scientific reports

OPEN

CLUMondo-BNU for simulating land system changes based on many-to-many demand–supply relationships with adaptive conversion orders

Peichao Gao[1,2], Yifan Gao[2], Xiaodan Zhang[2], Sijing Ye[1,2] & Changqing Song[1,2 ✉]

Land resources are fundamentally important to human society, and their transition from one macroscopic state to another is a vital driving force of environment and climate change locally and globally. Thus, many efforts have been devoted to the simulations of land changes. Among all spatially explicit simulation models, CLUMondo is the only one that simulates land changes by incorporating the multifunctionality of a land system and allows the establishment of many-to-many demand–supply relationships. In this study, we first investigated the source code of CLUMondo, providing a complete, detailed mechanism of this model. We found that the featured function of CLUMondo—balancing demands and supplies in a many-to-many mode—relies on a parameter called conversion order. The setting of this parameter is a manual process and requires expert knowledge, which is not feasible for users without an understanding of the whole, detailed mechanism. Therefore, the second contribution of this study is the development of an automatic method for adaptively determining conversion orders. Comparative experiments demonstrated the validity and effectiveness of the proposed automated method. We revised the source code of CLUMondo to incorporate the proposed automated method, resulting in CLUMondo-BNU v1.0. This study facilitates the application of CLUMondo and helps to exploit its full potential.

The sustainable management and conservation of land resources have been central to human society[1,2], as the resources are limited but provide the ultimate basis for "more than 95% of human food supplies, the greater part of clothing, and all needs for wood, both for fuel and construction"[3]. A critical focus of the management and conservation is on land-use and land-cover change, or land change for short e.g.,[4–6]. The land change represents the transition of land resources from one macroscopic state to another. More importantly, this transition is a crucial driving force of environmental and climate change locally and globally, which in turn affects land resources[7,8]. As a result, many efforts have been devoted to estimating future land changes in different scenarios and employing these estimates to inform management and conservation policies e.g.[9–11].

Given the importance of future land change estimates, tools have been actively developed for their generation. These tools are called land change simulation models, classified into spatially aggregated and spatially explicit. Spatially aggregated models estimate future land changes in terms of quantity (i.e., composition). Such models usually serve as an essential component of integrated models for simulating coupled human and natural systems[12]. A typical example is the Global Change Assessment Model[13,14], a maker model for the famous Shared Socioeconomic Pathways[15,16]. Its land use component produces future areas of more than 60 land types (e.g., rainfed cornland with high fertilizer or irrigated rice land with low fertilizer) at the spatial resolution of 235 water basins worldwide. Spatially explicit models generate the estimates of future land changes in configuration (if the composition information is an output of another model), or sometimes both configuration and composition. Examples of such models include cellular-automata-based models—e.g., Future Land Use Simulation (FLUS) model[17] and Land Use Scenario Dynamics-urban (LUSD-urban) model[18]—and suitability-based models, e.g., Conversion of Land Use and its Effects at Small regional extent (CLUE-S) model[19] and its latest version

[1]State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China. [2]Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China. ✉email: songcq@bnu.edu.cn

CLUMondo[20]. The output format of such models is usually a raster dataset, whose spatial resolution can be as fine as that of input data. Therefore, spatially explicit models are more specialized in land change simulation and are widely used.

Among all spatially explicit models, CLUMondo is the only one that simulates land change with many-to-many demand–supply relationships. Specifically, spatially explicit models balance a pre-defined, aggregated demand and the sum of corresponding, spatially explicit supply, although with different simulation strategies and techniques. Usually, the aggregated demand is specified as the areas of different land types e.g.[21–23]. In this case, the model adjusts the original types of land grid cells (hereafter cells), according to some mechanism, to supply the same areas of land types. The resultant relationship between the pre-defined demand and the corresponding supply is one-to-one; in other words, the demand for the area of a specific land type can only be met by supplying that type (i.e., by allocating that type of cells). Sometimes, the demand also involves the amount of goods or services, e.g., population, food production, or ecological/economic benefits. In practical simulations, however, such non-area demands are transformed into the area demands for different land types to achieve one-to-one demand–supply balances e.g.[24,25]. The only exception is CLUMondo, where the balance can be achieved in terms of not only land type areas but also the amount of goods or services e.g.[26,27]. The demand for goods or services can be employed by this model without being transformed into areas, and each land type can be designated a capability to supply the goods or services in need. Because the demand–supply relationships can be modeled in a many-to-many mode, CLUMondo accepts diverse demand/supply settings and allows a more realistic simulation of land changes. It has found increasing applications to simulate land change at local, regional, and global scales, as shown in Fig. 1.

In this case, the effectiveness of CLUMondo is crucial and should be improved if possible. Accordingly, this study was focused on its central mechanism, which is the transition potential of each basic unit in simulation (i.e., a cell). This transition potential is a parameter determining the future land system type of a cell. Once the transition potentials of all cells are calculated, a simulation result of CLUMondo can be immediately determined. In this study, we investigated the detailed mechanism of the transition potential. The investigation found that a key parameter in the mechanism is called conversion order, whose setting requires both expert knowledge and fine-tuning. More importantly, the values of conversion orders should vary with studies areas and land system characteristics, making the determination of conversion orders rather sophisticated. These facts are probably the reason why this key parameter should be manually set by users.

