frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 18 June 2025
DOI 10.3389/fenvs.2025.1517085

OPEN ACCESS

EDITED BY
Shuisen Chen,
Guangzhou Institute of Geography, China

REVIEWED BY
Yahui Guo,
Central China Normal University, China
Guan Xiaoke,
Zhengzhou University of Light Industry, China
Yaohui Liu,
Shandong Jianzhu University, China

*CORRESPONDENCE
Jun Zhou,
zhoujun200208@163.com

RECEIVED 25 October 2024
ACCEPTED 02 June 2025
PUBLISHED 18 June 2025

CITATION
Zhou H, Zhou J, Lu K, Niu M, Wang C, Zhang G and Kou J (2025) Marginal land identification and grain production capacity prediction of the coverage area of western route of China's South-to-North Water Diversion Project. *Front. Environ. Sci.* 13:1517085. doi: 10.3389/fenvs.2025.1517085

Marginal land identification and grain production capacity prediction of the coverage area of western route of China's South-to-North Water Diversion Project

Heng Zhou[1], Jun Zhou[2,3]*, Kunming Lu[1], Minghui Niu[1], Chenyi Wang[2], Gaofeng Zhang[1] and Jiawei Kou[1]

[1]Northwest Engineering Corporation Limited, Xi'an, China, [2]College of Land Science and Technology, China Agricultural University, Beijing, China, [3]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, China

The western route of the South-to-North Water Diversion Project (SNWDP) provides opportunities to improve agricultural production by altering regional water availability. This study identifies and evaluates marginal land—defined as undeveloped reserve cultivated land and low-quality and inefficiently-utilized farmland—within provinces along the SNWDP route. Using ecological, topographic, climatic, and soil indicators, we identified 145,062 km^2 of marginal land, including 3,626 km^2 of reserve cultivated land and 141,436 km^2 of low-quality and inefficiently-utilized farmland, mainly concentrated in northwestern Xinjiang, with Qinghai having the least. To assess the grain production potential of these lands, we used maize and wheat as representative crops. Three modeling approaches—random forest regression, gradient boosted regression trees, and two-point machine learning (TPML)—were compared for their predictive accuracy. The TPML model showed the best performance. For maize, the model yielded a root mean square error (RMSE) of 48.94, a mean absolute error (MAE) of 34.01, and a mean absolute percentage error (MAPE) of 7.65%. For wheat, the RMSE was 23.92, MAE 17.67, and MAPE 6.31%. Results reveal that maize has a higher production capacity than wheat, and that grain yields are higher in the west and lower in the east, with Xinjiang showing the highest average yields on marginal land. These findings provide a scientific basis for optimizing land use, improving food self-sufficiency, and supporting regional sustainable development and national food security.

KEYWORDS

marginal land, grain production capacity, corn, wheat, machine learning, the South-to-North Water Diversion Project

1 Introduction

In the face of the uneven distribution of water resources in China, the western route of the South-to-North Water Transfer Project, as a major inter-basin water transfer project, has potential benefits that cannot be ignored. Marginal lands have low agricultural production capacity, economic efficiency, and fragile ecology. This is due to significant soil barriers, severe

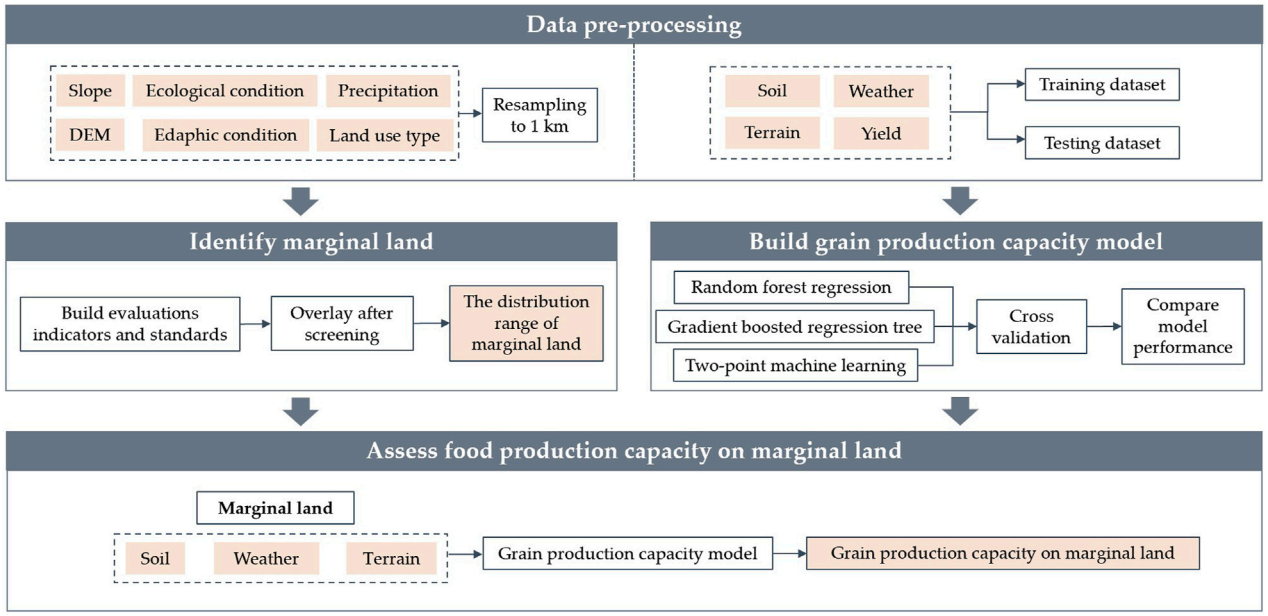

FIGURE 1
Technology roadmap framework.

water and heat resource constraints, and topographical limitations. They include cultivated reserves to be developed and marginal cultivated lands with poor quality and low use efficiency (Cao et al., 2021; Shortall, 2013). With population growth and economic development, the issue of food security has become increasingly important. Given the constraint of limited cultivated land resources, the exploitation of marginal lands and the enhancement of food production capacity have emerged as pivotal strategies for safeguarding national food security. However, most existing studies focus on improving yields on currently cultivated land, with limited attention paid to marginal land that remains underutilized due to environmental and infrastructural limitations. In particular, few studies have examined the production potential of marginal land in the northwest region under complex terrain and resource constraints.

Corn and wheat are currently the most widely used grain yield data in various studies, and they are not only widely grown and distributed crops in China (Wang et al., 2014), but also important food and feed crops. Moreover, the data of corn and wheat are easy to obtain and have high accuracy. Therefore, corn and wheat were selected to represent grain in this study. Grain yield data is a direct result of agricultural production activities, which directly reflects production capacity and can be used to measure production efficiency. Yield was used to characterize the actual production capacity of the land. In current studies, corn and wheat yield focus on predicting the production potential of large-scale arable land and analyzing the factors influencing yield (Ren et al., 2008; Zhang et al., 2014; Han et al., 2020; Huang et al., 2015; Song et al., 2016; Cheng et al., 2022), providing references for formulating agricultural policies, optimizing cropping structures and improving production efficiency. However, relatively few studies have been carried out on marginal lands that have development potential but have not been fully exploited, particularly in the north-western region.

Due to natural constraints such as arid climate, wind erosion, and soil salinization, a significant portion of marginal land in the wind-blown and arid/semi-arid regions of Northwest China remains underutilized and undeveloped, resulting in unfulfilled agricultural potential and an untapped reserve of arable land. The development of marginal land plays a critical role in safeguarding national food security. This study aims to fill this research gap by focusing on marginal land in the northwest region of China—specifically along the western route of the South-to-North Water Diversion Project—and assessing its grain production potential using machine learning approaches.

Grain capacity modeling is a quantitative analysis used to predict grain yields in a given region or over a given period. Machine learning methods are now widely used to predict yields (Rashid et al., 2021; Fei et al., 2023; Fu et al., 2021; Guo et al., 2021; Guo et al., 2022; Guo et al., 2023). Researchers used the similarity of covariates between points to build models of grain production capacity. Shrestha et al. derived a linear regression model between the curve of Normalized Difference Vegetation Index and the yield of corn (Shrestha et al., 2016); Wang et al. used remote sensing data, meteorological data and soil data as characteristic variables, analyzed the importance of the variables based on the Random Forest (RF) algorithm and built a wheat yield prediction model (Wang L. et al., 2022). Sun et al. predicted winter wheat yield from the perspective of county-level yield prediction, combining a convolution neural network and a backpropagation neural network to predict winter wheat yield (Sun et al., 2022). These methods can handle high-dimensional variables, but ignore spatial neighbors. Two-point machine learning (TPML) approach makes full use of spatial autocorrelation and attribute correlation, which can alleviate the problem of dimensional catastrophe in local modeling. It avoids the common factor covariance problem in regression prediction models and is able to improve more accurate spatial prediction results (Gao et al., 2022; Wang Y. et al., 2022). To our knowledge, TPML has not yet been applied to the evaluation of grain productivity on marginal land in northwest China, making this study a novel application of the method in this context.

FIGURE 2
Overview of the study area.

