

1 **ChinaSoyArea10m: a dataset of soybean planting areas**
2 **with a spatial resolution of 10 m across China from 2017**
3 **to 2021**

4 Qinghang Mei^{1,2,3}, Zhao Zhang^{1,2}, Jichong Han^{1,2,4}, Jie Song^{1,2,4}, Jinwei Dong^{5,6},
5 Huaqing Wu^{1,2,3}, Jialu Xu^{1,2}, Fulu Tao^{5,6}

6 ¹ Joint International Research Laboratory of Catastrophe Simulation and Systemic Risk Governance,
7 Beijing Normal University, Zhuhai 519087, China

8 ² School of National Safety and Emergency Management, Beijing Normal University, Beijing 100875 /
9 Zhuhai 519087, China

10 ³ Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

11 ⁴ School of Systems Science, Beijing Normal University, Beijing 100875, China

12 ⁵ Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural
13 Resources Research, Chinese Academy of Sciences, Beijing, 100101, China

14 ⁶ College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049,
15 China

16 *Correspondence to:* Zhao Zhang (zhangzhao@bnu.edu.cn)

17

18 **Abstract**

19 Soybean, an essential food crop, has witnessed a steady rise in demand in recent years. There is a lack of
20 high-resolution annual maps depicting soybean planting areas in China, despite China being the world's
21 largest consumer and fourth largest producer of soybeans. To address this gap, we developed a novel
22 Regional Adaptation Spectra-Phenology Integration method (RASP) based on Sentinel-2 remote sensing
23 images from the Google Earth Engine (GEE) platform. We utilized various auxiliary data (e.g., cropland
24 layer, detailed phenology observations) to select the specific spectra and indices that differentiate
25 soybeans most effectively from other crops across various regions. These features were then input for an
26 unsupervised classifier (K-means), and the most likely type was determined by a cluster assignment
27 method based on dynamic time warping (DTW). For the first time, we generated a dataset of soybean
28 planting areas across China, with a high spatial resolution of 10 meters, spanning from 2017 to 2021
29 (ChinaSoyArea10m). The R^2 values between the mapping results and the census data at both county- and
30 prefecture-level were consistently around 0.85 in 2017-2020. Moreover, the overall accuracy of mapping
31 results at the field level in 2017, 2018, and 2019 were 77.08%, 85.16% and 86.77%, respectively.
32 Consistency with census data was improved at the county level (R^2 increased from 0.53 to 0.84),
33 compared to the existing 10-m crop-type maps in Northeast China (Crop Data Layer, CDL) based on
34 field samples and supervised classification methods. ChinaSoyArea10m is spatially consistent well with
35 the two existing datasets (CDL and GLAD maize-soybean map). ChinaSoyArea10m provides important
36 information for sustainable soybean production and management, as well as agricultural system modeling
37 and optimization. ChinaSoyArea10m can be downloaded from an open-data repository (DOI:
38 <https://zenodo.org/doi/10.5281/zenodo.10071426>, Mei et al., 2023).

39 **1 Introduction**

40 Soybean, one of the most important crops around the world, plays an important role in diet and livestock
41 breeding (Hartman et al., 2011). As the global demand for protein and meat increases, China's demand
42 for soybeans has been keeping rising nowadays. In the past decade, China has averagely accounted for
43 over 30% of the world's total soybean consumption (Liu and Fan, 2021). Despite being the fourth-largest

44 producer of soybeans after Brazil, the United States, and Argentina, China's self-sufficiency rate is low
45 (FAOSTAT, 2023; Wang et al., 2023). Given the rapid growth of demand and the shortages of domestic
46 supply due to lower yield and self-sufficiency, mapping soybean planting areas across China is crucial
47 for sustainable soybean production and management (Cui and Shoemaker, 2018; Liu et al., 2021).

48 Soybean planting area in some regions of China was mapped in previous studies (You et al., 2021;
49 Huang et al., 2022; Chen et al., 2023), but long-term soybean maps over all major producing areas in
50 China have not been available. A decision tree method based on phenological and near-infrared
51 reflectance differences was applied in the state of Parana in Brazil to produce corn-soybean maps with a
52 resolution of 500 m (Zhong et al., 2016). However, this study was limited to one state and a simple
53 planting pattern (including soybeans and corn only) at a medium resolution. The field size in China is
54 generally small, and 500 m-resolution maps will inevitably bring pixel mixing problem (Lowder et al.,
55 2016). More recently, 20-year soybean-corn maps with 30 m resolution across the US Midwest have been
56 generated by collecting a large number of samples and using green chlorophyll vegetation index (GCVI)
57 time series features, which is a large-scale, high-precision soybean mapping attempt (Wang et al., 2020).
58 Similarly, high-precision soybean maps in China were also made by collecting major crop samples and
59 utilizing spectral reflectance and vegetation indexes characteristics, for 2017-2019 in Northeast China
60 (You et al., 2021). Some studies have utilized unique canopy water content and chlorophyll content to
61 produce soybean maps in the three provinces of Northeast China from 2017 to 2021 (Huang et al., 2022).
62 Other studies made laudable efforts to craft a comprehensive national maize-soybean map for China in
63 2019 by combining field data and regression estimators (Li et al., 2023). However, these studies were
64 confined in some degrees because of the specific region or a single year, despite prior attempts to
65 accurately map soybean cultivation areas. Long-term annual soybean maps over mainly planting areas
66 in China with a higher spatial resolution have not been available so far.

67 Mapping crops by remote sensing can be categorized into four methods : 1) supervision classification
68 based on a large number of field samples or high quality training labels (Song et al., 2017; You et al.,
69 2021; Shangguan et al., 2022; Li et al., 2023); 2) developing some composite indexes based on the feature
70 bands and determining the binary classification using appropriate thresholds (Huang et al., 2022; Chen
71 et al., 2023; Zhou et al., 2023); 3) threshold segmentation based on prior knowledge such as phenology
72 or spectra (Zhong et al., 2016); 4) combining unsupervised classification with cluster assignment (Wang

73 et al., 2019; You et al., 2023). Supervision classification methods relied on ground samples heavily, while
74 the 2nd and 3rd methods are both based on reliable and accurate thresholds. However, mapping soybean
75 by these methods was mainly applied in small areas, very few covering over a larger region. Because of
76 sufficient field samples, supervision classification can achieve maps with a higher accuracy, which is
77 relatively mature method used widely. However, collecting sufficient field samples is extremely time,
78 money, and labor consumed, and unsuitable for long-term years over larger areas (Luo et al., 2022).
79 Furthermore, the threshold-based methods (the 2nd and 3rd) have been applied into large areas, however,
80 determining the thresholds will inevitably bring significant uncertainty, especially for the areas with high
81 heterogeneity in climate, environment, and planting patterns. Thus, these methods show low
82 reproducibility, further hindering their application across diverse geographic areas. As for mapping
83 soybean, it is still a big challenge due to their similar growth characteristics with many other summer
84 crops (Wang et al., 2020; Di Tommaso et al., 2021). The thresholds that work well in some areas did not
85 perform well in other areas (Graesser and Ramankutty, 2017; Guo et al., 2018). These limitations restrict
86 accurate soybean maps available, especially over large regions in China. Given the challenges of
87 collecting sufficient field samples over larger region and the limited adaptability to environmental
88 variations of threshold-based method, previous researches have yet to achieve multi-year, high-resolution
89 soybean maps nationwide.

90 Along this line, the adaptive classification approach tailored to distinct areas, i.e., method (4), is a
91 highly effective for accurately mapping crops over a larger region. Such unsupervised classification can
92 effectively address the above issues such as insufficient samples and limited spatial scalability by training
93 classifiers separately in different areas (Ma et al., 2020; Wang et al., 2022). Remarkable successes have
94 been achieved when applying the approach into the United States in mapping soybean and maize (Wang
95 et al., 2019). Due to the different climatic and environmental conditions, together with huge differences
96 in cultivating patterns over various areas, crop phenological information has become an important
97 reference for crop classification. For example, the phenological observations at the agricultural
98 meteorological stations were employed as a reference to detect the critical phenological dates of pixels
99 through inflexion- and threshold-based methods, thereby generating planting areas for three major crops
100 in China with R^2 greater than 0.8 compared to county statistics (Luo et al., 2020). The time-weighted
101 dynamic time warping method based on the similarity of phenological curves of Normalized Difference

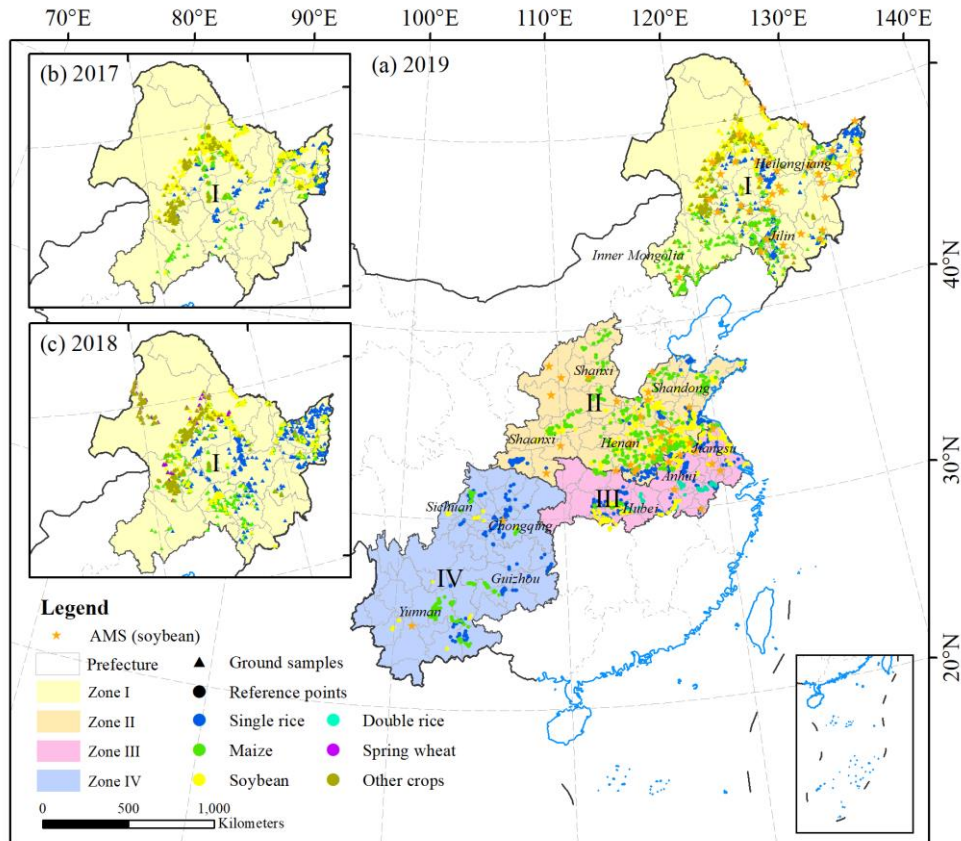
102 Vegetation Index (NDVI) has successfully estimated the planting area of maize in China, with provincial
103 averages for producer's and user's accuracies at 0.76 and 0.82, respectively (Shen et al., 2022).
104 Phenological-based Vertical transmit Horizontal receive (VH) polarized time series accurately captured
105 temporal characteristics of soybeans, thus were used for an unsupervised classifier to map the seasonal
106 soybeans, achieving an overall accuracy over 80% in Ujjain district (Kumari et al., 2019). By integrating
107 unsupervised classification's regional scalability with specific local soybean growth signs from
108 phenological data, we fully leverage soybean's characteristic spectra and vegetation indices during key
109 growth periods across different areas. Through training the local unsupervised classifier to accommodate
110 the crop growth variability across regions, and avoiding extensive jobs on collecting samples, the
111 approach provides an effective solution for regional adaptive large-area crop mapping.

