https://doi.org/10.5194/essd-2024-147
Preprint. Discussion started: 9 July 2024
© Author(s) 2024. CC BY 4.0 License.

CCD-Rice: A long-term paddy rice distribution dataset in China at 30 m resolution

Ruoque Shen [1], Qiongyan Peng [1], Xiangqian Li [1], Xiuzhi Chen [1], and Wenping Yuan [2]

[1]School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai 519082, Guangdong, China

[2]Institute of Carbon Neutrality, Sino-French Institute for Earth System Science. College of Urban and Environmental Sciences, Peking University, Beijing 100871.China

Correspondence to: Wenping Yuan (yuanwp@pku.edu.cn)

Abstract. As one of the most widely cultivated grain crops, paddy rice is a vital staple food in China and plays a crucial role in ensuring food security. Over the past decades, the planting area of paddy rice in China has shown substantial variability. Yet, there are no long-term high-resolution rice distribution maps in China, which hinders our ability to estimate greenhouse gas fluxes and crop production. This study developed a new optical satellite-based rice mapping method using a machine learning model and appropriate data preprocessing strategies to address the challenges of cloud contamination and missing data in optical remote sensing observations. This study produced CCD-Rice (China Crop Dataset-Rice), the first high-resolution rice distribution dataset in China from 1990 to 2016. Based on 391,659 validation samples, the overall accuracy of the distribution maps in each provincial administrative region averaged 90.26 %. Compared with 20,759 county-level statistical data, the coefficients of determination (R^2) of single- and double-season rice in each year averaged 0.84 and 0.80, respectively. The distribution maps can be obtained at https://doi.org/10.57760/sciencedb.15865 (Shen et al., 2024a).

Keywords: rice; single-season rice; double-season rice; Landsat; crop mapping

1. Introduction

Paddy rice (*Oryza sativa*) is one of the most critical crops in the world, accounting for 8 % of global food production in 2021, and is a staple food for more than 50 % of the world population (FAO, 2023; Elert, 2014). However, rice cultivation consumes large amounts of freshwater and emits methane, a significant greenhouse gas (Bouman et al., 2007; Mohammadi et al., 2020; Zhang et al., 2020). In addition, the spatial distribution of rice cultivation has changed significantly over the past few decades (Liu et al., 2013; Jiang et al., 2019). Therefore, the long-term identification of paddy rice is very important for food security, water resource management, and climate change research.

There are two main approaches for rice mapping: phenology-based and machine-learning methods. Phenology-based methods include approaches like dynamic time warping (DTW) and others (Xiao et al., 2005, 2006; Nguyen et al., 2016; Phan et al., 2018; Pan et al., 2021b; Han et al., 2021). These methods exploit the unique phenological characteristics of rice. For instance, Xiao et al. (2005, 2006) proposed comparing the land surface water index (LSWI) with vegetation indices during the rice transplantation period. The signal where LSWI exceeds the vegetation index is used to distinguish rice from other land cover types. This method has achieved good results in Southern China, Southeast Asia, and South Asia. Pan et al. (2021) utilized the time-weighted dynamic time warping (TWDTW) algorithm to compare the time series of backscattering coefficients from Sentinel-1 of unknown pixels during rice transplantation with the typical rice time series and achieved double-season rice mapping in China from 2016 to 2020. Additionally, Xu et al. (2023) suggested identifying rice by comparing the backscattering coefficients from Sentinel-1 of unknown pixels with those of typical water bodies and vegetation and achieved satisfactory results at the regional level (Zhang et al., 2023b). Machine learning methods include approaches such as support vector machine (SVM), random forest (RF), and deep learning (Mansaray et al., 2020; You et al., 2021; Waleed et al., 2022; Tian et al., 2023; Sun et al., 2023). For example, You et al. (2021) used RF to map rice and two other major crops in Northeast China and achieved high overall accuracies, ranging from 0.81 to 0.86. Machine learning methods can yield high accuracy but typically require a large volume of training samples and face challenges in transferring models between different years (Valero et al., 2016). In contrast, phenology-based methods usually require none or few samples but have higher requirements for the quality of satellite observations than machine learning method. Missing values in the time series are likely to result in the incorrect identification of crucial phenological periods, thereby affecting the accuracy of classification (Dong et al., 2016; Shen et al., 2023a).

The satellite data used for rice mapping can be categorized into two types: optical remote sensing data and synthetic aperture radar (SAR) data. Optical remote sensing data are typically obtained from satellites such as Moderate-resolution Imaging Spectroradiometer (MODIS), Landsat, and Sentinel-2. Valid optical observations are limited in cases of cloud cover, preventing the acquisition of the true reflectance of the ground surface. MODIS has a relatively high revisit frequency, with two spacecrafts, Terra and Aqua, scanning the Earth's surface every one to two days, making it possible to yield relatively dense observations. Many large-scale rice mapping efforts have been accomplished using MODIS data (Xiao et al., 2005, 2006; Clauss et al., 2016; Han et al., 2022). However, MODIS has a low spatial resolution of only 250 to 1000 m, leading to significant confusion in regions dominated by small-scale fields due to the issue of mixed pixels. There are also some studies using Landsat series or Sentinel-2 for rice mapping (Dong et al., 2016; Mansaray et al., 2020; Hu et al., 2023). These satellites offer images at a higher spatial resolution (10 to 30 m), but they have longer revisit periods of 16 days and 5 days, respectively. Due to its high water needs, rice cultivation is often done in regions with high precipitation, such as southern China, and South and Southeast Asia. Statistics indicated that in some of these regions, such as southern China, the annual average number of cloud-free Landsat observations was less than eight between 1984 and 2017 (Zhou et al., 2019). Such sparse observations pose challenges to rice mapping studies, especially when employing phenology-based methods, for which the impact of clouds on the classification accuracy cannot be ignored (Dong et al., 2016; Shen et al., 2023a). In recent years, several studies have utilized SAR data from Sentinel-1 for rice mapping (Nguyen et al., 2016; Pan et al., 2021b; Sun et al., 2023; Zhang et al., 2023b). SAR signals have the ability to penetrate clouds and provide images in all weather conditions (Oguro et al., 2001). However, SAR images have significant salt-and-pepper noise, which results in lower data quality compared to optical images (Oliver and Quegan, 2004; Veloso et al., 2017). Furthermore, the availability of open-access SAR satellite data was limited prior to the launch of Sentinel-1A in 2014. This limitation hinders its applicability for long-term rice mapping. Therefore, the challenge posed by the poor quality of optical remote sensing data still needs to be addressed to achieve long-term rice mapping.

China is the world's largest rice producer, and, until 2017, rice was the most widely cultivated grain crop in the country (FAO, 2023; National Bureau of Statistics of China, 2023). Rice is also one of the most important staple foods in China, consumed by more than two-thirds of the population, especially in southern China, where it can account for more than 80 % of cereal intake (Zhao et al., 2023). Although there have been many previous studies on mapping rice in China, a nationwide, long-term, high-resolution rice map is still lacking. Some studies, such as those by Pan et al., (2021b) and Shen et al., (2023a), have produced nationwide distribution maps of double- and single-season rice in China, respectively. However, due to limitations in the quality of the remote sensing data, both studies covered only recent years (2016–2020 and 2017–2022, respectively). To address this gap, this study focuses on mapping rice distribution before 2017 and tackling the challenge of poor-quality remote sensing data. Specifically, this study intends to (1) develop a new optical satellite-based rice mapping method; (2) produce high-resolution distribution maps of single- and double-season rice in China from 1990 to 2016; (3) evaluate the accuracy of the results and analyze changes in rice cultivation patterns.

2. Data and methods

2.1 Study area

Rice is cultivated in most of the provincial administrative regions of China. The study area for this research was selected to include 25 provincial administrative regions in the eastern monsoon region of mainland China. The proportion of rice planting area in these 25 provincial administrative regions ranged from 99.60 % to 99.74 % of the total rice planting area in mainland China from 1990 to 2016 (https://data.stats.gov.cn). Due to differences in cloud cover and rice calendar in each provincial administrative region, the study area was divided into four subregions (Fig. 1). Subregion I is located in northern China and includes Heilongjiang, Jilin, Liaoning, Hebei, Inner Mongolia, Ningxia, and Tianjin. Here, only single-season rice is cultivated, and optical satellite images are less affected by cloud cover due to relatively low precipitation. Subregion II, includes Jiangsu, Sichuan, Yunnan, Chongqing, Guizhou, Henan, Shanghai, Shaanxi, and Shandong. Subregion II is also planted with only single-season rice, but experiences more precipitation, leading to poorer quality of optical remote sensing data than in subregion I. Subregion III, which includes Hunan, Jiangxi, Hubei, Anhui, and Zhejiang, cultivates both single-

and double-season rice. Subregion IV, which is warmer, allows for early rice cultivation than in subregion III and includes Guangxi, Fujian, Guangdong, and Hainan. Among these, Guangxi and Fujian cultivate both single- and double-season rice, while Guangdong and Hainan cultivate only double-season rice.

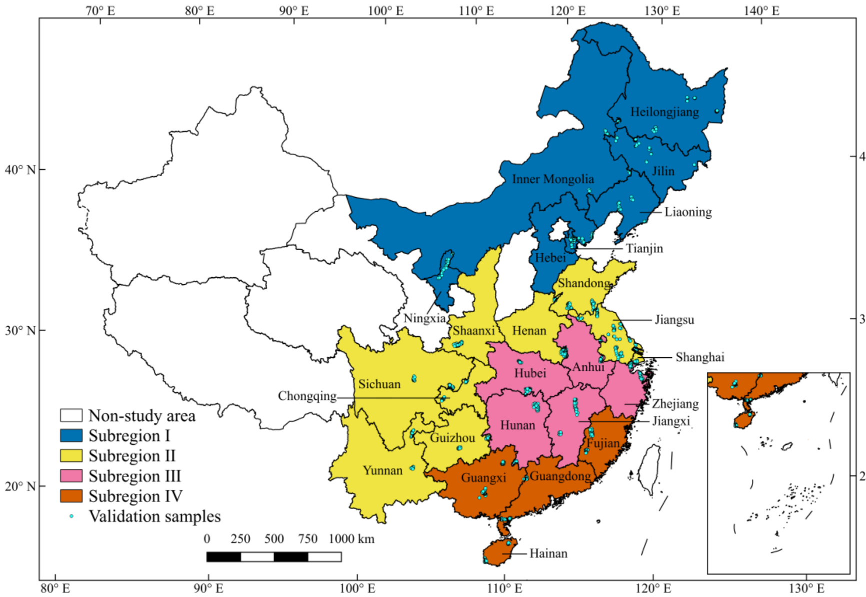


Figure 1: Study area and validation samples. The study area is divided into four subregions (shaded areas). The green dots indicate the centers of validation sample polygons.

2.2 Data

2.2.1 Satellite data and landcover data

The satellite data used for rice mapping in this study were sourced from the Landsat Collection 2 Level-2 Science Products, distributed by the United States Geological Survey (USGS). This product represents atmospherically corrected surface reflectance for the Landsat series. This study used band B5 of Landsat 5 and 7 as well as band B6 of Landsat 8, which both correspond to shortwave infrared 1 (SWIR1), with wavelength ranges of 1.55 to 1.75 μm and 1.566 to 1.651 μm, respectively. SWIR1 is sensitive to land surface water, and as such can capture the unique flooding signal during rice transplanting. Previous studies have demonstrated its effectiveness for rice mapping (Shen et al., 2023a). In addition, Landsat 7 data were only used from 1999 to 2002 and in 2012 due to the failure of its Scan Line Corrector (SLC) on 31 May 2003 and the lack of Landsat 5 and 8 images in 2012. The SWIR1 reflectance used for model training in this study was obtained from Landsat 8 and 9 from Landsat Collection 2 Level-2 Science Products as well as Sentinel-2 data provided by the European Space Agency (ESA).