The marginal land evaluation indices and criteria are first established on the basis of data relating to four aspects: ecology, topography, climate and soil. Factors influencing grain yield were selected from soil, meteorology and topography, and grain yield models were built using RF, gradient-enhanced regression tree (GBRT) and TPML respectively, comparing the performance of the three models and selecting the optimal model for predicting corn and wheat yield on marginal lands. By obtaining the grain yield of marginal lands in the area covered by the western route of the south-to-north water diversion, this study makes it possible to identify marginal lands and assess the grain production capacity in the region, assisting in optimizing the allocation of land resources and improving grain self-sufficiency (Figure 1).

2 Materials and methods

2.1 Overview of the study area

The South-to-North Water Diversion Project constructs dams in the upper reaches of the Tongtian River, a tributary of the Yangtze River, and the Yalong River and the Dadu River. The water transfer tunnels through the Ba Yan Ka La Mountain, which is the watershed between the Yangtze River and the Yellow River, are excavated to transfer water from the Yangtze River to the upper reaches of the Yellow River. It will solve the water shortage problem in the Northwest China and the upper and middle reaches of the Yellow River. The study area spans several provinces, including Xinjiang Uygur Autonomous Region, Qinghai province, Gansu province, Ningxia province, Shaanxi province, Shanxi province and the western region of Inner Mongolia (Figure 2). The study area is the main extent of water transfer from the western route of the South-North Water Diversion. The project plays an important role in ensuring regional water security and promoting sustainable agricultural development.

The study area is located in northwestern China and has a variety of climate types, including temperate continental climate and alpine plateau climate. The terrain is complex, including mountains, plateaus, basins, grasslands and other types of terrain, with a large number of ups and downs in the terrain (Figure 3). Although water resources are relatively scarce, some areas still have rivers and lakes, and the average annual precipitation ranges from 0 to 1,200 mm (Figure 4). It has a wide geographic space and rich natural resources, which are important for irrigated agriculture, industrial production and ecological protection. The population is unevenly distributed

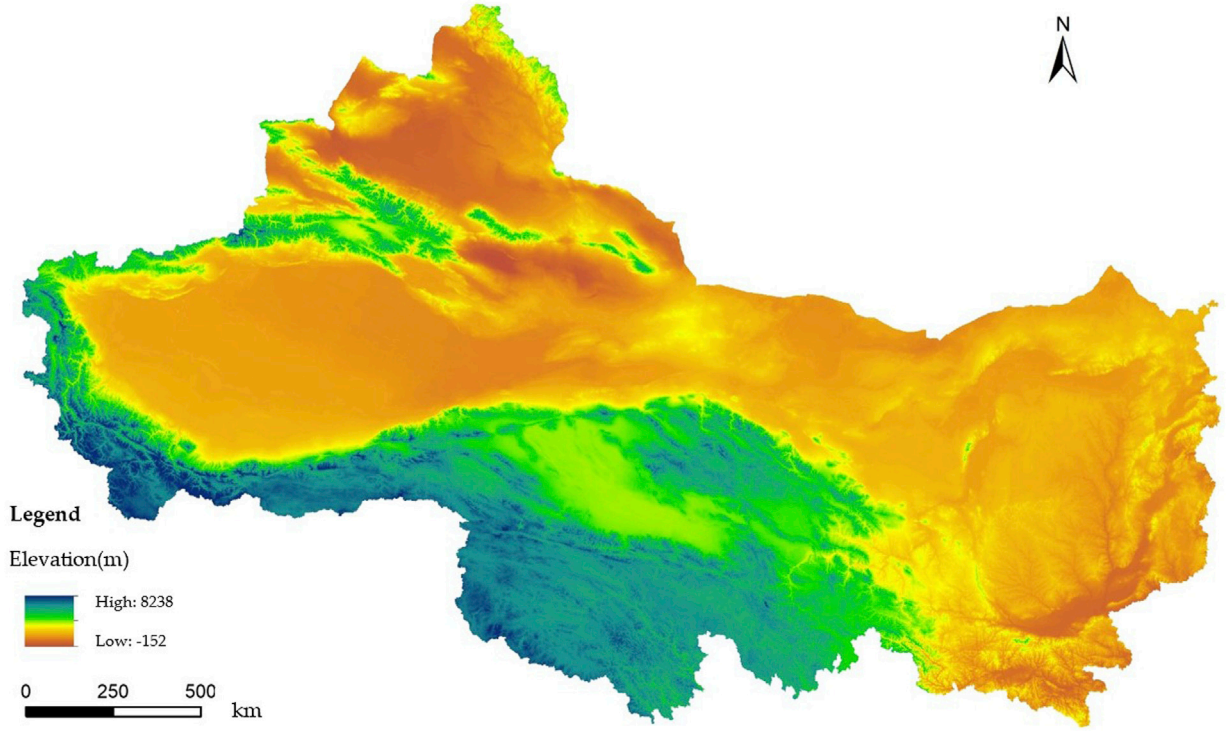

FIGURE 3
Altitude conditions in the study area.

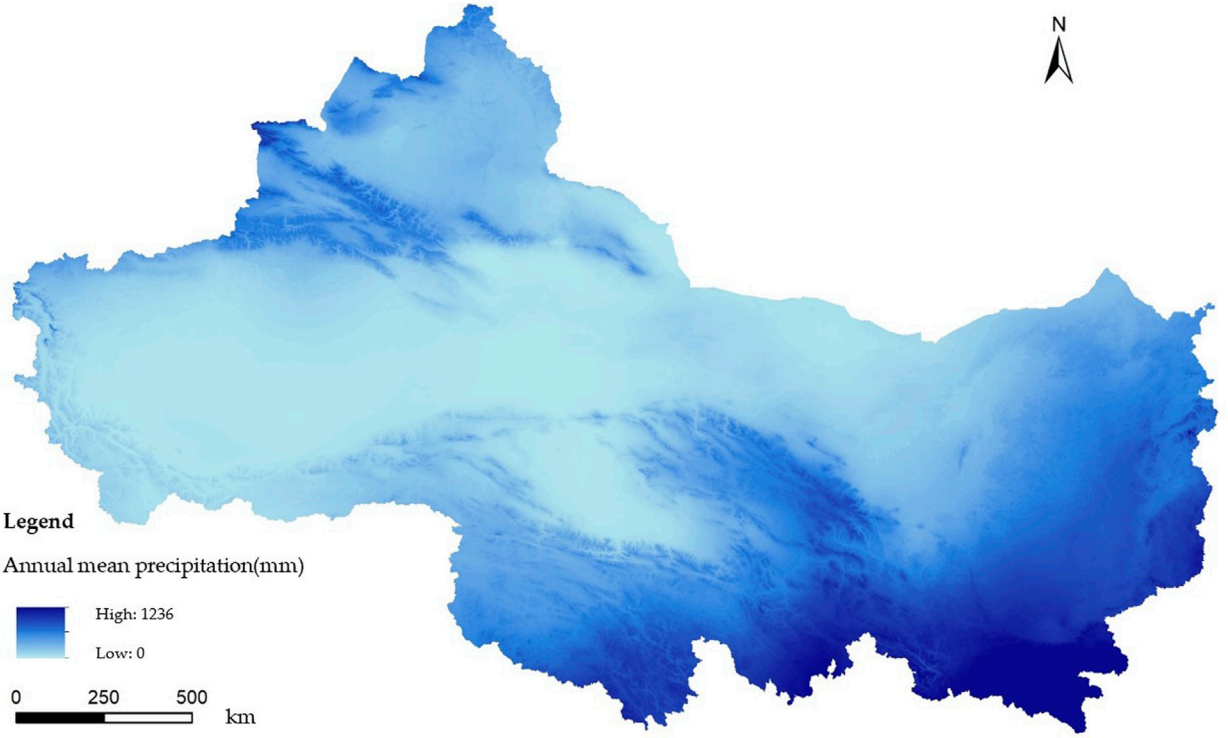

FIGURE 4
Precipitation conditions in the study area.

TABLE 1 Types and sources of data.

Data name	Type	Year	Data source
Land use type	CD	2015	https://www.resdc.cn/
DEM	LD	2015	https://www.resdc.cn/
Slope	LD	2015	Calculated by DEM
Soil data	LD	2015	The second national soil survey data
Mean annual precipitation	LD	—	https://www.worldclim.org/
Mean annual temperature	LD	—	https://www.worldclim.org/
Yield	LD	2015	Yield investigation

[a]CD is a categorical variable and LD is a continuous variable.

TABLE 2 Evaluation indicators and criteria for undeveloped reserve cultivated land resources.

Evaluating indicator	Reserve cultivated land resources to be developed
Slope	≤15°
Altitude	≤1,500 m
Annual precipitation	≥350 mm
Ecological condition	Not in nature reserves, high biodiversity
Soil pollution	The soil is pollution-free
Soil pH value	5–7.5
Organic matter content	≥20 g. kg^{-1}
Degree of salinity	Light
Soil texture	Clay or loam
Land area	≥10 km^2
Land use type	Grassland, saline-alkali land, sandy land, bare land

but rich in labor resources, and it is mainly planted with wheat, corn and other grain crops, making it an important base for grain crop production. It is an important region in western development strategy in China and the "One Belt, One Road" initiative, and the marginal land involved in the project is an area with high potential for agricultural and food production. At the same time, the region faces challenges such as fragile ecological environment and water shortage.

