Spatially Explicit Reconstruction of Cropland Using the Random Forest: A Case Study of the Tuojiang River Basin, China from 1911 to 2010

Qi Wang, Min Xiong, Qiquan Li, Hao Li, Ting Lan, Ouping Deng, Rong Huang, Min Zeng and Xuesong Gao *

College of Resources, Sichuan Agricultural University, Chengdu 611130, China; wq94@stu.sicau.edu.cn (Q.W.); xiongmin@stu.sicau.edu.cn (M.X.); liqq@lreis.ac.cn (Q.L.); lihao@sicau.edu.cn (H.L.); tlan@sicau.edu.cn (T.L.); ouping@sicau.edu.cn (O.D.); 14624@sicau.edu.cn (R.H.); 13596@sicau.edu.cn (M.Z.)

* Correspondence: xuesonggao@sicau.edu.cn

Abstract: A long-term, high-resolution cropland dataset plays an essential part in accurately and systematically understanding the mechanisms that drive cropland change and its effect on biogeochemical processes. However, current widely used spatially explicit cropland databases are developed according to a simple downscaling model and are associated with low resolution. By combining historical county-level cropland archive data with natural and anthropogenic variables, we developed a random forest model to spatialize the cropland distribution in the Tuojiang River Basin (TRB) during 1911–2010, using a resolution of 30 m. The reconstruction results showed that the cropland in the TRB increased from $1.13 \times 10^4$ km$^2$ in 1911 to $1.81 \times 10^4$ km$^2$. In comparison with satellite-based data for 1980, the reconstructed dataset approximated the remotely sensed cropland distribution. Our cropland map could capture cropland distribution details better than three widely used public cropland datasets, due to its high spatial heterogeneity and improved spatial resolution. The most critical factors driving the distribution of TRB cropland include nearby-cropland, elevation, and climatic conditions. This newly reconstructed cropland dataset can be used for long-term, accurate regional ecological simulation, and future policymaking. This novel reconstruction approach has the potential to be applied to other land use and cover types via its flexible framework and modifiable parameters.

Keywords: historical land use; random forest algorithm; reconstruction; spatial distribution

1. Introduction

The land use and cover change (LUCC) process has undergone drastic changes due to the population boom over the past century, during which rising demands for food and fiber have led to expansion of agricultural lands. As one of the predominant land use types [1], cropland serves as an important carbon budget component and has significant impacts on biogeochemical processes, such as food production, global change, regional ecosystem services, and biodiversity [2–4]. The long-term evolution of the cropland distribution plays a key role in understanding agricultural development trajectories and long-term environmental and economic strategies [5,6]. However, most currently used cropland datasets are developed at the large scale with medium- or low-resolution, and thus lack high-precision spatial information. This produces an important research gap because long-term, spatially explicit land-use maps are critical input data for many biogeochemical land-surface models (e.g., ORCHIDEE, DLEM) [7,8], which are used for the IPCC report and the annual global carbon budget estimate [9]. Therefore, there is an urgent need for a long-term, accurate cropland dataset that can improve ecological simulation accuracy and support detailed ecological analyses.

Although the remote sensing method provides a fairly accurate data source for mapping land change, it is resource-intensive and has only been available for the past four
decades [10,11]. Current widely used spatially explicit cropland databases are developed according to a simple downscaling model, make little or no use of available statistics and archives, and cannot accurately characterize the long-term change. One representative global-scale product is the History Database of the Global Environment (HYDE) [12], which is used widely in global climate change and carbon budget simulations due to its long time-span (10,000 BCE to 2017 CE). However, derived from socio-economic parameters such as population and consumption, it leads to substantial uncertainties and low resolution (5′ × 5′ resolution) [13]. By employing a top-down model based on land suitability and socio-economic variables, a series of regional to continental scales cropland datasets have been reconstructed at various times and spatial resolutions (60 km × 60 km, 10 km × 10 km, 5 km × 5 km, 1 km × 1 km) [14–17]. Another approach is to utilize a bottom-up approach that considers cell states and human land use activities. At the local and regional scales, the historical LUCC has been simulated using cellular automata models [18] and multi-agent system models that consider the decision-making behaviors of agents [19].

However, the existing widely used allocation method has many limitations. Due to the high dimensionality of the input variables, traditional allocation methods cannot handle high-dimensional data without encountering subjective effects. One needs to employ normalization of factors as well as abstract functions or a series of models to use these methods [20]. Machine learning algorithms have recently become a robust tool for land use change simulation. They can consider multiple land use change factors, as well as linear and non-linear phenomena [21]. The random forest (RF) algorithm is proven to be efficient at processing high-dimensional input datasets. Some advantages are that it does not consider multiple collinearities and therefore avoids overfitting and that it offers high operation speed and reliability [21–23]. It has been implemented successfully in short-term LUCC forecasting [20,24]. New approaches that make the best use of available official archives and high-dimensional ancillary data are therefore crucial to improving the robustness of long-term, spatially explicit reconstructed data.

Here, we fill this gap by developing a random forest classifier approach based on historical official cropland archives and available natural and anthropogenic ancillary data. We implement this approach for the case of the Tuojiang River Basin (TRB) in China, and reconstruct a 30 m-resolution time-series gridded cropland dataset that describes the TRB during 1911–2010. The major objectives of this paper are: (1) to develop an RF classifier workflow for historical cropland reconstruction; (2) to reconstruct a cropland dataset that describes the TRB during 1911–2010 with 30 m resolution; (3) to validate the reconstruction results based on remotely sensed land use data and three public datasets; and (4) to examine the proximate drivers of cropland distribution. The century-long spatially explicit cropland dataset establishes a good foundation for the more accurate simulation of climatic and ecological effects over long time scales.