The Quality Assessment (QA) band of Landsat data was used to eliminate the effect of clouds on Landsat images, while the Sentinel-2 Cloud Probability (S2C) product (https://developers.google.com/earthengine/datasets/catalog/COPERNICUS_S2_CLOUD_PROBABILITY) was used to exclude cloud-covered pixels from Sentinel-2 images. Pixels with a cloud probability greater than 50 % in the S2C product were considered to include cloud cover and were subsequently removed. The cloud-removed images were further composited to a median 8-day temporal resolution. For missing values in the time series due to cloud cover or observation frequency, this study did not use interpolation or other approaches to fill them, but uniformly set them to zero. Cloud removal and compositing were performed on the Google Earth Engine (GEE) platform (Gorelick et al., 2017).

In this study, the China land cover dataset (CLCD) product produced by Yang and Huang (2021, 2023) was used to exclude non-cultivated pixels. This product maps the land cover of China from 1985 to 2022 at a spatial resolution of 30 m and was produced using random forest (RF). The user's accuracy of cultivated land is 78.43 %.

2.2.2 Rice distribution maps of recent years

The training samples used in this study were extracted from two rice distribution maps of recent years: the distribution map of single-season rice in China from 2017 to 2022 produced by Shen et al. (2023a, b) and the distribution map of double-season rice in China from 2016 to 2020 produced by Pan et al. (2021a, b). The average overall accuracies of these two products over their studied provincial administrative regions were 85.23 % and 91.17 %, respectively. For provinces where only single-season rice or only double-season rice was cultivated, this study used the distribution maps for all years of two products, respectively. For provinces where both single- and double-season rice were cultivated, this study used only the common years of the two products, i.e., 2017 to 2020.

2.2.3 Validation sample and agricultural statistical data

The validation data used in this study consisted of validation samples and agricultural statistical data. The validation samples were visually interpreted from the very high-resolution images of Google Earth. The availability of imagery suitable for visual interpretation is limited by the scarcity of historical images from earlier years on Google Earth in China and the fact that early images tend to be for urban areas rather than rural areas. Therefore, instead of collecting validation samples across all study years, this study selected data from only two to four years in each provincial administrative region. We collected a total of 3420 polygons including 3671 rice field polygons and 749 polygons of other cover types (non-rice crops, natural vegetation, built-up areas, water bodies etc.) from 2002 to 2016, and further converted them into a total of 391,659 validation samples with a 30-m spatial resolution, including 201,891 rice samples and 189,768 samples of other cover types (Fig. 1 and Table 1). In provincial administrative regions where both single- and double-season rice are cultivated, the date of image capture can be used to determine whether the rice sample is single-season, early, or late rice. However, the sparse availability of Google Earth images, which do not always coincide with the appropriate time of year, means that this study does not differentiate between single- and double-season rice samples during validation.

Table 1: Years and number of validation samples in each provincial administrative region

Region	Years of samples	Number of samples	
		rice	others
Heilongjiang	2010, 2011, 2016	67329	41571
Jilin	2007, 2015	69472	40611
Liaoning	2011, 2015	2484	5621
Hebei	2008, 2015	1550	2492
Inner Mongolia	2013, 2015	2559	2543
Ningxia	2010, 2015	3082	3769
Tianjin	2002, 2006, 2011, 2014	4888	7304
Jiangsu	2004, 2011, 2013, 2014	6396	8656
Sichuan	2005, 2015	1896	1950
Yunnan	2005, 2013	2568	2190
Chongqing	2005, 2016	524	776
Guizhou	2012, 2015	899	1018
Henan	2010, 2016	1506	2937
Shanghai	2004, 2014	13380	26330
Shaanxi	2005, 2014, 2015	2853	2260
Shandong	2010, 2013	3409	6157
Hunan	2013, 2015	553	1570
Jiangxi	2006, 2012	1788	961
Hubei	2010, 2015	3104	1462
Anhui	2003, 2013	830	798
Zhejiang	2003, 2013	2875	3079
Guangxi	2011, 2016	1458	2382
Fujian	2013, 2015	707	725
Guangdong	2010, 2012	5291	21922
Hainan	2009, 2014	490	684

In this study, the planting area of single- and double-season rice in each provincial administrative region was collected from the following website: https://data.stats.gov.cn. This study also collected the rice planting area at the county level from the statistical yearbooks of provinces or cities. However, since it is difficult to trace statistical yearbooks back to 1990, and in some places the statistical yearbooks did not record the rice planting area, we were unable to collect complete statistics in all the years for all the county-level administrative regions. Furthermore, due to discrepancies between administrative divisions and statistical reporting, as well as changes in administrative divisions, some statistical data recording planting areas do not align with current or actual jurisdiction, making them unusable. Ultimately, this study was able to collect a total of 20,759 records at the county level from 1991 to 2016 (Fig. 2).

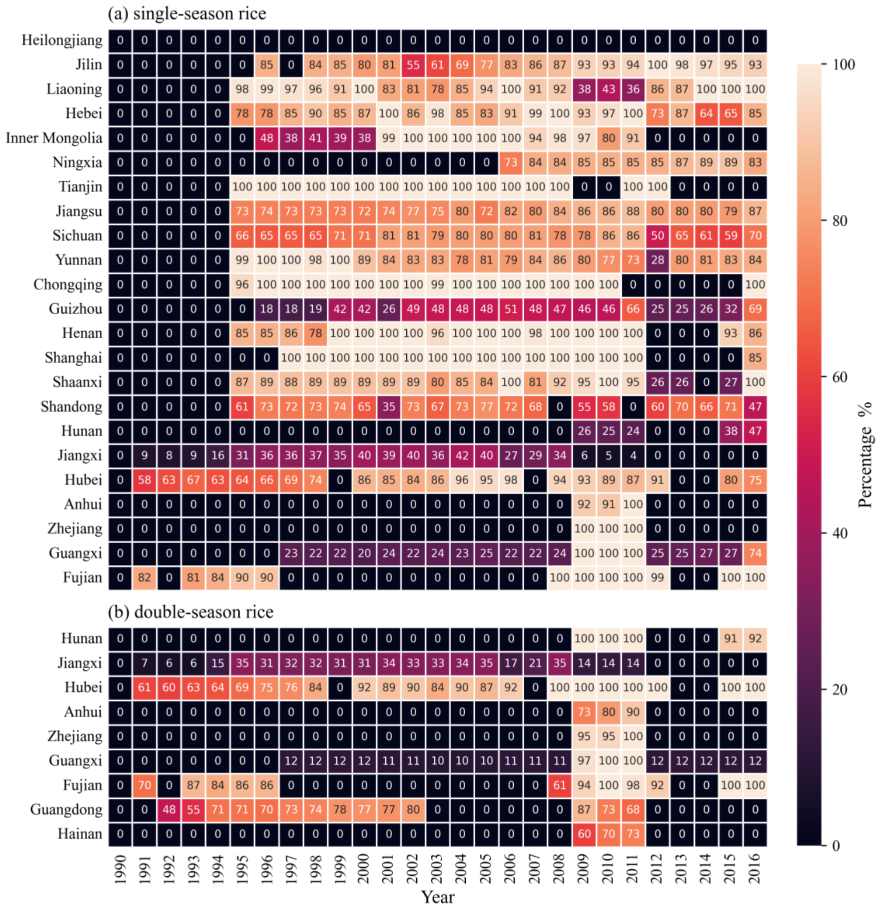


Figure 2: Percentage of the collected statistical planting areas relative to the province-level planting area in each provincial administrative region from 1990 to 2016. 150

2.2.4 Existing products

There are other studies that have produced long-term distribution maps of rice at regional, national, or continental scales. Our product will be compared with the following four open-access products: (1) *NEAsia_Rice* product, which produced the rice distribution in Northeast China from 2000 to 2017 at 500-m resolution using MODIS imagery and a phenology-based method (Xin et al., 2020). (2) *China three staple crops 1km* product. This product used the GLASS (Global LAnd Surface Satellite) product and a phenology-based method to produce distribution maps of three staple crops in China from 2000 to 2015 at 1-km resolution (Luo et al., 2020). (3) *APRA500* product, which produced rice distribution maps covering the entire Asian monsoon region from 2000 to 2021 at 500-m resolution using MODIS imagery and a phenology-based method (Han et al., 2022). (4) *Heilongjiang rice map* product, which produced rice distribution maps of Heilongjiang Province every five years from 1990 to 2020 at 30-m resolution using Landsat imagery and a phenology-assisted machine learning method (Zhang et al., 2023a).

2.3 Method

2.3.1 Training sample selection and preprocessing

We extracted training samples from two recent rice distribution map products mentioned in section 2.2.2. For provinces cultivating only single- or double-season rice, this study randomly extracted 5000 rice pixels and 10,000 non-rice cropland pixels each year from the distribution map. For provinces where both single- and double-season rice were cultivated, this study randomly extracted 5000 single-season rice pixels, 5000 double-season rice pixels and 5000 non-rice cropland pixels each year from the distribution map. We further obtained the SWIR1 time series for these pixels on the GEE platform.

Figure 3a–d shows the extracted time series of rice and non-rice crops in four provincial administrative regions in four subregions. Notably, in Jilin Province of subregion I, the SWIR1 reflectance of rice and non-rice crops differs significantly during the transplanting period (DOY 97 to 217) (Fig. 3a). The rice time series first decreased and then increased, showing a "V" shape, while the time series of other crops remained high. In Shanghai City of subregion II, the "V" shape of the rice time series is not as noticeable as in Jilin Province, with smaller differences between the time series of rice and non-rice crops. However, during the entire growing period of rice (DOY 97 to 289), the averaged SWIR1 of rice pixels is always lower than for non-rice crops (Fig, 3b). In Zhejiang and Guangdong of subregions III and IV, the differences between rice and non-rice crops are even smaller, but the SWIR1 is still relatively lower than in non-rice crops (Fig. 3c and d).

Figures 3e–h show the percentages of good observations of the extracted training time series. As Landsat 8 and 9 and Sentinel-2 were used to composite the 8-day training time series, the percentage of good observations was relatively high, averaging 86.38 %, 67.41 %, 56.20 %, and 47.41 % in these four provinces. Considering that the percentage of good observations was much higher in subregion I than in the other three subregions, and combined with the significant differences between rice and non-rice crops during the transplanting period, this study used the time series from DOY 97 to 217 as a model input in subregion I. Since the percentage of good observations in the other three subregions was lower, and the differences between rice and non-rice crops were more evenly distributed throughout the rice growing season, this study used time series of the entire growing season as model input in these three subregions, specifically DOY 97 to 289, DOY 73 to 289, and DOY 33 to 289, respectively.

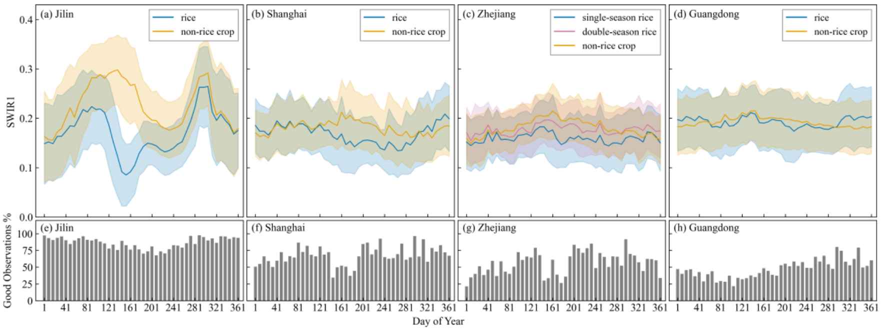


Figure 3: SWIR1 time series of training pixels and the percentage of good quality observations within the time series for Jilin, Shanghai, Zhejiang, and Guangdong. Solid lines indicate the average time series, and the shaded error bands represent the standard deviations.