2.2 Data sources

The experience and results of previous researchers in related fields are reviewed using the literature search method. Through in-depth analyses and summaries of existing literature, data on topography, meteorology, soil, land use and food in the study area are collected using a combination of methods such as Internet and field surveys, and data are processed with missing and outlier data, and data are visualized and processed using ArcGIS software.

Soil, land use and topography data are the basis for the extraction of marginal land extent and studies on capacity potential (Csikós and Tóth, 2023). Average annual precipitation and temperature were calculated for the period 2010–2020 (Table 1). Grain yield data were used for corn and wheat yields in kg/acre. With the exception of yield, which is a point data, all other data were resampled so that the resolution after resampling was 1 km.

2.3 Methods of the research

2.3.1 Random forest regression (RF)

RF is an integrated learning algorithm. It predicts continuous values by constructing several decision trees. Each tree is built independently of the original data based on the autonomous sampling method, and features are randomly selected for splitting until each decision tree has reached its maximum size (Breiman, 2001). When a new data point is to be predicted, it passes through all the trees to obtain several predicted values, and the final prediction is the average of these values.

RF algorithm is capable of handling a large number of input variables and assessing the importance of variables to analyze the extent of influence of different factors on grain yield (Archer and Kirnes, 2008). It is insensitive to missing values and capable of

TABLE 3 Evaluation indicators and criteria for low-quality and inefficiently-utilized farmland.

Evaluating indicator	Reserve cultivated land resources to be developed
Slope	>15°
Annual precipitation	<350 mm
Ecological condition	Low biodiversity
Soil pollution	The soil is polluted
Soil pH value	pH < 5.0 or pH > 7.5
Organic matter content	<20 g. kg^{-1}
Soil texture	Sand
Land use type	Cultivated land

handling unbalanced data sets, and can make full use of grain yield data (Jeong et al., 2016). Finally, RF has better generalization performance and can effectively reduce the risk of over-fitting.

2.3.2 Gradient boosting regression tree (GBRT)

GBRT progressively optimize the predictive power of a model by sequentially building several decision trees. During the iterative learning process, each tree in the sequence is learned from the residuals of the previous tree (Elith et al., 2008). It is trained in the direction of the negative gradient of the loss function, and a strong learner is generated by linearly combining weak learners over several training sessions.

GBRT can be implemented for categorical and numerical data by optimizing different loss functions and offering multiple hyperparameter tuning options, making function fitting more flexible. In addition, GBRT can handle missing data and avoid over-fitting by building simple trees at each iteration.

2.3.3 Two-point machine learning (TPML)

TPML unifies spatial autocorrelation and attribute similarity in a high-dimensional space, making full use of information from spatial neighbors and high-dimensional covariates to improve prediction accuracy. The algorithm first calculates the differences between target variables and covariates between different pairs of points, uses the differences to build a model for predicting the differences between target variables between a particular observation point and the point to be observed, and then combines information from nearest neighbors to obtain the final prediction value for the point to be observed (Gao et al., 2022; Wang Y. et al., 2022).

Unlike traditional supervised learning methods, TPML solves the problem of dimensionality catastrophe in local machine learning modeling, avoids the problem of factor covariance in regression models, and its standard error deviation can provide uncertainty estimates for prediction results (Wang Y. et al., 2022).

2.3.4 Model validation

Tenfold cross-validation is a method for evaluating the performance of machine learning models (Lei, 2020). It evaluates the performance of a model by dividing the dataset into ten parts and using nine of these as both the training set and the other as the test set. This method makes efficient use of limited data resources for multiple experiments, improving the accuracy and reliability of model evaluation (Hengl et al., 2017). Accordingly, three measures were selected to compare the effectiveness of model fitting, namely, root mean square error RMSE, mean absolute error MAE and mean relative error MAPE, with the following formulae (Equations 1-3):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3)$$

where n is the number of samples, y_i is the true value of the ith sample, and $\hat{y}_i$ is the predicted value of the ith sample.

3 Marginal land extraction and spatial distribution

3.1 Marginal land identification

The selection of evaluation indicators is grounded in extensive literature and national land evaluation standards. For instance, soil pH, organic matter content, and texture directly correlate with crop yield and soil fertility, while ecological and topographical factors, such as altitude and slope, are critical in determining land suitability for cultivation. The classification standards for each indicator are established in accordance with existing studies on the growth conditions of major crops like corn and wheat, ensuring the scientific rigor and practicality of the evaluation process.

3.1.1 Reserve cultivated land resources to be developed

Evaluation indicators are selected from four aspects: ecology, topography, climate and soil, and specific evaluation indicators include ecological conditions, topographic slope, altitude, annual precipitation, soil texture, soil pollution status by heavy metals, soil organic matter content, degree of salinization, soil pH value, plot area and land use type, a total of eleven indicators (Jiang et al., 2019; Yao et al., 2021; Lei et al., 2011). The evaluation indices and

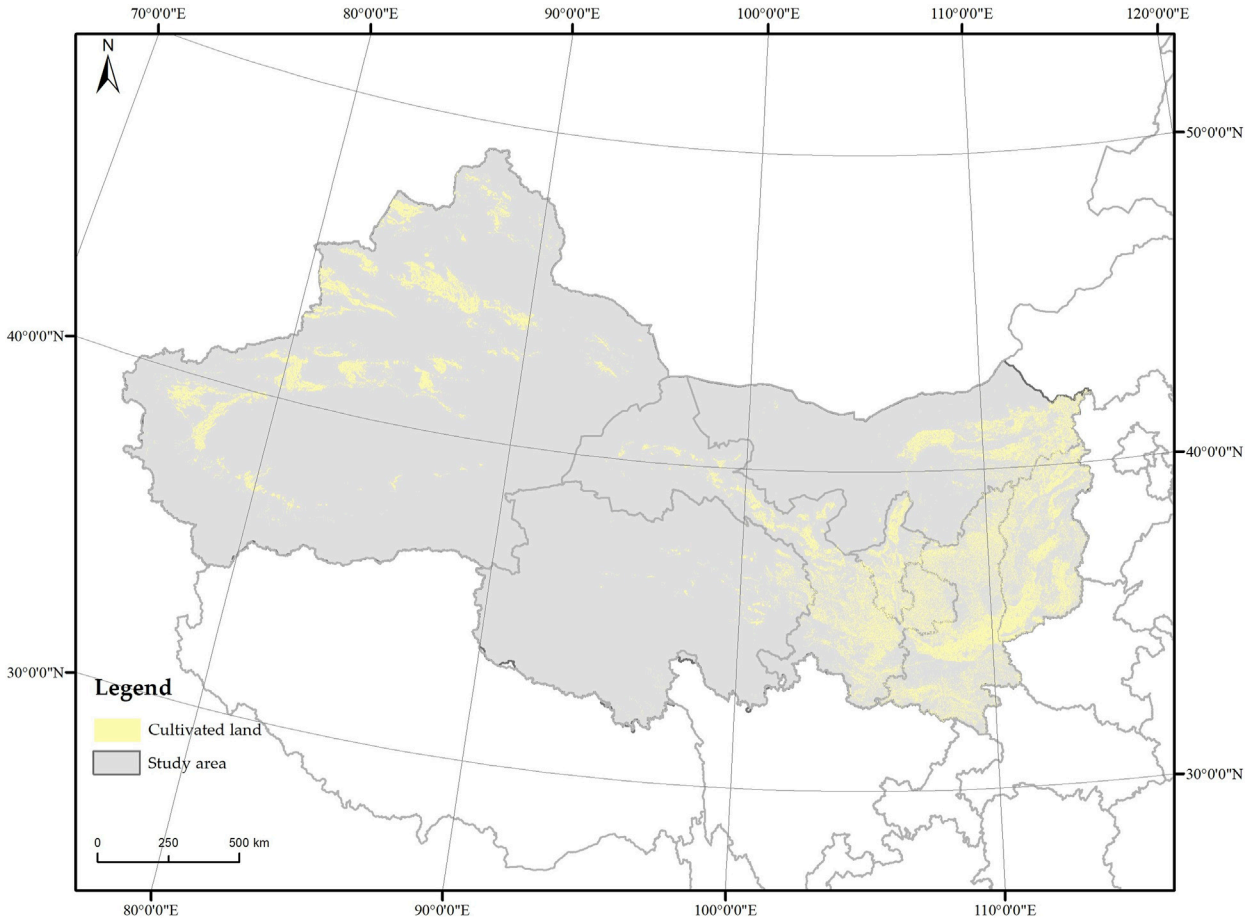

FIGURE 5
The distribution of existing cultivated land.

TABLE 4 Marginal land area of each province.

Province	Area of marginal land (km^2)	Share of marginal land area
Xinjiang	52,204	35.98%
Qinghai	792	0.54%
Gansu	30,539	21.05%
Ningxia	9,751	6.72%
Inner Mongolia	12,250	8.44%
Shaanxi	12,307	8.48%
Shanxi	27,747	19.12%

classification standards are shown in the table below (Table 2), and the experimental operation is carried out in ArcMap10.2.

