemian M K, Smith A, He W. 2021. Estimating the market effect of a trade war: The case of soybean tariffs. *Food Policy*, **105**, 102152.

Ashourloo D, Nematollahi H, Huete A, Aghighi H, Azadbakht M, Shahrabi H S, Goodarzdashti S. 2022. A new phenology-based method for mapping wheat and barley using time-series of Sentinel-2 images. *Remote Sensing of Environment*, **280**, 113206.

Barboza Martignone G M, Ghosh B, Papadas D, Behrendt K. 2024. The rise of Soybean in international commodity markets: A quantile investigation. *Heliyon*, **10**, e34669.

Belgiu M, Bijker W, Csillik O, Stein A. 2021. Phenology-based sample generation for supervised crop type classification. *International*

Journal of Applied Earth Observation and Geoinformation, **95**, 102264.
Cao X, Zeng W, Wu M, Li T, Chen S, Wang W. 2021. Water resources efficiency assessment in crop production from the perspective of water footprint. *Journal of Cleaner Production*, **309**, 127371.
Cavalett O, Ortega E. 2010. Integrated environmental assessment of biodiesel production from soybean in Brazil. *Journal of Cleaner Production*, **18**, 55–70.
Chai L, Liu A, Li X, Guo Z, He W, Huang J, Bai T, Liu J. 2024. Telecoupled impacts of the Russia–Ukraine war on global cropland expansion and biodiversity. *Nature Sustainability*, **7**, 432–441.
Chen H, Li H, Liu Z, Zhang C, Zhang S, Atkinson P M. 2023. A novel Greenness and Water Content Composite Index (GWCCI) for soybean mapping from single remotely sensed multispectral images. *Remote Sensing of Environment*, **295**, 113679.
Chen W, Zhang B, Kong X, Wen L, Liao Y, Kong L. 2022. Soybean production and spatial agglomeration in China from 1949 to 2019. *Land*, **11**, 734.
Dennison P E, Halligan K Q, Roberts D A. 2004. A comparison of error metrics and constraints for multiple endmember spectral mixture analysis and spectral angle mapper. *Remote Sensing of Environment*, **93**, 359–367.
Dong J, Fu Y, Wang J, Tian H, Fu S, Niu Z, Han W, Zheng Y, Huang J, Yuan W. 2020. Early-season mapping of winter wheat in China based on Landsat and Sentinel images. *Earth System Science Data*, **12**, 3081–3095.
Dong J, Xiao X, Kou W, Qin Y, Zhang G, Li L, Jin C, Zhou Y, Wang J, Biradar C, Liu J, Moore B. 2015. Tracking the dynamics of paddy rice planting area in 1986–2010 through time series Landsat images and phenology-based algorithms. *Remote Sensing of Environment*, **160**, 99–113.
Gao Z, Guo D, Ryu D, Western A W. 2023. Training sample selection for robust multi-year within-season crop classification using machine learning. *Computers and Electronics in Agriculture*, **210**, 107927.
Gocic M, Trajkovic S. 2013. Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia. *Global and Planetary Change*, **100**, 172–182.
Gong B. 2018. Agricultural reforms and production in China: Changes in provincial production function and productivity in 1978–2015. *Journal of Development Economics*, **132**, 18–31.
Guilpart N, Iizumi T, Makowski D. 2022. Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. *Nature Food*, **3**, 255–265.
Hu M, Wu W, Yu Q, Tang H, Wen Y, Zhao F. 2022. Spatial-temporal variations in green, blue and gray water footprints of crops: how do socioeconomic drivers influence? *Environmental Research Letters*, **17**, 124024.
Hu Q, Yin H, Friedl M A, You L, Li Z, Tang H, Wu W. 2021. Integrating coarse-resolution images and agricultural statistics to generate sub-pixel crop type maps and reconciled area estimates. *Remote Sensing of Environment*, **258**, 112365.
Huang C, You S, Liu A, Li P, Zhang J, Deng J. 2023. High-resolution national-scale mapping of paddy rice based on Sentinel-1/2 data. *Remote Sensing*, **15**, 4055.
Huang H, Wang J, Liu C, Liang L, Li C, Gong P. 2020. The migration of training samples towards dynamic global land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, **161**, 27–36.
Huang Y, Liu Z. 2024. Improving Northeast China's soybean and maize planting structure through subsidy optimization considering climate change and comparative economic benefit. *Land Use Policy*, **146**, 107319.
Jain A K. 2010. Data clustering: 50 years beyond *K*-means. *Pattern Recognition Letters*, **31**, 651–666.
Li H, Song X-P, Hansen M C, Becker-Reshef I, Adusei B, Pickering J, Wang L, Wang L, Lin Z, Zalles V, Potapov P, Stehman S V, Justice C. 2023. Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation. *Remote Sensing of Environment*, **294**, 113623.
Li X, Yang Q, Bao L, Li G, Yu J, Chang X, Gao X, Yu L. 2023. Temporal trends and future projections of accumulated temperature changes in China. *Agronomy*, **13**, 1203.
