Supplementary information

Supplementary Figures

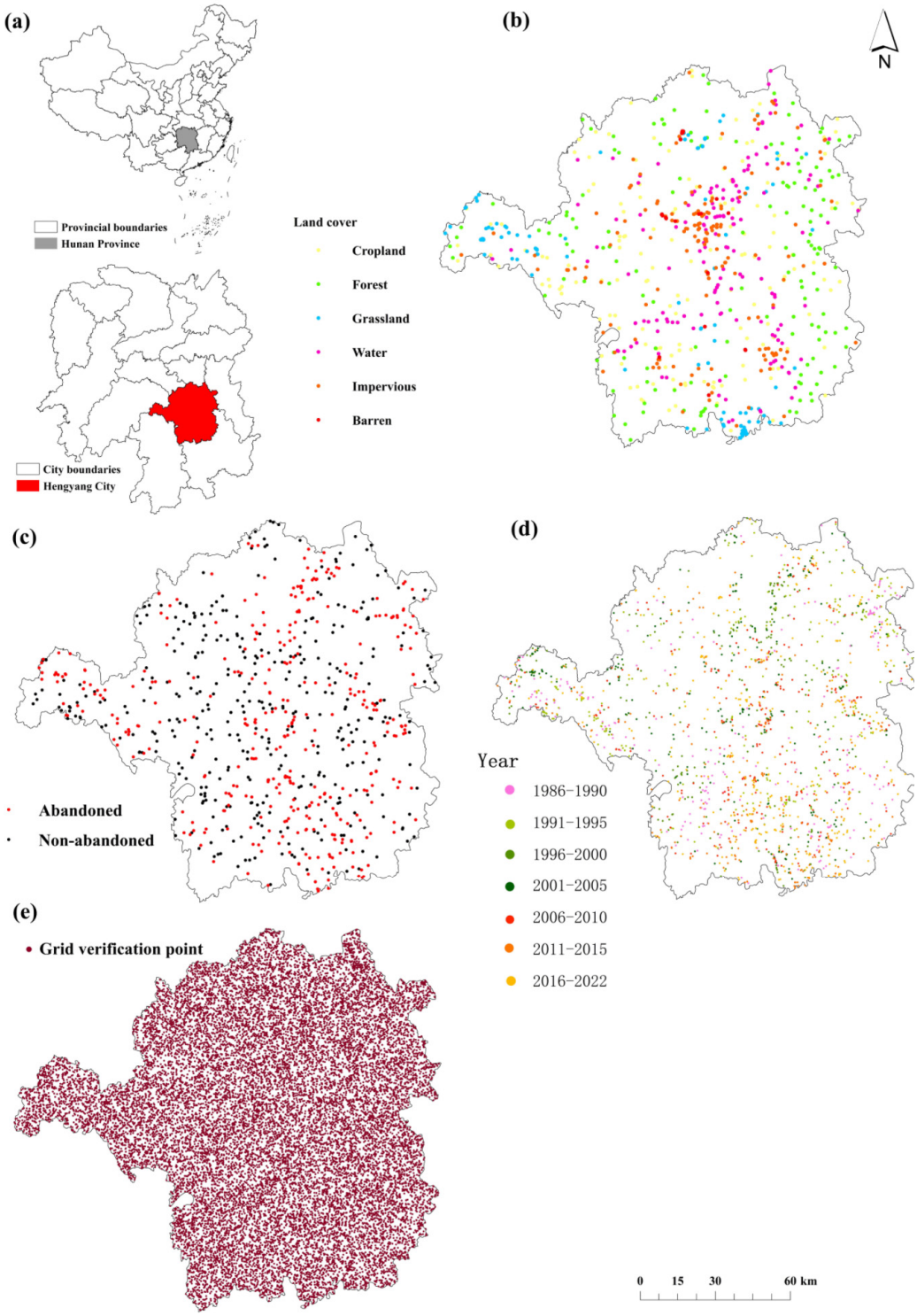
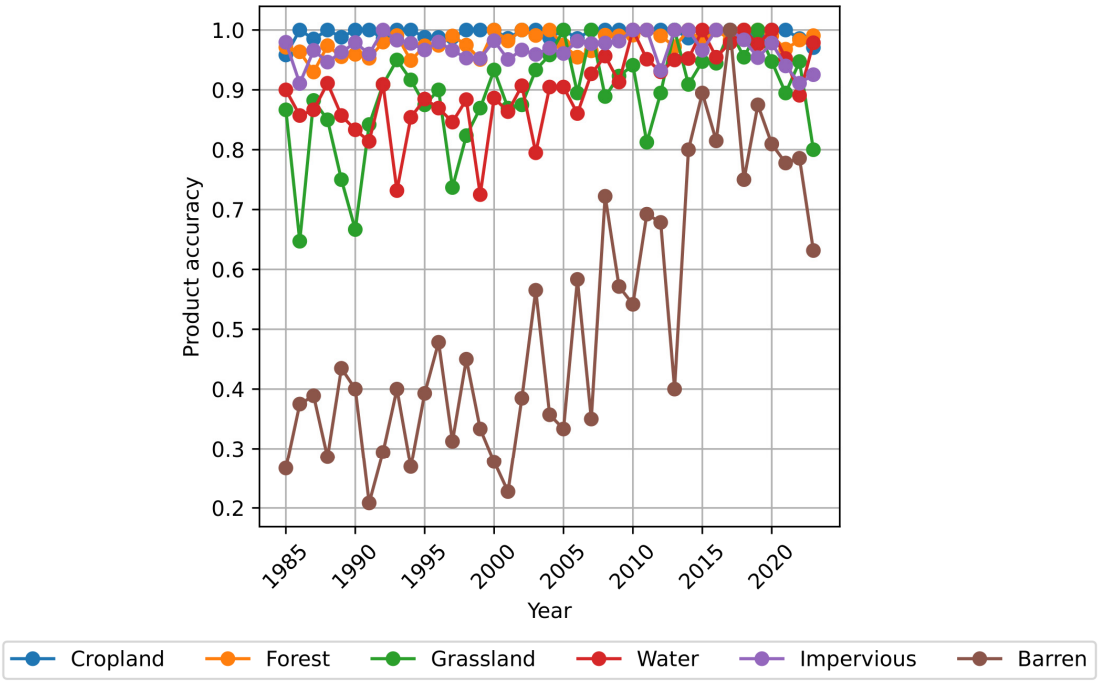