To facilitate the application of CLUMondo, we developed an automatic method for adaptively determining conversion orders. Evaluation results demonstrated that with this method, users could easily achieve a good simulation performance using CLUMondo. This method benefits not only non-expert but also expert users because its results can serve as a good starting point for fine-tuning conversion orders. We modified the source code of CLUMondo to integrate the proposed method as an option for users (who can still set conversion orders manually). To distinguish the modified CLUMondo from the official version, we referred to this modified one as CLUMondo-BNU v1.0 (where the abbreviation "BNU" stands for the university of the authors of this paper) and also released it for public use.

CLUMondo: simulating land system changes with many-to-many demand–supply relationships

Before explaining the mechanism of CLUMondo, we introduce two common concepts in the literature on CLUMondo e.g.[28–30], namely land system and land system services. In the context of CLUMondo literature, the concept of land system is synonymous with but broader than that of land use/cover. A land system can be simply a type of land use/cover; it can also represent a mixed type of land use/cover. In the latter complex case, land systems are defined "in terms of their land cover composition as well as land use intensity"[20]. For example, the land systems established by Jin, Jiang, Ma and Li[27] include low/medium/high-covered natural grassland, low/
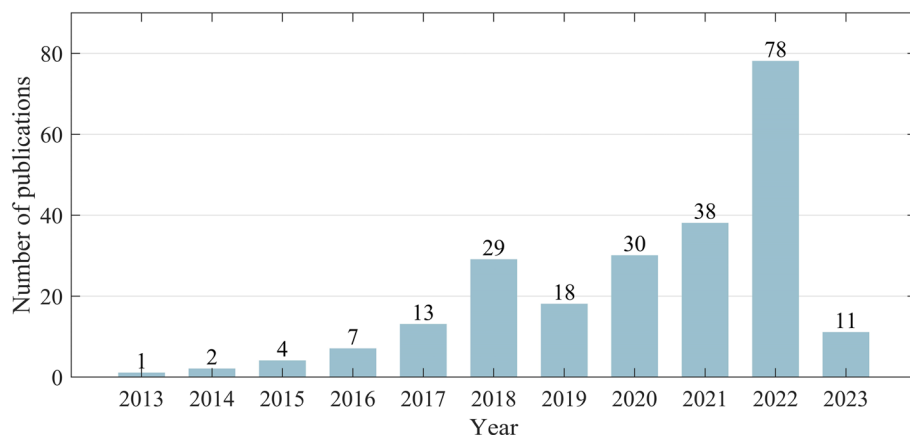

Figure 1. Increasing number of publications retrieved from Google Scholar with the keyword "CLUMondo" by 2 Feb 2023. When searching Google Scholar, we used double quotes to signify a phrase search and unchecked the "include citations" box.

medium/high-covered grassland with few livestock, low/medium/high-covered grassland with bovines, goats & sheep, extensive cropland, intensive cropland, sparse forest, and dense forest. As an extension to the concept of the land system, the concept of land system service was developed in parallel with that of ecosystem service; it refers to the area of specific land use/cover contained by a land system, or more generally, the goods or services that a land system provides for humans[31], e.g., various terrestrial ecosystem services.

CLUMondo simulates the changes of land systems in a predefined time step, which is usually one year. Each time step involves a large number of iterations, where the default maximum number is 20,000. In the i-th iteration of the t-th time step ($i, t \geq 1$), CLUMondo determines whether and to what the land system type of every cell is changed as follows:

$$T(c,t,i) = \begin{cases} T(c,t,0) & c \in \Phi \\ T(c,t,0) & \xi(T(c,t,0)) < \tau(T(c,t,0)) \\ \left\{ k | P_{c,k} = \max_j \left(P_{c,1}, P_{c,2}, \cdots, P_{c,j}, \cdots, P_{c,n} \right) \right\} & \begin{cases} \text{Con}(T(c,t,i),k) = 1 \\ c \notin \Phi \\ \xi(T(c,t,0)) \geq \tau(T(c,t,0)) \end{cases} \\ T(c,t,i-1) & else \end{cases}$$

(1)

where c denotes the c-th cell. k and j denote the k-th and j-th land system type, respectively ($1 \leq k, j \leq n$). $T(c,t,i)$ is land system type of c at the end of the current iteration (i.e., the i-th iteration of the t-th time step). $P_{c,j}$ is called the transition potential of c to the j-th land system type; in other words, $P_{c,j}$ is the probability of the c-th cell's land system type being converted into or maintained at the j-th land system type. This equation contains three "if" conditions:

- The first condition is a spatial restriction, where Φ is the restricted area where land system changes are not allowed.
- The second condition is a temporal restriction. $\xi(T(c,t,0))$ calculates how many time steps (usually years) c has been maintained at the initial land system type of this time step, namely $T(c,t,0)$. $\tau(T(c,t,0))$ is a non-negative integer representing the minimum time steps that the land system type $T(c,t,0)$ should be maintained. This condition requires that the initial land system type of this time step should be kept for a predefined number of time steps.
- The third condition is a conversion restriction. $\text{Con}(T(c,t,i),k)$ indicates whether the conversion from $T(c,t,i)$ to k is allowed according to users' settings, where one means allowed and zero means restricted.