112 The main objectives of this study are: 1) to develop a novel framework to map soybean planting area
113 over a larger region; 2) to test the generalization ability of the framework and assess the accuracy of maps
114 at different levels; and 3) to provide a new data product of soybean planting area across mainly planting
115 areas in China, for multi-years with a high spatial resolution.

116 **2 Materials and methods**

117 **2.1 Study area**

118 We selected 14 major soybean producing provinces (including Chongqing Municipality) as study area,
119 which cover over 90% of the total planting area in China (National Bureau of Statistics of China, 2023)
120 (Fig. 1). The soybean planting areas were classified into four agro-ecological zones (AEZs) based on
121 their diverse geographical environment and planting habits, including Northeast single cropping eco-
122 region (NE, Zone I), Huang-Huai-Hai double cropping eco-region (HH, Zone II), Middle-Lower Yangtze
123 River double cropping eco-region (MLY, Zone III) and Southwest double cropping eco-region (SW, Zone
124 IV) (Wang and Gai, 2002). In particular, Zone I and Zone II are the main soybean producer in China,
125 accounting for more than 70% of the national soybean planting area.



126
 127 **Figure 1.** The study area including 14 provinces (including Chongqing Municipality) and spatial distribution
 128 of ground samples and reference points across China in (a) 2019, (b) 2017, and (c) 2018. The 14 provinces
 129 include Heilongjiang, eastern Inner Mongolia, Anhui, Henan, eastern Sichuan, Jilin, Hubei, Guizhou, Jiangsu,
 130 Yunnan, Shandong, Shaanxi, Shanxi, and Chongqing. Stars, triangles, and dots represent the locations of
 131 soybean agricultural meteorological stations (AMSs), ground samples, and reference points, respectively.

132 2.2 Data

133 2.2.1 Remote sensing data

134 We used Sentinel-2A/B Multi-Spectral Instrument (MSI) Level-1C top-of-atmosphere (TOA) reflectance
 135 data during 2017-2021 (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2,
 136 last access: September 2023). Because of the longer-term coverage of Sentinel-2 Level-1C TOA
 137 reflectance data, and the nearly identical spectral profile time series extracted from both products, we opt
 138 to use L1C products instead of L2A, considering that TOA images fully meet the crop classification
 139 requirements (You and Dong, 2020; Han et al., 2021; Luo et al., 2022). Sentinel-2 sensors provide
 140 observations in 13 spectral bands at 10 m or 20 m resolution. The red-edge bands and shortwave infrared

141 bands equipped with sentinel-2 play a great role in enhancing the accuracy of crop classification (Luo et
142 al., 2021; Marshall et al., 2022). In addition, the S2 cloud probability dataset provided by the official can
143 identify cloud pollution areas and be used as cloud removal processing.

144 **2.2.2 In-situ phenological observations**

145 The soybean phenology observations in study area from 2017 to 2020 were obtained from 76 agricultural
146 meteorological stations (AMSs) governed by the CMA (<https://data.cma.cn/>, last access: May 2022).
147 Phenology information of each AMS is observed on alternate days or once a day, and key phenological
148 events such as sowing, emergence, three-true-leaves, branching, flowering, podding, full-seeding, and
149 maturity are noted by technicians to ensure accuracy. We defined the period from sowing to flowering as
150 the vegetative growth period (VGP), and the period from flowering to maturity as the reproductive
151 growth period (RGP) of soybeans (Gong et al., 2021). In cases of missing observation for a specific year,
152 we inserted the average of two closest observations before and after the year. For instance, if there was
153 missing data of flowering date in 2017, we filled it with the average of flowering records in 2016 and
154 2018 at the same station.

155 **2.2.3 Cropland data**

156 GLAD cropland product with a 30-m resolution in China was used as cropland masks
157 (<https://glad.umd.edu/dataset/croplands>, last access: September 2023) (Potapov et al., 2022). The crop
158 layer was conducted every four years from 2000 to 2019. We used the file for the 2016-2019 interval
159 which is closest to the study years. GLAD's overall accuracy of pixel-wise validation is 0.88 in China,
160 consistent well with the census data. The accuracy of the product is higher than that of similar products,
161 making it a reliable for crop mapping (Zhang et al., 2022).

162 **2.2.4 Census data and ground samples**

163 To determine the number of clusters at prefecture-level and validate the accuracy of the soybean maps at
164 county (2017-2018) or prefecture (2019-2020) level, we utilized agricultural census data obtained from
165 the statistical yearbook of each county or province by accessing National Bureau of Statistics of China
166 (<http://www.stats.gov.cn/>, last accessed: June 2023).

167 We used both ground samples and reference points based on available datasets to determine soybean
 168 standard curves and assess the reliability of the soybean maps (Fig. 1). All points were randomly divided
 169 in a 3:7 ratio for standard curve calculation and accuracy validation, respectively (Dong et al., 2020). We
 170 collected ground samples from field surveys from 2017 to 2019 in Heilongjiang (HLJ), Inner Mongolia
 171 (NMG), Anhui (AH), Henan (HN), and Jilin (JL), which account for more than 70% of the country's total
 172 soybean planting area (Table 1). Crop types (soybean, maize, rice, wheat, others) and other land cover
 173 types were recorded. To ensure the impartiality of verification results, we only selected crop samples for
 174 validation. In provinces without ground samples, we manually selected reference points on large soybean
 175 plots based on GLAD (<https://glad.earthengine.app/view/china-crop-map>, last access: March 2024)
 176 soybean layer. The criteria selected are: (1) located in large plots; (2) false color composite image (R:
 177 NIR, G: SWIR2, B: SWIR1) at the peak of growing season (Song et al., 2017; You and Dong, 2020); (3)
 178 phenological characteristics similar to local observations. Additionally, the reference points of maize,
 179 single-cropping rice and double-cropping rice in 2019 were selected based on GLAD maize layer, high
 180 resolution single-season rice map (<https://doi.org/10.57760/sciencedb.06963>, last access: March 2024),
 181 and double-season rice map (<https://doi.org/10.12199/nesdc.ecodb.rs.2022.012>, last access: March 2024)
 182 with the same principle to explore the spectral characteristics of crops in each sub-zone of the studied
 183 areas. The overall accuracy of all available maps in 2019 is above 85% (Pan et al., 2021; Li et al., 2023;
 184 Shen et al., 2023).

185 **Table 1. Summary of ground samples for validation.**

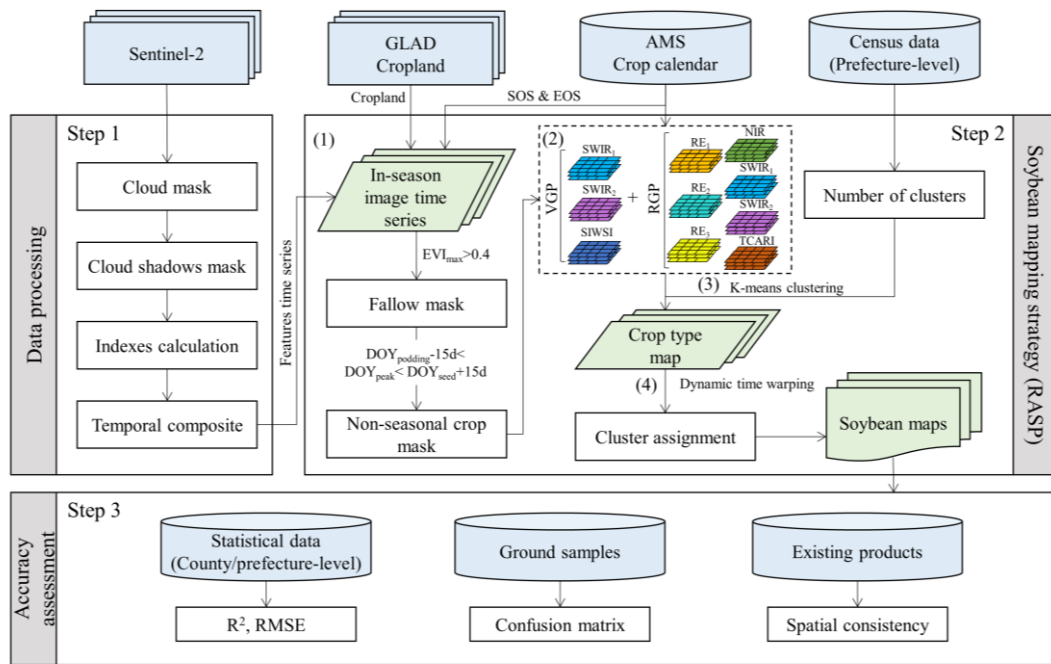
		HLJ	NMG	AH	HN	JL
2017	Soybean	1013	451	-	-	0
	Maize	1061	146	-	-	11
	Rice	513	38	-	-	13
	Other crops	124	459	-	-	0
2018	Soybean	525	746	72	15	117
	Maize	764	479	73	20	217
	Rice	587	42	0	0	71
	Wheat	10	141	0	0	0
	Other crops	70	1069	0	0	0
2019	Soybean	901	562	51	-	26
	Maize	468	463	53	-	197
	Rice	392	36	0	-	148
	Other crops	62	445	0	-	36

186 2.2.5 Existing products

187 We utilized the crop map CDL of Northeast China from 2017 to 2019
188 ([https://figshare.com/articles/figure/The_10-m_crop_type_maps_in_Northeast_China_during_2017-](https://figshare.com/articles/figure/The_10-m_crop_type_maps_in_Northeast_China_during_2017-2019/13090442)
189 [2019/13090442](https://figshare.com/articles/figure/The_10-m_crop_type_maps_in_Northeast_China_during_2017-2019/13090442), last access: September 2023) for consistency comparison with census data, and the
190 2019 GLAD maize-soybean map as a reference for spatial detail comparison with ChinaSoyArea10m.
191 CDL is a 10m resolution crop map dataset of Northeast China from 2017 to 2019 that was created
192 using Sentinel-2 key spectral bands and vegetation indices, multi-year field samples, and random forest
193 classifiers (You et al., 2021). The maps include three crop types: rice, maize, and soybeans. The GLAD
194 maize-soybean Map is a national classification map for 2019 that was produced using random forests,
195 based on field surveys and area estimates (Li et al., 2023). The agreement (R^2) between GLAD and the
196 statistics is higher than 0.9, and the overall mapping accuracy is greater than 90%, making it a reliable
197 reference for comparing spatial details. We extracted the soybean layers from all the existing products.

198 2.3 Methods

199 Mapping soybean consists of three main steps (Fig.2): data processing, soybean mapping, and accuracy
200 assessment. It is important to note that the Regional Adaption Spectra-Phenology Integration (RASP)
201 soybean mapping strategy involves several key steps, including potential area identification, feature
202 selection, unsupervised learning, and cluster assignment. Finally, we conducted multi-comparisons
203 between our soybean products with others, including census data, ground samples, and existing datasets,
204 to evaluate the accuracy of our data product.



205

206

Figure 2. The Regional Adaption Spectra-Phenology Integration methodology for retrieving soybean planting

207

area. AMS, agricultural meteorological station; $DOY_{podding}$, the podding date recorded by the nearest AMS;

208

EVI: Enhanced Vegetation Index; DOY_{peak} , the date when EVI reached peak; DOY_{seed} , the full-seed date

209

recorded by the nearest AMS; SOS, start of growing season; EOS, end of growing season; SWIR₁, Short Wave

210

Infrared band 1; SWIR₂, Short Wave Infrared band 2; SIWSI, shortwave Infrared Water Stress Index; RE₁,

211

Red Edge band 1; RE₂, Red Edge band 2; RE₃, Red Edge band 3; NIR, Near-infrared band; TCARI,

212

Transformed Chlorophyll Absorption in Reflectance Index; VGP: vegetative growing period; RGP:

213

reproductive growing season.