Figure S1. Location of the study area of (a), land cover samples of (b), abandoned and non-abandoned samples of (c), samples of abandonment year detection of (d), and grid sample points of (e).



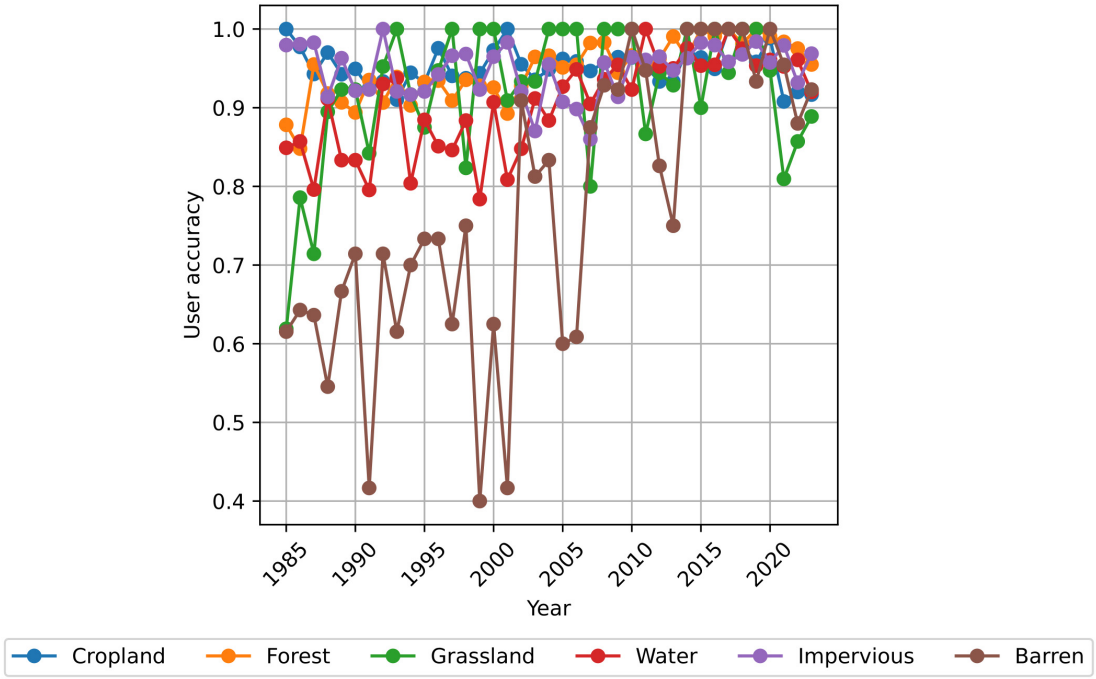


Figure S2. Accuracy of producers and users based on random forest models for different land cover in the study area from 1985 to 2023.

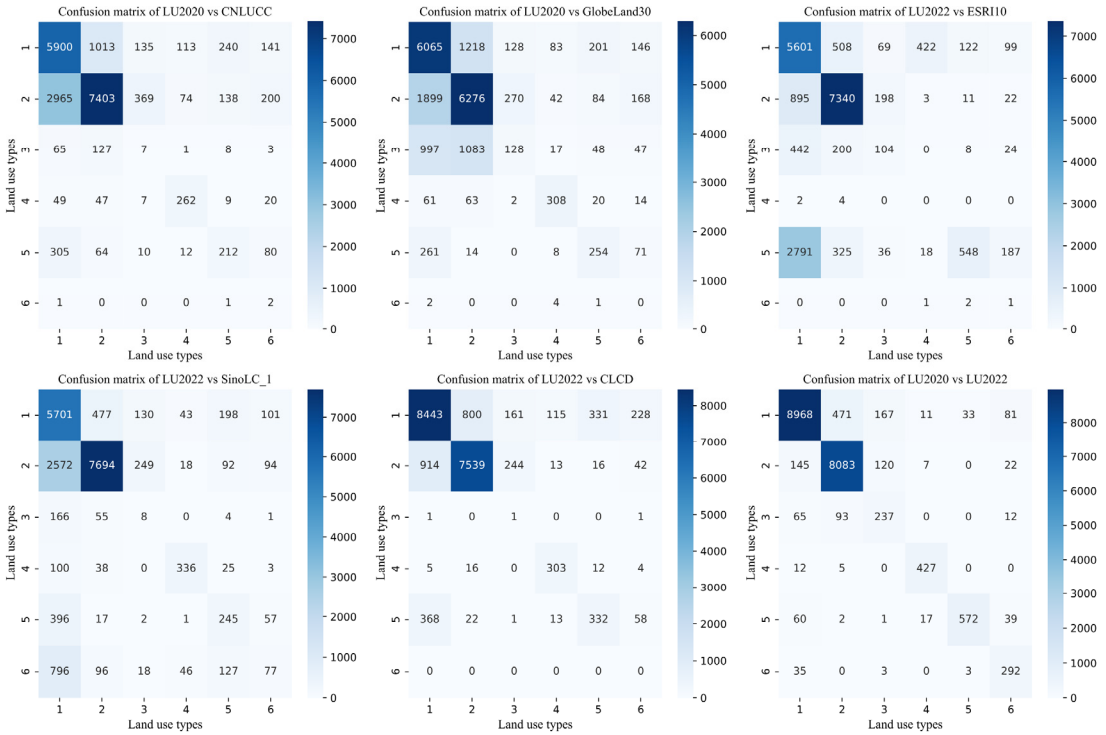


Figure S3. Confusion matrix between land cover in this study (LU2020 and LU2022) and several widely used land cover products (CNLUCC, GlobeLand30, ESRI10, SinoLC_1, and CLCD).