Li X, Yu L, Peng D, Gong P. 2021. A large-scale, long time-series (1984□2020) of soybean mapping with phenological features: Heilongjiang Province as a test case. *International Journal of Remote Sensing*, **42**, 7332–7356.
Li Z, Shen H, Weng Q, Zhang Y, Dou P, Zhang L. 2022. Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects. *ISPRS Journal of Photogrammetry and Remote Sensing*, **188**, 89–108.
Liang J, Pan S, Xia N, Chen W, Li M. 2024. Threshold response of the agricultural modernization to the open crop straw burning CO_2 emission in China's nine major agricultural zones. *Agriculture, Ecosystems & Environment*, **368**, 109005.
Liu S, Zhang P, Marley B, Liu W. 2019. The factors affecting farmers' soybean planting behavior in Heilongjiang Province, China. *Agriculture*, **9**, 188.
Liu W, Li S, Tao J, Liu X, Yin G, Xia Y, Wang T, Zhang H. 2024. CARM30: China annual rapeseed maps at 30 m spatial resolution from 2000 to 2022 using multi-source data. *Scientific Data*, **11**, 356.
Liu W, Zhang H. 2023. Mapping annual 10 m rapeseed extent using multisource data in the Yangtze River Economic Belt of China (2017–2021) on Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, **117**, 103198.
Liu Z, Ying H, Chen M, Bai J, Xue Y, Yin Y, Batchelor W D, Yang Y, Bai Z, Du M, Guo Y, Zhang Q, Cui Z, Zhang F, Dou Z. 2021. Optimization of China's maize and soy production can ensure feed sufficiency at lower nitrogen and carbon footprints. *Nature Food*, **2**, 426–433.
Luo Y, Zhang Z, Cao J, Chen Y, Zhang L. 2020. Drivers of planting area and yield shifts for three staple crops across China, 1950–2013. *Climate Research*, **80**, 73–84.
Mafanya M, Tsele P, Zengeya T, Ramoelo A. 2022. An assessment of image classifiers for generating machine-learning training samples for mapping the invasive *Campuloclinium macrocephalum* (Less.) DC (pompom weed) using DESIS hyperspectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, **185**, 188–200.
Maxwell A E, Warner T A, Fang F. 2018. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, **39**, 2784–2817.
Mei Q, Zhang Z, Han J, Song J, Dong J, Wu H, Xu J, Tao F. 2024. ChinaSoyArea10m: a dataset of soybean-planting areas with a spatial resolution of 10 m across China from 2017 to 2021. *Earth System Science Data*, **16**, 3213–3231.
Pan Y, Yang R, Qiu J, Wang J, Wu J. 2023. Forty-year spatio-temporal dynamics of agricultural climate suitability in China reveal shifted major crop production areas. *Catena*, **226**, 107073.
Qiu B, Hu X, Yang P, Tang Z, Wu W, Li Z. 2023. A robust approach for large-scale cropping intensity mapping in smallholder farms from vegetation, brownness indices and SAR time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, **203**, 328–344.
Roy D P, Kovalskyy V, Zhang H K, Vermote E F, Yan L, Kumar S S, Egorov A. 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sensing of Environment*, **185**, 57–70.
Savage S L, Lawrence R L, Squires J R, Holbrook J D, Olson L E, Braaten J D, Cohen W B. 2018. Shifts in forest structure in Northwest Montana from 1972 to 2015 using the Landsat archive from multispectral scanner to operational land imager. *Forests*, **9**, 157.
Shen R, Pan B, Peng Q, Dong J, Chen X, Zhang X, Ye T, Huang J, Yuan W. 2023. High-resolution distribution maps of single-season rice in China from 2017 to 2022. *Earth System Science Data*, **15**, 3203–3222.
Singh P, Kumar R, Sabapathy S N, Bawa A S. 2008. Functional and Edible Uses of Soy Protein Products. *Comprehensive Reviews in Food Science and Food Safety*, **7**, 14–28.
Song X-P, Hansen M C, Potapov P, Adusei B, Pickering J, Adami M, Lima A, Zalles V, Stehman S V, Di Bella C M, Conde M C, Copati E J, Fernandes L B, Hernandez-Serna A, Jantz S M, Pickens A H, Turubanova S, Tyukavina A. 2021. Massive soybean expansion in South America since 2000 and implications for conservation. *Nature Sustainability*, **4**, 784–792.
Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society* (Series B: Statistical Methodology), **63**, 411–423.
Verburg P H, Neumann K, Nol L. 2011. Challenges in using land use and land cover data for global change studies. *Global Change Biology*, **17**, 974–989.

Vogeler J C, Braaten J D, Slesak R A, Falkowski M J. 2018. Extracting the full value of the Landsat archive: Inter-sensor harmonization for the mapping of Minnesota forest canopy cover (1973–2015). *Remote Sensing of Environment*, **209**, 363–374.

Wang J, Zhang Z, Liu Y. 2018. Spatial shifts in grain production increases in China and implications for food security. *Land Use Policy*, **74**, 204–213.

Wang S, Azzari G, Lobell D B. 2019. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sensing of Environment*, **222**, 303–317.