214 2.3.1 Data processing

215

We employed the simple cloud score algorithm (Oreopoulos et al., 2011), QA60 band, cirrus band, and

216

cloud probability dataset to identify cloud masks. The following isolated cloud masks are created: (1)

217

Cloud and cirrus identified by QA60 band; (2) Cirrus identified by cirrus band in Level-1C products; (3)

218

Pixels with cloud score less than 0.9; and (4) Pixels with cloud probability more than 70. Each algorithm

219

has its own strengths and limitations. For example, QA60 band removes a large number of thin cirrus

220

clouds while ignoring small clouds with thicker resolution, and the fixed threshold values of cloud score

221

and cloud probability may introduce uncertainties. Therefore, we masked the pixels identified as clouds

222

by at least two methods to achieve better cloud removal effects. Then, we used Temporal Dark Outlier

223 Mask (TDOM) method to eliminate cloud shadows (Housman et al., 2018). We calculated the SIWSI
 224 and TCARI indices based on the Sentinel-2 image set processed above (see 2.3.2(2)). To fill the data gaps
 225 caused by cloud removal and smooth anomalies, Sentinel-2 time series was reconstructed by moving
 226 median composite method, resulting in a 10-day interval composite time series. We set the half-window
 227 size for the moving median methods to 10 days considering the 5-day revisit cycle of Sentinel-2 and
 228 computational efficiency. In areas with notably limited clear observations, a gap-filling method was
 229 conducted on the composite time series. This method involves substituting any given observation with
 230 the median value from three neighboring observations (i.e., previous, current, and subsequent
 231 observations) to maximize the continuity and completeness of time series.

232 **2.3.2 Regional Adaptation Spectra-Phenology Integration (RASP) soybean mapping strategy**

233 (1) Potential area identification

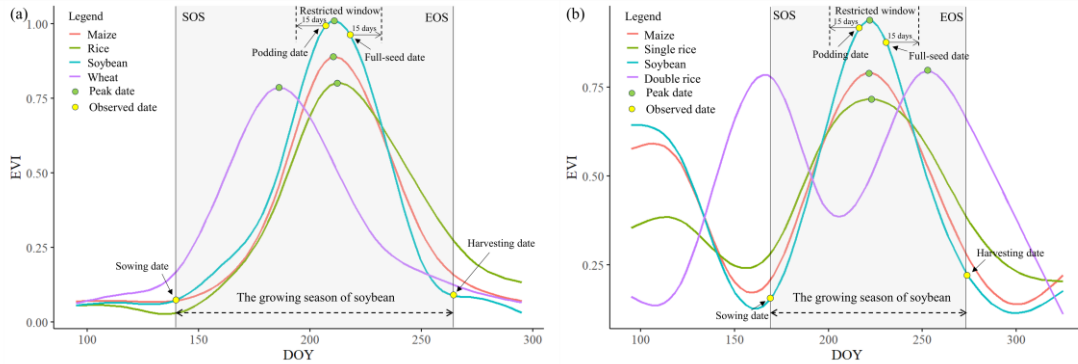
234 To minimize the impact from non-croplands, we firstly determine the potential cropping areas by
 235 masking GLAD cropland layer over study area. Sentinel-2 images within growing season were extracted
 236 by taking the sowing date and harvesting date recorded at the nearest agricultural meteorological station
 237 (AMS) as the starting and ending dates of the growing season, respectively. Based on the cropland
 238 extracted, we filtered out the pixels exhibiting an Enhanced Vegetation Index (EVI) maximum value
 239 during the growing season less than 0.4 to remove fallow land according to the analysis of ground
 240 samples (Fig. S1) and previous studies, which found that almost all crops had maximum EVI values
 241 above 0.4 (Li et al., 2014; Zhang et al., 2017; Han et al., 2022). EVI is a vegetation index with high
 242 sensitivity in biomass:

$$243 \quad EVI = G \times \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + C_1 \times \rho_{Red} - C_2 \times \rho_{Blue} + L} \quad (1)$$

244 Where ρ_{NIR} , ρ_{Red} , and ρ_{Blue} represented the reflectance of the Near-infrared (835.1nm (S2A) / 833nm
 245 (S2B)), Red (664.5nm (S2A) / 665nm (S2B)), Blue (496.6nm (S2A) / 492.1nm (S2B)), respectively.

246 The greenest period of soybean typically occurs between the podding date and the full-seed date, with a
 247 difference of more than a month from the peak date of non-seasonal crops, such as wheat (Fig. 4a). We
 248 obtained the phenological observations recorded by the nearest AMS as reference and set the restricted
 249 time window from 15 days before the podding date ($DOY_{podding}$) to 15 days after the full-seed date
 (DOY_{seed}) (Fig. 3). We generated the potential area by eliminating pixels whose EVI maximum occurs

250 outside the given time window because the phenological difference of soybeans in adjacent areas
 251 generally does not exceed one month. Moreover, the impacts of cloud-covered pixels appearing in the
 252 proposed period is minimized since we have reconstructed the original EVI time series.



253
 254 **Figure 3. Schematic diagram of seasonal crop identification for (a) single - and (b) double - cropping systems.**

255 (2) Feature selection

256 By exploring the spectral characteristics of crop field samples, we identified reflectance bands and
 257 vegetation indices that are significantly associated with soybeans but different from other crops. We
 258 selected six bands and two spectral indices for crop mapping, including Near-infrared (NIR) band, Red
 259 edge band 1 (RE1), Red edge band 2 (RE2), Red edge band 3 (RE3), Short Wave Infrared band 1
 260 (SWIR1), Short Wave Infrared band 2 (SWIR2), Shortwave Infrared Water Stress Index (SIWSI),
 261 Transformed Chlorophyll Absorption in Reflectance Index (TCARI). SIWSI is an indicator of canopy
 262 water content that reflects soil moisture variations and canopy water stress better than Normalized
 263 Difference Vegetation Index (NDVI) (Fensholt and Sandholt, 2003; Olsen et al., 2015). TCARI is an
 264 indicator which is sensitive to chlorophyll concentration (Sobejano-Paz et al., 2020). The two spectral
 265 indices were calculated as follows:

$$SIWSI = \frac{\rho_{SWIR1} - \rho_{NIR}}{\rho_{SWIR1} + \rho_{NIR}} \quad (2)$$

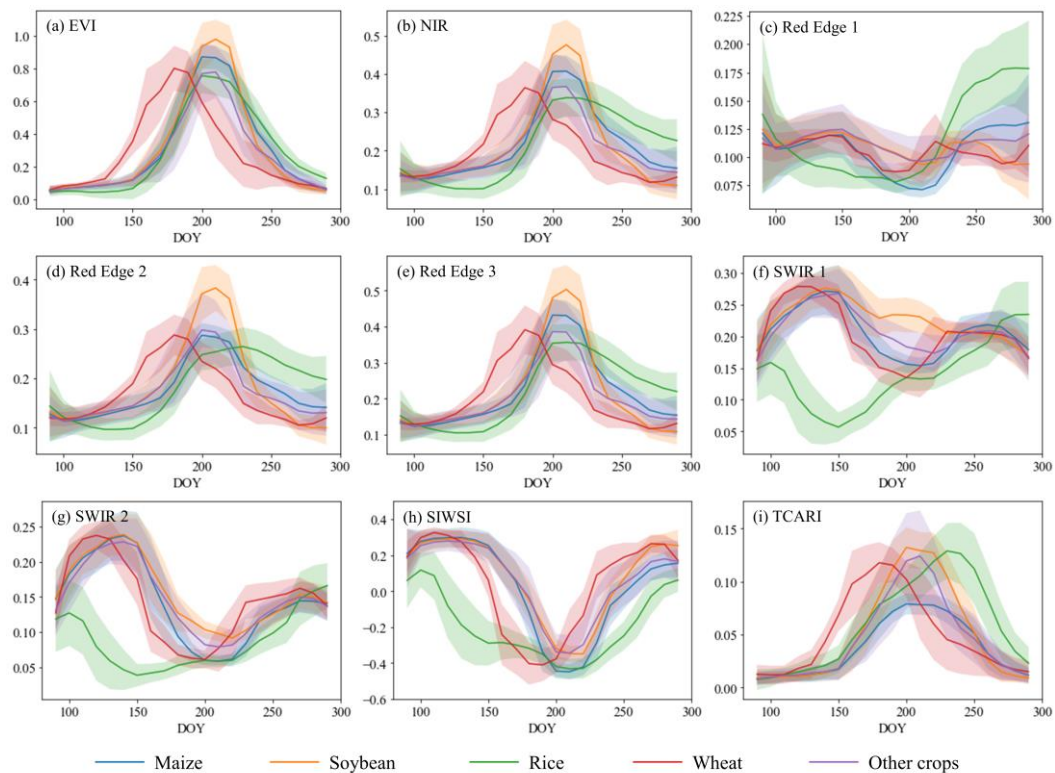
$$TCARI = 3 \times ((\rho_{VRE1} - \rho_{Red}) - 0.2 \times (\rho_{VRE1} - \rho_{Green}) \times \rho_{VRE1} / \rho_{Red}) \quad (3)$$

266 Where ρ_{SWIR1} , ρ_{NIR} , ρ_{VRE1} , ρ_{Red} and ρ_{Green} represented the reflectance of the Short Wave Infrared
 267 band1 (SWIR1, 1613.7nm (S2A) / 1610.4nm (S2B)), Near-infrared (835.1nm (S2A) / 833nm (S2B)),
 268 Red Edge1 (VRE1, 703.9nm (S2A) / 703.8nm (S2B)), Red (664.5nm (S2A) / 665nm (S2B)), Green
 269 (560nm (S2A) / 559nm (S2B)), respectively.

270 During early growing season of soybean (~DOY 120-190 in Zone I), the flooding signal of rice was
 271 obvious due to the transplanting period. This resulted in a significantly lower SWIR reflectance and

272 SIWSI index for rice compared to those of soybean (Fig. 4f-h). SWIR bands and SIWSI index during the
 273 vegetative growing period (VGP) of soybean can effectively distinguish dryland crops (such as soybean,
 274 maize) from paddy crops (such as rice).

275 Soybean has a lower water content during the middle and later growing season (~DOY 190-220 in
 276 Zone I) than maize, resulting in higher reflectivity in SWIR bands (Fig. 4b, 4f, 4g) (Chen et al., 2005). It
 277 has been demonstrated that SWIR and red-edge bands can effectively differentiate soybean and maize
 278 (Fig. 4c-g) (Zhong et al., 2016; You and Dong, 2020; Liu et al., 2018b). Additionally, the chlorophyll
 279 content of soybean in the middle and late growth period was lower than that of maize, leading to
 280 significantly higher TCARI values. Meanwhile, the timing of TCARI reaching saturation significantly
 281 differs among soybean, rice, and wheat (Fig. 4i). All these spectral-phenological characteristics are also
 282 applicable to soybeans planted in other sub-zones (Fig. S2-S4). Based on these findings, we selected NIR,
 283 red-edge bands, short-wave infrared bands, and TCARI index during soybean reproductive growing
 284 season (RGP) as key features.



285

286 **Figure 4. Temporal profiles of (a-i) for major crops in Northeast China and (j) key soybean phenological**
287 **periods by region based on ground samples. Lines depict the mean values of different crops and shaded areas**
288 **depict error bars with one positive/negative standard deviation. The number at the bottom represents the key**
289 **phenological periods of soybean: 1 – Sowing, 2 – Flowering, 3 – Seed fulling, 4 – Maturity.**

290 (3) Unsupervised learning

291 We utilized K-means algorithm to classify potential area data by using the wekaKMeans Clusterer
292 provided by Google Earth Engine (GEE). The m samples are divided into k clusters by alternately
293 assigning the samples to the nearest cluster centroid measured by Euclidean distance or the Manhattan
294 distance and updating the cluster centroid to the mean of the samples assigned to the cluster. This
295 approach had been widely used in land-cover classification and crop mapping (Xiong et al., 2017; Wang
296 et al., 2019). We used the detailed phenological records at AMSs to identify soybean growth periods and
297 selected the spectra and vegetation indices within specific growth periods (VGP, RGP) as input features.
298 The classifier was trained individually on each prefecture based on the number of clusters k input. The
299 cluster number k is defined as the number of “major crops” that constituting 95% of the total area for
300 seasonal crops (including rice, maize, soybean, cotton, peanuts, sesame, sweet potato, and sorghum)
301 according to prefecture-level statistics, and plus one for “other crops”.