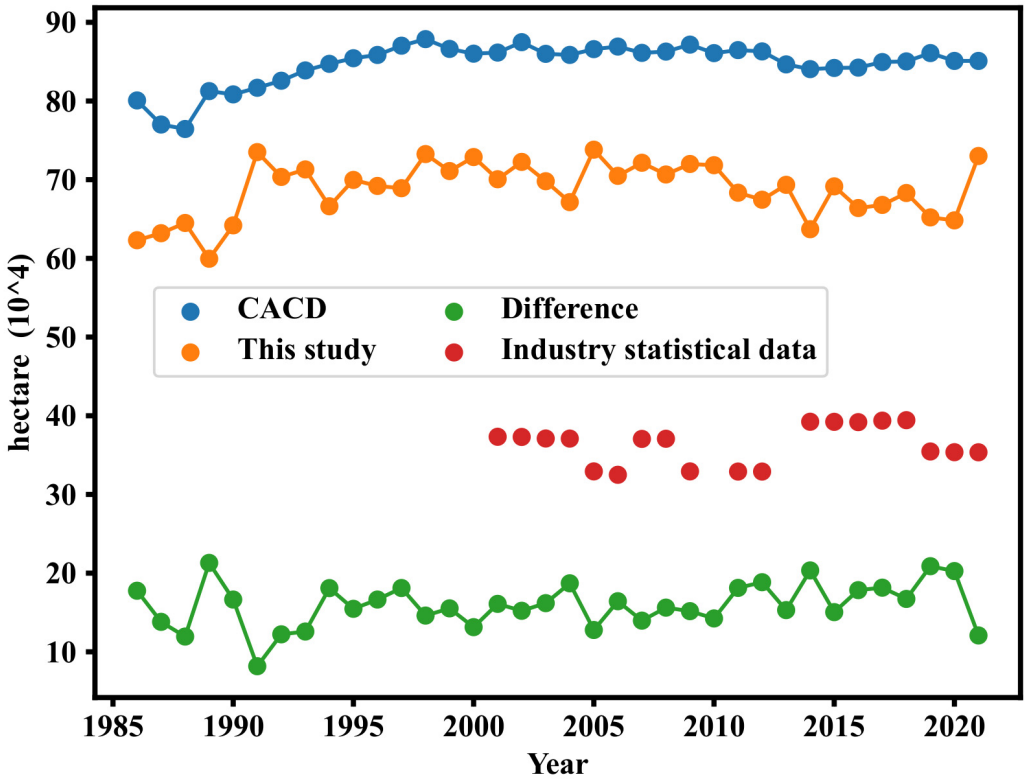


Figure S4. Cropland area comparison and differences between our land cover data and CACD.