In addition to these samples, this study also trained the model with some edge pixels to improve its performance. Edge detection was performed for two recent rice distribution products using the Canny algorithm. Pixels at the edges (i.e., pixels adjacent to other cover types) were randomly selected for training. The number of selected edge training pixels was one-tenth of the former pixel sample. Specifically, in provinces where only single- or double-season rice is cultivated, 500 edge rice pixels and 1000 edge non-rice cropland pixels were extracted annually. In provinces where both single- and double-season rice were cultivated, 500 edge pixels were extracted annually for each of category (single-season rice, double-season rice, and non-rice cropland).

During the 1990-2016 study period, most years had observations from only one Landsat satellite, resulting in a lower percentage of good observations in the time series compared to the time series used for training. Due to differences in the frequency of good observations between the study years and the training time series, the training time series could not be used directly in model training. While the traditional solution was to use approaches such as interpolation to fill the gap in the time series, this study chose to reduce the good quality observation in the training time series. Specifically, this study calculated the distribution of the percentage of good observations on cropland pixels in each province for the years 1990 to 2016 and randomly deleted valid observations from the training time series to make the distribution of the percentage of good observations in the training time series consistent with that for the years 1990 to 2016. Finally, in this study, both the original training time series and the reduced training time series were used as the model input.

2.3.2 Classification method

This study used the RF model from Scikit-learn for image classification (Pedregosa et al., 2011). RF is a versatile decision tree-based classification algorithm. It uses an ensemble of trees, each trained on a subset of the data, to improve accuracy and reduce overfitting. By aggregating predictions through majority voting, RF is effective for a wide range of classification tasks.

In addition to being able to output classification results directly, RF can also provide the probability of each class. For binary classification problems, the class with the highest probability (greater than 0.5) is the model's default classification result. This study used the default RF parameters to train the model for rice classification from 1990 to 2016, using the training data generated in the previous section for each provincial administrative region. This study did not directly use the model output for classification, as the direct output varies greatly from year to year. Instead, we used the probability provided by the RF model. Specifically, in provinces where only single- or double-season rice is cultivated, we used the rice probabilities provided by the model. Instead of adopting the default rice probability threshold of 0.5, we used the statistical rice planting area of the province for the year to determine a new threshold of rice probability. The total area of pixels with a rice probability greater than the threshold was consistent with the province-level statistical area, and these pixels were identified as rice. In provinces with both single- and double-season rice, we transformed the multi-classification problem into two binary classification problems. First, we classified rice and non-rice crops, and then we classified single- and double-season rice. Again, the province-level statistical area was used to determine the probability threshold to obtain the classification maps. Note that two rice map products in recent years did not include Tianjin and Hebei due to low rice planting areas. Therefore, this study applied the rice classification model developed for Liaoning Province, which has a similar cultivation context to the two provincial administrative regions. In addition, in some years there were a few pixels that had no high-quality observations during the study period. Given that the model we trained was applied to all years, the probabilities of the model outputs were to some extent comparable between years. Therefore, we filled these pixels with the probabilities of the neighboring years and then used the threshold to generate the rice maps after the filling was complete.

2.3.3 Post-processing of distribution maps

To eliminate fragmented pixels in the classification results, post-processing of the results was performed. Unlike dryland crops such as maize or soybean, which may rotate, once a plot of land is converted to paddy, it is typically used to cultivate rice over a long time. Therefore, isolated pixels that are classified as rice only a few years are likely to be inaccuracies, as they do not conform to the planting norm. This study set a threshold of five years, overlayed the preliminary rice maps to get the rice planting years on each pixel, eliminated the pixels with planting years less than five years, and regenerated the rice maps on the remaining pixels using the province-level planting area and the probability of the model output. However, for the first four years (1990 to 1993) and the last four years (2013 to 2016), some drylands may have been converted to paddy fields or paddy fields may have been converted to drylands in that year. These new paddy fields could be planted with rice, but could not meet the 5-year requirement, so in these years we relaxed the requirement and only required that the planting years be greater than or equal to the number of years between that year and 1990 or 2016.

In addition, the spatial variation of the planting area of the result fluctuated considerably from year to year. Since the annual rice area in an area the size of a county is generally stable, this study further post-processed the results. Specifically, each province was divided into several grids of 1000 pixels by 1000 pixels, and the rice area within each grid was calculated from each year's result. Then, a low-pass filter (five-year sliding average) was applied to the time series of rice area within each grid. The filtered rice areas eliminated unreasonable fluctuations in rice area from year to year. We used the filtered rice areas to redetermine the threshold of rice probability of each year within each grid, following the method mentioned in section 2.3.2, and regenerated the rice maps of each year within each grid using these thresholds.

2.3.4 Accuracy assessment

This study used validation samples and agricultural statistical data mentioned in section 2.2.3 to validate the results. We compared the result with the validation samples and calculated the confusion matrices and three metrics, user's accuracy (UA), producer's accuracy (PA), and overall accuracy (OA) to evaluate the accuracy of the result. UA indicates the percentage of correctly classified rice samples among all rice samples, PA indicates the percentage of correctly classified rice samples among all samples identified as rice, and OA indicates the percentage of correctly classified samples among all samples.

Since the validation sample could not cover too many years, especially before 2002, we further compared the result with the agricultural statistical data. A linear regression method was used to measure the relationship between the identified area and the statistical area, and the coefficient of determination (R^2) and the relative mean absolute error (RMAE) were calculated to measure the accuracy. The calculation of the RMAE was as follows:

$$\mathrm{RMAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} |SA_i - IA_i|}{\sum_{\mathrm{i}=1}^{\mathrm{n}} SA_i} \tag{1}$$

where SA_i and IA_i are the statistical and identified areas of the ith county, respectively. n indicates the number of counties in the investigated province.

2.3.5 Sensitivity analysis

In order to demonstrate the effectiveness of the data preprocess strategies mentioned in section 2.3.1, we conducted several experiments comparing them to the current preprocess strategies. We conducted these experiments in Jiangsu Province, using the current preprocessing strategies as the control group. The following four experimental groups were designed. Experimental group I used more accurate training sample points. Specifically, we overlayed the six-year distribution maps from Shen et al. (2023a) and randomly selected training points on pixels identified as rice in all six years and pixels identified as non-rice in all six years. Pixels identified as a certain type in all six years are less likely to be misidentified, which can be considered as more accurate training samples. Experimental group II did not delete data from the training sample and trained the model directly using the time series on the training points from 2017 to 2022. Experimental groups III–V all filled in missing values in the time series. The filling was done by linear interpolation and the time series were smoothed with a

Savitzky-Golay (SG) filter (Savitzky and Golay, 1964). Experimental group III performed the filling directly on the training samples and the time series used for prediction. Experimental group IV, on the other hand, first deleted observations from the training time series using the same method as the control group, and then filled in the missing values in both the training time series and the time series used for prediction. Experimental group V randomly selected 50 time series from the filled rice time series to synthesize standard rice time series and used the TWDTW method described in Shen et al. (2023a) to generate rice distribution maps. All other steps and post-processing for all experimental groups were kept consistent with the control group.

3. Results

3.1 Accuracy of rice distribution maps

The validation samples were used to verify the accuracy of the distribution maps. The distribution maps achieved high accuracy in almost all provincial administrative regions (Table 2). Specifically, the user's accuracy, producer's accuracy, and overall accuracy for rice averaged 88.40 %, 89.10 %, and 90.26 %, respectively, across all 25 provincial administrative regions. Liaoning had the highest user's accuracy at 98.55 %, while Anhui had the lowest at 61.08 %. Ningxia had the highest producer's accuracy at 99.40 %, while Hainan had the lowest at 70.00 %. The highest overall accuracy was in Liaoning at 96.63 %, while the lowest was in Anhui at 75.49 %.

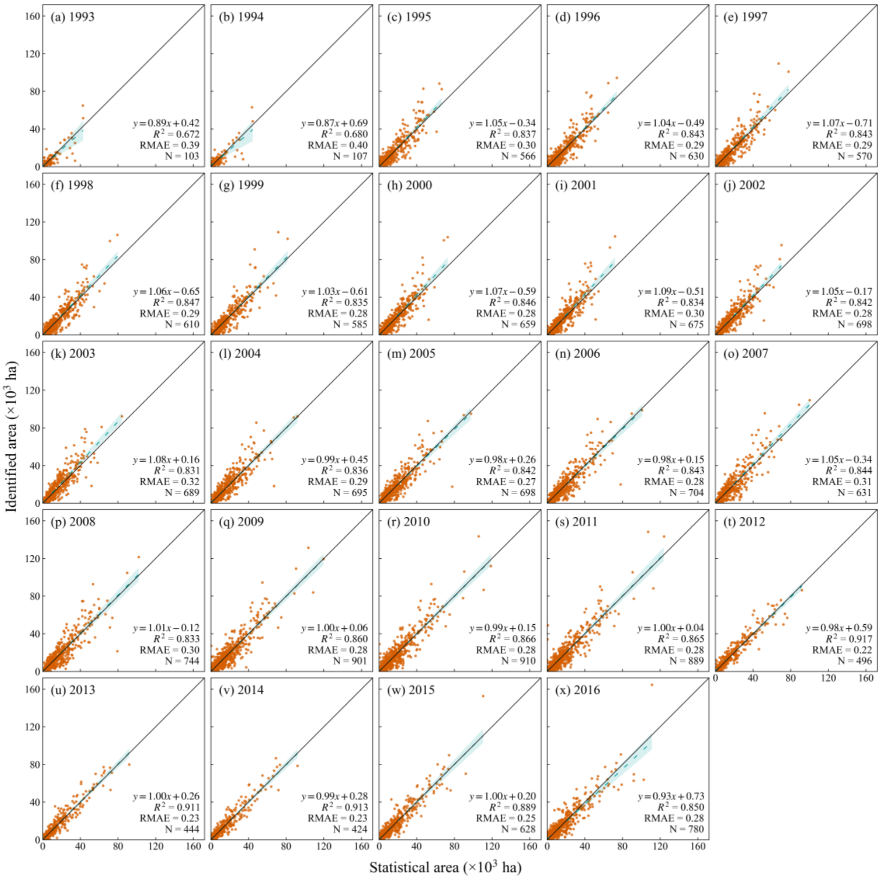
Table 2: Confusion matrices of the rice distribution maps in 25 provincial administrative regions.