2. Materials and Methods
2.1. Study Area

Located in the upper reaches of the Yangtze River, the TRB extends between 103°40′–105°45′ E and 28°52′–31°42′ N and covers an area of 2.79 × 10⁴ km², which includes 27 complete counties and districts (Figure 1). It has diverse altitude features with relatively high in the northwest and low in the southeast. With favorable climatic conditions, fertile soil, and dense water networks, the TRB is an important agribusiness area with the highest population density in southwestern China. The soil types are diverse in the region, and mainly include Feleachi-Stagnic Anthrosols, Purpli Cambosols, and Ali-Perudic Argosols (Chinese soil taxonomic classification) [25]. Because of its excellent climatic conditions and rapid social development, the TRB has been one of the major traditional agricultural regions and socio-economic centers in the southwest of China. The cropland area has expanded at an unprecedented speed over the past century due to the population boom [26]. In addition, it has many problems, including over-intensive
agriculture, human–land contradiction, and eco-environmental problems [27]. Thus, it is an ideal testbed for century-long cropland reconstruction research.

Figure 1. The location of the Tuojiang River Basin.

2.2. Input Factors and Data Resources

Given the availability of variables for model training and prediction, ten representative natural and anthropogenic variables (elevation, slope, relief amplitude, climatic productive potential, nearby cropland, distance to rural settlements, distance to a river, flood risk, soil erosion, and soil fertility) were used as independent variables in an effort to describe the key factors that affect cropland probability. Population density, waterbody, and elevation were selected as three constraint factors.

Multi-source data were used in this study: (1) county-level cropland census data, (2) remotely sensed land use data, (3) topography data, (4) climate data, (5) waterbody data, (6) demographic data, (7) administrative division data, and (8) open access public cropland datasets were all used (Table 1).
Table 1. List of ancillary data used in this study.

| Data                        | Description                                                                 | Resolution | Year          | Source                                                                 |
|-----------------------------|-----------------------------------------------------------------------------|------------|---------------|------------------------------------------------------------------------|
| **Raster data (remote sensing data)** |                                                                               |            |               |                                                                         |
| FROM-GLC10                  | Cropland layer data                                                         | 10 m       | 2017          | http://data.ess.tsinghua.edu.cn/ (accessed on 17 May 2021)             |
| GlobeLand30                 | Cropland layer data                                                         | 30 m       | 1980          | National Earth system Science Data Center, China (accessed on 20 May 2021) |
| ASTER-GDEM Version 3        | Digital Elevation Model                                                     | 30 m       | 2019          | National Aeronautics and Space Administration, USA (accessed on 20 May 2021) |
| Solar radiation             | Annual solar radiation data in China                                       | 1 km       | 1950–1980     | National Earth system Science Data Center, China (accessed on 22 March 2021) |
| frost-free period data      | Annual frost-free period data in China                                      | 1 km       | 1951–2012     | National Earth system Science Data Center, China (accessed on 22 March 2021) |
| Soil physicochemical data   | Soil organic matter, total nitrogen, total phosphorus and total potassium contents (surface soil 0–20 cm) | 1 km       | 1990          | National Earth system Science Data Center, China (accessed on 25 March 2021) |
| Soil erosion data           | Chinese soil erosion modulus dataset                                       | 1 km       | 2010          | National Earth system Science Data Center, China (accessed on 25 March 2021) |
| Population density          | Population density distribution in the TRB                               | 1 km       | 1911–2010     |                                                                         |
| **Vector data (social sensing data)** |                                                                               |            |               |                                                                         |
| Administrative boundary maps| County-level administrative boundary maps                                  | 1:1,000,000| 1911, 1953, 1980, 2010| China Historical Geographic Information System Resource Discipline Innovation Platform, China (accessed on 25 May 2021) |
| Rural settlement            | Rural settlement locations                                                  | -          | 1911, 2010    | National Earth System Science Data Center (accessed on 25 May 2021)    |
| waterbody                   | River networks, lakes, ponds and reservoirs data                          | 1:1,000,000| 1906, 2010    | China Historical Geographic Information System (accessed on 25 May 2021) |

2.2.1. County-Level Cropland Census

Based on historical cropland statistics and archives, a county-level TRB cropland area dataset was reconstructed for 1911–2010 via revision, calibration, and verification. In the Republic of China era, the data sources used for cropland reconstruction mainly included statistics and survey data from government departments. In addition to the local gazetteers, cropland data were obtained from the Republic of China Agricultural and Commercial Statistics (1912–1921) (Ministry of Agriculture and Commerce), China Land Use Statistics [28], and Sichuan Province Statistics Summary (1945). During the People’s Republic of China period after 1949, the archives were superior to those that came before in terms of quantity and accuracy. Cropland data were obtained from the Sichuan Provincial Digital Local Gazetteers, Sichuan Provincial Statistical Yearbook, China’s Rural Economic Statistics by Counties (1980–1987) [29], and China Regional Economic Statistics Yearbook [30]. National Land Survey results were used as inventory data during this study. Finally, we get the dynamics of the statistical cropland area in the Tuojiang River Basin from 1911 to 2010 (Supplementary Materials Figure S1).
2.2.2. Remotely Sensed Land Use Data

Cropland distribution data from 2017 were obtained from the FROM-GLC10 product (Table 1), which was developed based on Sentinel-2 satellite imagery in conjunction with a random forest classifier [31]. The FROM-GLC10 dataset provided the most detailed spatial information regarding various land uses and offered 10 m spatial resolution in 2017. The TRB cropland distribution was clipped, resampled, and derived from the FROM-GLC10 product, and was later used as training data. Cropland layer data with a 30 m resolution in 1980 were derived from the GlobeLand30 dataset.