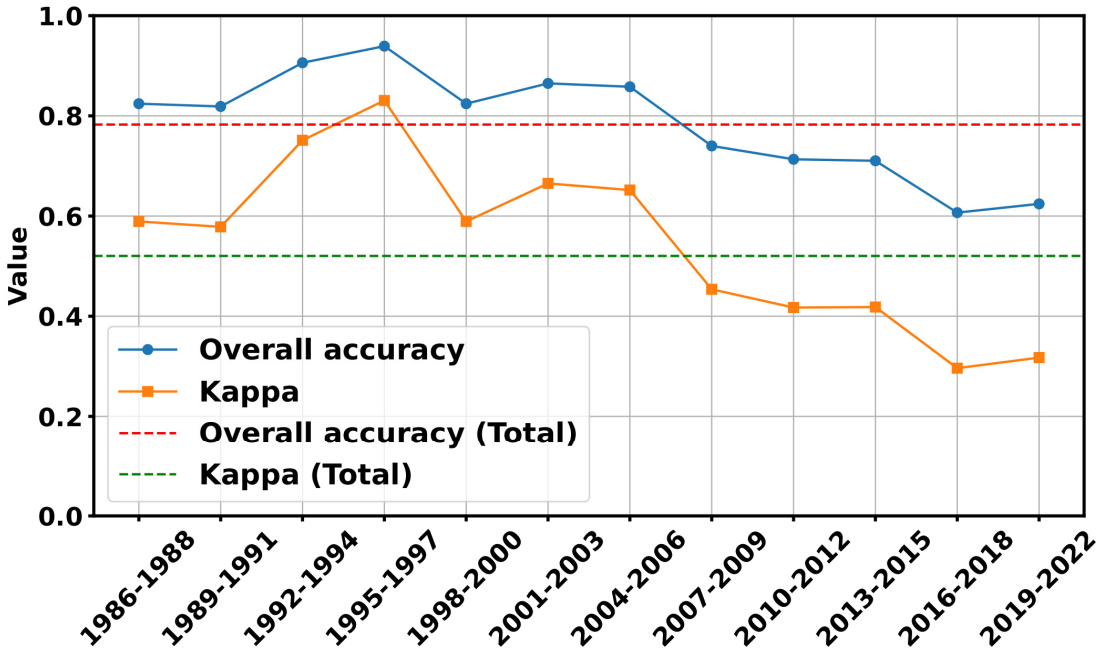


Figure S5. Overall accuracy and kappa coefficient for the timing of cropland abandonment in the study area.

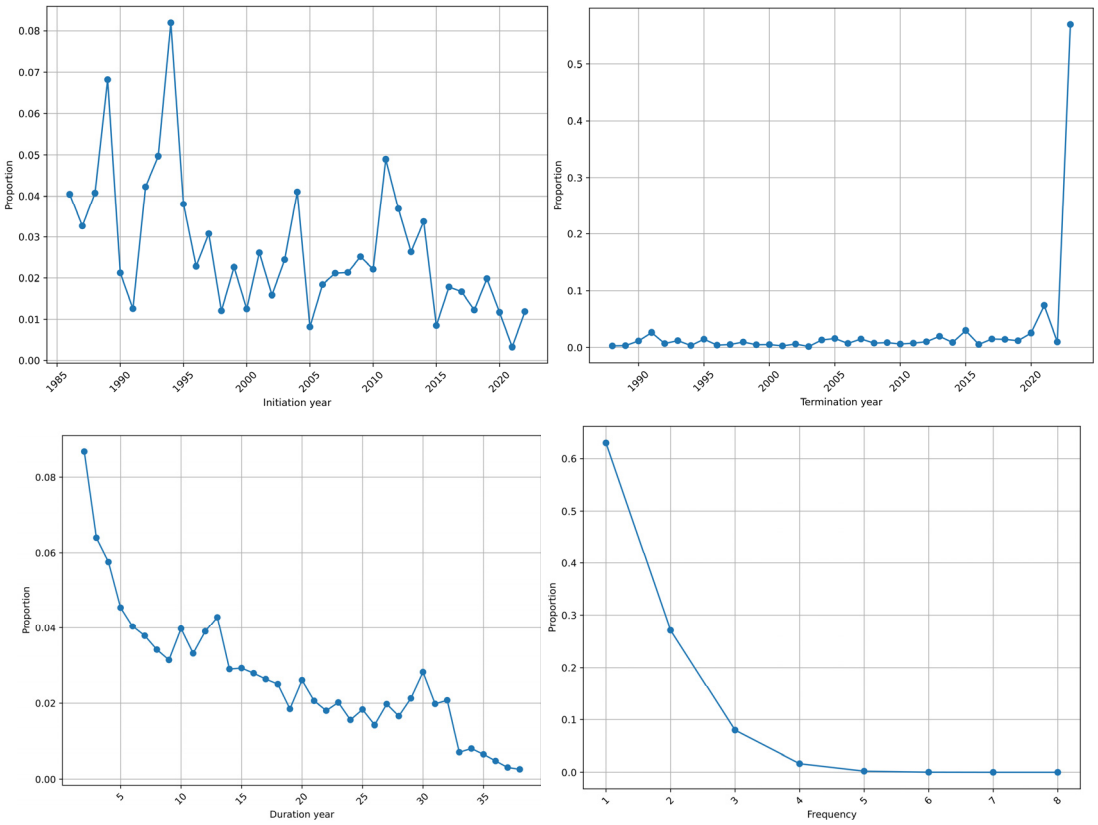


Figure S6. The proportion of the initiation year, termination year, duration year, and frequency of cropland abandonment.