Province	Class	Rice[a]	Other	UA (%)	PA (%)	OA (%)
Heilongjiang	Rice[b]	66083	5602	98.15	92.19	93.71
	Other	1246	35969	86.52	96.65	
Jilin	Rice	61877	2380	89.07	96.30	90.94
	Other	7595	38231	94.14	83.43	
Liaoning	Rice	2448	237	98.55	91.17	96.63
	Other	36	5384	95.78	99.34	
Hebei	Rice	1407	296	90.77	82.62	89.14
	Other	143	2196	88.12	93.89	
Inner Mongolia	Rice	2246	64	87.77	97.23	92.61
	Other	313	2479	97.48	88.79	
Ningxia	Rice	2480	15	80.47	99.40	90.99
	Other	602	3754	99.60	88.79	
Tianjin	Rice	4023	51	82.30	98.75	92.49
	Other	865	7253	99.30	89.34	
Jiangsu	Rice	5775	926	90.29	86.18	89.72
	Other	621	7730	89.30	92.56	
Sichuan	Rice	1399	349	73.79	80.03	78.00
	Other	497	1601	82.10	76.31	

Yunnan	Rice	2376	78	92.52	96.82	94.33
	Other	192	2112	96.44	91.67	
Chongqing	Rice	502	65	95.80	88.54	93.31
	Other	22	711	91.62	97.00	
Guizhou	Rice	821	46	91.32	94.69	93.53
	Other	78	972	95.48	92.57	
Henan	Rice	1241	212	82.40	85.41	89.26
	Other	265	2725	92.78	91.14	
Shanghai	Rice	9325	1989	69.69	82.42	84.78
	Other	4055	24341	92.45	85.72	
Shaanxi	Rice	2703	279	94.74	90.64	91.61
	Other	150	1981	87.65	92.96	
Shandong	Rice	3109	265	91.20	92.15	94.09
	Other	300	5892	95.70	95.16	
Hunan	Rice	526	130	95.12	80.18	92.60
	Other	27	1440	91.72	98.16	
Jiangxi	Rice	1726	118	96.53	93.60	93.45
	Other	62	843	87.72	93.15	
Hubei	Rice	2882	148	92.85	95.12	91.90
	Other	222	1314	89.88	85.55	
Anhui	Rice	507	76	61.08	86.96	75.49
	Other	323	722	90.48	69.09	
Zhejiang	Rice	2608	240	90.71	91.57	91.48
	Other	267	2839	92.21	91.40	
Guangxi	Rice	1441	140	98.83	91.14	95.91
	Other	17	2242	94.12	99.25	
Fujian	Rice	656	119	92.79	84.65	88.13
	Other	51	606	83.59	92.24	
Guangdong	Rice	4406	1127	83.27	79.63	92.61
	Other	885	20795	94.86	95.92	
Hainan	Rice	441	189	90.00	70.00	79.73
	Other	49	495	72.37	90.99	

[a] number of visually interpreted samples. [b] number of identified samples.

Compared with county-level statistical data, the distribution maps also achieved high performances in both single- and double-season rice. Specifically, the identified area of single-season rice in this study showed a strong linear correlation with the statistical area, with scatters all close to the 1:1 line in all years included in the comparison (Fig. 4). The slopes of the regression lines between identified and statistical areas ranged from 0.87 to 1.09, with an average slope of 1.01, and the R^2 values ranged from 0.67 to 0.92, with an average of 0.84. The distribution maps also accurately represent the spatial variation of double-season rice. There are strong linear correlations between the identified double-season rice area and the county-level statistical area for all years (Fig. 5). The slopes ranged from 0.63 to 1.06, with an average of 0.87, and the R^2 values ranged from 0.67 to 0.91, with an average of 0.80.



295 **Figure 4: Comparison between identified single-season rice area and county-level statistics for each year. Solid lines are 1:1 lines and dashed lines are regression lines. The confidence intervals are shaded in blue. N indicates the number of counties included in the comparison.**

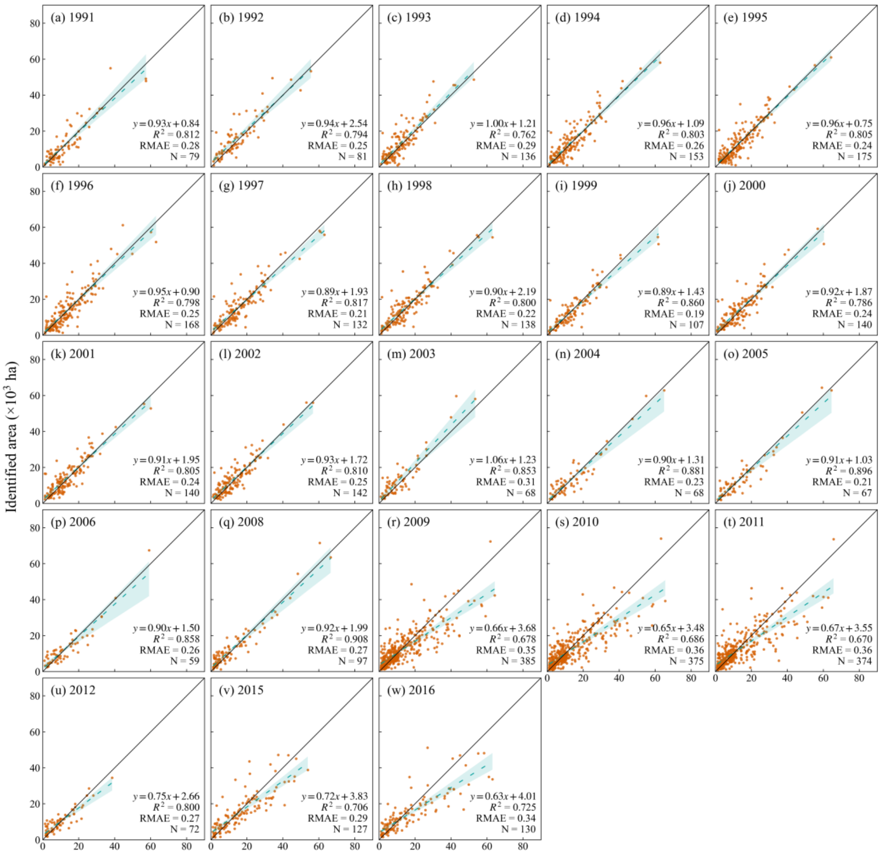
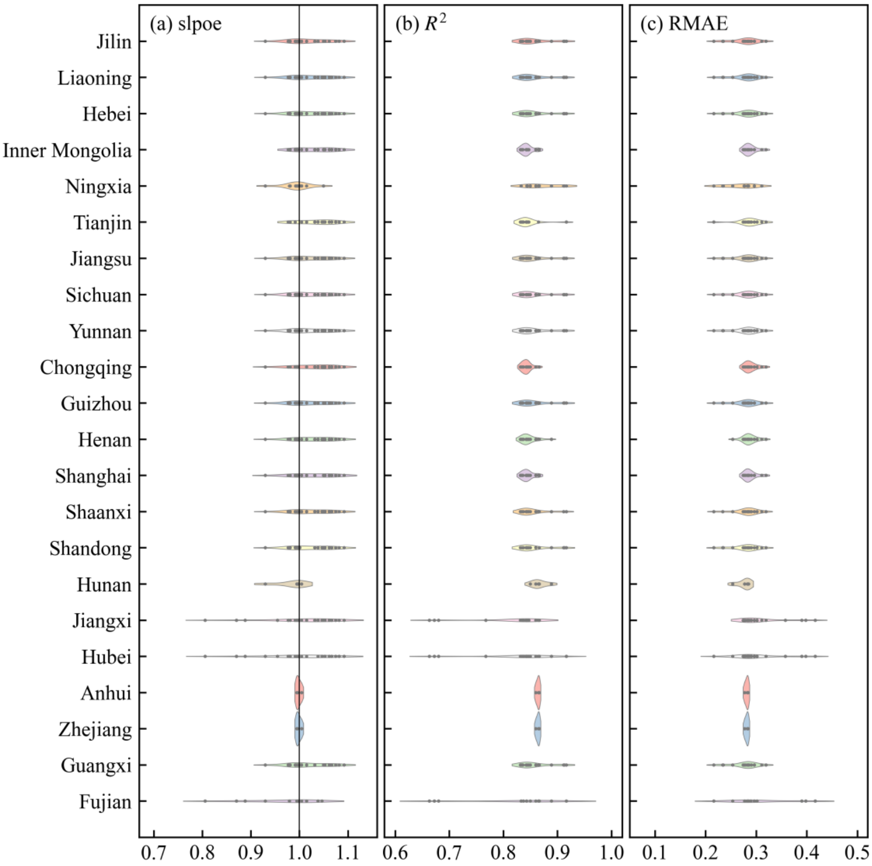


Figure 5: Comparison between identified double-season rice area and county-level statistics for each year. Solid lines are 1:1 lines and dashed lines are regression lines. The confidence intervals are shaded in blue. N indicates the number of counties included in the comparison.

The distribution maps achieved high accuracy in all provincial administrative regions. For single-season rice, the average slopes of all years in each province ranged from 0.96 to 1.03, the averaged R^2 values ranged from 0.81 to 0.87, and the averaged RMAE ranged from 0.27 to 0.31 (Fig. 6). The identification accuracies in Jiangxi, Hubei, and Fujian varied considerably from year to year, with standard deviations of R^2 values of 0.06, 0.07, and 0.09, respectively. For double-season rice, the accuracies were slightly lower than those for single-season rice. The average slopes in each province ranged from 0.69 to 0.91, the R^2 values ranged from 0.71 to 0.83, and RMAE ranged from 0.27 to 0.37 (Fig. 7).

https://doi.org/10.5194/essd-2024-147
Preprint. Discussion started: 9 July 2024
© Author(s) 2024. CC BY 4.0 License.



310 **Figure 6: Comparison between identified single-season rice planting area and county-level statistics by provincial administrative region each year.**

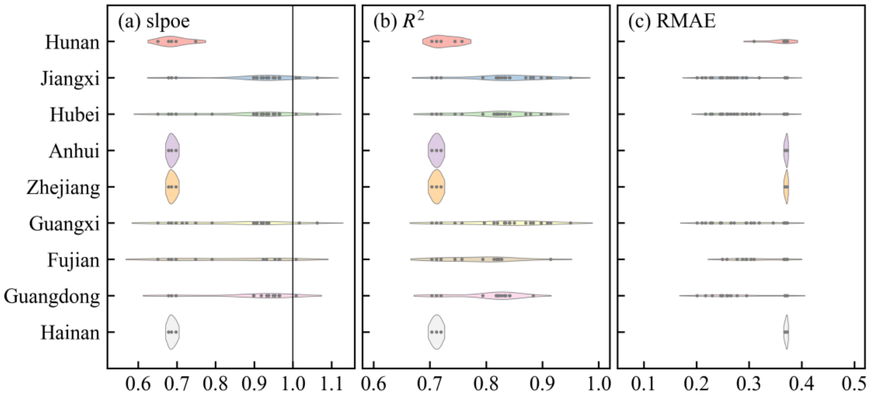


Figure 7: Comparison between identified double-season rice planting area and county-level statistics by provincial administrative region each year.

3.2 Planting frequency

In this study, the rice maps produced for 25 provincial administrative regions in mainland China from 1990 to 2016 accurately reflect the distribution of rice cultivation in China during the 27-year period (Fig. 8). Rice cultivation in the Northeast, Yangtze-Huaihe, and Southwest was dominated by single-season rice, and the planting frequency is lower than that in the Southeast provinces, where double-season rice is cultivated. The lowest average planting frequency was 11.21 in Chongqing, and the highest average planting frequency was 30.89 in Jiangxi (Fig. 8). For single-season rice, the highest average planting frequency of single-season rice was 19.31 years, in Liaoning, and the lowest was in Guangxi, only 5.21 years (Fig. 9). For double-season rice, Jiangxi was the province with the highest planting frequency, at 14.29 years, while Anhui had the lowest planting frequency, at 7.24 years (Fig. 10).