2.2.3. Other Ancillary Data

Ancillary raster data used for cropland spatialization included terrain, climate, soil, rural settlement, river, and population data. Topography data with a 30 m spatial resolution were obtained from NASA’s ASTER GDEM version 3 (https://search.earthdata.nasa.gov/) (accessed on 20 May 2021), and the slope was calculated from Digital Elevation Model data. Meteorological data for the study region from 1951 to 2018 were provided by the National Meteorological Science Data Center in China (http://data.cma.cn/) (accessed on 22 March 2021). The climatic productive potential was calculated using photosynthetic, temperature, and moisture correction coefficients. These were used to represent comprehensive agricultural production conditions [32]. Soil physicochemical and erosion data were downloaded from the National Earth System Science Data Center (NESSDC) (http://www.geodata.cn) (accessed on 25 March 2021).

The administrative divisions from 1911 were provided by the China Historical Geographic Information System (CHGIS) (https://dataverse.harvard.edu/dataverse/chgis_v6) (accessed on 25 May 2021). The administrative divisions from 1953 and 1980 were provided by the Resource Discipline Innovation Platform of China (http://www.data.ac.cn/) (accessed on 25 May 2021). Administrative division data from 2010 (scale of 1:1,000,000) were provided by the National Basic Information Center (http://www.webmap.cn) (accessed on 25 May 2021). Rural TRB settlements from 1911 were identified by CHGIS (accessed on 25 May 2021). Information regarding rural settlements from 2010 was downloaded from the National Earth System Science Data Center (http://www.geodata.cn) (accessed on 25 May 2021). River and lake data from 1906 were downloaded from the Chinese Historical Geographic Information System (https://dataverse.harvard.edu/dataverse/chgis_v6) (accessed on 25 May 2021). Data regarding contemporary river networks and lakes in 2010 (scale of 1:1,000,000) were downloaded from the National Geomatics Center of China (http://www.webmap.cn) (accessed on 25 May 2021). TRB population data were developed for 1911 to 2010 [33]. All raster data were resampled to a 30 m resolution using bilinear spatial interpolation.

2.3. Data Preprocessing

2.3.1. Climatic Productive Potential

The climatic productive potential was calculated using the annual average solar radiation, photosynthetic correction coefficient, temperature correction coefficient, and moisture correction coefficient from meteorological datasets developed since 1951, from 40 stations in the TRB and surrounding areas [34].

\[
Climate = Q \times f(q) \times f(t) \times f(w)
\]

where \(Climate\) is the climatic productive potential; \(Q\) is the total solar radiation; \(f(q)\) is the photosynthetic correction coefficient, which is set to 21.9; \(f(t)\) is the temperature correction coefficient, which is the number of frost-free days in the year; and \(f(w)\) is the moisture correction coefficient, which is the ratio of precipitation to evaporation.
2.3.2. Soil Fertility

The soil organic matter, total nitrogen, total phosphorus, and total potassium contents were used to calculate the soil fertility indicators. After the normalization of each factor, the comprehensive soil fertility index of each grid unit was calculated using the linearity synthetic method [18].

\[
Soil_i = \sum_{j=1}^{n} \mu_j \times Soil_j
\]  

(2)

where \(Soil_i\) is the soil fertility index of grid \(i\); \(Soil_j\) is the normalized soil nutrient content; and \(\mu_j\) is the weight coefficient of an individual single soil factor determined via the analytic hierarchy process.

2.3.3. Flood Risk

Flood disasters may have a certain degree of impact on agricultural activity due to the dense river networks in the TRB. The flood risk index was calculated from four indicators, including the multi-year average rainfall during flood season; altitude and terrain relief factors; river network distribution characteristics; and historical flood frequency [35]. Using the analytic hierarchy process [36], the weight coefficients of each factor were determined to be 0.239, 0.239, 0.433, and 0.089, respectively. Finally, the flood risk index of the TRB area was calculated.

2.3.4. Distance to Rural Settlements

The distance between the cropland fields and rural settlements is an important reference for farmers when selecting sites for land cultivation. The cultivation probability decreases as the distance increases, and vice versa. Therefore, the rural settlements present in 1911 were used to represent the rural settlements in the Republic of China period, and the rural settlements present in 1980 were used to represent settlements from 1949 to 1980.

2.3.5. Time Slice Selection and Cropland Area Calibration

We chose eight time slices, 1911, 1933, 1945, 1957, 1960, 1980, 2000, and 2010, based on data availability and representativeness to create cropland maps. Cropland area calibration mainly involves the unification and merging of administrative units from different historical periods, raster resampling, and unification of coordinate systems. In light of changes in administrative divisions, we calibrated the cropland area of each county using administrative maps from 2010. Adjacent counties that were involved in complex changes were merged and reallocated. Jintang county and Qingbaijiang district were merged into the Jinqing region, Jianyang city and Longquanyi district were merged into the Jianlong region, Dongxing and Shizhong districts were merged into the Neijiang region, Luxian and Longmatan were merged into the Longlu region, and the districts and counties in Zigong were merged into the Zigong region. All raster spatial covariates were projected onto WGS_1984_UTM_Zone_48 and resampled to produce a 30 m-resolution grid.

2.3.6. Cropland Data Reconciliation

There is a gap between the cropland archive area data and the cropland remote sensing area data. We used a reconciliation algorithm [37] to reconcile the archive data to the remote sensing data level. The reconciled cropland area from 2017 was derived from FROM-GLC10 data. The reconciliation algorithm is calculated as follows:

\[
A^R_{t1}(k) = a(k) \left[ A^R_{t1}(k) \frac{A^T_{C}(k)}{A^R_{C}(k)} \right] + (1 - a(k)) \left( A^R_{t1}(k) - (A^T_{C}(k) - A^R_{C}(k)) \right)
\]

(3)

where \(t1\) and \(t2\) are the current and previous years, respectively. \(A^R_{t1}(k)\) is the cropland area from the archive in year \(t1\) for county \(k\); \(A^T_{C}(k)\) is the cropland area from the archive in year \(t2\) for county \(k\); \(A^R_{C}(k)\) is the reconciled cropland area from the archive in year \(t1\) for county

The whole data preparation process is shown in a flowchart in Figure 2 and processed based on ArcGIS Pro 2.5.