Topographic slope is one of the main factors influencing land use. It affects soil moisture loss and the ease of development and land use. The study area has a fragile ecological environment, where soil erosion and water loss are common, and plots are classified with 15° as the slope limit. Annual precipitation is one of the indicators that influence agricultural production, and an appropriate volume of precipitation favors the production and development of foodstuffs and improves production capacity and quality. Soil texture is classified into sand, loam and clay according to the proportion of particle composition, which is one of the criteria for measuring soil fertility and grain production capacity, and directly affects the soil's ability to retain water and fertilizers. Corn and wheat can be grown in loam and clay. Soil pH affects the growth of the crop's root system, and the most favorable conditions for grain growth are those where the pH is more than neutral, and the standard is formulated according to the growing conditions of corn and wheat. Soil

TABLE 5 The influence factors of grain production capacity.

Data classification	Field name	Implication
Soil	pH	pH value of the soil
	T_SAND	Sand content in 0–30 cm soil layer
	T_CLAY	Clay content in 0–30 cm soil layer
	S_SAND	Sand content in 30–100 cm soil layer
	S_CLAY	Clay content in 30–100 cm soil layer
Weather	tem_a	Average annual precipitation
	pre_a	Average annual temperature
Terrain	DEM	Altitude
	Slope	Slope
Yield	Yield	Grain production

TABLE 6 Comparison of the performance of corn capacity models.

Method	RMSE (kg/μ)	MAE (kg/μ)	MAPE (%)
RF	124.16	85.67	16.26
GBRT	150.66	116.31	18.93
TPML	48.94	34.01	7.65

TABLE 7 Comparison of the performance of wheat capacity models.

Method	RMSE (kg/μ)	MAE (kg/μ)	MAPE (%)
RF	135.58	91.68	16.46
GBRT	90.03	71.63	24.77
TPML	23.92	17.67	6.31

organic matter is one of the main sources of soil fertility. Soils with a high organic matter content are generally more fertile, which favors crop growth and production capacity; to maintain soil structure stability, soil organic matter content should be at least 20 g/kg (Or et al., 2021). The development of reserve arable land resources must also take into account the cost of development, which is why plots of at least ten square kilometers are selected for development. The types of arable reserve land resource use mainly include grassland, saline land, sandy land and bare land. Grassland needs to be transformed and restored before it can be converted to arable land, and saline-alkaline land is currently unused but can be converted to arable land after treatment and improvement. The arable land reserves to be developed are an important support for the sustainable development of Chinese agriculture. Through scientific planning and rational development, their potential can be fully exploited, contributing to the country's food security and economic development.

3.1.2 Cultivated land of low quality and ineffective utilization

Specific evaluation indicators include ecological conditions, topographic slope, annual precipitation, soil texture, soil heavy metal pollution status, soil organic matter content, soil pH value, and land use type, totaling eight indicators (Fan et al., 2012; Wang et al., 2021). Each index layer is classified according to the standard, and the range of cultivated land with low quality and low utilization efficiency is obtained through superposition (Table 3). The experimental operation was carried out in ArcMap10.2.

Large slopes not only lead to lower soil fertility and nutrient loss, but also increase the difficulty and cost of farming, reducing the efficiency of arable land use. The climate of the study area is more arid, and lower rainfall will limit agricultural production. The need for more irrigation will also increase the cost of agricultural production, affecting farmers' income. Reduced biodiversity is generally accompanied by a decline in soil fertility, making it more vulnerable to pests and diseases, upsetting the ecological balance and lowering the quality of agricultural produce. An imbalance between soil acidity and alkalinity forces farmers to invest more in improving soil conditions, increasing production costs and reducing economic efficiency. Due to the large inter-particle voids and weak capillary action in sandy soils, nutrient content is low and prone to leaching. As a result, sand have a low water and fertilizer retention capacity, and although good aeration and permeability promote respiration and crop root growth, water evaporates easily, leading to drought and the need for more water and timely irrigation.

3.2 Spatial distribution and characteristics of marginal land

According to the land use type data, 343,299 square kilometers of cultivated land are available in the study area (Figure 5). It is mainly concentrated in the northwestern part of Xinjiang Uygur Autonomous Region, the southeastern part of Gansu province, Ningxia province, Shaanxi province and Shanxi province. Most of the existing cultivated land was found to be of low quality and inefficiently utilized.

The marginal land is mainly concentrated in the northwestern region of Xinjiang Uygur Autonomous Region, Gansu province,

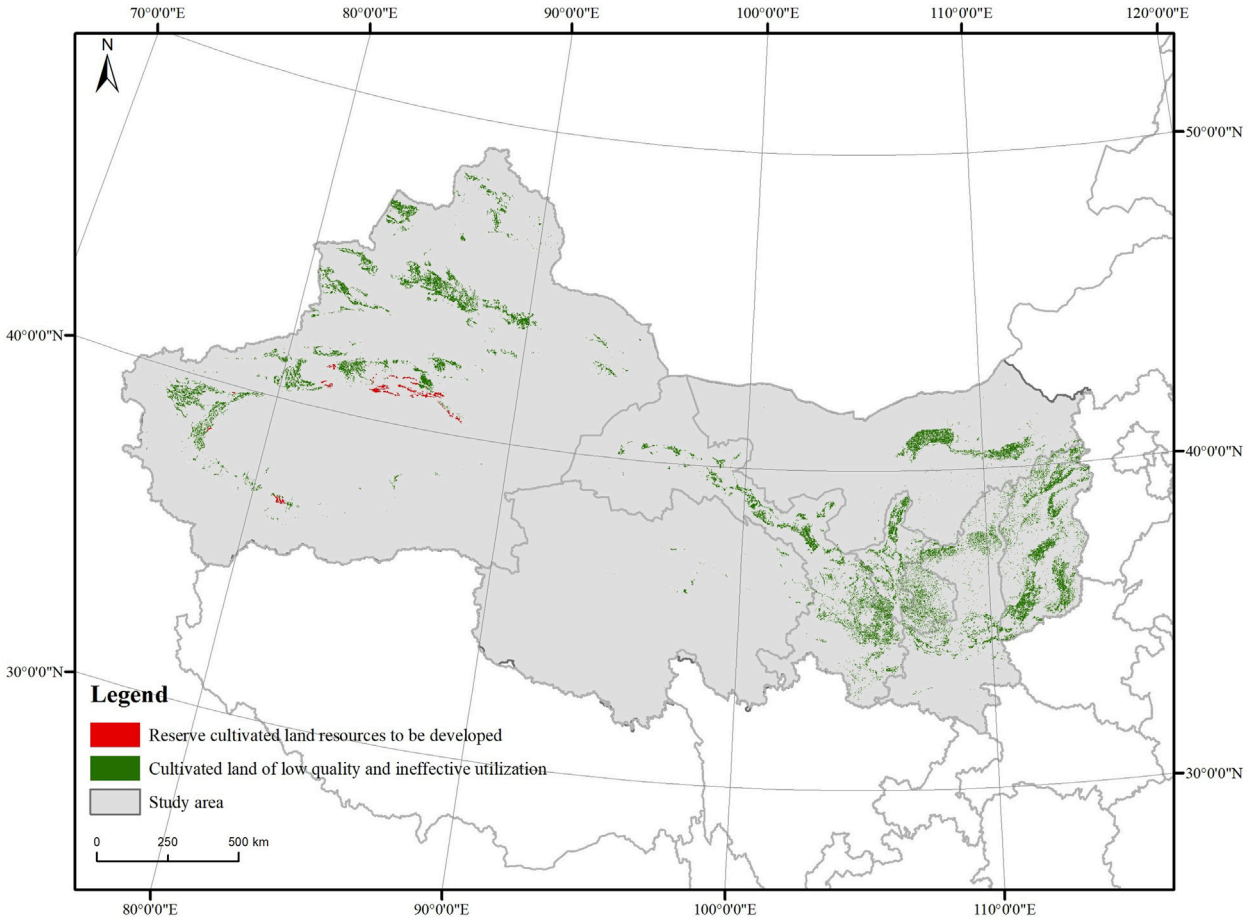

FIGURE 6
The distribution of marginal land.

Ningxia province and Shanxi province, with a total area of 145,062 square kilometers. All of the reserve cultivated land resources to be developed are in Xinjiang Uygur Autonomous Region, totaling 3,626 square kilometers (Figure 6). Cultivated land of low quality and inefficient utilization totaled 141,436 square kilometers. Marginal land is unevenly distributed, mainly in Xinjiang Uygur Autonomous Region, Gansu province and Shanxi province, with Qinghai province having the smallest percentage of marginal land (Table 4).

Xinjiang Uygur Autonomous Region is deep inland in China and has low precipitation, but it has sufficient light and a large temperature difference between day and night, which is favorable for the accumulation of sugar in fruit and sugar crops. The northern region of Xinjiang Uygur Autonomous Region has better conditions for agricultural production than eastern and western parts of Xinjiang Uygur Autonomous Region in terms of light, temperature and precipitation. The high topography and varied slopes of Qinghai province tend to result in soil and nutrient loss. Crop options are limited, and those adapted to this environment tend to be cold- and drought-tolerant crops. These crops are generally less productive and of poorer quality, which explains the small amount of marginal land in Qinghai province.

4 Grain production capacity prediction of marginal land

Capacity forecasting plays an essential role in optimizing the distribution of agricultural production, allocating resources and monitoring crop growth in real time. The spatial distribution of grain capacity is influenced by spatio-temporal conditions such as soil properties, climatic conditions, terrain topography and spatial heterogeneity (Cheng et al., 2022). The resolution of the predicted food production capacity is 1 km grid. This section is implemented using RStudio.