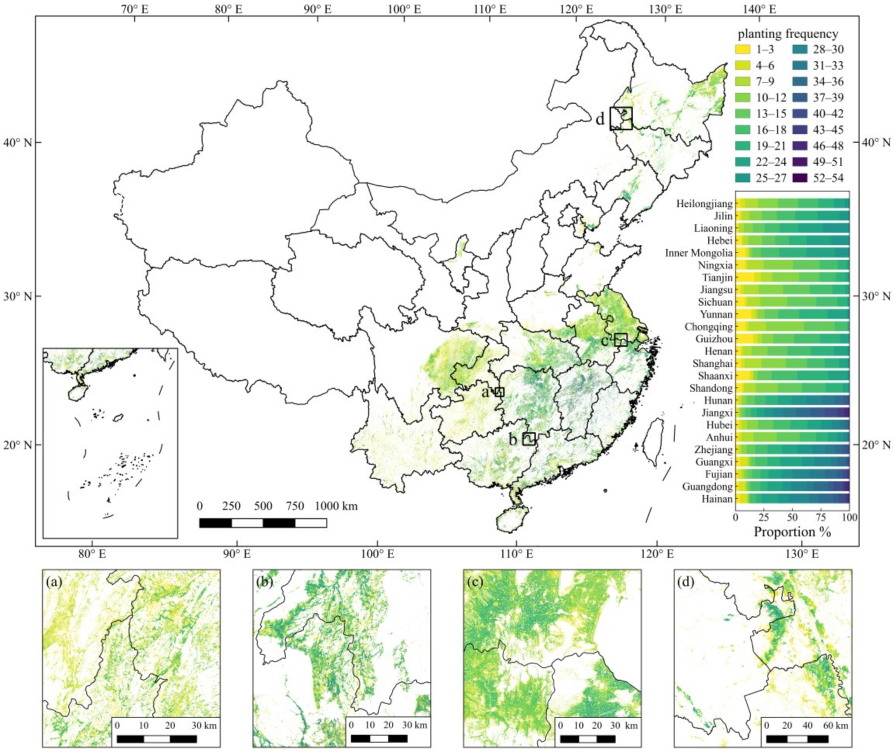


Figure 8: Planting frequency of rice from 1990 to 2016. Panels a–d on the bottom are the zoomed-in maps.

https://doi.org/10.5194/essd-2024-147
Preprint. Discussion started: 9 July 2024
© Author(s) 2024. CC BY 4.0 License.

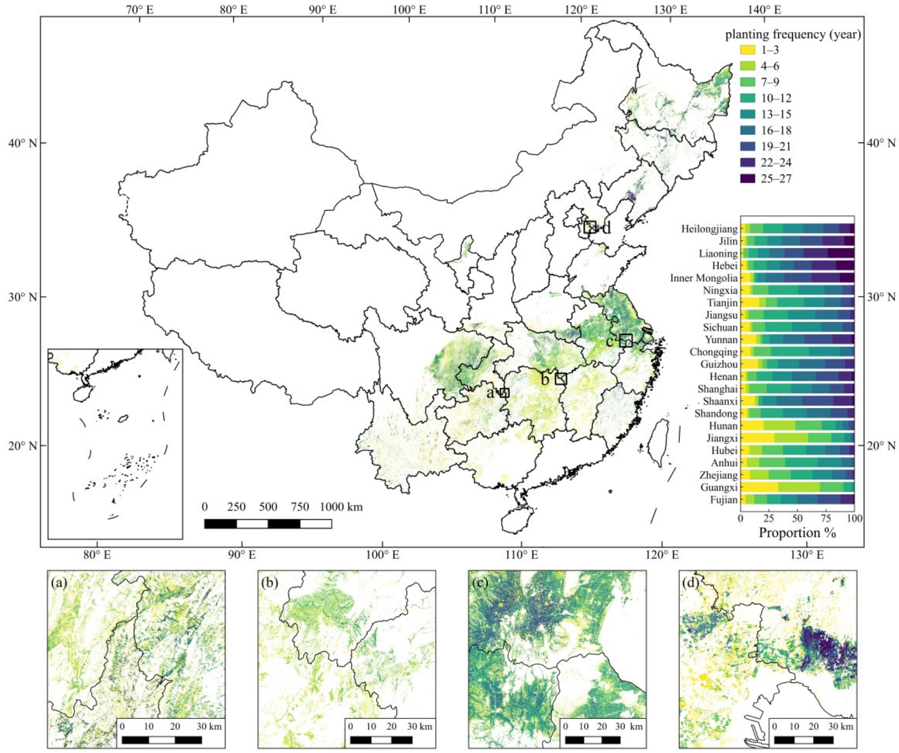
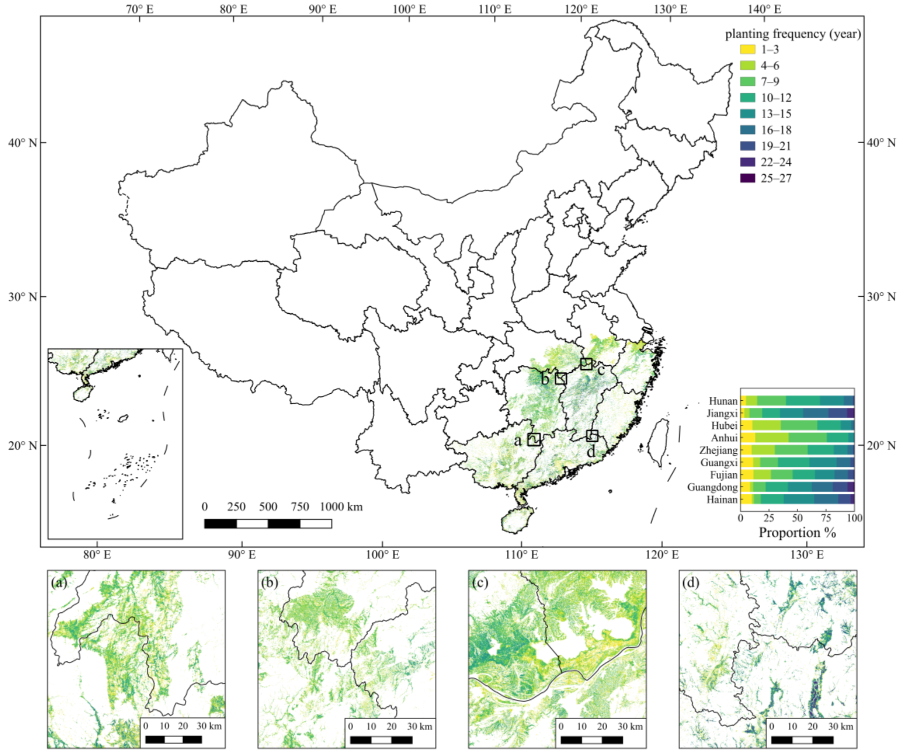


Figure 9: Planting frequency of single-season rice from 1990 to 2016. Panels a–d on the bottom are the zoomed-in maps.



© Author(s) 2024. CC BY 4.0 License.

Figure 10: Planting frequency of double-season rice from 1990 to 2016. Panels a–d on the bottom are the zoomed-in maps.

3.3 Comparison with existing products

To demonstrate the ability of our products to depict the details of rice fields, very high-resolution images obtained from Google Earth were used to compare the actual distribution of rice with the distribution map in four small areas in Heilongjiang, Jilin, Shanghai, and Guangdong. The images were taken in 2010, 2007, 2004, and 2010, respectively. This study compared the four existing products mentioned in section 2.2.4 on these four small areas (Fig. 11). To facilitate the comparison, we visually interpreted the images and labeled the rice fields (Fig. 11a2–d2). The result showed high performance in all four small areas, accurately reflecting rice cultivation patterns (Fig. 11a3–d3). The *NEAsia_Rice* product was able to roughly reflect the distribution of rice cultivation in both small areas but was limited in its ability to portray the details of paddy fields due to its spatial resolution (Fig. 11a4–b4). The *China three staple crops 1km* product differs significantly from the actual rice field distribution in all four small areas (Fig. 11a5–d5). The *APRA500* product roughly reflects the rice planting distribution in the first three study areas but fails to do so in the fourth area (Fig. 11a6–d6). In contrast, the *Heilongjiang rice map* product provides a detailed portrayal of rice field distribution (Fig. 11a7).

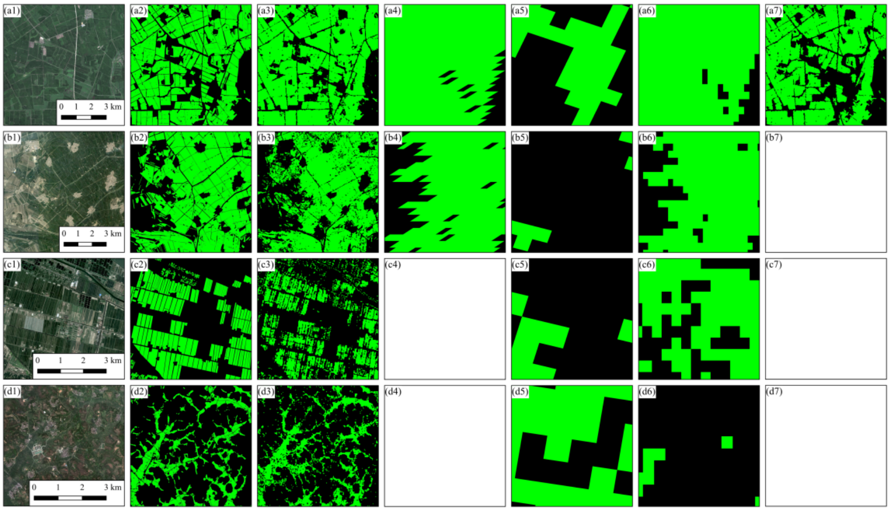
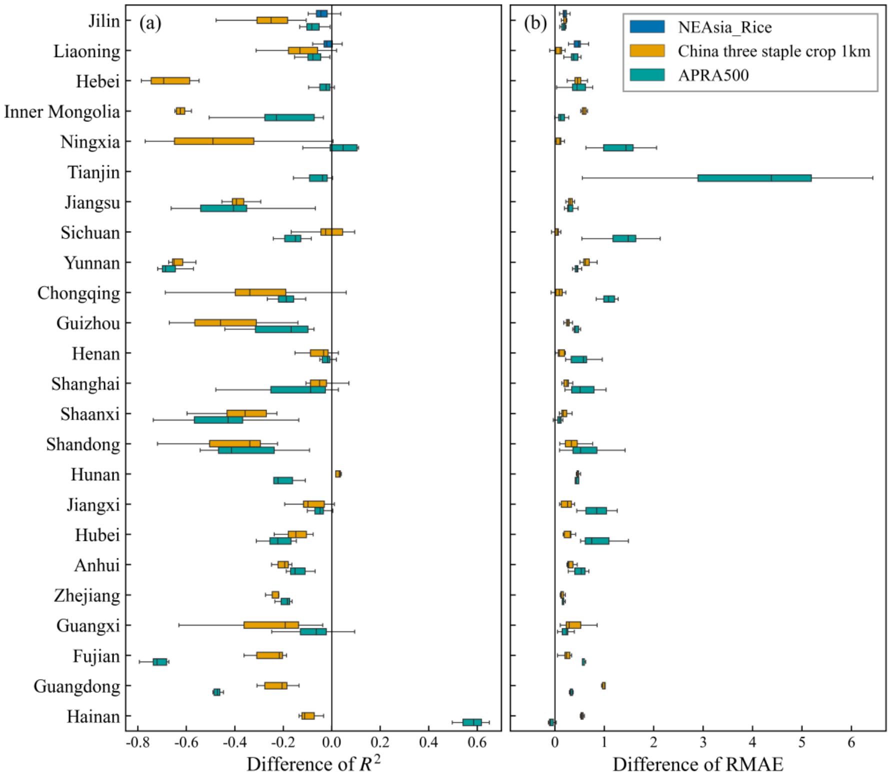


Figure 11: Comparison of this study with four other studies on four small areas located in Heilongjiang (45°52′11″ N, 132°52′16″ E), Jilin (45°17′52″ N, 124°37′9″ E), Shanghai (31°42′46″ N, 121°28′24″ E), and Guangdong (21°25′30″ N, 110°36′0″ E), respectively. The first column shows very high-resolution imagery obtained from © Google Earth, with image acquisition dates of 24 July 2010, 11 June 2007, 21 July 2004, and 2 September 2010. The second column shows visually interpreted results. The third to seventh columns show the classification maps from this study, the NEAsia_Rice product, the China three staple crop 1km product, the APRA500 product, and the Heilongjiang rice map product, respectively. Blank panels indicate that the product did not have a classification map for that area.