From Eq. (1), it can be seen that $P_{c,j}$ is the key component. According to the allocation procedure outlined by van Asselen and Verburg[7], the standard determination of $P_{c,j}$ is a linear combination process involving three basic factors:

$$P_{c,j} = \begin{cases} P\_loc_{c,j} + P\_res_{T(c,t,0)} + P\_cmp_{i,j} & if\, T(c,t,0) = j \\ P\_loc_{c,j} + P\_cmp_{i,j} & if\, T(c,t,0) \neq j \end{cases}$$

(2)

where $P\_loc_{c,j}$, $P\_res_{T(c,t,0)}$, and $P\_cmp_{i,j}$ are referred to as local suitability, conversion resistance, and competitive advantage, respectively. The functions and determinations of these three basic factors are as follows:

- The local suitability $P\_loc_{c,j}$ refers to the suitability that the j-th land system type occurs at the c-th cell. According to Eq. (2), only $P\_loc_{c,j}$ is a spatial parameter because it varies with location (i.e., the c-th cell). It is by default calculated using a logistic regression based on a series of driving factors (i.e., biophysical and/or socioeconomic conditions): where $X_{1,c}, X_{2,c}, \cdots, X_{m,c}$ are the values of different driving factors at the location of c-th cell, and $\beta_{1,j}, \beta_{2,j}, \cdots, \beta_{m,j}$ are coefficients. $\beta_{0,j}$ is a constant. The value range of $P\_loc_{c,j}$ is $(0, 1)$, where a greater value indicates higher suitability.

$$\ln \left(\frac{P\_loc_{c,j}}{1 - P\_loc_{c,j}} \right) = \beta_{0,j} + \beta_{1,j} X_{1,c} + \beta_{2,j} X_{2,c} + \cdots + \beta_{m,j} X_{m,c}$$

(3)

- The conversion resistance $P\_res_{T(c,t,0)}$ reflects the difficulty (e.g., cost) of converting the land system type $T(c,t,0)$ to another, or equivalently, the ease of remaining unchanged for the land system type $T(c,t,0)$. Note that $P\_res_{T(c,t,0)}$ changes along with t (i.e., time step) but not i (i.e., iteration). The allowed value range for $P\_res_{T(c,t,0)}$ is $[0, 1]$; the greater the value, the higher the difficulty, the higher probability of keeping $T(c,t,0)$, and the lower probability of converting $T(c,t,0)$ to j. In practice, the value of $P\_res_{T(c,t,0)}$ is usually determined according to expert knowledge or historical land system changes. For the latter case, the determination can be mathematically expressed as follows: where c' denotes the c'-th cell. $h1$ and $h2$ denote two historical years, and $h1 < h2$.

$$P\_res_{T(c,t,0)} = \sum_{c'} y_{c',T(c,t,0)}^{h1,h2} / \sum_{c'} y_{c',T(c,t,0)}^{h1}$$

(4)

$$y_{c',T(c,t,0)}^{h1,h2} = \begin{cases} 1 & if\, T(c',h2,0) = T(c',h1,0) = T(c,t,0) \\ 0 & else \end{cases}$$

(5)

$$y^{h1}_{c',\mathrm{T}(c,t,0)} = \begin{cases} 1 & if\ \mathrm{T}(c',h1,0) = \mathrm{T}(c,t,0) \\ 0 & else \end{cases}$$

- The competitive advantage $P\_cmp_{i,j}$ characterizes the capability of j, relative to other land system types, for filling the gap between the aggregated demand for land system services and the corresponding supply in the i-th iteration. According to van Vliet and Verburg[20], $P\_cmp_{i,j}$ has the following properties:where $P\_cmp_{i,j,d}$ is the $P\_cmp_{i,j}$ specified for the d-th kind of land system service, $S_{j,d}$ is the capability of the j-th land system type to supply the d-th kind of land system service, and $d_{i,d}$ is the gap in the supply of the d-th kind of land system service.

$$\begin{cases} P\_cmp_{i,j} = \sum_d P\_cmp_{i,j,d} \\ P\_cmp_{i,j,d} \propto S_{j,d}, d_{i,d} \end{cases} \tag{7}$$

Investigated mechanism and novel method

Detailed mechanism of the competitive advantage. In this study, we investigated the detailed mechanism of the competitive advantage ($P\_cmp_{i,j}$) by exploring and testing the source code for CLUMondo (https://github.com/vueg/clumondo). The detailed mechanism is mathematically expressed in this study as Eqs. (8)–(10).

$$P\_cmp_{\mathrm{T}(c,t,0),j,(t,i)} = \sum_d \mathrm{sign}\left(L_{j,d} - L_{\mathrm{T}(c,t,0),d}\right) \cdot \omega_d \cdot diff_{d,(t,i)} \tag{8}$$

where $L_{j,d}$ and $L_{\mathrm{T}(c,t,0),d}$ are the so-called "conversion orders" of the land system types j and $\mathrm{T}(c,t,0)$ when supplying the d-th land system service, respectively. The values of a conversion order can be $-1, 0, 1, 2, \cdots$. The greater conversion order a land system type is assigned against a land system service, the higher priority the land system type will be given in allocation for filling the gap between the demand and supply of the land system service. In particular, the value of -1 denotes that a land system type is of no use in filling the gap. $\mathrm{sign}(x - y)$ is a sign function (also called signum function); it returns 1 if $x > y$, 0 if $x = y$, and -1 if $x < y$. ω_d is a weight parameter indicating the importance of the d-th land system service. The greater value (with 1 as the default value) ω_d has, the more important the d-th land system service is.