4.1 Influence factors for selecting capacity

The production capacity of corn and wheat is affected by a variety of factors such as soil pH, soil texture, climatic conditions and topographic conditions, so the influence factors of grain production capacity are selected from soil, meteorology and topography, and the grain production capacity model is constructed separately (Table 5).

Correlation analysis of the factors was carried out on corn and wheat data respectively, and different degrees of correlation

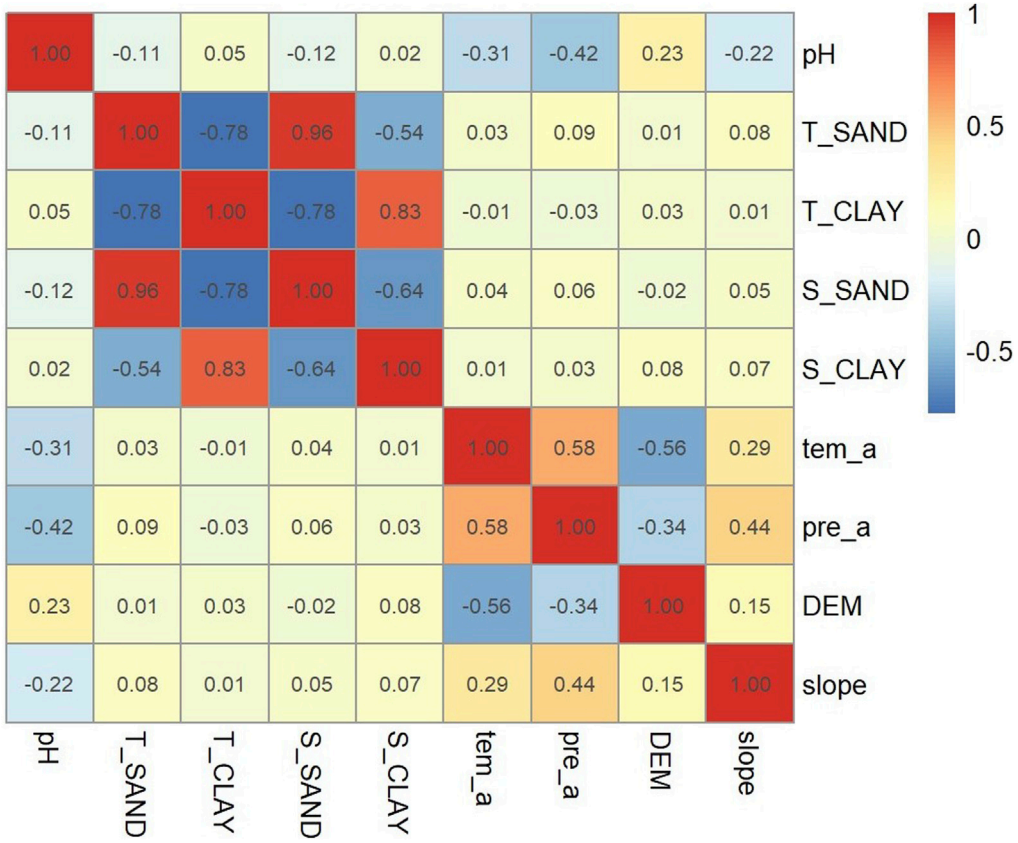

FIGURE 7
Correlation between independent variables of corn.

were found between these factors (Figures 7, 8). Soil top layer sand content and soil bottom layer sand content are highly positively correlated, soil top layer clay content and soil bottom layer clay content are highly positively correlated, the degree of positive correlation between average annual temperature and average annual precipitation is high at around 0.5, and the degree of correlation of all other influencing factors is very low.

4.2 Build grain production capacity model

Two hundred sample points were randomly selected from all the sampling points, for which RF, GBRT and TPML methods were used to model food production capacity. The ten-fold cross-validation method was used to evaluate the performance of the model. All the samples were divided into ten groups, and one group was sequentially selected as the test data and the remaining nine groups were used as the training data. Comparing the performance of the three models it is found that the error of the model constructed by TPML is much smaller than that of the model constructed by the other two methods (Table 6, 7). Therefore, a dataset with a sample size of 200 is chosen to construct a grain production capacity model using TPML to estimate the production capacity of corn and wheat on marginal land.

4.3 Predict production capacity on marginal land

In general, the production capacity of corn is higher than that of wheat, which may be related to the strong resistance and adaptability of corn, which can grow under various climatic and soil conditions. The production capacity of both shows a spatial trend higher in the west and lower in the east, with higher production capacity for corn and wheat in the western regions of Gansu province and Xinjiang Uygur Autonomous Region than in other regions (Figures 9, 10). The corn production capacity was higher in the eastern part of Gansu province and Ningxia province than in Shaanxi province and Shanxi province, while wheat production capacity showed the opposite trend, probably due to more abundant precipitation in Shaanxi province and Shanxi province.

The average grain production capacity of marginal land in each province in the study area is not much different, and the total production capacity is mainly affected by the size of marginal land area (Figures 11, 12). Except that the average grain production capacity of Xinjiang Uygur Autonomous Region is slightly higher, the average production capacity of corn in other regions is between 500–600 kg/μ, and the average production capacity of wheat is about 300 kg per mu. The total grain production capacity of Xinjiang Uygur Autonomous Region is much higher than that of other regions. It is mainly due to its vast cultivated land resources and abundant light conditions (Shi et al., 2014). It is

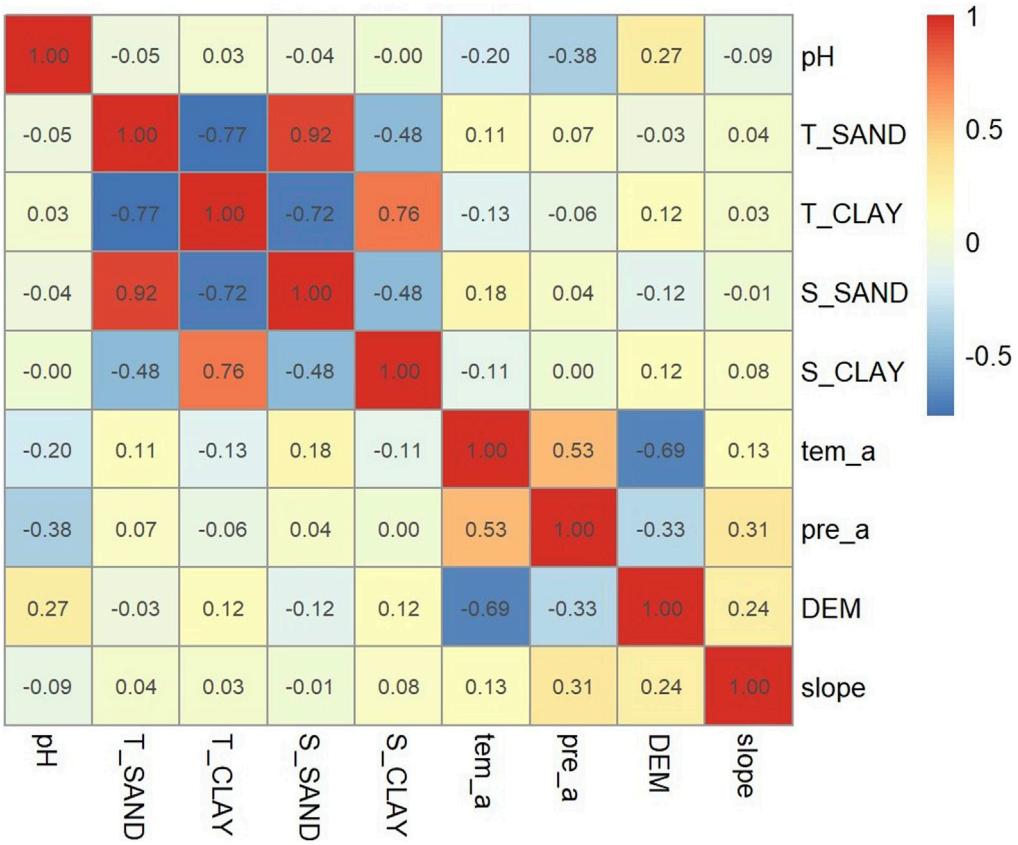

FIGURE 8
Correlation between independent variables of wheat.

mainly based on plains and basins and is suitable for the cultivation of various grain crops. The total grain production capacity of Qinghai province is the lowest, because the marginal land area is small, the terrain is undulating, and the precipitation is small. It is difficult to form a large-scale grain production base, and the production capacity is naturally low, which also increases the cost and risk of agricultural production.

In Xinjiang Uygur Autonomous Region and Gansu province, the climate and soil advantages of the region should be further utilized, the planting structure should be optimized, and the productivity and quality of crops should be improved. In Xinjiang Uygur Autonomous Region, Gansu province and the eastern region of Ningxia province, the support and management of corn planting should be strengthened to improve the production capacity and market competitiveness of corn. At the same time, in view of the low productivity of wheat, it is possible to explore the cultivation of other crops adapted to local conditions or take improvement measures to increase the productivity of wheat. In contrast, in Shaanxi province and Shanxi province, the planting of wheat should be strengthened.