Figure 2. Flowchart of data preparation process.

2.4. Cropland Reconstruction Method

2.4.1. Random Forest Classifier Algorithm

An RF algorithm is an ensemble machine learning and data mining approach composed of a set of decision trees. Created based on bootstrap resampling and a simple decision tree, the algorithm can be used for classification and regression and can achieve significant improvements in classification accuracy [38]. RF can handle high-dimensional data effectively, achieve high tolerance for noise and outliers, and reduce generalization error [22]. This approach is used to explore the occurrence probability associated with a specific land use or cover type in a grid pixel. The cropland distribution can be interpreted as binary data. Thus, contemporary cropland can be used as a binary dependent variable (cropland = 1, non-cropland = 0), and the driving factors that affect cropland can be used as independent variables. The RF algorithm is able to estimate variable importance via out-of-bag error measurements [39].

2.4.2. RF Model Training

A flowchart that describes model building and result analysis is presented in Figure 3. The natural and anthropogenic factors and cropland distribution map from FROM-GLC10 in 2017 were put into the RF model as independent variables and a dependent variable, respectively, for training. The RF model was executed using the “scikit-learn” package under the Python 3.6 environment in ArcGIS Pro 2.5. The RF model was used to model cropland probability. The training and testing sample sets were divided at an 8:2 ratio. Then, the predicted value of the random forest cropland gridding model after training was taken as the weight of the cropland grid allocation in the historical period. The number of decision trees (n_estimators) and the number of features (max_features) are the two critical parameters that affect RF classification accuracy. The grid search method was used to test each possibility in each parameter selection opportunity via loop traversal. Then, the best parameters were used in the latter analysis as the model approached a steady state. In this

\[
\alpha(k) = \min\left(1, \exp\left(-0.5 \left(\frac{A_R^{12}(k)}{A_R^{11}(k)} - 1.1\right) \right)\right)
\]
study, n_estimators and max_features were set to 150 and 3, respectively. At runtime in each county, the model was evaluated between the prediction and the validation using the receiver operating characteristic (ROC) curve and the area under curve (AUC) value. The results indicate good predictive accuracy performance and an average AUC value of 0.92 (Supplementary Materials Figure S2).

![Flowchart for reconstruction and assessment of cropland maps.](image)

**Figure 3.** Flowchart for reconstruction and assessment of cropland maps.

### 2.4.3. Spatial Allocation of Cropland

We performed the following steps under a Python 3.6 environment in ArcGIS Pro 2.5. The reconciled cropland area was combined with the adjacent year cropland change and cropland probability layers generated by the RF model, and was allocated to pixels in each county iteratively. During cropland area increase years, the newly added cropland grid was determined according to the distribution probability from high to low. During cropland area decrease years, the cropland grid converted to non-cropland was determined according to the cropland probability from low to high. Several constraint factors were used to exclude impossible cropland areas. No cultivated land could be distributed in an area with population density of over 3000 people/km², elevation over 2000 m, or waterbody grid. The model results for each year were aggregated into one dataset. Finally, we generated a spatio-temporal cropland dataset with 30 m resolution that described the past 100 years of the TRB.

### 2.4.4. Accuracy Assessment

We used a cropland cover map derived from satellite-based data from 1980 to perform the accuracy assessment. The comparison formula is as follows:

\[
\text{Differences}(i) = C_m(i) - C_s(i)
\]  

where Differences(i) is the absolute difference between the modeled and satellite-based data in grid i, and C_m(i) and C_s(i) are the cropland fractions from grid i in this study and the satellite-based data, respectively.

We also compared our results to three public global and regional datasets reported in previous studies (Table 2). The three datasets were the (1) widely used HYDE 3.2 dataset [12], (2) Yang-dataset, a historical cropland dataset that covers 300 years of the traditional
cultivated regions of China [18], and (3) ChinaCropland dataset, a cropland dataset from China that is based on multiple data sources [26]. Our dataset was up-scaled to a spatial resolution of 5 km for the grid-by-grid comparison. Notably, the cropland fraction was compared using the following formula:

\[ \text{Differences}(i) = C_m(i) - C_p(i) \]  

(6)

where \( \text{Differences}(i) \) is the absolute difference between the modeled data and the public dataset in grid \( i \), and \( C_m(i) \) and \( C_p(i) \) are the cropland fractions of grid \( i \) in this study and the public dataset, respectively.

Table 2. Global and national public cropland datasets used in the comparison.

| Dataset            | Temporal Coverage | Spatial Resolution | Source |
|--------------------|-------------------|--------------------|--------|
| HYDE 3.2           | 10,000 BCE–2015 CE| 5'                 | [12]   |
| Yang-dataset       | 1661–1980         | 1 km               | [18]   |
| ChinaCropland      | 1900–2016         | 5 km               | [26]   |

3. Results

3.1. Changes in Cropland Area during 1911–2010

Overall, the cropland area in the TRB increases rapidly before 1980, and then decreases slowly after 1980 (Figure 4). The cropland area increases steadily from \( 1.13 \times 10^4 \) km\(^2 \) in 1911 to \( 1.74 \times 10^4 \) km\(^2 \) in 1957. This indicates an average annual growth rate of 1.17%. There is a sharp decrease in cropland area from 1957 to 1960, which indicates an annual decrease rate of 1.81%. From 1960 to 1980, the amount of cropland increases rapidly from \( 1.64 \times 10^4 \) km\(^2 \), to a peak of \( 2.13 \times 10^4 \) km\(^2 \) in 1980. This indicates an annual growth rate of 1.46%. After 1980, the cropland area decreases to \( 1.81 \times 10^4 \) km\(^2 \) via a \(-0.51\%\) average annual growth rate.