302 (4) Cluster assignment

303 To identify the most likely cluster that represents soybean, we randomly selected 100 points per cluster
304 and extracted feature series. We then used dynamic time warping (DTW) method to measure the
305 similarity between each cluster’s eight features involved in classification and the soybean standard curves.
306 We averaged the data of 30% samples in each sub-zone to establish the standard curves, reducing the
307 impact of regional phenological variations. The time coverage of Zone I-IV was set to April-September,
308 May-October, June-October, and August-November, respectively, which are corresponding with the
309 soybean growing season. The cluster with the minimal average of 8 DTW values was identified as the
310 soybean cluster. DTW is a flexible algorithm that allows for deviations in time between two sequences,
311 and it calculates the minimum distance between them by finding misalignment matches between
312 elements. This approach is widely used in land cover and crop identification due to its ability to handle
313 time distortions associated with seasonal changes (Guan et al., 2016; Dong et al., 2020).

314 2.3.3 Accuracy assessment

315 To assess the accuracy of the soybean maps we generated, we validated and compared the results using
316 1) county- and prefecture-level census data, 2) ground samples, and 3) existing products. Since the
317 county-level statistics after 2019 were not fully collected, we used the county-level statistics for 2017-
318 2018 and the prefecture-level statistics for 2019-2020 to calculate the R^2 and RMSE of the mapped area
319 with the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^n (s_i - y_i)^2}{\sum_{i=1}^n (s_i - \bar{s})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (s_i - y_i)^2}{n}} \quad (5)$$

320 where s_i and y_i are the statistical and mapped soybean area for county (prefecture) i , \bar{s} is the average
321 statistical area, and n represents the total number of counties (prefectures). We calculated the local crop
322 mapping area based on the Universal Transverse Mercator (UTM) projection corresponding to the
323 location of the province.

324 We also used ground samples during 2017-2019 to verify the authenticity of the soybean maps.
325 Confusion matrices were calculated as follows:

$$PA = \frac{N_i}{R_i} \quad (6)$$

$$UA = \frac{N_i}{C_i} \quad (7)$$

$$OA = \frac{N_c}{A} \quad (8)$$

$$F1 = 2 \times \frac{UA \times PA}{UA + PA} \quad (9)$$

326 where N_i is the number of correctly identified validation samples of class i , R_i is the number of
327 ground validation samples of class i , C_i is the number of validation samples classified as class i , C_i
328 is the number of validation samples classified as class i , N_c is the total number of correctly identified
329 validation samples, A is the total number validation samples. PA , UA , and OA represent producer's
330 accuracy, user's accuracy, and overall accuracy, respectively.

331 To ensure that the products are accurate not only in quantity but also in space, we further compared
332 the ChinaSoyArea10m with existing products in detail space.

333 **3 Results**

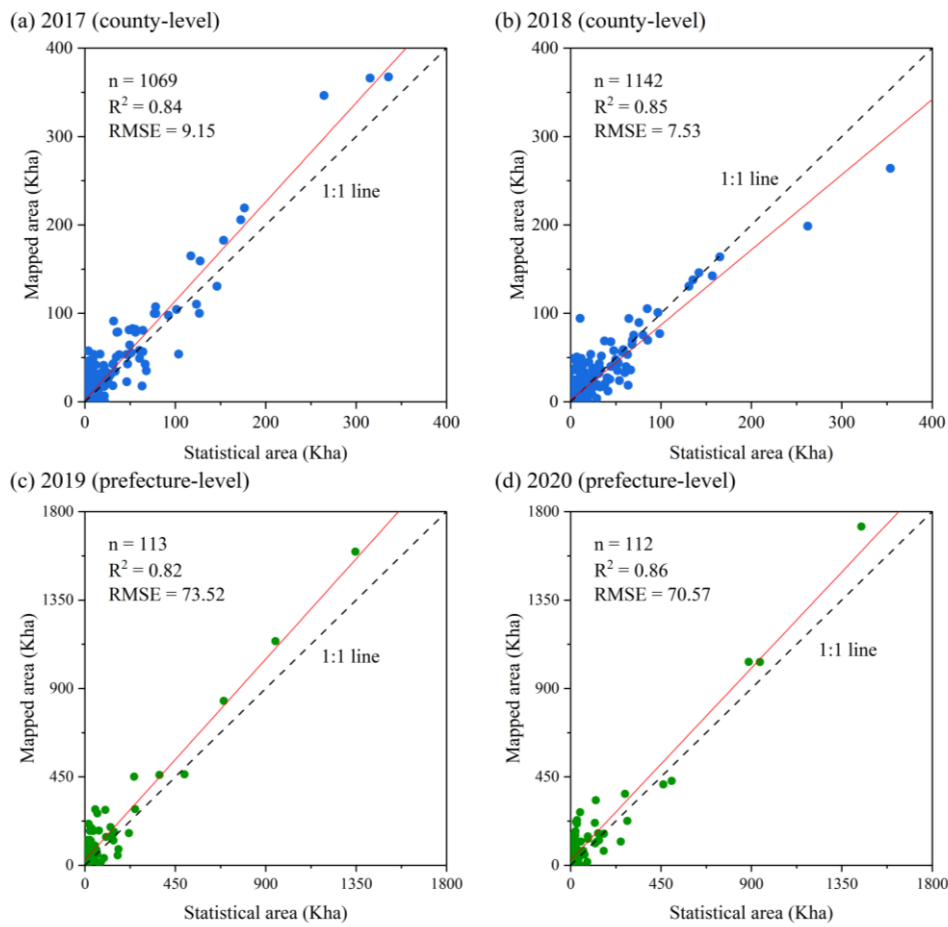
334 **3.1 Accuracy assessment**

335 We utilized the available census data from 2017-2020 (at county-level in 2017-2018 and prefecture-level
336 in 2019-2020) to verify the accuracy of the soybean maps across the entire studied area. Annual
337 ChinaSoyArea10m is consistent well with the census data ($R^2 > 0.8$), with an R^2 value of 0.84, 0.85, 0.82,
338 and 0.86 for 2017, 2018, 2019, and 2020, respectively (Fig. 5). These results demonstrate that our RASP
339 method is inter-annual robustness and can accurately capture annual dynamics of soybean planting areas.
340 The scattered points are generally distributed around 1:1 line, without large overestimations or
341 underestimations. However, the areas are overestimated for counties with planting area < 20 kha, or
342 prefectures with planting area < 100 kha (Fig. 5). This uncertainty, particularly overestimation, could be
343 caused by the low proportion of soybean cultivation. If maize or other same-season crops are planted in
344 a much higher proportion than soybeans there, distinctly recognizing soybeans (as a less prevalent crop)
345 as a separate category will be a big challenge for classifiers, consequently resulting in misclassified
346 clusters including maize or other crops.

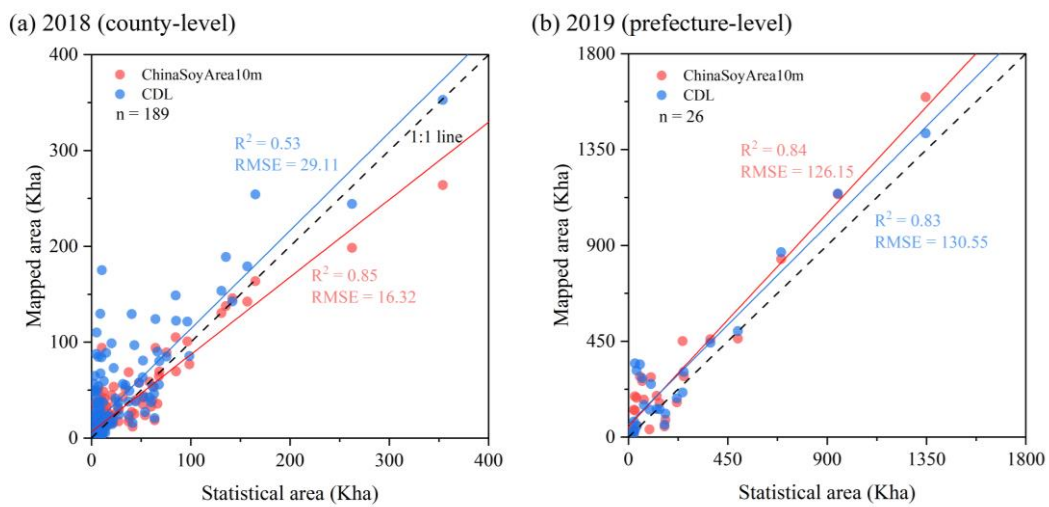
347 The mapping accuracy in Zone I closely matched county-level statistics, showing high consistency
348 ($R^2=0.86$). Zones II-IV also demonstrated reasonable agreement ($R^2=0.50\sim 0.69$), despite relatively lower
349 accuracy due to the scarcer planted areas (Fig. S5). No significant trend deviation from statistics was
350 indicated for the mapping area in Zone I, with slight overestimates for Zone II and III, and underestimates
351 for Zone IV (Fig. S5). These accuracy variations are acceptable, given the challenges in accurately
352 identifying soybeans in regions where they are planted less prevalently. Specifically, maize is more
353 dominant than soybeans in Zone II, while Zone III is characterized by diverse crops and complex planting
354 patterns. Underestimation in Zone IV is possibly due to fewer clear observations in the southwest.
355 Nevertheless, the overall accuracy across the zones is acceptable.

356 ChinaSoyArea10m is consistent well with census data compared to the existing product (CDL) (You et
357 al., 2021), using both the county level in 2018 and prefecture level in 2019 (Fig. 6). CDL's results are
358 consistent with census data at the prefecture scale, with more overestimations at the county level (Fig.
359 6), implying the comparison at finer scale would reveal more details. ChinaSoyArea10m is consistent

360 with statistics at the both levels ($R^2 \sim 0.85$), with R^2 increases 0.31 compared with CDL in county level
 361 (Fig. 6a).



362
 363 **Figure 5. Comparison of soybean areas with statistics in (a) 2017 at county-level, (b) 2018 at county-level, (c)**
 364 **2019 at prefecture-level, (d) 2020 at prefecture-level.**



365
 366 **Figure 6. Comparison of soybean areas of ChinaSoyArea10m and CDL with statistics in (a) 2018 at county-**
 367 **level, (b) 2019 at prefecture-level.**

368 Furthermore, we used ground samples in 2017-2019 to validate the reliability of the soybean maps.
 369 Since the soybean planting area maps are 0-1 binary images, we categorized the ground samples into
 370 soybean and non-soybean (maize, rice, wheat, and other crops). The verification results based on ground
 371 samples indicated that the overall accuracy of soybean maps during 2017-2019 was in the range of 77.08%
 372 to 86.77%. The F1 scores of soybeans increased from 2017 to 2019 (0.69, 0.75 and 0.84, respectively)
 373 (Table 2). The variance in accuracy among years could be attributed to the quality of Sentinel-2 images,
 374 which had been indicated in previous studies (Liu et al., 2020; Han et al., 2021). The overall accuracy
 375 for each sub-zone in 2019 varied from 83.58% to 90.67% (Table S1). Specifically, Zone I demonstrated
 376 the highest producer's accuracy for soybean at 88.31%, aligning with its high consistency with statistics.
 377 Zone III achieved the highest overall accuracy at 90.67%, attributed to its superior user's accuracy for
 378 soybean, indicating fewer misclassifications, and effective differentiation from non-soybean crops (Table
 379 S1). The producer's accuracy in Zone IV was relatively lower at 63.89%, possibly due to the limited
 380 samples, high heterogeneity, and fewer clear observations (Table S1).