The difference in grain production capacity between maize and wheat is influenced by a combination of factors, including length of fertility, photosynthetic efficiency, threat of pests and diseases, root development, level of mechanization, and ecological protection policies. Maize is more efficient in photosynthesis under high temperatures and intense light, has a well-developed root system and is tolerant of barrenness, and is suitable for growth in arid and semi-arid areas of the Northwest, while wheat has a more advantageous production capacity in mild and humid environments. The frequency and extent of pests and diseases also have a significant impact on crop productivity, and regions with less precipitation help reduce the threat of corn pests and diseases. In addition, higher levels of mechanization on contiguous land in the Northwest help boost corn production capacity, while wheat cultivation is relatively weak in hilly areas due to lower levels of mechanization. Ecological conservation policies that limit the scale of maize cultivation in some areas and encourage the cultivation of drought-tolerant crops such as wheat further exacerbate the difference in production capacity between the two. Therefore, in the process of marginal land development, it is necessary to take into account crop growth characteristics, environmental conditions and policy guidance in order to optimize crop layout, maximize food production capacity and achieve sustainable development of the ecological environment.

In the process of agricultural production, economic benefits should also be considered. The benefits of planting corn and wheat depend on many factors, including climatic conditions, soil quality, planting technology and food supply. Corn usually has higher productivity potential and is widely used in feed, industrial raw materials and food. Wheat is mainly used for flour processing and

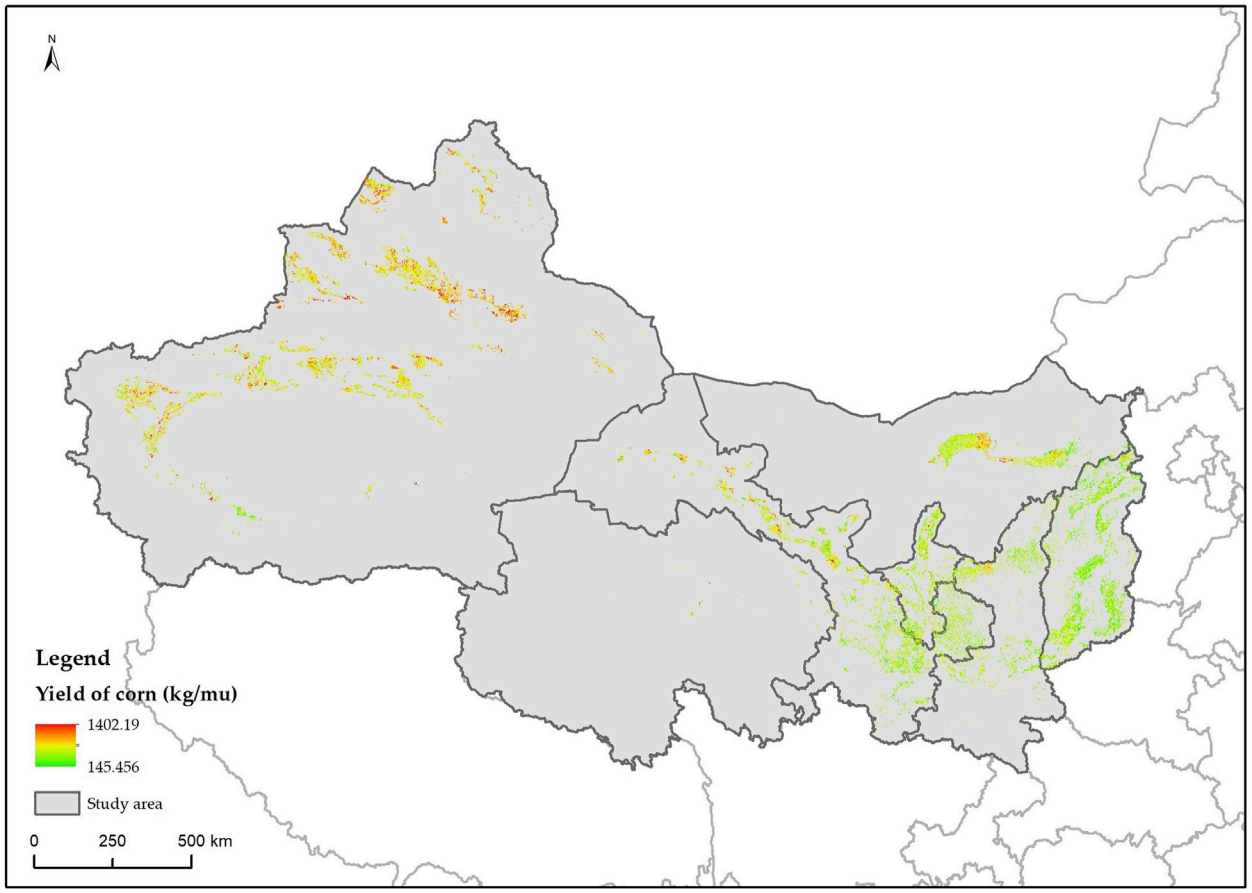

FIGURE 9
Production capacity of corn on marginal land.

food production. The price of grain crops is affected by many factors such as market supply and demand, international trade policy and so on, and the price fluctuates. At the same time, it is necessary to consider the planting cost, including the input of seeds, fertilizers, pesticides and labor. The planting cost will vary greatly in different regions with different climate, topography and soil conditions (Wang and Tian, 2017). In terms of the production cost of corn planting, the northwest region is often higher than the national average. The corn production in Gansu province is the highest in the country, and the corn production cost in Shanxi province is relatively high. In terms of the production cost of wheat planting, the production costs of Xinjiang Uygur Autonomous Region, Qinghai province, Shaanxi province and Shanxi province are lower than the national average, and the production costs of Gansu province and Inner Mongolia are the highest.

Corn planting can also improve soil structure, reduce soil erosion and soil erosion. Corn straw can be used as organic fertilizer to increase soil organic matter content (Zhang et al., 2010). Wheat has a relatively small demand for water resources, which helps to maintain soil health and fertility and maintain biodiversity. Interplanting wheat with corn can improve the utilization rate of land resources, strengthen the light transmittance, reduce the use of chemical fertilizers and pesticides, and help protect the ecological environment.

Therefore, in the actual agricultural production, the monitoring and analysis of environmental factors such as climate and soil should be strengthened. According to the local natural conditions, economic conditions and social conditions, the suitable crop varieties and planting methods are selected. Scientific planting plans and management measures should be adopted to achieve optimal production capacity and economic benefits and promote the sustainable development of agricultural production.

5 Discussion

We determine exploitable cultivated land reserves and low-quality, low-utilization-efficiency cultivated lands in the coverage areas of the western route of the South-to-North Water Diversion Project through the identification of marginal lands. TPML is used to predict the production capacity of different food crops on marginal lands, providing a scientific basis for agricultural production.

However, there are limitations due to the insufficient selection of factors in this study. While we focused on marginal lands with suitable cultivation conditions, factors like land accessibility, economic viability, and farming conditions were not considered, as actual yield data for marginal lands was lacking. This means the results are based solely on natural conditions. In practice, factors