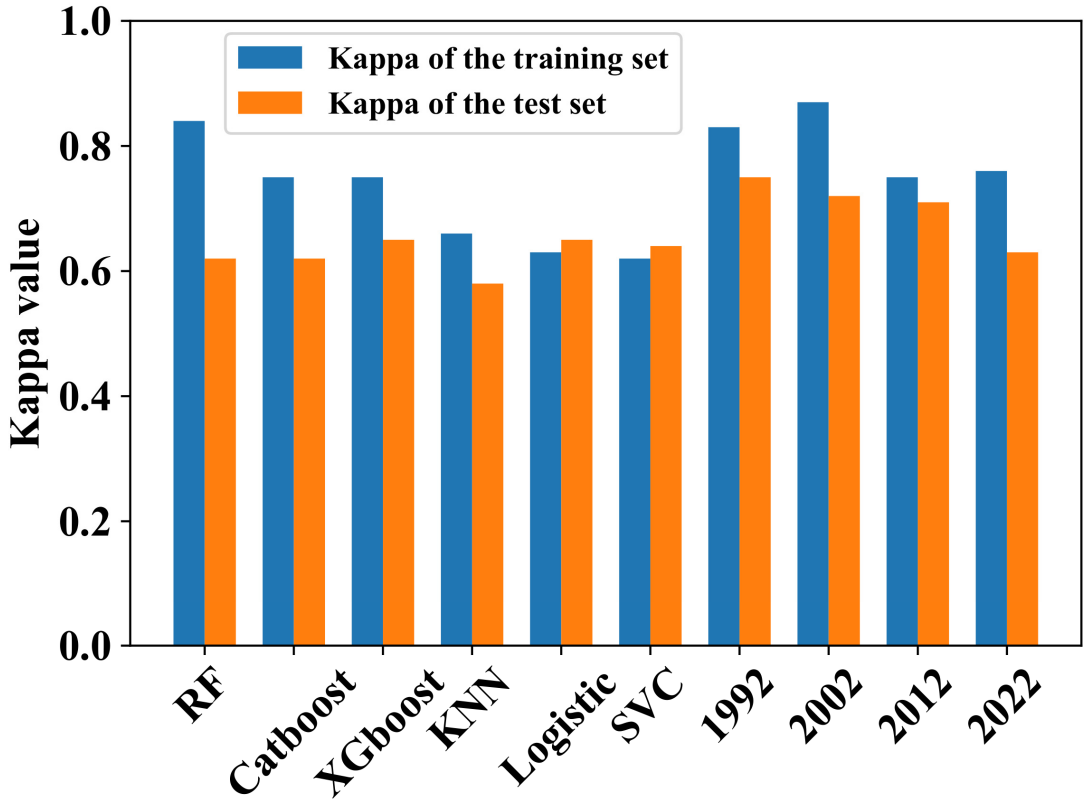


Figure S7. Kappa values of training and testing sets for different machine learning models and different time periods.