In addition to the higher spatial resolution, the accuracy of the distribution maps of this study was also superior to that of existing products. We validated the existing products using statistical data on rice field area, which is calculated as the sum of

the planting area of single- and double-season rice, as not all products distinguished between single- and double-season rice. However, the Heilongjiang rice map product could not be validated due to the unavailability of statistics in Heilongjiang. Compared with the statistical rice planting area, the distribution maps of this study had a higher R^2 value and a lower RMAE than three existing products in most provinces and years (Fig. 12). Specifically, the R^2 values of the maps from this study with
355 statistical data were higher than the *NEAsia_Rice* product in 70.59 % of the years and provinces, higher than the *China three staple crop 1km* product in 93.44 % of the provinces and years, and higher than the *APRA500* product in 92.41 % of the years and provinces. Meanwhile, the RMAE of this study is lower than the *NEAsia_Rice* product in all years and provinces, lower than the *China three staple crop 1km* product in 95.75 % of provinces and years, and lower than the *APRA500* product in 97.93 % of years and provinces.



360

Figure 12: Differences in R^2 and RMAE of the comparison with the statistical data between three exist maps and our distribution maps, respectively.

Additionally, we compared the accuracy of our map with that of existing products using validation samples. We only compared our map with the *Heilongjiang rice map* product due to the lower spatial resolution of the other products, which
365 made them unsuitable for validation with 30 m resolution samples. The comparison in Heilongjiang in 2010 showed that our map achieved similar accuracy to the *Heilongjiang rice map* product (Table 3). The UA of our map was higher, while PA and

OA were slightly lower than those of the *Heilongjiang rice map* product.

Table 3: Confusion matrices of the distribution map of our map and *Heilongjiang rice map* product in Heilongjiang 2010.

Product	Class	Rice[a]	Other	UA (%)	PA (%)	OA (%)
Our map	Rice[b]	58064	5373	98.83	91.53	93.06
	Other	687	23251	81.23	97.13	
Heilongjiang rice map	Rice	55221	1661	93.99	97.08	94.06
	Other	3530	26963	94.20	88.42	

[a] number of visually interpreted samples. [b] number of identified samples.

4. Discussion

4.1 Superiority of CCD-Rice dataset

Although rice is one of the most important crops in China, few long-term rice map products are available due to various challenges. The main difficulties in long-term rice mapping stem from two factors: mapping methods and quality of remote sensing data. Machine learning methods require a large volume of training samples; transferring the model between years is an additional challenge (Millard and Richardson, 2015; Belgiu and Csillik, 2018). The most accurate approach is to collect samples for model training every year. However, the precise planting situations in past years are difficult to collect through field surveys, even if farmers are asked about their past plantings. On the other hand, knowledge-based methods such as phenology-based approaches may require no or very little training data. However, while these methods are not limited by transferability between years, they are more affected by the quality of observations (Dong et al., 2016; Shen et al., 2023a). Southern China is cloudier and rainier, resulting in less reliable observations of optical remote sensing data (Li and Chen, 2020). High spatial resolution satellites, such as Landsat, typically have lower temporal resolutions, which can exacerbate the effects of missing data (Li et al., 2024). This data limitation hinders the application of both methods, especially for phenology-based methods that rely on irrigation signals during the transplanting period. Therefore, existing high-resolution long-term rice distribution maps were limited to less cloudy and rainy areas such as Northeastern China (You et al., 2021; Zhang et al., 2023a). Medium-resolution optical satellites usually have high temporal resolution and can largely reduce the probability of not having cloud-free observations during rice transplantation. As a result, many studies use medium-resolution optical satellites to produce rice distribution products (Xiao et al., 2005; Zhang et al., 2017; Luo et al., 2020; Han et al., 2022). However, their low spatial resolution precludes these products from accurately depicting the details of rice cultivation. When compared with statistical data, the accuracies of these products are lower than those in this study (Fig. 11 and 12).

Although rice differs most from other crops during the transplanting period, there are spectral characteristics that distinguish it from other crops during other stages of rice growth (Fig. 3). These characteristics have been observed in some previous studies, and some of the studies have also utilized images of the entire growing season (Shen et al., 2023a; Xuan et al., 2023; Zhang et al., 2023a). This study also utilized remote sensing images throughout the entire growing season of rice in southern China, rather than just during the transplanting period, resulting in more usable images for rice classification. Such a strategy allowed this study to achieve high-resolution rice mapping in southern China using only Landsat data.

In conclusion, previous studies have failed to achieve long-term, high-resolution rice mapping in China. Compared with previous products, our product has the advantages of wide coverage (all of China), high resolution (30 m), long-term (27 years), and differentiation of cropping systems (single- and double-season rice). Furthermore, this product contributes to the China Crops Dataset (CCD), following CCD-Maize and CCD-Wheat (Shen et al., 2022; Peng et al., 2023; Dong et al., 2020, 2024). Together, the three datasets form a long-term, high-resolution distribution dataset of the three major staple crops in China, providing crucial data support for crop research in China.

4.2 Sensitivity analysis of the classification method

In this study, unique strategies for training sample selection and preprocessing differed from common practice. Specifically, to overcome the limitation of insufficient training samples required for machine learning methods, this study obtained a large volume of training samples from two recent rice maps mentioned in section 2.2.2. The samples obtained from the recent rice maps are more evenly distributed in the study area than those from the field surveys, and cover both rice and non-rice land covers throughout the region. In this study, valid data were randomly deleted from the training data to simulate the effect of cloud contamination on the observations and improve the ability of the model to be transferred to previous years. However, it has not yet been demonstrated whether these two strategies are effective. Firstly, there is some uncertainty in the two recent rice maps, and random sampling may result in many mistakenly labeled training samples. This study also selected some training samples at the edges of the rice maps, which may include erroneous data. Additionally, instead of filling the missing values in the time series through interpolation as done in previous studies, valid observations were randomly deleted. However, it is unclear whether this deletion strategy effectively aids in the model's transferability to other years.

To test the effectiveness of our current preprocessing strategies, we designed several experiments, as elaborated in section 2.3.5, and validated the identification results for each experimental group using county-level statistical data. The average R^2 for the control group and for the five experimental groups were 0.85, 0.75, 0.49, 0.74, 0.81, and 0.56, respectively (Fig. 13a). The average RMAE for the control group and the five experimental groups were 0.16, 0.21, 0.40, 0.25, 0.20, and 0.26, respectively (Fig. 13c). In almost all years, the control group had the highest R^2 and the lowest RMAE (Fig. 13b and d). The results of validation using the validation sample were the same. The overall accuracy of the control group was higher than that of any of the experimental groups (Table 3).

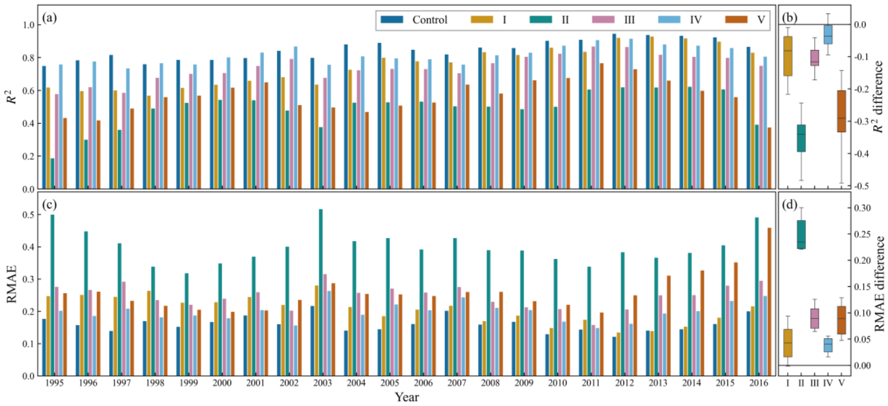


Figure 13: Comparison between identified single-season rice planting area and county-level statistics of the control group and five experimental groups each year. Panels a and c are the R^2 and RMAE of the comparison, respectively. Panels b and d are the differences in R^2 and RMAE between the five experimental groups and the control group, respectively.

Table 3: Confusion matrices of the distribution maps of the control group and five experimental groups.

Groups	Class	Rice[a]	Other	UA (%)	PA (%)	OA (%)
Control group	Rice[b]	5775	926	90.29	86.18	89.72
	Other	621	7730	89.30	92.56	
Experimental group I	Rice	5778	1110	90.34	83.89	88.52
	Other	618	7546	87.18	92.43	
Experimental group II	Rice	2719	1842	42.51	59.61	63.33
	Other	3677	6814	78.72	64.95	
Experimental group III	Rice	4714	1113	73.70	80.90	81.43
	Other	1682	7543	87.14	81.77	
Experimental group IV	Rice	5087	1283	79.53	79.86	82.78
	Other	1309	7373	85.18	84.92	
Experimental group V	Rice	2238	1043	34.99	68.21	65.45
	Other	4158	7613	87.95	64.68	

[a] number of visually interpreted samples. [b] number of identified samples.

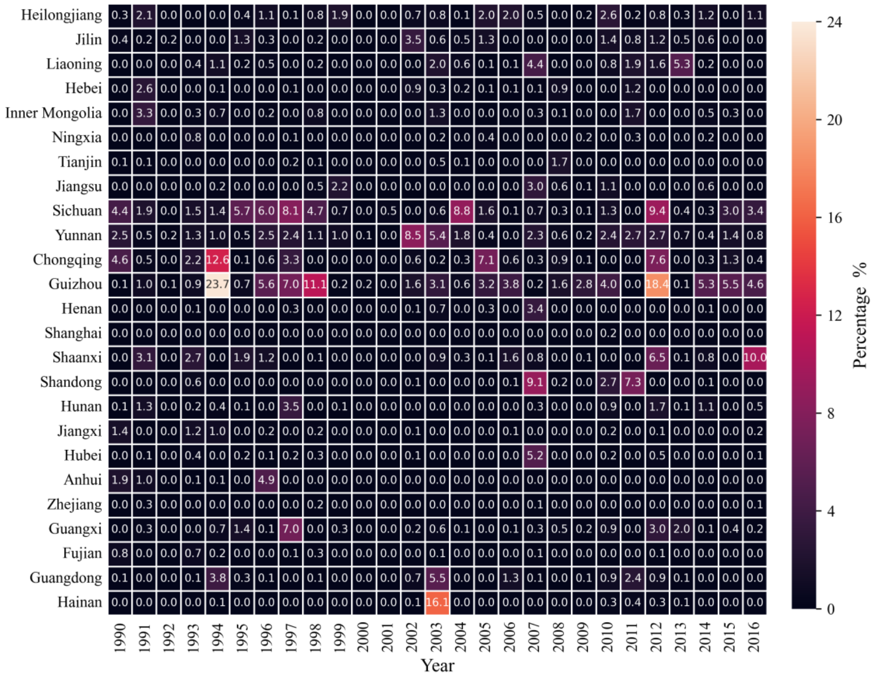
Some previous studies have adopted strategies to improve the accuracy of training samples during the selection process (Zhang and Roy, 2017; Wen et al., 2022). Some of these strategies are: selecting only pixels that remain constant across years, selecting only pixels whose neighboring pixels are all the same type, or selecting only pixels at the center of patches. These strategies improve the accuracy of the samples to a great extent and avoid including erroneous samples. However, this sampling method may reduce sample diversity to some extent. Pixels that have undergone land cover changes or are situated at the edges are excluded from model training, which weakens the ability of the model for such pixels. There are also some studies that suggest that more diverse samples help to improve the accuracy of the model when selecting training samples (Fu et al., 2023). The comparison with Experimental Group I indicates that more diverse training samples improve the performance of the classification model (Fig. 13 and Table 3). This improvement may be because pixels located at the image edges are more likely to have features in the feature space that are close to the classification decision boundary.