![Cropland area variation](image)

Figure 4. Variation in the reconciled cropland area in the Tuojiang River Basin from 1911–2010.

3.2. Spatial Distribution of Cropland Patterns

The cropland maps in the TRB from 1911 to 2010 (including 1911, 1933, 1945, 1960, 1980, 2000, and 2010) were reconstructed using the RF model (Figure 5). In 1911, croplands are concentrated primarily in the upper reaches of the TRB with higher cropland fraction levels, particularly the Chengdu Plain area. In 1933, there is a cropland expansion in the middle of the basin (Figure 6). From then on, spatial agglomeration occurs in the upper reaches and spreads to the middle of the basin. In 1957, grids with cropland area fractions...
of 0–10%, more than 60%, and more than 80% account for 9.84%, 55.66%, and 25.18%, respectively, of the total. During 1957–1960, the cropland area decreases sharply in most counties. After 1960, dramatic growth occurs in the middle reaches of the basin. Between 1960 and 1980, a sharp increase in the cropland area is observed in both the middle and lower reaches of the river basin. In the 1980s, the cropland area reaches its maximum level in all regions. At this point, 72.42% of grids have a cropland fraction of more than 60%. Between 1911 and 1980, there is a great expansion in the amount of intensive cropland from the upper reaches to the middle and lower reaches of the TRB, as the cropland coverage fraction increases continuously.

Figure 5. Spatial distributions of 30 m × 30 m cropland grids in the Tuojiang River Basin from 1911–2010 (a–h).

Grids with cropland area fractions of 0–10% account for 18.57% of the total and those with fractions of more than 60% represent 20.95% of the total (Figure 7). In 2000, the percentage of grids with cropland fractions of 0–10% increases to 6.57%, and only 64.68% have more than 60% cropland. The largest quantities of abandoned cropland are detected in the Chengdu Plain areas of the TRB, such as in the Jinqing and Jianlong areas. The spatial distribution of cropland in 2010 is similar to that in 2000. Only a slight decrease in the cropland fraction is noted in the middle and lower parts of the basin. Grids with cropland fractions of 0–10% and more than 60% decrease to 9.33% and 56.71% of the total, respectively. Our spatial cropland reconstruction results reveal that the most intensive cropland area over the past century is distributed in the upper reaches of the TRB basin, mainly in the Chengdu Plain area, because of its fertile soil, flat terrain, and developed socio-economic conditions.
Figure 6. Cropland changes in the Tuojiang River Basin from 1911–2010 (a–g).

Figure 7. Cropland coverage changes in the Tuojiang River Basin in (a) 1911, (b) 1933, (c) 1945, (d) 1957, (e) 1960, (f) 1980, (g) 2000, and (h) 2010.
3.3. Verification of Simulation Results

We used remotely sensed cropland data from 1980 to validate the newly reconstructed cropland dataset (Figure 8). Comparison of the cropland fractions produced by this study and GlobeLand30 (1 km resolution) shows that 63.05% of grids have differences of between −20% and 20%. The greater the difference rate, the lower the proportion of the grid. The difference is close to a Gaussian distribution (Figure 9). This indicates that our reconstructed result captures the actual cropland pattern. Further spatial details are generated under more restrictive factors (Supplementary Materials Figure S3). However, there remains misinterpretations of two places in the GlobeLand30, which affected the results comparison. Firstly, we can see there almost no reclamation in Zigong county area (Figure 8b), which is contrary to the fact. However, in reality, this area is an important agricultural area, and clear evidence could be found in the agricultural survey data. On the other hand, other land use categories may be misinterpreted as cropland in GlobeLand30, especially in the mountain area of Anyue, Renshou and Weiyuan Counties [40], which caused a higher reclamation in these areas. The other noticeable difference existed in the northern mountain areas, which indicated that the cropland fraction was overestimated in the mountainous area in the study.

![Figure 8](image1.png)

**Figure 8.** Comparison of reconstructed and remote sensing data. (a) Cropland reconstruction for 1980; (b) remotely sensed cropland data in 1980; (c) difference between (a,b).

![Figure 9](image2.png)

**Figure 9.** Difference in the reconstruction result and the remotely sensed cropland data from 1980.
4. Discussion

4.1. Comparison to Public Cropland Datasets

To analyze the reasonableness of the reconstructed dataset further, three widely used regional and global datasets, ChinaCropland, the Yang-dataset, and HYDE 3.2, were selected for comparison with our results (Figure 10). For each year, the HYDE cropland is much smaller than the cropland area from our reconstructed result. This is consistent with the conclusion presented by Yu that the HYDE dataset underestimates the cropland area in the Sichuan Plain [26]. This difference may be caused by differences in data sources. The HYDE dataset aggregates statistical data at the country scale with no provincial-level inventory data, and allocates a cropland area directly based on demographic and socio-economic parameters [12]. Since the Yang dataset only collects provincial cropland archives in Sichuan and Chongqing, it overestimates the total cropland area in the TRB. The cropland area in our study most closely matches that from the ChinaCropland dataset, because they are both derived from National inventory data after 1980. Both datasets show that the basin cropland area is maximized in 1980.