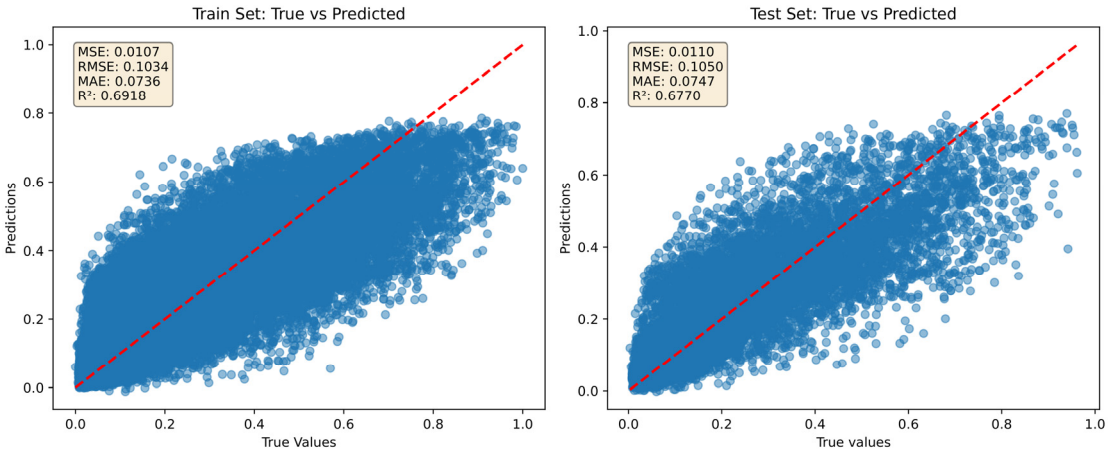


Figure S8. Prediction of scatter plots using the training and testing sets of the long short-term memory network model

Supplementary Tables

Table S1 Transformation rules of different land cover products.

CNLUCC	**GlobeLand30**	**ESRI10**	**Unified land cover**
● Paddy fields dry land	Cropland	Crops	Cropland
● Forest ● Shrubland ● sparse Forest Other forest	Forest	Trees	Forest
● High coverage grassland ● Medium coverage grassland Low coverage grassland	Grassland	Rangelands areas	Grassland
● Rivers ● Canals ● Reservoirs and ponds Beaches	Water	Water areas	Water
● Urban land ● Rural residential areas, Other construction land	Artificial surface	Built areas	Impervious
Bare land	● Bare land ● Wetlands	● Bare groud areas ● Clouds	Barren
SinoLC_1		**CLCD**	**Unified land cover**
Cropland		Cropland	Cropland
● Tree cover ● Shrubland		● Forest ● Shrub	Forest
Grassland		Grassland	Grassland
Water		Water	Water
● Traffic route ● Building		Impervious	Impervious
● Barren and sparse vegetation ● Snow and ice ● Wetland ● Moss and lichen		● Sonw/Ice ● Barren ● Wetland	Barren

Table S2 Different model hyperparameter settings based on grid search method.

Models	Parameters	Possible values for parameters		
Logistic	C	0.1	1	10
	penalty	l1	l2	
	solver	liblinear	saga	
SVC	C	0.1	1	10
	kernel	linear	rbf	
	gamma	scale	auto	
KNN	n_neighbors	3	5	7
	weights	uniform	distance	
	p	1	2	
XGBoost	n_estimators	50	100	150
	learning_rate	0.01	0.1	0.2
	max_depth	3	5	7
CatBoost	iterations'	50	100	150
	learning_rate	0.01	0.1	0.2
	depth	3	5	7
Random forest	n_estimators	50	100	150
	max_depth	3	5	7
	min_samples_split	2	5	10

Table S3 Spatial and temporal accuracy of cropland abandonment.

Temporal accuracy	Overall accuracy	Kappa	Spatial accuracy	Overall accuracy	Kappa
Total	0.7826	0.5203	Total	0.8811	0.7622
1986-1988	0.8243	0.5892			
1989-1991	0.8187	0.5784			
1992-1994	0.906	0.7512			

1995-1997	0.9391	0.8307			
1998-2000	0.8243	0.5892			
2001-2003	0.8648	0.6651			
2004-2006	0.8581	0.6518			
2007-2009	0.7400	0.4537			
2010-2012	0.7133	0.4173			
2013-2015	0.7103	0.4183			
2016-2018	0.6068	0.2962			
2019-2022	0.6243	0.3174			

Table S4 Proportion of cropland abandonment under different landforms.