381 **Table 2. Confusion matrix of the soybean maps during 2017-2019.**

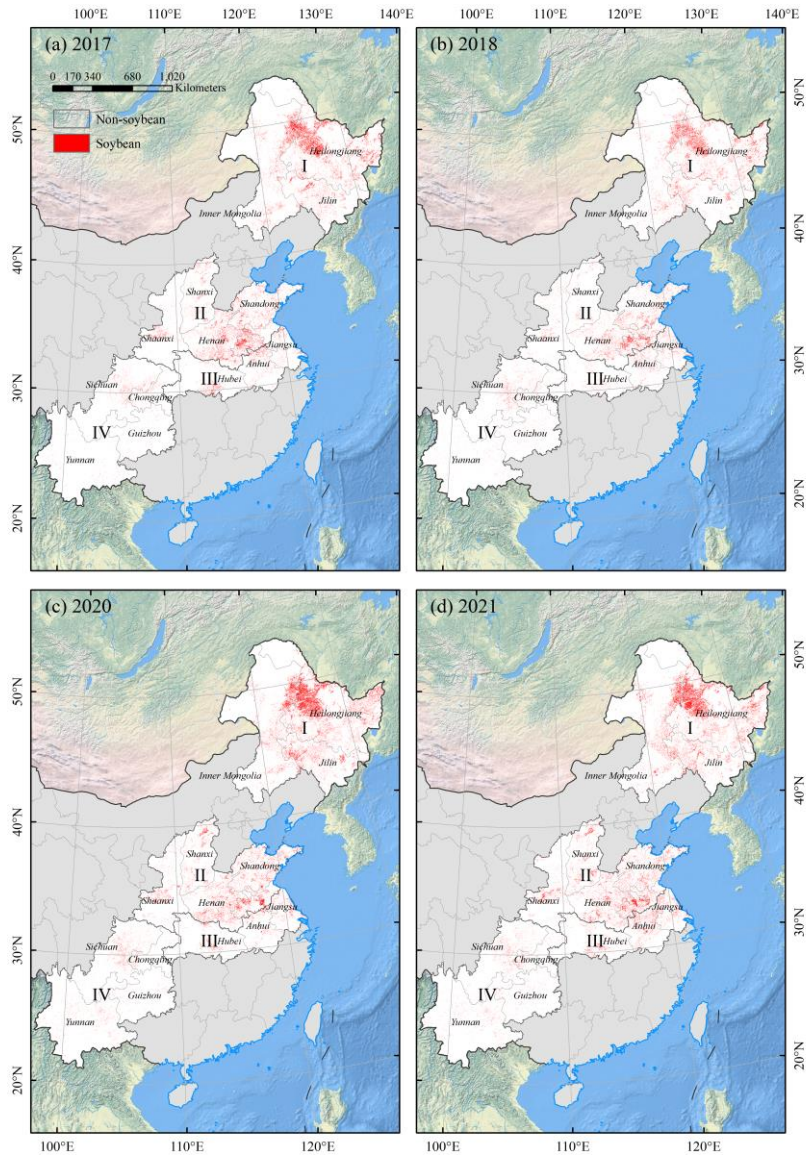
	Reference	Map		Producer's Accuracy	User's Accuracy	F1 Score	Overall Accuracy
		Soybean	Non-Soybean				
2017	Soybean	679	352	65.86%	72.47%	0.69	77.08%
	Non-Soybean	258	1372	84.17%	79.58%	0.82	
2018	Soybean	799	246	76.46%	74.19%	0.75	85.16%
	Non-Soybean	278	2208	88.82%	89.98%	0.89	
2019*	Soybean	1279	235	84.48%	83.32%	0.84	86.77%
	Non-Soybean	256	1940	88.34%	89.20%	0.89	

382 * Including ground samples and nationwide reference points based on existing datasets.

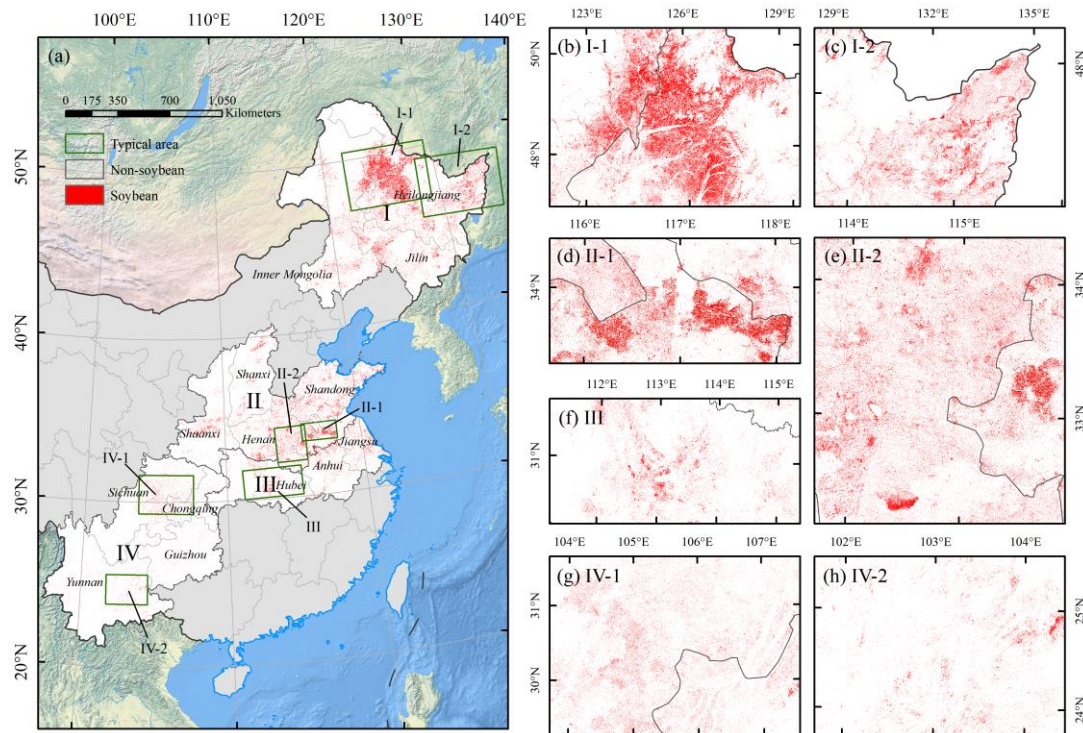
383 3.2 Spatial distributions of soybean planting areas

384 Based on the soybean maps, we further analyzed the spatial patterns of soybean distribution in China
 385 during 2017-2021. There were small changes in the spatial distribution of soybean in China in recent
 386 years (Fig. 7-8). Several hot spots were obviously observed in Heilongjiang Province, eastern Inner
 387 Mongolia, and northern Anhui, especially for eastern Inner Mongolia and western Heilongjiang,
 388 extensively and densely distributed by soybean fields (Fig. 8b-c). In Region II, soybean was planted at a
 389 larger scale, mainly concentrated in northern Anhui (Fig. 8d), and extensively distributed in Henan and

390 Shandong (Fig. 8e). Soybeans in other provinces of Region II, III, and IV were scattered distribution,
391 especially in the southwestern mountainous region (Fig. 8f-h).



392
393 **Figure 7. Spatial distribution of soybean areas at 10 m resolution across China in (a) 2017, (b) 2018, (c) 2020**
394 **and (d) 2021.**

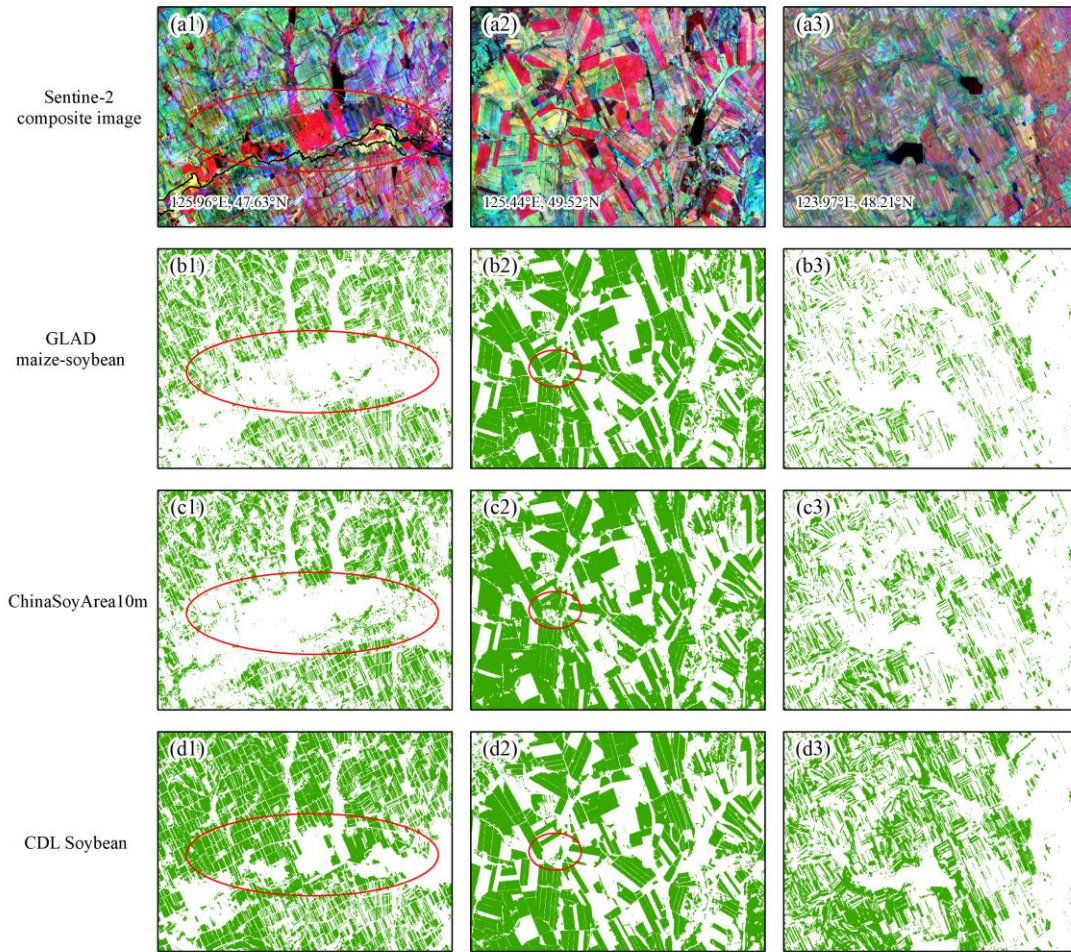


395

396 **Figure 8. Spatial distribution of soybean areas at 10 m resolution across China (a) and zoom-in maps of each**
 397 **region (b-h) in 2019.**

398

399 To further compare soybean maps in detail, we compared ChinaSoyArea10m with GLAD maize-
 400 soybean map and CDL data products in space. The GLAD product is a 10-m resolution maize-soybean
 401 map of China in 2019, and their R^2 values with provincial and prefecture statistics were reported by 0.93
 402 and 0.94 (Li et al., 2023). Arable land near waterbodies is often misclassified as soybean plots by CDL,
 403 which has not occurred by GLAD and ChinaSoyArea10m, implying other crop types are possibly
 404 misclassified as soybeans by CDL (Fig. 9 a1-d1). As for the second case (Fig. 9 a2), our extraction results
 405 are similar to those of GLAD, while small plots failed to be identified by CDL (Fig. 9 a2-d2). In areas
 406 where banded soybeans are planted less concentrated, CDL tended to overestimate the soybean area (Fig.
 407 9 a3-d3), further substantiating the above limitations (Fig. 6). Conversely, our mapping results behaved
 408 similarly as GLAD did (Fig. 9 a3-d3). The overall accuracy of GLAD map based on pure samples reaches
 409 95.4% (Li et al., 2023), so GLAD can be regarded as a reliable reference. From the three cases, therefore,
 410 ChinaSoyArea10m has behaved more similarly with GLAD than CDL does, indicated by less
 411 underestimation, less overestimation, and higher accuracy in details.