The parameter $diff_{d,(t,i)}$ in Eq. (8) can be intuitively understood as the gap between the demand and supply of the d-th land system service in the i-th iteration of the t-th time step. However, its calculation in CLUMondo is more complex than this intuition, as shown in Eqs. (9)–(10).

$$diff_{d,(t,i)} = \begin{cases} 0 & i = 1 \\ diff_{d,(t,i-1)} - \left(\frac{Supply_{d,(t,i-1)} - Demand_{d,t}}{Demand_{d,t}}\right)/\left(Speed_i \times R_i\right) & i \geq 2 \end{cases} \tag{9}$$

$$Speed_i = \begin{cases} 0.05 & i = 1 \\ Speed_{i-1} + 0.0002 & i \geq 2 \end{cases} \tag{10}$$

where $Demand_{d,t}$ is the demand for the d-th land system service at the beginning of the t-th time step. $Supply_{d,(t,i-1)}$ is the supply of the d-th land system service by all land systems at the end of the $(i - 1)$-th iteration within the t-th time step. According to Eq. (9), the value of $diff_{d,(t,i)}$ increases if $Demand_{d,t} > Supply_{d,(t,i-1)}$, whereas it decreases if $Demand_{d,t} < Supply_{d,(t,i-1)}$. $Speed_i$ and R_i are two dynamic variables changing along with the iteration process to accelerate its convergence, using the following convergence conditions:

$$if\ i > 20,000\ or \begin{cases} \sum_d \frac{Supply_{d,(t,i-1)} - Demand_{d,t}}{Demand_{d,t}}/n_d < 0.5\% \\ \frac{Supply_{d,(t,i-1)} - Demand_{d,t}}{Demand_{d,t}} < 1\%, \forall d \end{cases} \tag{11}$$

where n_d is the total number of land system services. By investigating the source code of CLUMondo, we found that $Speed_i$ had been set with an initial value of 0.05 and an increment of 0.0002 per iteration. We also found that R_i is a random number ranging from 322 to 365. The incorporation of $Speed_i$ and R_i gradually reduces the amount of change in $diff_{d,(t,i)}$ along with the iteration, further making the number of cells to be changed smaller and smaller in each iteration. This decreasing number facilitates the convergence of the iteration when minor changes to land systems are needed.

Difficulty in the manual setting of conversion orders. Having understood the detailed mechanism of the competitive advantage, one may realize the important role of the conversion orders therein and the importance of their determination. This is probably why the determination should be performed manually and carefully in CLUMondo. van Asselen and Verburg[7] illustrated one such determination, with the result shown in Table 1. On the explanation of this table, van Asselen and Verburg[7] noted that it "indicates the relative order of the land systems contribution to fulfilling a specific demand type" and also "ensures logical trajectories of land change" (p. 3651). They recommended determining conversion orders "differently by region, depending on the land system characteristics in the specific regions and the likely trajectories of fulfilling increasing (or decreasing) demands" (p. 3651), implying that the determination is sophisticated and requires fine tuning.

	Crop production (tons)		Land-based livestock (number)		Landless livestock (number)		Built-up area (km²)	
	Capability	Order	Capability	Order	Capability	Order	Capability	Order
Cropland extensive, few livestock (ls.)	8977	4	4658	2	9620	2	0.11	1
Cropland extensive, land-based ls	11,047	4	10,250	3	31,630	2	0.12	1
Cropland extensive, landless ls	11,110	4	3317	2	81,968	3	0.11	1
Cropland med. intensive, few ls	11,695	5	3704	2	11,899	2	0.36	1
Cropland med. intensive, land-based ls	13,421	5	14,282	4	79,960	2	0.40	1
Cropland med. intensive, landless ls	16,363	5	4387	2	102,894	4	0.43	1
Cropland intensive, few ls	24,076	6	2076	2	7934	2	0.69	1
Cropland intensive, land-based ls	37,740	6	23,949	5	339,985	5	1.24	1
Cropland intensive, landless ls	31,785	6	4264	2	172,779	5	0.67	1
Mosaic cropland and grassland, land-based ls	13,563	4	13,843	4	132,327	4	0.52	1
Mosaic cropland and grassland, landless ls	16,080	4	4005	2	122,532	4	0.46	1
Mosaic cropland (ext.) and grassland, few ls	3871	2	4736	2	8152	2	0.09	1
Mosaic cropland (med.) and grassland, few ls	6504	3	4403	2	11,890	2	0.25	1
Mosaic cropland (int.) and grassland, few ls	10,984	4	3374	2	8815	2	0.47	1
Mosaic cropland and forest, landless ls	14,548	3	3815	2	112,431	4	0.33	1
Mosaic cropland (ext.) and open forest, few ls	6,104	2	3754	2	15,727	2	0.09	1
Mosaic cropland (med. int.) and forest, few ls	6752	3	3511	2	13,386	2	0.17	1
Mosaic cropland (intensive) and forest, few ls	9774	4	3127	2	12,680	2	0.32	1
Dense forest	1478	1	2368	−1	28,849	−1	0.07	1
Open forest, few ls	1459	1	2302	1	8976	1	0.09	1
Open forest, landless ls	4576	−1	3073	−1	89,483	3	0.13	1
Mosaic grassland and forest	3043	1	3441	1	38,943	1	0.14	1
Mosaic grassland and bare	381	1	2824	1	3294	1	0.08	1
Natural grassland	749	1	0	1	0	1	0	1
Grassland, few ls	1610	1	2720	1	18,250	1	0.13	1
Grassland, land-based ls	2059	−1	14,159	4	37,991	−1	0.23	1
Bare	18	−1	4	−1	4	−1	0	1
Bare, few ls	430	−1	2928	1	2948	1	0.04	1
Peri-urban and villages	22,056	−1	9110	−1	184,526	−1	8.97	2
Urban	17,796	−1	5010	−1	193,283	−1	37.60	3

Table 1. Capability and conversion orders determined by van Asselen and Verburg[7] for 30 different land systems in supplying four defined land system services: crop production (tons), land-based livestock (bovines, goats, and sheep; number), landless livestock (pigs and poultry; number), and built-up area (km²). *ls.* Livestock, *ext.* extensive, *med.* medium intensive, *int.* intensive.