Landform type	Classification basis (based on altitude)	Proportion
Plain	<=200	65.24%
Hills	200~500	12.01%
Mountainous	>=500	22.75%

Table S5 Proportion of cropland abandonment under different slopes.

Type of cropland	Classification basis (based on slope)	Proportion
Flat cropland	<=2°	26.80%
Gentle slope cropland	2°~6°	23.86%
Slope cropland	>=6°	49.35%

Table S6 Proportion of the risk range of cropland abandonment based on random forest model from 1986 to 2022 (average value 0.36).

Risk value	Proportion
0.14~0.5	85.43%
0.5~0.7	14.21%
>=0.7	0.37%

Supplementary Text

Formula for calculating vegetation index

- NDVI (Normalized Difference Vegetation Index) = (NIR - Red) / (NIR + Red)
- NDMI (Normalized Difference Moisture Index) = (NIR - SWIR1) / (NIR + SWIR1)
- EVI (Enhanced Vegetation Index) = 2.5 × (NIR - Red) / (NIR + 6 × Red - 7.5 × Blue + 1)
- MNDWI (Modified Normalized Difference Water Index) = (Green - SWIR1) / (Green + SWIR1)
- NDBI (Normalized Difference Built-up Index) = (SWIR1 - NIR) / (SWIR1 + NIR)

Where Blue is Blue band; Green is Green band; Red is Red band; NIR is Near-Infrared band; SWIR1 is Short-Wave Infrared Band 1; SWIR2 is Short-Wave Infrared Band 2.

Logistic regression

Logistic is a machine learning method that can solve a binary classification problem. It is supported by the theory of linear regression and introduces nonlinear factors through the sigmoid function to deal with classification problems. The sigmoid function is an *S*-shaped curve, which maps an actual number to the interval of [0, 1] and can be used to determine whether cropland has been abandoned. *C*, penalty, and solver are three critical parameters.

Support vector classification (SVC)

SVC is a classification model. In classification tasks, it separates data of different categories by finding an optimal hyperplane. For nonlinear problems, the data can be mapped to a high-dimensional space through a kernel function, which makes the data linearly separable. In this study, it attempts to find a suitable hyperplane based on all environmental covariates to distinguish whether cropland has been abandoned. *C*, kernel, and gamma are three critical parameters.

K-nearest neighbors (KNN)

KNN is a simple and intuitive non-parametric model. For new samples, it calculates the distance to the samples in the training set based on a certain distance metric, identifies the nearest *K* neighbors, and predicts the category of the new sample

based on the categories of these neighbors. In this study, it is based on the similarity of environmental covariates to determine whether cropland has been abandoned. *n*_neighbors, weights, and *p* are three critical parameters.

eXtreme gradient boosting (XGBoost)

XGBoost is a machine learning algorithm based on the gradient-boosting framework. It iteratively adds new decision trees and fits the residuals of the previous tree, gradually improving the predictive ability of the model. In this study, it can comprehensively consider the complex interactions between numerous environmental covariates to determine whether cropland has been abandoned. *n*_estimators, learning_rate, and max_depth are three critical parameters.

Categorical boosting (CatBoost)

CatBoost is a gradient-boosting decision tree algorithm. It adopts unique ordered boosting techniques and category feature encoding methods, reducing prediction bias and effectively capturing complex patterns in data, making it particularly suitable for processing datasets with category features. Similar to XGBoost, it can also fully utilize the characteristic information of all environmental covariates to determine whether cropland has been abandoned. iterations, learning_rate, and depth are three critical parameters.

Random forest

Random forest was proposed by Breiman in 2001 and is a widely used supervised machine learning method. It uses the bootstrap resampling technique to fit a large number of decision trees on the sub-samples of the training dataset and uses weighted averaging to improve the accuracy of the prediction results. Compared with single models such as decision trees, this model is less sensitive to overfitting, multicollinearity, and missing and unbalanced data. In this study, it can learn the relationship between environmental covariates and cropland abandonment from multiple perspectives, and then improve the accuracy of predictions. *n*_estimators, max_depth, and min_samples_split are three critical parameters.