Time series analysis generally requires complete series. Previous studies typically perform gap filling and filtering to preprocess time series of remote sensing images. This study diverged from previous studies by adding missing values into the time series. Comparisons between the control and experimental group II, as well as between experimental groups III and IV, demonstrate that adding missing values to the time series indeed improves model performance (Fig. 13 and Table 3). This improvement is attributed to the composite training time series of recent years using Landsat and Sentinel-2 data, which have significantly fewer missing values compared to previous years. The model trained with such training data could not make correct predictions for past time series with more missing values. The results of the control group comparing experimental groups III and IV show that using the time series after filling in the missing values resulted in lower accuracy (Fig. 13 and Table 3). This may be because improperly filling missing values may not introduce new information and instead inject noise into the training samples, which, in turn, reduces the model's performance. The comparison with the experimental group V demonstrates that the phenology-based rice classification method is not applicable in the case of poor optical observations (Fig. 13 and Table 3).

4.3 Uncertainties

The model results were post-processed. Pixels with no good optical observation during the study period were filled with values from neighboring years. For most years and most provinces, the percentage of filled pixels was less than 1 % (Fig. 14). In several years in Guizhou, Chongqing, and Sichuan, the quality of the optical observations was poor, and the percentage of filled pixels was high, exceeding 5 %, which would increase the error of the product to some extent (Fig. 14). Paddy fields do not have the same flexibility to grow a wide range of crops as drylands, so filling with results from neighboring years is considered a desirable solution.

© Author(s) 2024. CC BY 4.0 License.



	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Heilongjiang	0.3	2.1	0.0	0.0	0.0	0.4	1.1	0.1	0.8	1.9	0.0	0.0	0.7	0.8	0.1	2.0	2.0	0.5	0.0	0.2	2.6	0.2	0.8	0.3	1.2	0.0	1.1
Jilin	0.4	0.2	0.2	0.0	0.0	1.3	0.3	0.0	0.2	0.0	0.0	0.0	3.5	0.6	0.5	1.3	0.0	0.0	0.0	0.0	1.4	0.8	1.2	0.5	0.6	0.0	0.0
Liaoning	0.0	0.0	0.0	0.4	1.1	0.2	0.5	0.0	0.2	0.0	0.0	0.0	0.0	2.0	0.6	0.1	0.1	4.4	0.0	0.0	0.8	1.9	1.6	5.3	0.2	0.0	0.0
Hebei	0.0	2.6	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.9	0.3	0.2	0.1	0.1	0.1	0.9	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0
Inner Mongolia	0.0	3.3	0.0	0.3	0.7	0.0	0.2	0.0	0.8	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.3	0.1	0.0	0.0	1.7	0.0	0.0	0.5	0.3	0.0
Ningxia	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.4	0.0	0.0	0.0	0.2	0.0	0.3	0.0	0.0	0.2	0.0	0.0
Tianjin	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.5	0.1	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jiangsu	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.5	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.6	0.1	1.1	0.0	0.0	0.0	0.6	0.0	0.0
Sichuan	4.4	1.9	0.0	1.5	1.4	5.7	6.0	8.1	4.7	0.7	0.0	0.5	0.0	0.6	8.8	1.6	0.1	0.7	0.3	0.1	1.3	0.0	9.4	0.4	0.3	3.0	3.4
Yunnan	2.5	0.5	0.2	1.3	1.0	0.5	2.5	2.4	1.1	1.0	0.1	0.0	8.5	5.4	1.8	0.4	0.0	2.3	0.6	0.2	2.4	2.7	2.7	0.7	0.4	1.4	0.8
Chongqing	4.6	0.5	0.0	2.2	12.6	0.1	0.6	3.3	0.0	0.0	0.0	0.0	0.6	0.2	0.3	7.1	0.6	0.3	0.9	0.1	0.0	0.0	7.6	0.0	0.3	1.3	0.4
Guizhou	0.1	1.0	0.1	0.9	23.7	0.7	5.6	7.0	11.1	0.2	0.2	0.0	1.6	3.1	0.6	3.2	3.8	0.2	1.6	2.8	4.0	0.0	18.4	0.1	5.3	5.5	4.6
Henan	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.1	0.7	0.0	0.3	0.0	3.4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Shanghai	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Shaanxi	0.0	3.1	0.0	2.7	0.0	1.9	1.2	0.0	0.1	0.0	0.0	0.0	0.0	0.9	0.3	0.1	1.6	0.8	0.0	0.1	0.0	0.0	6.5	0.1	0.8	0.0	10.0
Shandong	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	9.1	0.2	0.0	2.7	7.3	0.0	0.0	0.1	0.0	0.0
Hunan	0.1	1.3	0.0	0.2	0.4	0.1	0.0	3.5	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.9	0.0	1.7	0.1	1.1	0.0	0.5
Jiangxi	1.4	0.0	0.0	1.2	1.0	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.2
Hubei	0.0	0.1	0.0	0.4	0.0	0.2	0.1	0.2	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	5.2	0.0	0.0	0.2	0.0	0.5	0.0	0.0	0.0	0.1
Anhui	1.9	1.0	0.0	0.1	0.1	0.0	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Zhejiang	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Guangxi	0.0	0.3	0.0	0.0	0.7	1.4	0.1	7.0	0.0	0.3	0.0	0.0	0.2	0.6	0.1	0.0	0.1	0.3	0.5	0.2	0.9	0.0	3.0	2.0	0.1	0.4	0.2
Fujian	0.8	0.0	0.0	0.7	0.2	0.0	0.0	0.1	0.3	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Guangdong	0.1	0.0	0.0	0.1	3.8	0.3	0.1	0.0	0.1	0.0	0.0	0.0	0.7	5.5	0.0	0.0	1.3	0.1	0.0	0.1	0.9	2.4	0.9	0.1	0.0	0.0	0.0
Hainan	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	16.1	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.4	0.3	0.1	0.0	0.0	0.0

Figure 14: Percentage of filled pixels to cropland pixels in each year in each provincial administrative region.

Rice mapping research has historically been constrained by the quality of optical remote sensing data. In this study, a new rice mapping method was developed to improve the temporal transferability of the classification model through suitable preprocessing and to enhance the robustness of the classification model against missing values in the time series. However, the method used in this study is still relatively simple and does not truly enable the model to understand the missing values in the time series. Several studies have pointed out that some deep-learning methods yield better results when handling time series data with missing values (Che et al., 2018). In addition, the method does not completely solve the influence of low-quality optical remote sensing data. There is a small proportion of pixels with zero good observations that need to be filled with neighboring years, which introduces uncertainty into the results. Many recently developed data fusion methods can combine the advantages of multi-source remote sensing data to provide more reliable time series with valid information for crop classification (Li et al., 2024; Meng et al., 2024). We hope that these advances will further address the limitations of optical remote sensing data quality and produce more accurate rice classification products.

5. Data availability

The distribution maps of rice in China from 1990 to 2016 (CCD-Rice) are publicly available on https://doi.org/10.57760/sciencedb.15865 (Shen et al., 2024a). The file format of the product is GeoTIFF with the spatial reference of WGS84 (EPSG:4326). Alternatively, the distribution maps can be viewed through a Google Earth Engine App using the following link: https://ee-shenrq.projects.earthengine.app/view/ccd-rice. The validation samples are available on https://doi.org/10.6084/m9.figshare.25515019.v1 (Shen et al., 2024b). The file format of the validation samples is GeoParquet, and the geometries are Polygons.

6. Code availability

The codes used to produce the CCD-Rice product is publicly available on https://github.com/shenrq/CCD-Rice.

7. Conclusions

In this study, a new optical satellite-based rice mapping method was developed by combining a machine learning model with appropriate data preprocessing strategies to address the challenges of cloud contamination and missing data in optical remote sensing observations. Using this method, this study produced the first long-term (1990–2016), high-resolution (30 m) paddy rice distribution dataset in China. The distribution maps captured the spatiotemporal changes of single- and double-season rice cultivation across 25 provincial administrative regions in mainland China. Validation using 391,659 validation samples and 20,759 agricultural statistical records showed high accuracy, with an average overall accuracy of 90.26 % and strong correlations between mapped and statistical areas, with an average R^2 of 0.84 and 0.80 for single- and double-season rice, respectively. This study also demonstrated the validity of the methodology by comparing different preprocessing strategies, including training sample selection strategies, and missing value filling strategies in the time series. Overall, the distribution maps produced in this study demonstrate good accuracy and provide a comprehensive and reliable dataset for monitoring long-term changes in rice cultivation in China, and provide strong data support for food security, sustainable agriculture, and other related studies.

Author contributions

RS and WY conceptualized the study. RS and QP performed the investigation. RS and XL developed the method. RS implemented the computer code, performed the formal analysis, validation, and visualized the results, and wrote the manuscript. WY, XC, and QP edited and revised the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Financial support

This study was supported by the National Natural Science Foundation of China (42141020).

References

Belgiu, M. and Csillik, O.: Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis, Remote Sensing of Environment, 204, 509–523, https://doi.org/10.1016/j.rse.2017.10.005, 2018.

Bouman, B. A. M., Humphreys, E., Tuong, T. P., and Barker, R.: Rice and Water, in: Advances in Agronomy, vol. 92, Elsevier, 187–237, https://doi.org/10.1016/S0065-2113(04)92004-4, 2007.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y.: Recurrent Neural Networks for Multivariate Time Series with Missing Values, Sci Rep, 8, 1–12, https://doi.org/10.1038/s41598-018-24271-9, 2018.

Clauss, K., Yan, H., and Kuenzer, C.: Mapping Paddy Rice in China in 2002, 2005, 2010 and 2014 with MODIS Time Series, Remote Sensing, 8, 434, https://doi.org/10.3390/rs8050434, 2016.

Dong, J., Xiao, X., Menarguez, M. A., Zhang, G., Qin, Y., Thau, D., Biradar, C., and Moore, B.: Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine, Remote Sensing of Environment, 185, 142–154, https://doi.org/10.1016/j.rse.2016.02.016, 2016.

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

Dong, J., Pang, Z., Fu, Y., Peng, Q., Li, X., and Yuan, W.: Winter wheat spatial-temporal dynamics in China from 2001 to 2020, ISPRS Journal of Photogrammetry and Remote Sensing, in review, 2024.

Elert, E.: Rice by the numbers: A good grain, Nature, 514, S50–S51, https://doi.org/10.1038/514S50a, 2014.