We recognize the necessity for manually determining the conversion orders, but we note the difficulty in the determination by non-expert users, especially beginners. To overcome the difficulty, we will propose an automatic method for adaptively determining the conversion orders for different land systems. This automatic method will be incorporated into CLUMondo as an option for non-expert users, as well as for expert users to find a good starting point for fine-tuning.

A method for automatically determining conversion orders. This study presents a method for automatically determining the conversion orders of different land systems based on their capability for supplying a specific service. The method is powerful in that it is effective in improving the simulation accuracy of CLUMondo, efficient in operation, and widely applicable.

Before developing the method, we rethink the functionality of the conversion order as a parameter of the competitive advantage. As noted in Section "Detailed mechanism of the competitive advantage", the conversion order was initially not included as a parameter of the competitive advantage, which was designed to be proportional to $S_{j,d}$ (the capability of the j-th land system type to supply the d-th kind of land system service) in concept. However, as shown in Section "Difficulty in the manual setting of conversion orders", the conversion order was included in implementing CLUMondo, whereas $S_{j,d}$ is not used in practice. The conversion order is employed as a proxy of $S_{j,d}$, to avoid the competition in filling the demand–supply gap of a land system service between two land systems with similar capabilities for supplying that service. For example, as shown in Table 1,

the "extensive cropland system with few livestock" was assigned the same conversion order (i.e., 1) as the "intensive cropland system with few livestock" in order not to promote the conversion from the former type to the latter type when filling the demand–supply gap of the "built-up area" service, although the former type has a lower capability in supplying the "built-up area" service than the later type (i.e., 0.11 vs. 0.69). Essentially, this functionality of the conversion order is achieved by transforming $\{S_{j,d}\}_j$ from a series of ratio values to categorized, ordinal ones refer to[32] for the nominal, ordinal, interval, and ratio scales of measurement.

From this understanding of the functionality, we propose to automatically determine the conversion orders of different land systems using the classification of univariate data. To this end, we adopted the time-tested and overwhelmingly popular classification algorithm for univariate data, namely Natural Breaks[33–35]. Natural Breaks is to find a classification of univariate data by maximizing the total difference between every two classes and minimizing the total difference within each class. The general algorithm of Natural Breaks is an enumeration of all possible classifications (Fig. 2a), from which the one with the largest goodness of variance fit [GVF, Eq. (12)] is selected.

$$GVF = 1 - \frac{\text{SDCM}}{\text{SDAM}} = 1 - \sum_x \sum_y \left(Z_{x,y} - M_x\right)^2 / \sum_x \sum_y \left(Z_{x,y} - M\right)^2 \qquad (12)$$

where SDCM denotes the sum of squared deviations from the class means, and SDAM denotes the sum of squared deviations from the array mean (here, the array means all values of the univariate data). $Z_{x,y}$ is the y-th value in the x-th class, M_x is the mean of all values in the x-th class, and M is the mean of all values in all classes.

However, there is a practical problem in adopting Natural Breaks. As shown in Fig. 2a, Natural Breaks works with a user-specified number of classes. In the case of CLUMondo, this number should not be static and should be capable of varying with different applications, or more specifically, with different sets of $\{S_{j,d}\}_j$. A straightforward approach to address this problem is to slightly alter the algorithm to make it perform a complete enumeration. The so-called complete enumeration aims to select the classification scheme with the largest GVF by enumerating all possible classification schemes under all possible numbers of classes. But such a straightforward approach is infeasible for two reasons. First, this approach is inefficient as it substantially increases the number of possibilities. Second and more important, the largest GVF in theory (i.e., 1 when SDCM = 0) will be achieved only if the number of classes equals the total number of values, meaning that there is no classification at all.