412

413 **Figure 9. Visual comparison of our soybean maps and existing products in typical regions in 2019: (a1-a3)**

414 **RGB composite images comprise NIR (Band 8), SWIR 2 (Band 12), and SWIR 1 (Band 11) bands from**

415 **Sentinel-2 median composite images during the peak growth period of soybean; (b1-b3) soybean layer**

416 **extracted from GLAD maize-soybean map; (c1-c3) ChinaSoyArea10m map; (d1-d3) soybean layer extracted**

417 **from CDL.**

418 **4 Discussion**

419 **4.1 Our advantages and potential applicability**

420 We proposed a new framework (RASP) to identify annual dynamic of soybean planting areas over

421 larger regions and produced the longer-term series of soybean maps (ChinaSoyArea10m) across

422 mainly planting areas in China from 2017 to 2021 at the first time. The accuracy of

423 ChinaSoyArea10m is acceptable ($R^2 \sim 0.85$) at both county- and prefecture-level, with relatively less

424 R^2 than GLAD ($R^2 = 0.93$ at prefecture-level), but higher than CDL ($R^2 = 0.53$ at county-level).
425 Compared with existing products, ChinaSoyArea10m accurately depict the soybean with more
426 spatial and temporal details as well.

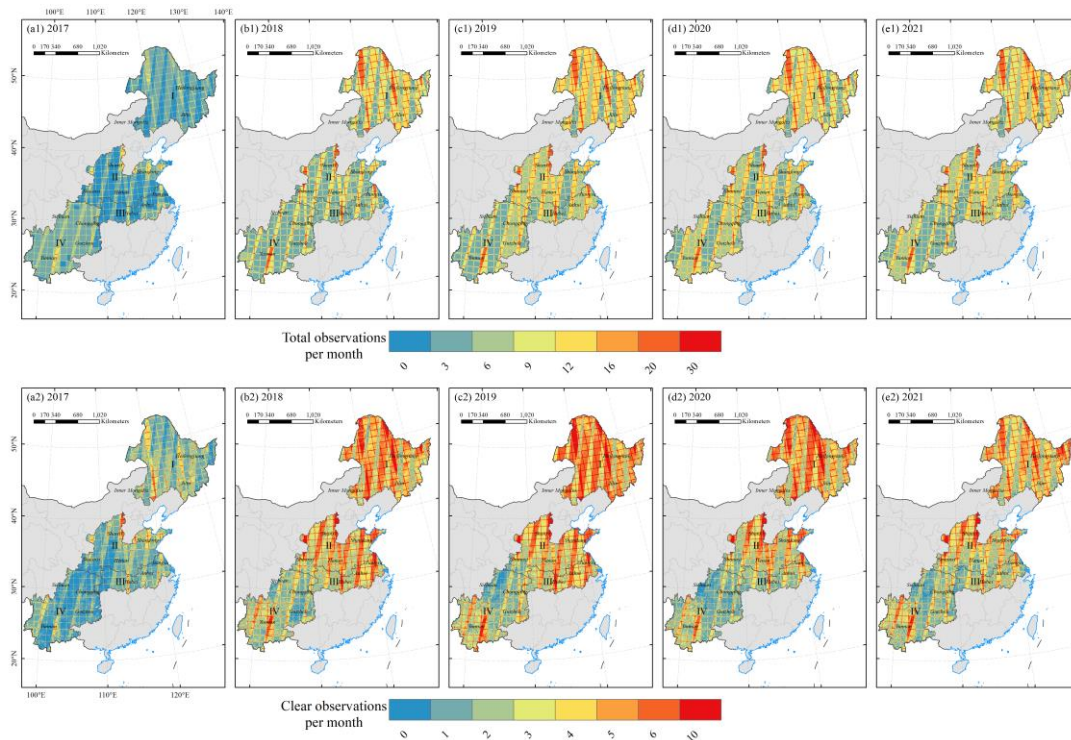
427 The methodology developed for identifying soybean planting areas indicate several notable
428 strengths that make it an attractive option for wide application. Firstly, it operates independently,
429 without extensive ground samples required. The conventional supervised approaches like random
430 forest (RF) and long short-term memory (LSTM) depend on quantities of observations, with much
431 money, time, and labor consumed. In this context, both transferable learning model and our RASP
432 methods (combing unsupervised learning with statistics) indeed provide huge potential for crop
433 mapping. However, transferable models are suitable for areas or years with similar cropping patterns.
434 In areas with diverse and complex cropping patterns, it is a challenge to apply the supervised model
435 trained in limited areas or limited years into others (Wang et al., 2019; Ma et al., 2020). In contrast,
436 our strategy leverages a specific, pre-existing set of samples to stably differentiate soybean
437 characteristics from other crops, which can accurately map annual dynamics without updated
438 requirement in annual samples. Consequently, this method significantly weakens limitations in crop
439 classification during years without specific samples, enabling crop mapping consistently and
440 continually.

441 Another key advantage of our spectra-phenology integration approach is its quick applicability
442 over larger areas, coupled with excellent spatial scalability. It can self-adopt to different
443 environments by considering phenology information. Compared to methods that rely on composite
444 indicators and specific thresholds, our approach simplifies the requirements for inputs and
445 experienced judgements. The only inputs required are the phenological information of soybeans and
446 the number of other primary crops during the same growing season in the targeted area. This allows
447 to classify crop swiftly and efficiently without additional inputs for background knowledge or
448 setting complex thresholds. The input of phenological information in each prefecture enhanced the
449 zonal adaptive assessment of soybean growth status across various areas, thereby facilitating crop
450 classification. This innovative approach ensures its applicability into other soybean-producing areas,
451 showcasing its potential for broader implementation.

452 4.2 The uncertainty from image quality

453 The method we proposed (RASP) is strongly dependent on remote sensing images and subregional
454 unsupervised classification by considering the bands and vegetation indices, which are all sensitive to
455 the unique characteristics of soybeans. Therefore, the accuracy of soybean maps inevitably is associated
456 with the quality of remote sensing images. By using ground samples to validate the mapping results, we
457 found that the accuracy of 2017 is lower than that of 2018 and 2019, with an overall accuracy is less than
458 80% (Table 2).

459 We extracted cloud-free images in different regions during the soybean growing season and calculated
460 the monthly average number of clear observations. In general, the monthly averages of clear observations
461 in Northeast region and Huang-Huai-Hai region (Zone I and Zone II) are relatively higher than the
462 southern zones (Zone III and IV) (Fig. 10a2-e2). In areas with quite lower clear observations, despite
463 a gap-filling method was conducted to generate complete 10-day composite time series, higher
464 uncertainty is inevitable. The gap-filling time series might contain duplicate values, which cannot
465 accurately reflect the crop growth process in reality. Obviously, the total number of images available
466 in 2017 over the study areas was significantly fewer than those of other years, because the second
467 satellite Sentinel-2B only commenced operations and started providing data after March of 2017
468 (Fig. 10a1-e1). Removing the cloudy pixels has left ever fewer clear images available (upper vs. down
469 layer in Fig. 10). During the growing season, the average number of clear observations per month was 0-
470 2 in partial regions, lower than the requirements of 10-day time series composite we mentioned in 2.3.1.
471 This might explain the lower user's accuracy of soybean in Zone IV compared to other sub-zones
472 (Table S1) and low overall accuracy based on sample verification in 2017 (Table 2).



473

474 **Figure 10. Total (a1-e1) and clear (a2-e2) observations per month during soybean growing season.**

475 **4.3 Limitations in small-scale planting areas**

476 Validation based on statistics shows that ChinaSoyArea10m reached a high consistency ($R^2 \sim 0.85$) across
 477 China. However, in areas with soybean sparsely planted, the consistency is lower than that in densely
 478 planted areas, with more overestimations observed in the sparse areas. Such overestimations are caused
 479 by the limitations of unsupervised classification algorithm. K-means is difficult to accurately capture
 480 small plots of crops in a complex cropping system, although it can make up for the shortage of crop
 481 mapping in some areas with limited training samples (Kwak and Park, 2022). Studies have proved that
 482 the classifier performs inferiorly where dominant crop phenotypes are similar, and crop diversity is higher
 483 (Wang et al., 2019; Konduri et al., 2020). Therefore, the classifier is challenged in areas where soybean
 484 is not the dominant type due to the small plot size and spectral overlap between different crops (Chabalala
 485 et al., 2022). In southern China, cropland plots are typically small (<0.04 ha in most regions) and the
 486 crop diversity is high. The growth periods of soybean, peanut, potato, and maize are similar, dominantly
 487 indicated by a mixed planting pattern, which has contributed to the low accuracy of non-main soybean
 488 producing areas in southern China (Liu et al., 2020). Additionally, soybeans are intercropped with maize
 489 or other crops in some areas, where the strip width is less one meter (Yang et al., 2014; Du et al., 2018).

490 This planting pattern will introduce the mixed pixels problem as well under the background of 10 m
491 resolution crop mapping.

492 The lower accuracy in soybean area sparsely planted could be explained by the characteristics of K-
493 means algorithm. K-means algorithm is developed to minimize the distance between each point within a
494 cluster and the cluster's centroid. When the sample size in a particular category substantially exceeds
495 those of others, the algorithm might preferentially optimize the cohesion of the larger category, and would
496 neglect the accurate clustering for smaller categories (Tan et al., 2016). The effectiveness of K-means
497 classification is highly dependent on the selection of initial clustering centers. In scenarios of unbalanced
498 categories, initial centers randomly selected might inadequately represent the minor categories, resulting
499 in inaccurate results (Tan et al., 2016). Additionally, K-means assumes that each cluster is spherical;
500 therefore, it does not perform well when clusters are non-spherical and uneven in size and density. Hence,
501 in areas with unbalanced crop categories, the algorithm faces challenge to assign each crop to a
502 corresponding cluster precisely (Tan et al., 2016; Wang et al., 2019).

503 Our regional adaptive large-area crop mapping method in future will further be improved by the
504 follows: (1) Classification on a finer scale by specifying a more precise number of target clusters can
505 reduce spatial heterogeneity and emphasize the relative importance of non-dominant categories, and
506 increase classification accuracy consequently (Li and Yang, 2017). (2) Optimizing data preprocessing
507 methods. Outliers can interrupt classification because the unsupervised methods is highly sensitive to
508 anomalies (Raykov et al., 2016; Wang et al., 2019). Therefore, eliminating outliers can further improve
509 the classification validity. In addition, since K-means weights all dimensions equally, minimizing the
510 features' correlation and reducing irrelevant variables are also important means to enhance the
511 classification effect (Hastie et al., 2009). (3) Improving algorithm performance. A variety of algorithms
512 have been proposed to address the inherent defects of K-means (Ahmed et al., 2020), such as by
513 optimizing the initial clustering center (e.g., K-means++), weighting classes (e.g., Weighted k-means),
514 and non-spherical clustering assumptions (e.g., DBSCAN, Spectral Clustering) (Ester et al., 1996; Bach
515 and Jordan, 2003; Kerdprasop et al., 2005; Arthur and Vassilvitskii, 2007). The improved algorithms will
516 address the issues on complex and highly diverse crop classification in some degrees (Li et al., 2022;
517 Rivera et al., 2022). (4) Better post-processing of data. Misclassification of field ridges and image
518 speckles is inevitable during mapping crops over large areas. With the progress of computing power,

519 auxiliary data and image processing algorithms can further eliminate these issues (Liu et al., 2018a; Li
520 and Qu, 2019; Hamano et al., 2023). We are sure that integrating cloud computing platforms with
521 advanced algorithms will provide substantial potential for accurate crop identification covering larger
522 areas in future.

523 **5 Data availability**

524 The soybean planting area product for China during 2017-2021 (ChinaSoyArea10m) is available at
525 <https://zenodo.org/doi/10.5281/zenodo.10071426> (Mei et al., 2023). We encourage users to
526 independently verify data products for special study areas before using them.