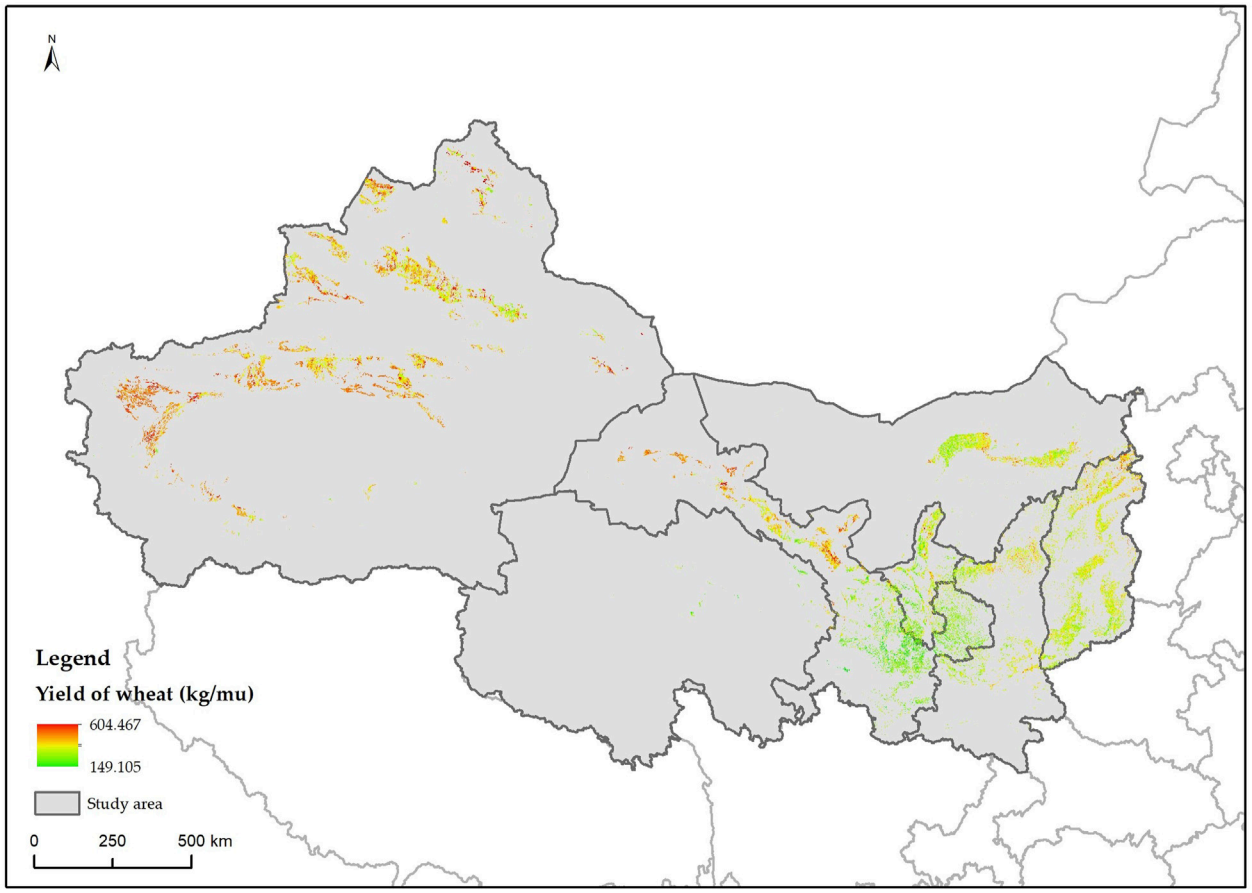

FIGURE 10
Production capacity of wheat on marginal land.

such as economic returns, population distribution, and ecological conditions should be considered.

To improve the accuracy of predictions, more evaluation factors such as irrigation conditions, proximity to water sources, and roads could be incorporated into the model. Additionally, integrating multiple methods or adopting more advanced machine learning approaches could enhance the model's performance. It is also essential to develop a more refined evaluation system for different types of marginal land, considering their specific characteristics and constraints.

Research on marginal land development and utilization in the coverage area of the western route of the South-to-North Water Diversion Project is still limited. Future work should focus on analyzing limiting factors more deeply, which will guide the formulation of targeted strategies for land development. A more detailed approach involving regional zoning and exploration of crop systems suited to local conditions will improve productivity and sustainability. Ecological protection must be prioritized, and risks related to ecological, economic, and social factors must be properly managed.

In this context, it is also important to understand the relationship between land degradation and the development of unused land in northwest China. This should not be seen simply as expanding cultivated land, but as a strategic response to land degradation, population outmigration, and resource constraints. Development in northwest China must be based on comprehensive assessments of resources, ecological sensitivity, and agricultural feasibility, ensuring that it aligns with sustainability and food security goals.

Future research could incorporate dynamic simulations based on evolving water resource development plans to assess the long-term benefits of the South-to-North Water Diversion Project on regional grain productivity. This would provide valuable insights into the relationship between water resources and agricultural production. Additionally, the impact of policies, particularly those related to the South-to-North Water Diversion Project, should be considered in future studies to understand how changes in infrastructure, water allocation, and regulations may influence the development of marginal land.

6 Conclusion

The study identified 145,062 square kilometers of marginal land in the Western Route of the South-to-North Water Diversion Project area, including 3,626 square kilometers of reserve arable land and 141,436 square kilometers of low-quality and inefficiently-utilized farmland. The marginal land

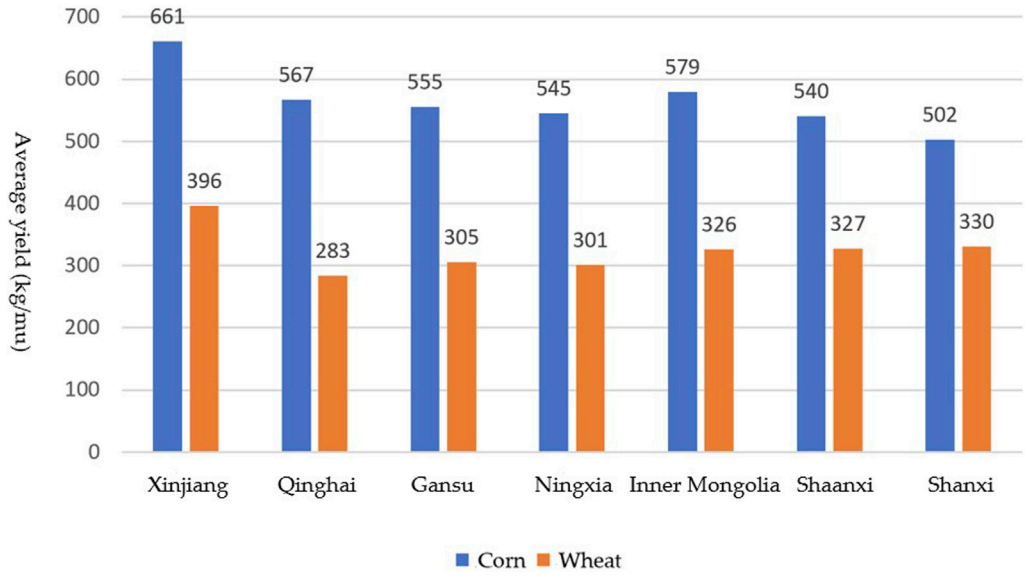

FIGURE 11
Average grain production capacity of marginal land.

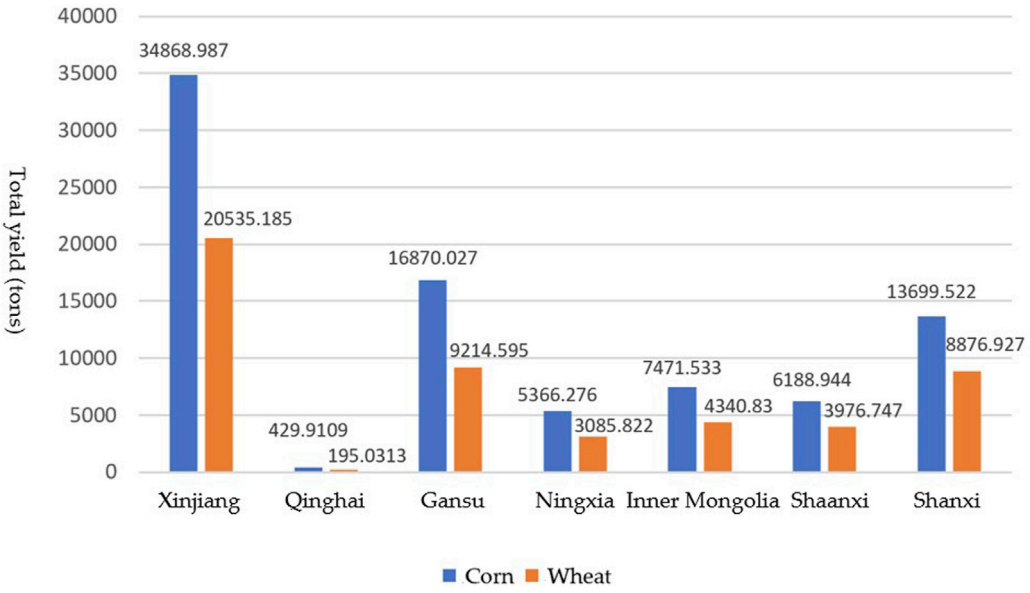

FIGURE 12
Total grain production capacity of marginal land.

is primarily concentrated in the northwestern part of Xinjiang Uygur Autonomous Region and Gansu Province, with the smallest areas found in Qinghai Province. The productivity of maize on marginal land is generally higher than that of wheat, with a spatial distribution pattern of higher yields in the west and lower yields in the east. The average grain production capacity on marginal land in Xinjiang is higher than in other regions, with maize yields typically ranging from 500 to 600 kg per mu, while wheat yields average around 300 kg per mu.

The results of the study provide some insights into future land management and agricultural production practices. In Xinjiang Uygur Autonomous Region, Gansu Province and Ningxia Hui Autonomous Region, where grain production capacity is high, it is recommended to further optimize the maize planting structure and combine irrigation and soil improvement measures to enhance yields. In Shaanxi and Shanxi provinces, suitable high-yielding wheat varieties are promoted to take full advantage of local natural conditions.

Future research could focus on the long-term impacts of marginal land development on ecosystems and water resources, explore the potential of multi-crop rotation or replanting patterns in enhancing marginal land productivity, and refine methods for identifying and assessing marginal land productivity through remote sensing monitoring and field surveys in order to improve the adaptability of the models and the accuracy of their predictions. Through locally adapted agricultural management and policy guidance, we will promote the rational development and sustainable utilization of marginal land and provide strong support for guaranteeing national food security and promoting regional economic development.

This study employs the TPML model to evaluate the grain production capacity of marginal land. The model demonstrates its effectiveness in enhancing the accuracy of spatial distribution predictions. However, the assumptions and limitations of the model may introduce uncertainties that affect the reliability and generalizability of the results. The TPML model relies on the assumptions of spatial autocorrelation and attribute similarity, suggesting that points in close proximity or with similar environmental conditions exhibit similar attributes. In regions with high spatial heterogeneity or complex environmental conditions, these assumptions may not fully hold. Furthermore, the model is highly sensitive to the selection of auxiliary variables, meaning that inaccuracies or biases in key variable data can significantly impact prediction outcomes. As a result, the applicability of this study's findings to other regions or under different environmental conditions remains to be further validated. In areas characterized by diverse land use types and complex terrain, adaptive adjustments to model parameters and structures may be necessary to improve its generalization capability and predictive accuracy.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The raw data supporting the conclusions of this article will be made available by the authors on request. Requests to access these datasets should be directed to JZ, zhoujun200208@163.com.