In this study, we propose to solve the preceding problem by modifying the Natural Breaks algorithm. Our core idea is to incorporate a threshold of GVF into the algorithm to stop complete enumerations. In this way, users no longer need to specify the number of classes; more importantly, this number will be adaptively determined. Specifically, the modified algorithm iterates all possible numbers of classes, i.e., from the smallest (i.e., 2) to the largest one (i.e., the total number of values). Within each iteration (i.e., under each number of classes), the modified algorithm further enumerates all possible classification schemes. Note that different classification schemes have the same number of classes at this stage. Each classification scheme corresponds to a GVF. A
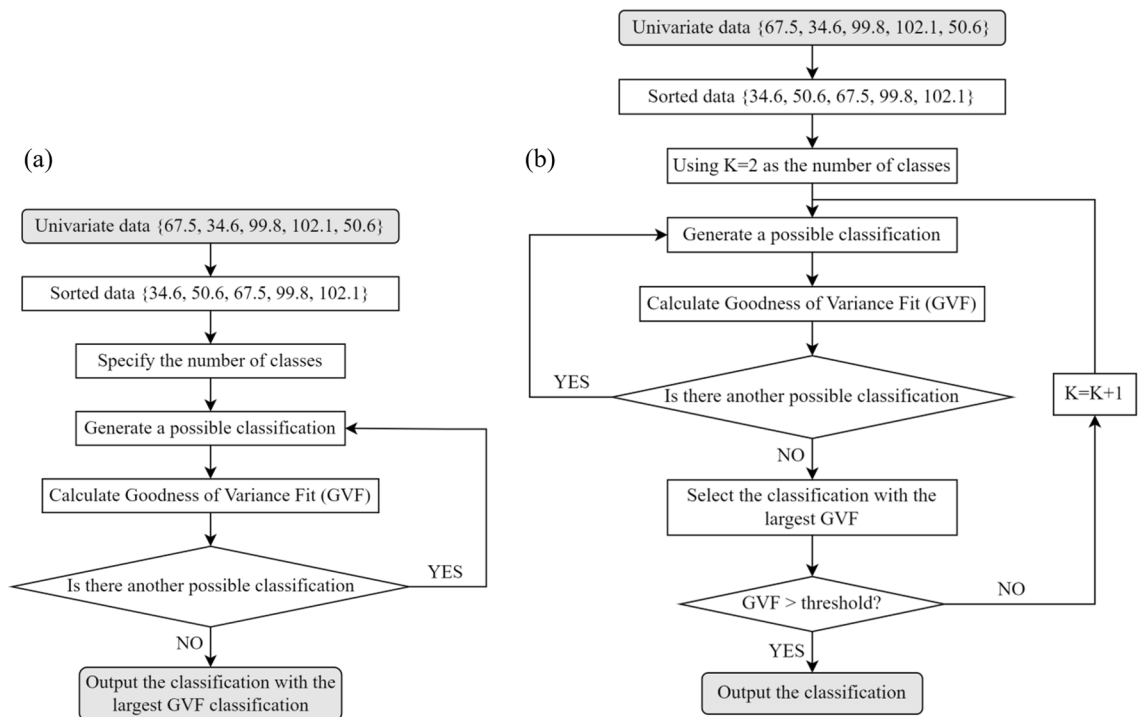

Figure 2. General algorithm of Natural Breaks (**a**) and the modified version proposed in this study (**b**).

comparison will be made between the largest GVF observed and the threshold. If that GVF is greater, then the enumeration will be stopped, and the classification scheme corresponding to that GVF will finally be adopted. Otherwise, the number of classes will be increased by one to start the next iteration. The workflow of the modified algorithm is summarized in Fig. 2b.

In practically utilizing this modified algorithm of Natural Breaks, we set the threshold of GVF as 0.8, which is an empirical value e.g.[36] and indicates an excellent classification. To automatically determine the conversion orders of different land systems in supplying the d-th service, we apply the modified algorithm to $\{S_{j,d}\}_j$ and obtain a resultant classification scheme as follows: $\{\Phi_1 < \Phi_2 < \cdots < \Phi_K\}$, where K is the number of classes, and $\Phi_{\vartheta 1} < \Phi_{\vartheta 2}$ means that the average of the ϑ1-th class is smaller than that of the ϑ2-th class. This resultant classification scheme is translated into conversion orders according to the following mapping: $L_{j,d} = \kappa - 1$ where $S_{j,d} \in \Phi_\kappa$.

Experimental evaluation

Study areas and raw data.
To select study areas, we consider the following three criteria. First, the study area should not be too small to ensure the complexity of land system changes. For example, a study area of a small city is not desirable accordingly because its land system changes are probably monotonous. Second, there should be more than one study area, to avoid the coincidence of evaluation results. Ideally, study areas should have distinct structures of land use, in terms of the composition of land system types and/or their spatial patterns. The third criterion is a practical issue: data availability and sufficiency. The data used for experimental evaluation should include land system data with a fine spatial resolution for two historical years and various potential driving factors with the same spatial resolution as the land system data.

According to the preceding criteria, we selected two study areas, namely the Sichuan and Henan provinces of China. Their geographic locations are shown in Fig. 3. Sichuan has a large area of 486,000 km², ranking fifth among the 34 Chinese provinces (or equivalent administrative units). The province covers the western part of a lowland region called the Sichuan Basin, surrounded by the Himalayas to the west, the Qinling range (i.e., the Qin Mountains) to the north, and the mountainous areas of Yunnan Province to the south. The topography of Sichuan is characterized by a considerable decrease in elevation from west to east, as shown in Fig. 3a. Dominant types of land use/cover of Sichuan include forests (40.4%), grasslands (30.7%), and cultivated lands (24.1%), and the proportion of urban areas is noticeably tiny (0.49%), according to the 2010 dataset of GlobeLand30[37].