Wang S, Di Tommaso S, Deines J M, Lobell D B. 2020. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. *Scientific Data*, **7**, 307.

Wang W, Wei L. 2021. Impacts of agricultural price support policy on price variability and welfare: Evidence from China's soybean market. *Agricultural Economics*, **52**, 3–17.

Weiss M, Jacob F, Duveiller G. 2020. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, **236**, 111402.

Whitcraft A K, Vermote E F, Becker-Reshef I, Justice C O. 2015. Cloud cover throughout the agricultural growing season: Impacts on passive optical earth observations. *Remote Sensing of Environment*, **156**, 438–447.

Wu F, Geng Y, Zhang Y, Ji C, Chen Y, Sun L, Xie W, Ali T, Fujita T. 2020. Assessing sustainability of soybean supply in China: Evidence from provincial production and trade data. *Journal of Cleaner Production*, **244**, 119006.

Wu S, Ding S. 2021. Efficiency improvement, structural change, and energy intensity reduction: Evidence from Chinese agricultural sector. *Energy Economics*, **99**, 105313.

Xiao G, Huang J, Song J, Li X, Du K, Huang H, Su W, Miao S. 2024. A novel soybean mapping index within the global optimal time window. *ISPRS Journal of Photogrammetry and Remote Sensing*, **217**, 120–133.

Xiao X, Zhang J, Liu Y. 2024. Impacts of crop type and climate changes on agricultural water dynamics in Northeast China from 2000 to 2020. *Remote Sensing*, **16**, 1007.

Xin F, Xiao X, Dong J, Zhang G, Zhang Y, Wu X, Li X, Zou Z, Ma J, Du G, Doughty R B, Zhao B, Li B. 2020. Large increases of paddy rice area, gross primary production, and grain production in Northeast China during 2000–2017. *Science of the Total Environment*, **711**, 135183.

Xu S, Wang J, Altansukh O, Chuluun T. 2024. Spatiotemporal evolution and driving mechanisms of desertification on the Mongolian Plateau. *Science of the Total Environment*, **941**, 173566.

Xuan F, Dong Y, Li J, Li X, Su W, Huang X, Huang J, Xie Z, Li Z, Liu H, Tao W, Wen Y, Zhang Y. 2023. Mapping crop type in Northeast China during 2013–2021 using automatic sampling and tile-based image classification. *International Journal of Applied Earth Observation and Geoinformation*, **117**, 103178.

Yang G, Yu W, Yao X, Zheng H, Cao Q, Zhu Y, Cao W, Cheng T. 2021. AGTOC: A novel approach to winter wheat mapping by automatic generation of training samples and one-class classification on Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, **102**, 102446.

Yang J, Huang X. 2021. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth System Science Data*, **13**, 3907–3925.

Yang M, Guo B, Wang J. 2024. A novel and robust method for large-scale single-season rice mapping based on phenology and statistical data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **213**, 14–32.

Yin X, Olesen J E, Wang M, Öztürk I, Zhang H, Chen F. 2016. Impacts and adaptation of the cropping systems to climate change in the Northeast Farming Region of China. *European Journal of Agronomy*, **78**, 60–72.

You N, Dong J. 2020. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, **161**, 109–123.

You N, Dong J, Huang J, Du G, Zhang G, He Y, Yang T, Di Y, Xiao X. 2021. The 10-m crop type maps in Northeast China during 2017–2019. *Scientific Data*, **8**, 41.

Zaaboul F, Zhao Q, Xu Y, Liu Y. 2022. Soybean oil bodies: A review on composition, properties, food applications, and future research aspects. *Food Hydrocolloids*, **124**, 107296.

Zhang C, Zhang H, Tian S. 2023. Phenology-assisted supervised paddy rice mapping with the Landsat imagery on Google Earth Engine: Experiments in Heilongjiang Province of China from 1990 to 2020. *Computers and Electronics in Agriculture*, **212**, 108105.

Zhang D, Pan Y, Zhang J, Hu T, Zhao J, Li N, Chen Q. 2020. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. *Remote Sensing of Environment*, **247**, 111912.

Zhang H, Liu W, Zhang L. 2022. Seamless and automated rapeseed mapping for large cloudy regions using time-series optical satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, **184**, 45–62.

Zhang H, Lou Z, Peng D, Zhang B, Luo W, Huang J, Zhang X, Yu L, Wang F, Huang L, Liu G, Gao S, Hu J, Yang S, Cheng E. 2024. Mapping annual 10-m soybean cropland with spatiotemporal sample migration. *Scientific Data*, **11**, 439.

Zhang Y, Qi Y, Shen Y, Wang H, Pan X. 2019. Mapping the agricultural land use of the North China Plain in 2002 and 2012. *Journal of Geographical Sciences*, **29**, 909–921.

Zhao H, Chang J, Havlík P, van Dijk M, Valin H, Janssens C, Ma L, Bai Z, Herrero M, Smith P, Obersteiner M. 2021. China's future food demand and its implications for trade and environment. *Nature Sustainability*, **4**, 1042–1051.

Executive Editor-in-Chief Keke Yi
Managing Editor Lujuan Sun