![Figure 10. Comparison with cropland data from public datasets.](image-url)

Due to its consistent cropland trends, our cropland dataset was aggregated to 5 km resolution for spatial comparison with the ChinaCropland dataset at five time points (Figure 11). In 1911 and 1933, the ChinaCropland dataset indicates that cropland area is scarce in the upper reaches of the basin. However, this is inconsistent with fact. For comparison, the cropland area around Chengdu Plain in the upper reaches was considered in this study. The Chengdu Plain region cropland fractions identified via this study are significantly greater than those in the ChinaCropland database. For 1911–1933, 8% of grids exhibit a difference of more than 50% in the upper reaches of the basin. This indicates underestimation of the cropland area in the upper reaches. The ChinaCropland dataset shows that the cropland in the middle and southeastern portions of the basin expands significantly between 1933 and 2010. However, this is inconsistent with the actual changes in cropland area, and substantially different from the cropland spatial patterns in the study. The many extensive negative differences of more than 50% in the eastern portion of the basin indicate that ChinaCropland overestimates the cropland area in the region.
4.2. Forces That Drive Cropland Dynamics

4.2.1. Policy Factors and Cropland Area Data

Cropland expansion consists of two phases. The first phase is the climax of cropland expansion after the application of the household responsibility system in the 1980s [41], and the other phase is mainly concentrated after 2010. Historical changes in the TRB cropland area might potentially be explained via association with social productive forces and government policies. In the early 1900s, Migration Reclamation Planning was used to guide refugees to reclaim land. The cropland area in the TRB increased, as did the amount of cropland throughout the Republic of China [42]. In the 1950s, the land reform movement greatly encouraged farmers to engage in agricultural production [43]. The establishment of agricultural cooperatives is conducive to the construction of new water conservation facilities and land improvements. The agricultural production capacity improved significantly during this stage. However, the great leap forward and three-year natural disasters during 1959–1961 delivered a two-punch blow to agricultural development in the TRB [44]. After 1980, urbanization drove cropland decline in most parts of the TRB (Figure 5), in which plenty of cropland is occupied by non-agricultural construction land. The amount of cropland area in economically active regions has decreased during recent decades, due to rapid urban expansion. Since the implementation of the Grain for Green Project in the 1990s, some unsuitable cropland is converted into forestland [45]. The cropland area in the basin stabilized gradually after the Dynamic Balance of Total Farmland Area policy [46] was enacted in 1998 [47].
4.2.2. Causes of Cropland Distribution

The RF machine learning method can handle complex correlations between dependent geographic variables effectively [48]. The feature importance represents the degree that the accuracy of the RF model is influenced when an independent variable is replaced with randomly distributed data. The higher the importance of the feature, the more important the independent variable, and vice versa. Nearby cropland, the elevation, and the climatic productive potential are the top three factors that influence the cropland distribution (Figure 12). The feature importance of nearby cropland is 0.46, which is consistent with the principle that land is more likely to be reclaimed around existing cropland [16]. It also confirms that more cropland tends to intensify under the trend of agricultural development [49]. This is followed by the elevation and climatic productive potential data, where the feature importance values are 0.09 and 0.08, respectively. Natural factors such as the relief amplitude, soil erosion, and soil fertility are also of little influence. In addition, the quantity of ancillary data is vital to cropland distribution modeling. Our study adds more comprehensive variables that affect cropland distribution than traditional allocation methods [8,12]. These additional variables provide more complete and reliable data.

![Figure 12. The feature importance values for the independent variables.](image)

4.3. Uncertainties and Prospects

First, the availability of temporally dynamic data limited the selection of influencing factors. Undoubtedly, the RF model performance may improve if more variables are accessible. Some of the anthropogenic variables were unavailable beyond recent decades, and remained unchanged in the study. Therefore, related driving factor datasets with high resolutions and long time-spans are urgently needed to support more precise modeling. In addition, urban areas without cropland are defined based on a population density threshold of 3000. However, historical areas with fewer than 3000 people might be similarly urban. The value used in the study may not be optimal for some regions. The altitude threshold is also set to the same value for mountainous area throughout the period, which may cause an overestimation of cropland in the history. As a result, it cannot be used as the basis for some local research. However, our flexible framework allows researchers to replace input data or revise model parameters to acquire the best results for their specific study areas. To improve the model further, one can seek to reduce the uncertainties of the most sensitive parameters, and include more process details and additional constraints as more data become available. Secondly, although inevitable misinterpretations exist in small local regions (Figure 8b), the remotely sensed data are still a reliable validation method. It also stresses the necessity of cropland reconstruction. In the future work, regional remotely sensed data with a higher precision can be used for verification.

This study can provide a method for the reconstruction of other land use types in a more extensive spatial range (continental to global scale), and at a longer time scale. In future work, we are challenged to incorporate more detailed indicators, which may improve...
the accuracy of mapping. There is also potential for improvement in the machine learning models. Even though RFs are used in this study, there are still many other machine learning methods, such as artificial networks, generalized linear models, deep learning models, and various ensemble learning algorithms. Notably, ensemble models have been shown to provide better modeling of land use change probabilities. They can improve the efficiencies of base classifiers that are used as ensemble classifiers and minimize bias, variance, and overfitting problems during the deforestation probability modeling process [50]. These simulation rules can be changed via the use of other programming languages that improve the simulation capabilities of the machine learning models [21].

5. Conclusions

This paper proposed an RF model workflow for historical cropland reconstruction. Based on calibrated historical cropland data, we developed a century-long historical TRB cropland dataset with 30 m resolution using a RF model. Our spatial cropland reconstruction result revealed that the most intensive cropland area over the past century was distributed in the upper reaches of the TRB. The model achieved precision superior to those of existing widely used public datasets and approximated actual cropland patterns. Nearby cropland, the elevation, and the climatic productive potential were the key factors that affected cropland distribution in the TRB. After accuracy assessments, the proposed method was validated as a promising way for the reconstruction of high-resolution cropland grids. This framework can be applied to larger areas with more complex conditions or other land use or cover changes via its flexible structure and parameters. In the future, we suggest that researchers reconstruct historical cropland patterns using ensemble machine learning algorithms, and compare the resulting data accuracy.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/land10121338/s1, Figure S1: Dynamics of the statistical cropland area in the Tuojiang River Basin from 1911 to 2010. Figure S2: Receiver operator characteristic curves and associated area under the curve for the random forest models for 20 counties. Figure S3: Comparison of remotely sensed cropland data and modeled cropland maps for the year 1980. (a) Cropland map generated by random forest in the Tuojiang River Basin, (b) Remotely sensed cropland data in the local area, and (c) Reconstructed cropland in the local area.