Henan is a province in the central part of China, covering a large part of the agriculturally fertile and densely populated North China Plain. It is an agricultural province with food production of 65.4 million tons per year, ranking second out of the 34 Chinese provinces or equivalent administrative units[38]. The population of Henan is 99.3 million[39], which ranks third in China and is greater than 94% of countries (or dependent territories) according to the data by the United Nations[40]. In comparison to Sichuan, the topography of Henan is dominated
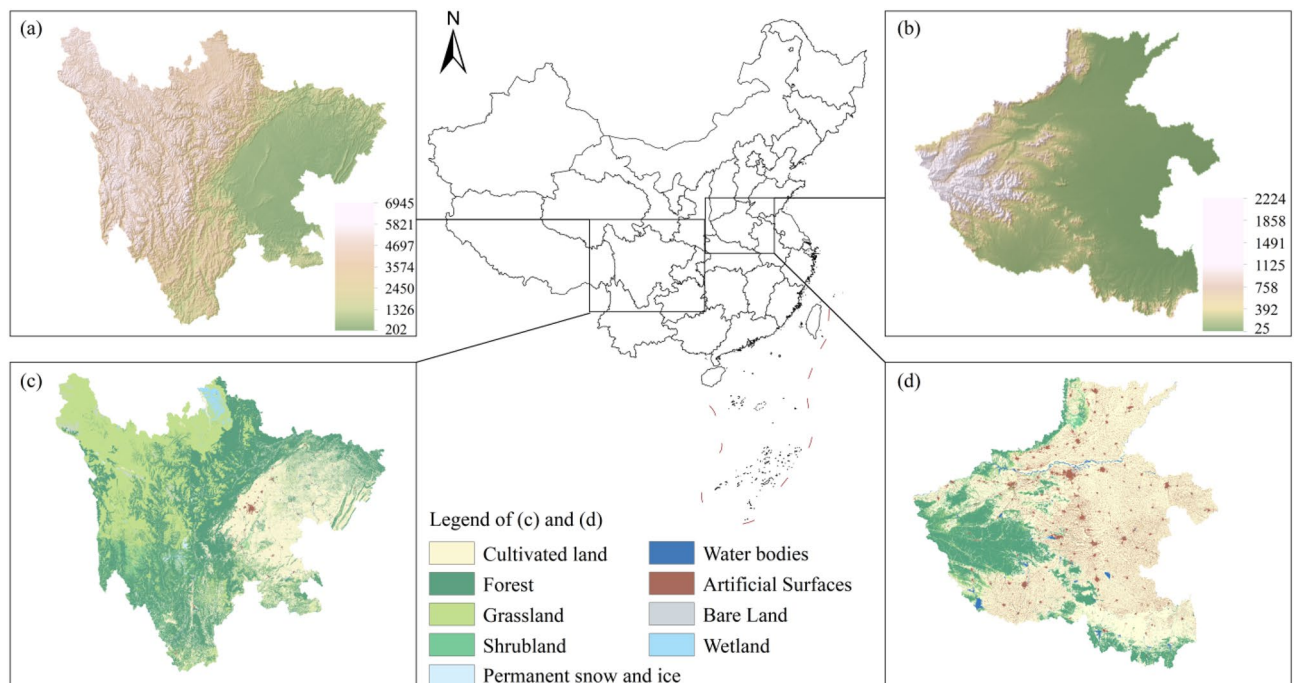

Figure 3. Study areas. (**a**) Topography of Sichuan; (**b**) Topography of Henan; (**c**) GlobeLand30 land use/cover of Sichuan; and (**d**) GlobeLand30 land use/cover of Henan. The land use/cover data can be obtained from the website of the GlobeLand30 (http://www.globallandcover.com/). The topography data can be obtained from the website of the WorldClim (https://www.worldclim.org/). The figure is generated using ArcGIS 10.7 (http://www.esri.com/software/arcgis/arcgis-for-desktop/).

by a flat plain with a few highlands, as shown in Fig. 3b. In addition, the structure of land use/cover in Henan is quite different from Sichuan. Cultivated lands, forests, and urban areas are the first three major types of land use/cover, occupying 64.9%, 19.4%, and 11.3% of Henan's total area, respectively.

The experimental evaluation relies on two types of raw data: land use/cover data and potential driving factors. This study set two criteria for preparing the land use/cover data. First, the data of Sichuan and Henan should be available for at least two periods. The data for the earlier period are used as the starting point of the simulation, whereas that for a later period is used as the benchmark for the simulation results. Second, the data should have a fine spatial resolution to facilitate the generation of land systems at a coarse scale. According to the two criteria, we employed the GlobeLand30 datasets[37], a 30-m resolution global land cover product released for 2000, 2010, and 2020. The thematic resolution of GlobeLand30 is ten types of land cover, i.e., cultivated land, forest, grassland, shrubland, wetland, water bodies, tundra, artificial surfaces, bare land, and permanent snow/ice. We extracted the 2010 and 2020 data for Sichuan and Henan from GlobeLand30. Note that the extracted GlobeLand30 are only the raw data of our 1-km resolution land system.

Potential driving factors should be prepared at the same spatial resolution of the land systems and as diverse as possible. Because we will produce land system data at the spatial resolution of 1 km, the expected spatial resolution of potential driving factors is 1 km. We collected or generated a total of 55 1-km potential driving factors, which can be classified into seven categories as shown in Table 2. Some of the potential driving factors are visualized in Fig. 4.

Establishment of multifunctional land systems. As noted in the introduction, CLUMondo features the capability of simulating land changes with many-to-many demand–supply relationships. Therefore, a comprehensive evaluation should be carried out to exploit such featured capability, where the key lies in establishing multifunctional land systems. The establishment involves two steps: generating a taxonomy of land systems and quantifying the services of different land systems.