FAO: World Food and Agriculture – Statistical Yearbook 2023, FAO, Rome, Italy, https://doi.org/10.4060/cc8166en, 2023.

Fu, Y., Shen, R., Song, C., Dong, J., Han, W., Ye, T., and Yuan, W.: Exploring the effects of training samples on the accuracy of crop mapping with machine learning algorithm, Science of Remote Sensing, 100081, https://doi.org/10.1016/j.srs.2023.100081, 2023.

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

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Cheng, F., Zhuang, H., Zhang, J., and Tao, F.: NESEA-Rice10: high-resolution annual paddy rice maps for Northeast and Southeast Asia from 2017 to 2019, Earth Syst. Sci. Data, 13, 5969–5986, https://doi.org/10.5194/essd-13-5969-2021, 2021.

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Zhuang, H., Cheng, F., Zhang, J., and Tao, F.: Annual paddy rice planting area and cropping intensity datasets and their dynamics in the Asian monsoon region from 2000 to 2020, Agricultural Systems, 200, 103437, https://doi.org/10.1016/j.agsy.2022.103437, 2022.

Hu, J., Chen, Y., Cai, Z., Wei, H., Zhang, X., Zhou, W., Wang, C., You, L., and Xu, B.: Mapping Diverse Paddy Rice Cropping Patterns in South China Using Harmonized Landsat and Sentinel-2 Data, Remote Sensing, 15, 1034, https://doi.org/10.3390/rs15041034, 2023.

Jiang, M., Li, X., Xin, L., and Tan, M.: Paddy rice multiple cropping index changes in Southern China: Impacts on national grain production capacity and policy implications, J. Geogr. Sci., 29, 1773–1787, https://doi.org/10.1007/s11442-019-1689-8, 2019.

Li, J. and Chen, B.: Global Revisit Interval Analysis of Landsat-8 -9 and Sentinel-2A -2B Data for Terrestrial Monitoring, Sensors, 20, 6631, https://doi.org/10.3390/s20226631, 2020.

Li, X., Peng, Q., Zheng, Y., Lin, S., He, B., Qiu, Y., Chen, J., Chen, Y., and Yuan, W.: Incorporating environmental variables into spatiotemporal fusion model to reconstruct high-quality vegetation index data, IEEE Transactions on Geoscience and Remote Sensing, 1–1, https://doi.org/10.1109/TGRS.2024.3349513, 2024.

Liu, Z., Li, Z., Tang, P., Li, Z., Wu, W., Yang, P., You, L., and Tang, H.: Change analysis of rice area and production in China during the past three decades, J. Geogr. Sci., 23, 1005–1018, https://doi.org/10.1007/s11442-013-1059-x, 2013.

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

Mansaray, L. R., Wang, F., Huang, J., Yang, L., and Kanu, A. S.: Accuracies of support vector machine and random forest in rice mapping with Sentinel-1A, Landsat-8 and Sentinel-2A datasets, Geocarto International, 35, 1088–1108, https://doi.org/10.1080/10106049.2019.1568586, 2020.

Meng, L., Li, Y., Shen, R., Zheng, Y., Pan, B., Yuan, W., Li, J., and Zhuo, L.: Large-scale and high-resolution paddy rice intensity mapping using downscaling and phenology-based algorithms on Google Earth Engine, International Journal of Applied Earth Observation and Geoinformation, 128, 103725, https://doi.org/10.1016/j.jag.2024.103725, 2024.

Millard, K. and Richardson, M.: On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping, Remote Sensing, 7, 8489–8515, https://doi.org/10.3390/rs70708489, 2015.

Mohammadi, A., Khoshnevisan, B., Venkatesh, G., and Eskandari, S.: A Critical Review on Advancement and Challenges of Biochar Application in Paddy Fields: Environmental and Life Cycle Cost Analysis, Processes, 8, 1275, https://doi.org/10.3390/pr8101275, 2020.

National Bureau of Statistics of China: China Statistical Yearbook 2023, China Statistics Press, 2023.

Nguyen, D. B., Gruber, A., and Wagner, W.: Mapping rice extent and cropping scheme in the Mekong Delta using Sentinel-1A data, Remote Sensing Letters, 7, 1209–1218, https://doi.org/10.1080/2150704X.2016.1225172, 2016.

Oguro, Y., Suga, Y., Takeuchi, S., Ogawa, M., Konishi, T., and Tsuchiya, K.: Comparison of SAR and optical sensor data for monitoring of rice plant around Hiroshima, Advances in Space Research, 28, 195–200, https://doi.org/10.1016/S0273-1177(01)00345-3, 2001.

Oliver, C. and Quegan, S. (Eds.): Understanding synthetic aperture radar images, SciTech Publishing, Raleigh, NC, 479 pp., 2004.

Pan B., Zheng Y., Shen R., Ye T., Zhao W., Dong J., Ma H., and Yuan W.: A 10 m Resolution Distribution Dataset of Double-Season Paddy Rice in China from 2016 to 2020, National Ecosystem Science Data Center [data set] (in Chinese), https://doi.org/10.12199/nesdc.ecodb.rs.2022.012, 2021a.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Peng, Q., Shen, R., Li, X., Ye, T., Dong, J., Fu, Y., and Yuan, W.: A twenty-year dataset of high-resolution maize distribution in China, Sci Data, 10, 658, https://doi.org/10.1038/s41597-023-02573-6, 2023.

Phan, H., Le Toan, T., Bouvet, A., Nguyen, L., Pham Duy, T., and Zribi, M.: Mapping of Rice Varieties and Sowing Date Using X-Band SAR Data, Sensors, 18, 316, https://doi.org/10.3390/s18010316, 2018.

Savitzky, Abraham. and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Anal. Chem., 36, 1627–1639, https://doi.org/10.1021/ac60214a047, 1964.

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

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

Shen, R., Pan, B., Peng, Q., Dong, J., Chen, X., Zhang, X., Ye, T., Huang, J., and Yuan, W.: High-resolution distribution maps of single-season rice in China from 2017 to 2022, Science Data Bank [data set], https://doi.org/10.57760/sciencedb.06963, 2023b.

Shen, R., Peng, Q., Li, X., Chen, X., and Yuan, W.: CCD-Rice: A paddy rice distribution dataset in China from 1990 to 2016 at 30 m resolution, Science Data Bank [data set], https://doi.org/10.57760/sciencedb.15865, 2024a.

Shen, R., Peng, Q., and Yuan, W.: Several samples visually interpreted from the very high-resolution images from Google Earth for rice mapping research in China, figshare [data set], https://doi.org/10.6084/m9.figshare.25515019.v1, 2024b.

Sun, C., Zhang, H., Xu, L., Ge, J., Jiang, J., Zuo, L., and Wang, C.: Twenty-meter annual paddy rice area map for mainland Southeast Asia using Sentinel-1 synthetic-aperture-radar data, Earth System Science Data, 15, 1501–1520, https://doi.org/10.5194/essd-15-1501-2023, 2023.

Tian, X., Bai, Y., Li, G., Yang, X., Huang, J., and Chen, Z.: An Adaptive Feature Fusion Network with Superpixel Optimization for Crop Classification Using Sentinel-2 Imagery, Remote Sensing, 15, 1990, https://doi.org/10.3390/rs15081990, 2023.

Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., and Koetz, B.: Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions, Remote Sensing, 8, 55, https://doi.org/10.3390/rs8010055, 2016.

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., and Ceschia, E.: Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications, Remote Sensing of Environment, 199, 415–426, https://doi.org/10.1016/j.rse.2017.07.015, 2017.

Waleed, M., Mubeen, M., Ahmad, A., Habib-ur-Rahman, M., Amin, A., Farid, H. U., Hussain, S., Ali, M., Qaisrani, S. A., Nasim, W., Javeed, H. M. R., Masood, N., Aziz, T., Mansour, F., and EL Sabagh, A.: Evaluating the efficiency of coarser to finer resolution multispectral satellites in mapping paddy rice fields using GEE implementation, Sci Rep, 12, 13210, https://doi.org/10.1038/s41598-022-17454-y, 2022.

Wen, Y., Li, X., Mu, H., Zhong, L., Chen, H., Zeng, Y., Miao, S., Su, W., Gong, P., Li, B., and Huang, J.: Mapping corn dynamics using limited but representative samples with adaptive strategies, ISPRS Journal of Photogrammetry and Remote Sensing, 190, 252–266, https://doi.org/10.1016/j.isprsjprs.2022.06.012, 2022.

Xiao, X., Boles, S., Liu, J., Zhuang, D., Frolking, S., Li, C., Salas, W., and Moore, B.: Mapping paddy rice agriculture in southern China using multi-temporal MODIS images, Remote Sensing of Environment, 95, 480–492, https://doi.org/10.1016/j.rse.2004.12.009, 2005.

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., and Moore, B.: Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images, Remote Sensing of Environment, 100, 95–113, https://doi.org/10.1016/j.rse.2005.10.004, 2006.

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., and Li, B.: 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, https://doi.org/10.1016/j.scitotenv.2019.135183, 2020.

Xu, S., Zhu, X., Chen, J., Zhu, X., Duan, M., Qiu, B., Wan, L., Tan, X., Xu, Y. N., and Cao, R.: A robust index to extract paddy fields in cloudy regions from SAR time series, Remote Sensing of Environment, 285, 113374, https://doi.org/10.1016/j.rse.2022.113374, 2023.

Xuan, F., Dong, Y., Li, J., Li, X., Su, W., Huang, X., Huang, J., Xie, Z., Li, Z., Liu, H., Tao, W., Wen, Y., and Zhang, Y.: 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, https://doi.org/10.1016/j.jag.2022.103178, 2023.

Yang, J. and Huang, X.: The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019, Earth System Science Data, 13, 3907–3925, https://doi.org/10.5194/essd-13-3907-2021, 2021.

Yang, J. and Huang, X.: The 30 m annual land cover datasets and its dynamics in China from 1985 to 2022, Zenodo [data set], https://doi.org/10.5281/zenodo.8176941, 2023.

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

Zhang, C., Zhang, H., and Tian, S.: 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, https://doi.org/10.1016/j.compag.2023.108105, 2023a.

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

Zhang, G., Xiao, X., Dong, J., Xin, F., Zhang, Y., Qin, Y., Doughty, R. B., and Moore, B.: Fingerprint of rice paddies in spatial–temporal dynamics of atmospheric methane concentration in monsoon Asia, Nat Commun, 11, 554, https://doi.org/10.1038/s41467-019-14155-5, 2020.

Zhang, H. K. and Roy, D. P.: Using the 500m MODIS land cover product to derive a consistent continental scale 30m Landsat land cover classification, Remote Sensing of Environment, 197, 15–34, https://doi.org/10.1016/j.rse.2017.05.024, 2017.

Zhang, X., Shen, R., Zhu, X., Pan, B., Fu, Y., Zheng, Y., Chen, X., Peng, Q., and Yuan, W.: Sample-free automated mapping of double-season rice in China using Sentinel-1 SAR imagery, Frontiers in Environmental Science, 11, https://doi.org/10.3389/fenvs.2023.1207882, 2023b.

Zhao, Q., Ding, X., Zhu, C., Zhao, W., Fan, S., Zhao, L., and Yu, D.: Healthy Diets in China, in: 2023 China and Global Food Policy Report, 2023.

Zhou, Y., Dong, J., Liu, J., Metternicht, G., Shen, W., You, N., Zhao, G., and Xiao, X.: Are There Sufficient Landsat Observations for Retrospective and Continuous Monitoring of Land Cover Changes in China?, Remote Sensing, 11, 1808, https://doi.org/10.3390/rs11151808, 2019.