527 **6 Conclusions**

528 In this study, a Regional Adaption Spectra-Phenology Integration (RASP) method over large-scale was
529 developed and utilized to generate soybean planting area maps for major producing regions in China
530 from 2017 to 2021. By utilizing Sentinel-2 images, spectral features and vegetation indices that best
531 distinguish soybeans were extracted and input into an unsupervised classifier in each prefecture. The
532 DTW method was then employed to identify the soybean distribution. RASP does not rely on many
533 ground samples and considers the soybean phenology in various planting areas, suggesting a potential
534 way for long-term crop mapping over larger regions. Verification results demonstrated a high consistency
535 between the mapping results and census data at county or prefecture level (all > 0.82), with overall
536 accuracies of field samples reaching 77.08%~86.77%. These findings confirm the reliability of
537 ChinaSoyArea10m. Our data products fill the gap in regional long-term soybean maps in China, and
538 provide important information for sustainable soybean production and management, agricultural system
539 modeling, and optimization.

540 **Author contributions.**

541 ZZ and FT conceive this study. QM, JH, and JD collected datasets. QM implemented the research and
542 wrote the original draft of the paper. All authors discussed the results and revised the manuscript.

543 **Competing interests.**

544 The contact author has declared that neither they nor their co-authors have any competing interests.

545 **Financial support.**

546 This research was funded by the National Key Research and Development Program of China
547 (2020YFA0608201) and National Natural Science Foundation of China (42061144003, 41977405).

548 **References**

549 Ahmed, M., Seraj, R., and Islam, S. M. S.: The k-means Algorithm: A Comprehensive Survey and
550 Performance Evaluation, *Electronics*, 9, 1295, <https://doi.org/10.3390/electronics9081295>, 2020.

551 FAOSTAT: https://www.fao.org/faostat/en/#rankings/countries_by_commodity, last access: 10
552 October 2023.

553 National Bureau of Statistics of China: <http://www.stats.gov.cn/english/>, last access: 23 October
554 2023.

555 Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, in: *Soda*, 1027–
556 1035, 2007.

557 Bach, F. and Jordan, M.: Learning spectral clustering, in: *Advances in neural information processing*
558 *systems*, 2003.

559 Chabalala, Y., Adam, E., and Ali, K. A.: Machine Learning Classification of Fused Sentinel-1 and
560 Sentinel-2 Image Data towards Mapping Fruit Plantations in Highly Heterogenous Landscapes,
561 *Remote Sens.*, 14, 2621, <https://doi.org/10.3390/rs14112621>, 2022.

562 Chen, D., Huang, J., and Jackson, T. J.: Vegetation water content estimation for corn and soybeans
563 using spectral indices derived from MODIS near- and short-wave infrared bands, *Remote Sens.*
564 *Environ.*, 98, 225–236, <https://doi.org/10.1016/j.rse.2005.07.008>, 2005.

565 Chen, H., Li, H., Liu, Z., Zhang, C., Zhang, S., and Atkinson, P. M.: A novel Greenness and Water
566 Content Composite Index (GWCCI) for soybean mapping from single remotely sensed multispectral
567 images, *Remote Sens. Environ.*, 295, 113679, <https://doi.org/10.1016/j.rse.2023.113679>, 2023.

568 Cui, K. and Shoemaker, S. P.: A look at food security in China, *npj Sci. Food*, 2, 4,
569 <https://doi.org/10.1038/s41538-018-0012-x>, 2018.

570 Di Tommaso, S., Wang, S., and Lobell, D. B.: Combining GEDI and Sentinel-2 for wall-to-wall
571 mapping of tall and short crops, *Environ. Res. Lett.*, 16, 125002, <https://doi.org/10.1088/1748-9326/ac358c>, 2021.

573 Dong, J., Fu, Y., Wang, J., Tian, H., Fu, S., Niu, Z., Han, W., Zheng, Y., Huang, J., and Yuan, W.:
574 Early-season mapping of winter wheat in China based on Landsat and Sentinel images, *Earth Syst.
575 Sci. Data*, 12, 3081–3095, <https://doi.org/10.5194/essd-12-3081-2020>, 2020.

576 Du, J., Han, T., Gai, J., Yong, T., Sun, X., Wang, X., Yang, F., Liu, J., Shu, K., Liu, W., and Yang,
577 W.: Maize-soybean strip intercropping: Achieved a balance between high productivity and
578 sustainability, *J. Integr. Agric.*, 17, 747–754, [https://doi.org/10.1016/S2095-3119\(17\)61789-1](https://doi.org/10.1016/S2095-3119(17)61789-1), 2018.

579 Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters
580 in large spatial databases with noise, in: *kdd*, 226–231, 1996.

581 Fensholt, R. and Sandholt, I.: Derivation of a shortwave infrared water stress index from MODIS
582 near- and shortwave infrared data in a semiarid environment, *Remote Sens. Environ.*, 87, 111–121,
583 <https://doi.org/10.1016/j.rse.2003.07.002>, 2003.

584 Gong, L., Tian, B., Li, Y., and Wu, S.: Phenological Changes of Soybean in Response to Climate
585 Conditions in Frigid Region in China over the Past Decades, *Int. J. Plant Prod.*, 15, 363–375,
586 <https://doi.org/10.1007/s42106-021-00145-5>, 2021.

587 Graesser, J. and Ramankutty, N.: Detection of cropland field parcels from Landsat imagery, *Remote
588 Sens. Environ.*, 201, 165–180, <https://doi.org/10.1016/j.rse.2017.08.027>, 2017.

589 Guan, X., Huang, C., Liu, G., Meng, X., and Liu, Q.: Mapping Rice Cropping Systems in Vietnam
590 Using an NDVI-Based Time-Series Similarity Measurement Based on DTW Distance, *Remote
591 Sens.*, 8, 19, <https://doi.org/10.3390/rs8010019>, 2016.

592 Guo, W., Ren, J., Liu, X., Chen, Z., Wu, S., and Pan, H.: Winter wheat mapping with globally
593 optimized threshold under total quantity constraint of statistical data, *J. Remote Sens.*, 22, 1023–
594 1041, <https://doi.org/10.11834/jrs.20187468>, 2018.

595 Hamano, M., Shiozawa, S., Yamamoto, S., Suzuki, N., Kitaki, Y., and Watanabe, O.: Development
596 of a method for detecting the planting and ridge areas in paddy fields using AI, GIS, and precise

597 DEM, *Precision Agric*, 24, 1862–1888, <https://doi.org/10.1007/s11119-023-10021-z>, 2023.

598 Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Zhang, J., and Li, Z.: The RapeseedMap10 database:
599 annual maps of rapeseed at a spatial resolution of 10 m based on multi-source data, *Earth Syst. Sci.*
600 *Data*, 13, 2857–2874, <https://doi.org/10.5194/essd-13-2857-2021>, 2021.

601 Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Zhuang, H., Cheng, F., Zhang, J., and Tao, F.: Annual
602 paddy rice planting area and cropping intensity datasets and their dynamics in the Asian monsoon
603 region from 2000 to 2020, *Agric. Syst.*, 200, 103437, <https://doi.org/10.1016/j.agry.2022.103437>,
604 2022.

605 Hartman, G. L., West, E. D., and Herman, T. K.: Crops that feed the World 2. Soybean—worldwide
606 production, use, and constraints caused by pathogens and pests, *Food Secur.*, 3, 5–17,
607 <https://doi.org/10.1007/s12571-010-0108-x>, 2011.

608 Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: *The elements of statistical learning:*
609 *data mining, inference, and prediction*, Springer, 2009.

610 Housman, I. W., Chastain, R. A., and Finco, M. V.: An Evaluation of Forest Health Insect and
611 Disease Survey Data and Satellite-Based Remote Sensing Forest Change Detection Methods: Case
612 Studies in the United States, *Remote Sens.*, 10, 1184, <https://doi.org/10.3390/rs10081184>, 2018.

613 Huang, Y., Qiu, B., Chen, C., Zhu, X., Wu, W., Jiang, F., Lin, D., and Peng, Y.: Automated soybean
614 mapping based on canopy water content and chlorophyll content using Sentinel-2 images, *Int. J.*
615 *Appl. Earth Obs.*, 109, 102801, <https://doi.org/10.1016/j.jag.2022.102801>, 2022.

616 Kerdprasop, K., Kerdprasop, N., and Sattayatham, P.: Weighted k-means for density-biased
617 clustering, in: *International conference on data warehousing and knowledge discovery*, 488–497,
618 https://doi.org/10.1007/11546849_48, 2005.

619 Konduri, V. S., Kumar, J., Hargrove, W. W., Hoffman, F. M., and Ganguly, A. R.: Mapping crops
620 within the growing season across the United States, *Remote Sens. Environ.*, 251, 112048,
621 <https://doi.org/10.1016/j.rse.2020.112048>, 2020.

622 Kumari, M., Murthy, C. S., Pandey, V., and Bairagi, G. D.: Soybean Cropland Mapping Using Multi-
623 Temporal Sentinel-1 Data, *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-3-W6, 109–
624 114, <https://doi.org/10.5194/isprs-archives-XLII-3-W6-109-2019>, 2019.

625 Kwak, G.-H. and Park, N.-W.: Unsupervised Domain Adaptation with Adversarial Self-Training for

626 Crop Classification Using Remote Sensing Images, *Remote Sens.*, 14, 4639,
627 <https://doi.org/10.3390/rs14184639>, 2022.

628 Li, B. and Yang, L.: Clustering accuracy analysis of building area in high spatial resolution remote
629 sensing images based on k-means algorithm, in: 2017 2nd International Conference on Frontiers of
630 Sensors Technologies (ICFST), 2017 2nd International Conference on Frontiers of Sensors
631 Technologies (ICFST), 174–178, <https://doi.org/10.1109/ICFST.2017.8210497>, 2017.

632 Li, H., Song, X.-P., Hansen, M. C., Becker-Reshef, I., Adusei, B., Pickering, J., Wang, L., Wang, L.,
633 Lin, Z., Zalles, V., Potapov, P., Stehman, S. V., and Justice, C.: Development of a 10-m resolution
634 maize and soybean map over China: Matching satellite-based crop classification with sample-based
635 area estimation, *Remote Sens. Environ.*, 294, 113623, <https://doi.org/10.1016/j.rse.2023.113623>,
636 2023.

637 Li, L., Friedl, M. A., Xin, Q., Gray, J., Pan, Y., and Frohling, S.: Mapping Crop Cycles in China
638 Using MODIS-EVI Time Series, *Remote Sens.*, 6, 2473–2493, <https://doi.org/10.3390/rs6032473>,
639 2014.

640 Li, T., Johansen, K., and McCabe, M. F.: A machine learning approach for identifying and
641 delineating agricultural fields and their multi-temporal dynamics using three decades of Landsat
642 data, *ISPRS J. Photogramm. Remote Sens.*, 186, 83–101,
643 <https://doi.org/10.1016/j.isprsjprs.2022.02.002>, 2022.

644 Li, Y. and Qu, H.: LSD and Skeleton Extraction Combined with Farmland Ridge Detection, in:
645 *Advances in Intelligent, Interactive Systems and Applications*, Cham, 446–453,
646 https://doi.org/10.1007/978-3-030-02804-6_59, 2019.

647 Liu, H., Zhang, J., Pan, Y., Shuai, G., Zhu, X., and Zhu, S.: An Efficient Approach Based on UAV
648 Orthographic Imagery to Map Paddy With Support of Field-Level Canopy Height From Point Cloud
649 Data, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 11, 2034–2046,
650 <https://doi.org/10.1109/JSTARS.2018.2829218>, 2018a.

651 Liu, J., Wang, L., Yang, F., Yao, B., and Yang, L.: Recognition ability of red edge and short wave
652 infrared spectrum on maize and soybean, *Chinese Agricultural Science Bulletin*, 34, 120–129,
653 2018b.

654 Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y., and Qiao, Z.: Mapping cropping intensity in

655 China using time series Landsat and Sentinel-2 images and Google Earth Engine, *Remote Sens.*
656 *Environ.*, 239, 111624, <https://doi.org/10.1016/j.rse.2019.111624>, 2020.