Author contributions

HZ: Investigation, Writing – review and editing. JZ: Writing – original draft, Writing – review and editing. KL: Investigation, Writing – review and editing. MN: Resources, Writing – review and editing. CW: Data curation, Writing – review and editing. GZ: Project administration, Writing – review and editing. JK: Funding acquisition, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded the Northwest Engineering Corporation Limited (grant number 2024110043003239). The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

Authors HZ, KL, MN, GZ, and JK were employed by Northwest Engineering Corporation Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Archer, K. J., and Kirnes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* 52, 2249–2260. doi:10.1016/j.csda.2007.08.015

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Cao, X., Sun, B., Chen, H., Zhou, J., Song, X., Liu, X., et al. (2021). Approaches and research progresses of marginal land productivity expansion and ecological benefit improvement in China. *Bull. Chin. Acad. Sci.* 36, 336–348. doi:10.16418/j.issn.1000-3045.20201228002

Cheng, M., Penuelas, J., McCabe, M. F., Atzberger, C., Jiao, X., Wu, W., et al. (2022). Combining multi-indicators with machine-learning algorithms for maize yield early prediction at the county-level in China. *Agric. For. Meteorology* 323, 109057. doi:10.1016/j.agrformet.2022.109057

Csikós, N., and Tóth, G. (2023). Concepts of agricultural marginal lands and their utilisation: a review. *Agric. Syst.* 204, 103560. doi:10.1016/j.agsy.2022.103560

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813. doi:10.1111/j.1365-2656.2008.01390.x

Fan, M. S., Shen, J. B., Yuan, L. X., Jiang, R. F., Chen, X. P., Davies, W. J., et al. (2012). Improving crop productivity and resource use efficiency to ensure food security and environmental quality in China. *J. Exp. Bot.* 63, 13–24. doi:10.1093/jxb/err248

Fei, S., Hassan, M. A., Xiao, Y., Su, X., Chen, Z., Cheng, Q., et al. (2023). UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precis. Agric.* 24, 187–212. doi:10.1007/s11119-022-09938-8

Fu, Y., Huang, J., Shen, Y., Liu, S., Huang, Y., Dong, J., et al. (2021). A satellite-based method for national winter wheat yield estimating in China. *Remote Sens.* 13, 4680. doi:10.3390/rs13224680

Gao, B., Stein, A., and Wang, J. (2022). A two-point machine learning method for the spatial prediction of soil pollution. *Int. J. Appl. Earth Observation Geoinformation* 108, 102742. doi:10.1016/j.jag.2022.102742

Guo, Y., Chen, S., Li, X., Cunha, M., Jayavelu, S., Cammarano, D., et al. (2022). Machine learning-based approaches for predicting SPAD values of maize using multi-spectral images. *Remote Sens.* 14, 1337. doi:10.3390/rs14061337

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., et al. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecol. Indic.* 120, 106935. doi:10.1016/j.ecolind.2020.106935

Guo, Y., Xiao, Y., Hao, F., Zhang, X., Chen, J., de Beurs, K., et al. (2023). Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. *Int. J. Appl. Earth Observation Geoinformation* 124, 103528. doi:10.1016/j.jag.2023.103528

Han, J. C., Zhang, Z., Cao, J., Luo, Y. C., Zhang, L. L., Li, Z. Y., et al. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sens.* 12, 236. doi:10.3390/rs12020236

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotic, A., et al. (2017). SoilGrids250m: global gridded soil information based on machine learning. *PLoS One* 12, 40. doi:10.1371/journal.pone.0169748

Huang, J. X., Tian, L. Y., Liang, S. L., Ma, H. Y., Becker-Reshef, I., Huang, Y. B., et al. (2015). Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. *Agric. For. Meteorology* 204, 106–121. doi:10.1016/j.agrformet.2015.02.001

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLoS One* 11, e0156571. doi:10.1371/journal.pone.0156571

Jiang, W., Jacobson, M. G., and Langholtz, M. H. (2019). A sustainability framework for assessing studies about marginal lands for planting perennial energy crops. *Biofuels Bioprod. Biorefining* 13, 228–240. doi:10.1002/bbb.1948

Lei, J. (2020). Cross-validation with confidence. *J. Am. Stat. Assoc.* 115, 1978–1997. doi:10.1080/01621459.2019.1672556

Lei, S., Hao, J., and Wang, L. (2011). Evaluation of exploitation suitability of unutilized arable lands in ecologically fragile areas - a case study of Datong City, Shanxi Province. *Zhongguo Shengtai Nongye Xuebao/Chin. J. Eco-Agriculture* 19, 1417–1423. doi:10.3724/SP.J.1011.2011.01417

Or, D., Keller, T., and Schlesinger, W. H. (2021). Natural and managed soil structure: on the fragile scaffolding for soil functioning. *Soil Tillage Res.* 208, 104912. doi:10.1016/j.still.2020.104912

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., and Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access* 9, 63406–63439. doi:10.1109/access.2021.3075159

Ren, J. Q., Chen, Z. X., Zhou, Q. B., and Tang, H. J. (2008). Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *Int. J. Appl. Earth Observation Geoinformation* 10, 403–413. doi:10.1016/j.jag.2007.11.003

Shi, P. J., Sun, S., Wang, M., Li, N., Wang, J. A., Jin, Y. Y., et al. (2014). Climate change regionalization in China (1961-2010). *Sci. China-Earth Sci.* 57, 2676–2689. doi:10.1007/s11430-014-4889-1

Shortall, O. K. (2013). "Marginal land" for energy crops: exploring definitions and embedded assumptions. *Energy Policy* 62, 19–27. doi:10.1016/j.enpol.2013.07.048

Shrestha, R., Di, L., Yu, E. G., Kang, L., Li, L., Rahman, M. S., et al. (2016). "Regression based corn yield assessment using MODIS based daily NDVI in Iowa state," in 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 1–5.

Song, R. Z., Cheng, T., Yao, X., Tian, Y. C., Zhu, Y., and Cao, W. X. (2016). "Evaluation of landsat 8 time series image stacks for predicitng yield and yield components of winter wheat," in 36th IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (Beijing, China: IEEE), 6300–6303.

Sun, S., Wu, M., Zhuang, L., He, Y., and Li, X. (2022). Forecasting winter wheat yield at county level using CNN and BP neural networks. *Trans. Chin. Soc. Agric. Eng.* 38, 151–160. doi:10.11975/j.issn.1002-6819.2022.11.017

Wang, L., Zheng, G., Guo, Y., He, J., and Cheng, Y. (2022a). *Prediction of winter wheat yield based on fusing multi-source spatio-temporal data*. Transactions of the Chinese Society for Agricultural Machinery 53 (1), 198–204.

Wang, S.-G., and Tian, X. (2017). Causes of the rising grain production cost in China: an empirical analysis of rice, wheat and corn. *Res. Agric. Mod.* 38, 571–580. doi:10.13872/j.1000-0275.2017.0065

Wang, X., Li, C., Liu, Y., Ji, Z., Li, L., and Yu, L. (2021). *Zoning and improving path of cultivated land use efficiency based on evaluation of cultivated land suitability*. Transactions of the Chinese Society for Agricultural Machinery 52 (5), 212–218.

Wang, Y., Yang, K., Gao, B., Feng, A., Tian, J., Jiang, C., et al. (2022b). Prediction of the spatial distribution of soil organic matter based on two-point machine learning method. *Trans. Chin. Soc. Agric. Eng.* 38, 65–73. doi:10.11975/j.issn.1002-6819.2022.12.008

Wang, L., Xiong, W., Wen, X., and Feng, L. (2014). Effect of climatic factors such as temperature, precipitation on maize production in China. *Trans. Chin. Soc. Agric. Eng.* 30, 138–146. doi:10.3969/j.issn.1002-6819.2014.21.017

Yao, M. L., Shao, D. G., Lv, C. H., An, R. H., Gu, W. Q., and Zhou, C. (2021). Evaluation of arable land suitability based on the suitability function-A case study of the Qinghai-Tibet Plateau. *Sci. Total Environ.* 787, 147414. doi:10.1016/j.scitotenv.2021.147414

Zhang, J., Wen, X., Liao, Y., and Liu, Y. (2010). Effects of different amount of maize straw returning on soil fertility and yield of winter wheat. *Plant Nutr. Fertitizer Sci.* 16, 612–619. doi:10.4028/www.scientific.net/AMM.37-38.1549

Zhang, Q. Q., Yang, J., Zhang, Y. C., and Ji, M. C. (2014). "The gray GM(1.1) prediction in wheat production of Shandong province based on the MATLAB," in 2nd International Conference on Soft Computing in Information Communication Technology (SCICT) (Taipei, Taiwan: Atlantis Press), 14–17. doi:10.2991/scict-14.2014.4