Author Contributions: Q.W.: Conceptualization, Methodology, Writing; M.X.: Methodology, Data curation, Writing—review and editing; Q.L.: Data curation, Drawing, Software, Validation; H.L.: Data curation, Drawing, Software, Validation; T.L.: Drawing, Reviewing and Editing; O.D.: Writing—review and editing; R.H.: Writing—review and editing; M.Z.: Writing—review and editing; X.G.: Conceptualization, Reviewing and Editing, Funding acquisition, Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sichuan Science and Technology Program (grant number 2021YFH0121).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments: The authors thank the anonymous reviewers for the helpful comments that improved this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Lambin, E.F.; Meyfroidt, P. Global land use change, economic globalization, and the looming land scarcity. Proc. Natl. Acad. Sci. USA 2011, 108, 3465–3472. [CrossRef] [PubMed]
2. Godfray, C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Robinson, S.; Thomas, S.M.; Toulmin, C.; Pretty, J. Food security: The challenge of feeding 9 billion people. Science 2010, 327, 812–818. [CrossRef] [PubMed]
3. Shi, X.; Wang, W.; Shi, W. Progress on quantitative assessment of the impacts of climate change and human activities on cropland change. *J. Geogr. Sci.* **2016**, *26*, 339–354. [CrossRef]

4. Houghton, R.A. Interactions between land-use change and climate-carbon cycle feedbacks. *Curr. Clim. Chang. Rep.* **2018**, *4*, 115–127. [CrossRef]

5. Molotoks, A.; Stehfest, E.; Doelman, J.; Albanito, F.; Fitton, N.; Dawson, T.P.; Smith, P. Global projections of future cropland expansion to 2050 and direct impacts on biodiversity and carbon storage. *Glob. Chang. Biol.* **2018**, *24*, 5895–5908. [CrossRef]

6. Zabel, F.; Delzeit, R.; Schneider, J.M.; Seppelt, R.; Mauser, W.; Václavík, T. Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity. *Nat. Commun.* **2019**, *10*, 2844. [CrossRef]

7. Gervois, S.; Ciais, P.; Noblet-Ducoudré, N.; Brisson, N.; Vuichard, N.; Viovy, N. Carbon and water balance of European croplands throughout the 20th century. *Glob. Biogeochem. Cycles.* **2008**, *22*. [CrossRef]

8. Wei, X.; Widgren, M.; Li, B.; Ye, Y.; Fang, X.; Zhang, C.; Chen, T. Dataset of 1 km cropland cover from 1690 to 1999 in Scandinavia. *Earth Syst. Sci. Data* **2021**, *13*, 3035–3056. [CrossRef]

9. Yu, Z.; Lu, C.; Cao, F.; Tian, H. Long-term terrestrial carbon dynamics in the Midwestern United States during 1850–2015: Roles of land use and cover change and agricultural management. *Glob. Chang. Biol.* **2018**, *24*, 2673–2690. [CrossRef]

10. Congalton, R.G.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. Global land cover mapping: A review and uncertainty analysis. *Remote Sens.* **2014**, *6*, 12070–12093. [CrossRef]

11. Meiyappan, P.; Jain, A.K. Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. *Front. Earth Sci.* **2012**, *6*, 122–139. [CrossRef]

12. Klein, G.K.; Beusen, A.; Doelman, J.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Toulmin, C. Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth Syst. Sci. Data* **2017**, *9*, 927–953. [CrossRef]

13. Yu, Z.; Lu, C.; Tian, H.; Canadell, J.G. Largely underestimated carbon emission from land use and land cover change in the contiguous United States. *Glob. Chang. Biol.* **2019**, *25*, 3741–3752. [CrossRef] [PubMed]

14. Yu, Z.; Lu, C. Historical cropland expansion and abandonment in the continental US during 1850 to 2016. *Glob. Ecol. Biogeogr.* **2018**, *27*, 322–333. [CrossRef]

15. Paudel, B.; Zhang, Y.; Li, S.; Liu, L. Spatiotemporal changes in agricultural land cover in Nepal over the last 100 years. *J. Geogr. Sci.* **2018**, *28*, 1519–1537. [CrossRef]

16. Wei, X.; Ye, Y.; Zhang, Q.; Fang, X. Reconstruction of cropland change over the past 300 years in the Jing-Jin-Ji area, China. *Reg. Environ. Chang.* **2016**, *16*, 2097–2109. [CrossRef]

17. Moulds, S.; Buytaert, W.; Mijic, A. A spatio-temporal land use and land cover reconstruction for India from 1960–2010. *Sci. Data* **2018**, *5*, 180159. [CrossRef]

18. Yang, X.; Jin, X.; Guo, B.; Ying, L.; Zhou, Y. Research on reconstructing spatial distribution of historical cropland over 300 years in traditional cultivated regions of China. *Glob. Planet. Chang.* **2015**, *128*, 90–102. [CrossRef]

19. Yang, X.; Jin, X.; Du, X.; Xiang, X.; Han, J.; Shan, W.; Fan, Y.; Zhou, Y. Multi-agent model-based historical cropland spatial pattern reconstruction for 1661–1952, Shandong Province, China. *Glob. Planet. Chang.* **2016**, *143*, 175–188. [CrossRef]

20. Samardžić-Petrović, M.; Kovačević, M.; Bajat, B.; Dragičević, S. Machine learning techniques for modelling short term land-use change. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 387. [CrossRef]