657 Liu, M. and Fan, Q.: Study on the Current Situation and Problems of Soybean Consumption,
658 Production and Import in China, *Grain Science And Technology And Economy*, 46, 28–35,
659 <https://doi.org/10.16465/j.gste.cn431252ts.20210606>, 2021.

660 Liu, Z., Ying, H., Chen, M., Bai, J., Xue, Y., Yin, Y., Batchelor, W. D., Yang, Y., Bai, Z., Du, M.,
661 Guo, Y., Zhang, Q., Cui, Z., Zhang, F., and Dou, Z.: Optimization of China's maize and soy
662 production can ensure feed sufficiency at lower nitrogen and carbon footprints, *Nat. Food*, 2, 426–
663 433, <https://doi.org/10.1038/s43016-021-00300-1>, 2021.

664 Lowder, S. K., Skoet, J., and Raney, T.: The Number, Size, and Distribution of Farms, Smallholder
665 Farms, and Family Farms Worldwide, *World Dev.*, 87, 16–29,
666 <https://doi.org/10.1016/j.worlddev.2015.10.041>, 2016.

667 Luo, C., Liu, H., Lu, L., Liu, Z., Kong, F., and Zhang, X.: Monthly composites from Sentinel-1 and
668 Sentinel-2 images for regional major crop mapping with Google Earth Engine, *J. Integr. Agr.*, 20,
669 1944–1957, [https://doi.org/10.1016/S2095-3119\(20\)63329-9](https://doi.org/10.1016/S2095-3119(20)63329-9), 2021.

670 Luo, Y., Zhang, Z., Li, Z., Chen, Y., Zhang, L., Cao, J., and Tao, F.: Identifying the spatiotemporal
671 changes of annual harvesting areas for three staple crops in China by integrating multi-data sources,
672 *Environ. Res. Lett.*, 15, 074003, <https://doi.org/10.1088/1748-9326/ab80f0>, 2020.

673 Luo, Y., Zhang, Z., Zhang, L., Han, J., Cao, J., and Zhang, J.: Developing High-Resolution Crop
674 Maps for Major Crops in the European Union Based on Transductive Transfer Learning and Limited
675 Ground Data, *Remote Sens.*, 14, 1809, <https://doi.org/10.3390/rs14081809>, 2022.

676 Ma, Z., Liu, Z., Zhao, Y., Zhang, L., Liu, D., Ren, T., Zhang, X., and Li, S.: An Unsupervised Crop
677 Classification Method Based on Principal Components Isometric Binning, *ISPRS Int. J. Geo-Inf.*,
678 9, 648, <https://doi.org/10.3390/ijgi9110648>, 2020.

679 Marshall, M., Belgiu, M., Boschetti, M., Pepe, M., Stein, A., and Nelson, A.: Field-level crop yield
680 estimation with PRISMA and Sentinel-2, *ISPRS J. Photogramm. Remote Sens.*, 187, 191–210,
681 <https://doi.org/10.1016/j.isprsjprs.2022.03.008>, 2022.

682 Mei, Q., Zhang, Z., Han, J., Song, J., Dong, J., Wu, H., Xu, J., and Tao, F.: ChinaSoyArea10m: a
683 dataset of soybean planting areas with a spatial resolution of 10 m across China from 2017 to 2021

684 (V1), Zenodo [data set], <https://doi.org/10.5281/zenodo.10071427>, 2023.

685 Olsen, J. L., Stisen, S., Proud, S. R., and Fensholt, R.: Evaluating EO-based canopy water stress
686 from seasonally detrended NDVI and SIWSI with modeled evapotranspiration in the Senegal River
687 Basin, *Remote Sens. Environ.*, 159, 57–69, <https://doi.org/10.1016/j.rse.2014.11.029>, 2015.

688 Oreopoulos, L., Wilson, M. J., and Várnai, T.: Implementation on Landsat Data of a Simple Cloud-
689 Mask Algorithm Developed for MODIS Land Bands, *IEEE Geosci. Remote. Sens. Lett.*, 8, 597–
690 601, <https://doi.org/10.1109/LGRS.2010.2095409>, 2011.

691 Pan, B., Zheng, Y., Shen, R., Ye, T., Zhao, W., Dong, J., Ma, H., and Yuan, W.: High Resolution
692 Distribution Dataset of Double-Season Paddy Rice in China, *Remote Sens.*, 13, 4609,
693 <https://doi.org/10.3390/rs13224609>, 2021.

694 Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens,
695 A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland
696 expansion in the twenty-first century, *Nat. Food*, 3, 19–28, [https://doi.org/10.1038/s43016-021-](https://doi.org/10.1038/s43016-021-00429-z)
697 00429-z, 2022.

698 Raykov Y. P., Boukouvalas A., Baig F., and Little M. A.: What to Do When K-Means Clustering
699 Fails: A Simple yet Principled Alternative Algorithm, *PLOS ONE*, 11, e0162259,
700 <https://doi.org/10.1371/journal.pone.0162259>, 2016.

701 Rivera, A. J., Pérez-Godoy, M. D., Elizondo, D., Deka, L., and del Jesus, M. J.: Analysis of
702 clustering methods for crop type mapping using satellite imagery, *Neurocomputing*, 492, 91–106,
703 <https://doi.org/10.1016/j.neucom.2022.04.002>, 2022.

704 Shangguan, Y., Li, X., Lin, Y., Deng, J., and Yu, L.: Mapping spatial-temporal nationwide soybean
705 planting area in Argentina using Google Earth Engine, *Int. J. Remote Sens.*, 43, 1724–1748,
706 <https://doi.org/10.1080/01431161.2022.2049913>, 2022.

707 Shen, R., Dong, J., Yuan, W., Han, W., Ye, T., and Zhao, W.: A 30 m Resolution Distribution Map
708 of Maize for China Based on Landsat and Sentinel Images, *J. Remote Sens.*, 2022, 2022/9846712,
709 <https://doi.org/10.34133/2022/9846712>, 2022.

710 Shen, R., Pan, B., Peng, Q., Dong, J., Chen, X., Zhang, X., Ye, T., Huang, J., and Yuan, W.: High-
711 resolution distribution maps of single-season rice in China from 2017 to 2022, *Earth Syst. Sci. Data*,
712 15, 3203–3222, <https://doi.org/10.5194/essd-15-3203-2023>, 2023.

713 Sobejano-Paz, V., Mikkelsen, T. N., Baum, A., Mo, X., Liu, S., Köppl, C. J., Johnson, M. S., Gulyas,
714 L., and García, M.: Hyperspectral and Thermal Sensing of Stomatal Conductance, Transpiration,
715 and Photosynthesis for Soybean and Maize under Drought, *Remote Sens.*, 12, 3182,
716 <https://doi.org/10.3390/rs12193182>, 2020.

717 Song, X.-P., Potapov, P. V., Krylov, A., King, L., Di Bella, C. M., Hudson, A., Khan, A., Adusei, B.,
718 Stehman, S. V., and Hansen, M. C.: National-scale soybean mapping and area estimation in the
719 United States using medium resolution satellite imagery and field survey, *Remote Sens. Environ.*,
720 190, 383–395, <https://doi.org/10.1016/j.rse.2017.01.008>, 2017.

721 Tan, P.-N., Steinbach, M., and Kumar, V.: Introduction to data mining, Pearson Education India,
722 2016.

723 Wang, S., Azzari, G., and Lobell, D. B.: Crop type mapping without field-level labels: Random
724 forest transfer and unsupervised clustering techniques, *Remote Sens. Environ.*, 222, 303–317,
725 <https://doi.org/10.1016/j.rse.2018.12.026>, 2019.

726 Wang, S., Di Tommaso, S., Deines, J. M., and Lobell, D. B.: Mapping twenty years of corn and
727 soybean across the US Midwest using the Landsat archive, *Sci. Data*, 7, 307,
728 <https://doi.org/10.1038/s41597-020-00646-4>, 2020.

729 Wang, Y. and Gai, J.: Study on the ecological regions of soybean in China II · Ecological
730 environment and representative varieties, *Chinese Journal of Applied Ecology*, 71–75, 2002.

731 Wang, Y., Feng, L., Sun, W., Zhang, Z., Zhang, H., Yang, G., and Meng, X.: Exploring the potential
732 of multi-source unsupervised domain adaptation in crop mapping using Sentinel-2 images, *Gisci.*
733 *Remote Sens.*, 59, 2247–2265, <https://doi.org/10.1080/15481603.2022.2156123>, 2022.

734 Wang, Y., Ling, X., Ma, C., Liu, C., Zhang, W., Huang, J., Peng, S., and Deng, N.: Can China get
735 out of soy dilemma? A yield gap analysis of soybean in China, *Agron. Sustain. Dev.*, 43, 47,
736 <https://doi.org/10.1007/s13593-023-00897-6>, 2023.

737 Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., Yadav,
738 K., and Thau, D.: Automated cropland mapping of continental Africa using Google Earth Engine
739 cloud computing, *ISPRS J. Photogramm. Remote Sens.*, 126, 225–244,
740 <https://doi.org/10.1016/j.isprsjprs.2017.01.019>, 2017.

741 Yang, F., Huang, S., Gao, R., Liu, W., Yong, T., Wang, X., Wu, X., and Yang, W.: Growth of soybean

742 seedlings in relay strip intercropping systems in relation to light quantity and red:far-red ratio, *Field*
743 *Crops Res.*, 155, 245–253, <https://doi.org/10.1016/j.fcr.2013.08.011>, 2014.

744 You, N. and Dong, J.: Examining earliest identifiable timing of crops using all available Sentinel
745 1/2 imagery and Google Earth Engine, *ISPRS J. Photogramm. Remote Sens.*, 161, 109–123,
746 <https://doi.org/10.1016/j.isprsjprs.2020.01.001>, 2020.

747 You, N., Dong, J., Huang, J., Du, G., Zhang, G., He, Y., Yang, T., Di, Y., and Xiao, X.: The 10-m
748 crop type maps in Northeast China during 2017–2019, *Sci. Data*, 8, 41,
749 <https://doi.org/10.1038/s41597-021-00827-9>, 2021.

750 You, N., Dong, J., Li, J., Huang, J., and Jin, Z.: Rapid early-season maize mapping without crop
751 labels, *Remote Sens. Environ.*, 290, 113496, <https://doi.org/10.1016/j.rse.2023.113496>, 2023.

752 Zhang, C., Dong, J., and Ge, Q.: Quantifying the accuracies of six 30-m cropland datasets over
753 China: A comparison and evaluation analysis, *Comput. Electron. Agr.*, 197, 106946,
754 <https://doi.org/10.1016/j.compag.2022.106946>, 2022.

755 Zhang, G., Xiao, X., Biradar, C. M., Dong, J., Qin, Y., Menarguez, M. A., Zhou, Y., Zhang, Y., Jin,
756 C., Wang, J., Doughty, R. B., Ding, M., and Moore, B.: Spatiotemporal patterns of paddy rice
757 croplands in China and India from 2000 to 2015, *Sci. Total Environ.*, 579, 82–92,
758 <https://doi.org/10.1016/j.scitotenv.2016.10.223>, 2017.

759 Zhong, L., Hu, L., Yu, L., Gong, P., and Biging, G. S.: Automated mapping of soybean and corn
760 using phenology, *ISPRS J. Photogramm. Remote Sens.*, 119, 151–164,
761 <https://doi.org/10.1016/j.isprsjprs.2016.05.014>, 2016.

762 Zhou, W., Wei, H., Chen, Y., Zhang, X., Hu, J., Cai, Z., Yang, J., Hu, Q., Xiong, H., Yin, G., and Xu,
763 B.: Monitoring intra-annual and interannual variability in spatial distribution of plastic-mulched
764 citrus in cloudy and rainy areas using multisource remote sensing data, *Eur. J. Agron.*, 151, 126981,
765 <https://doi.org/10.1016/j.eja.2023.126981>, 2023.

766