21. Aburas, M.M.; Ahamad, M.S.S.; Omar, N.Q. Spatio-temporal simulation and prediction of land use and cover change using conventional and machine learning models: A review. *Environ. Monit. Assess.* **2019**, *191*, 205. [CrossRef] [PubMed]

22. Breima, L. Random Forests. *Mach. Learn.* **2010**, *45*, 5–32. [CrossRef]

23. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]

24. Ray, D.K.; Pijanowski, B.C. A backcast land use change model to generate past land use maps: Application and validation at the Muskegon River watershed of Michigan, USA. *J. Land Use Sci.* **2010**, *5*, 1–29. [CrossRef]

25. Li, Q.; Li, S.; Xiao, Y.; Zhao, B.; Wang, C.; Li, B.; Gao, X.; Li, Y.; Bai, G.; Wang, Y. Soil acidification and its influencing factors in the purple hilly area of southwest China from 1981 to 2012. *Catera* **2019**, *175*, 278–285. [CrossRef]

26. Yu, Z.; Jin, X.; Miao, L.; Yang, X. A historical reconstruction of cropland in China from 1900 to 2016. *Earth Syst. Sci. Data* **2021**, *13*, 3203–3218. [CrossRef]

27. Peng, W.F.; Zhou, J.M. Change of cultivated land area in sichuan province. *Resour. Sci.* **2005**, *27*, 15–29. [CrossRef]

28. Buck, J.K. Land Utilization in China; Cheng City Press: Chengdu, China, 1941.

29. National Bureau of Statistics. *Summary of Rural Economic Statistics by County in China*; China Statistics Press: Beijing, China, 1991.

30. National Bureau of Statistics. *China Regional Economic Statistics Yearbook*; China Statistics Press: Beijing, China, 2005.

31. Gong, P.; Liu, H.; Zhang, M.; Li, C.; Wang, J.; Huang, H.; Clinton, N.; Ji, L.; Li, W.; Bai, Y.; et al. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. [CrossRef]

32. Qin, Y.; Liu, J.; Shi, W.; Tao, F.; Yan, H. Spatial-temporal changes of cropland and climate potential productivity in northern China during 1990–2010. *Food Secur.* **2013**, *5*, 499–512. [CrossRef]

33. Wang, Q.; Gao, X.; Li, Q.; Lan, T.; Huang, R.; Deng, O. Spatially explicit reconstruction of the population distribution in the Tuojiang River Basin during 1911–2010 using random forest regression. *Reg. Environ. Chang.* **2021**, accepted.

34. Huang, B. Chinese agricultural potential productivity-photosynthetic potential productivity. *Ann. Geogr.* **1985**, *17*, 15–22.
35. Cai, Q.; Sheng, Z. Temporal and spatial characteristics of flood disaster in Tuojiang River Basin in Ming and Qing Dynasties. *J. China Three Gorges Univ.* **2020**, *42*, 104–109.

36. He, B.; Zhang, S.; Du, Y.; Zhongshan, Y.; Li, B. Flood risk assessment of Hubei province. *J. Yangtze River Sci. Res.* **2004**, *21*, 21–25.

37. Ramankutty, N.; Foley, J.A. Estimating historical changes in land cover: North American croplands from 1850 to 1992. *Global Ecol. Biogeogr.* **1999**, *8*, 381–396. [CrossRef]

38. Marchese, R.L.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792. [CrossRef] [PubMed]

39. Calle, M.L.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **2011**, *12*, 86–89. [CrossRef]

40. Li, Y.; Ye, Y.; Zhang, C.; Li, J.; Fang, X. A spatially explicit reconstruction of cropland based on expansion of polders in the Dongting Plain in China during 1750–1985. *Reg. Environ. Chang.* **2019**, *19*, 2507–2519. [CrossRef]

41. Tian, G.; Duan, J.; Yang, L. Spatio-temporal pattern and driving mechanisms of cropland circulation in China. *Land Use Policy* **2021**, *100*, 105118. [CrossRef]

42. Guo, C. On the transition of Sichuan agriculture in the period of the Republic of China. *J. Sichuan Norm. Univ.* **1997**, *24*, 111–119.

43. Li, L. On the Land Reform Movement Immediately after the Foundation of People’s Republic of China. *J. Jiangsu Univ.* **2004**, *6*, 39–43.

44. Bi, Y.Y.; Zheng, Z.Y. The actual changes of cultivated area since the founding of new China. *Resour. Sci.* **2000**, *22*, 8–12.

45. Liu, F.; Zhang, Z.; Zhao, X.; Wang, X.; Zuo, L.; Wen, Q.; Yi, L.; Xu, J.; Hu, S.; Liu, B. Chinese cropland losses due to urban expansion in the past four decades. *Sci. Total Environ.* **2019**, *650*, 847–857. [CrossRef] [PubMed]

46. Xin, L.; Li, X. China should not massively reclaim new farmland. *Land Use Policy* **2018**, *72*, 12–15. [CrossRef]

47. He, X.; Yan, J.; Cheng, X. Household perspective on cropland expansion on the Tibetan Plateau. *Reg. Environ. Chang.* **2021**, *21*, 21. [CrossRef]

48. Liu, Y.; Cao, G.; Zhao, N.; Mulligan, K.; Ye, X. Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach. *Environ. Pollut.* **2018**, *235*, 272–282. [CrossRef] [PubMed]

49. Hu, Q.; Xiang, M.; Chen, D.; Zhou, J.; Song, Q. Global cropland intensification surpassed expansion between 2000 and 2010: A spatio-temporal analysis based on GlobeLand30. *Sci. Total Environ.* **2020**, *746*, 141035. [CrossRef]

50. Saha, S.; Saha, M.; Mukherjee, K.; Arabameri, A.; Paul, G.C. Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, REPTree: A case study at the Gumani River Basin, India. *Sci. Total Environ.* **2020**, *730*, 139197. [CrossRef]