In this study, we generated the taxonomy of land systems based on the scale transformation of the GlobeLand30 datasets, or more specifically, by transforming the spatial resolution of the GlobeLand30 datasets from 30 m to 1 km. First, we upscaled the GlobeLand30 datasets from the initial spatial resolution of 30 m to a coarser resolution of 990 m, by aggregating every 33 × 33 pixels (referred to as micro-pixels) of the raw data into new ones (referred to as macro-pixels). For more information on the aggregation technique, we refer the reader to materials on the multiscale representation of spatial data[50]. Then, we specified the land system type of each macro-pixel as the dominated type of corresponding micro-pixels, and we further distinguished three levels of dominance, namely high, medium, and low-density. Accordingly, we generated as many as 30 land systems, such as high/medium/low-density forests, as shown in Fig. 5. In particular, thresholds for the three levels of each dominated type of micro-pixel were determined using Natural Breaks with a designated classification number of three; values of these thresholds are also shown in Fig. 5. Finally, the land systems were slightly resampled to the spatial resolution of 1 km to match the resolution of most potential driving factors.

The services of each land system were defined as the area of each of the ten types of GlobeLand30 land use/cover. Under this definition, each land system would potentially become multifunctional to supply all ten services. Because our land systems were generated by transforming the spatial resolution of GlobeLand30, a pixel of any land system (a macro-pixel) corresponds to many GlobeLand30 pixels (i.e., micro-pixels) with usually diverse types. To quantify the capability of each land system in supplying every service (e.g., in 2010), we first performed an overlay analysis between the generated land systems and their corresponding raw data of land use/cover (i.e., the 2010 dataset of GlobeLand30). Based on the resultant overlaps, the capability can be determined using the following equation:

$$S_{j,d} = \Lambda_{j,d}/\Lambda_j \tag{13}$$

where Λ_j denotes the total area of the j-th land system in a given study area, and $\Lambda_{j,d}$ is the total area of the micro-pixels that overlap the j-th land system and have the d-th type of GlobeLand30 land use/cover (i.e., the d-th service). The units of $S_{j,d}$ are km^2/km^2. The aggregated demand for the d-th service in a year (e.g., 2020) was calculated as the total area of the d-th type of pixel (i.e., micro-pixels) in the GlobeLand30 dataset of that year.

Settings of other simulation parameters. In addition to establishing multifunctional land systems, some other parameters must be set before running CLUMondo, such as local suitability and conversion resistance. Since these parameters were not the objective of our experimental evaluation, we adopted default but reasonable settings, or setting methods, if possible.

The location suitability was calculated using the default method, i.e., the logistic regression based on a series of driving factors, as shown in Eq. (3). Note that not all of our potential driving factors (as previously shown in Table 2) were included in the logistic regression. We removed some potential driving factors to reduce the correlation among them. Specifically, we first measured the correlation between each pair of potential driving factors using Spearman's rank correlation coefficient (SRCC). Then, for each pair with an SRCC greater than 0.9, we removed from the pair the one that is more correlated with all other potential driving factors. To determine which is more correlated with others, we calculated the sum of SRCCs between the one potential driving factor and each other. Third, the first two steps were repeated until the SRCC of each pair of potential driving factors were less than 0.9. It is also noted that not all spatial locations within the study area were included in the logistic regression. We sampled the study area using an interval of one pixel; thus, only approximately 25% of locations were used for regression. Such a sampling strategy avoids the selection of neighboring locations, so it improves the independence of our samples.

Category	No.	Data	Year	Source
Soil	1	Bulk density	2017	[41]
	2	Cation exchange capacity	2017	
	3	Clay content	2017	
	4	Coarse fragments volumetric	2017	
	5	Derived available soil water capacity	2017	
	6	Organic carbon density	2017	
	7	pH in H_2O	2017	
	8	Sand content	2017	
	9	Silt content	2017	
	10	Texture class	2017	
Socio-economic	11	Market access index	2011	[42]
	12	Market influence index ($/person)	2011	
	13	Market density index	2011	
	14	Night-time lights	2010	DMSP-OLS Version 4
	15	Total GDP (based on purchasing power parity)	2015	[43]
	16	Gridded population of the world	2010	[44]
Accessibility	17	Time to nearest cities	2015	[45]
	18	Distance to the nearest river	N/A	Calculated in this study
	19	Distance to the nearest road	N/A	
	20	Distance to the nearest railway	N/A	
	21	Travel time one meter (motorized)	2019	[46]
	22	Travel time one meter (walking-only)	2019	
	23	Time to the nearest healthcare facility (motorized)	2019	
	24	Time to the nearest healthcare facility (walking-only)	2019	
Agriculture and vegetation	25	175 crops yield per hectare	2000	[47]
	26	Gross primary production-March/June/September/December	2010	[48]
	27	Normalized Difference Vegetation Index (NDVI)-March	2010	https://doi.org/10.5281/zenodo.3256275
	28	NDVI-June/September/December	2010	
Terrain	29	Elevation	2000	[49]
	30	Variance of elevation	N/A	Calculated in this study
	31	Slope	N/A	
	32	Aspect	N/A	
Climate	33	Annual mean precipitation	2007–2018 avg	https://doi.org/10.5067/modis/myd13a2.006
	34	Mean precipitation-March/June/September/December	2007–2018 avg	
	35	Annual mean temperature	2000–2017 avg	https://doi.org/10.5281/zenodo.1435938
	36	Mean temperature-March/June/September/December	2000–2017 avg	
Livestock	37	Buffaloes	2010	https://doi.org/10.7910/dvn/5u8mwi
	38	Cattle	2010	https://doi.org/10.7910/dvn/givq75
	39	Chickens	2010	https://doi.org/10.7910/dvn/sufasb
	40	Ducks	2010	https://doi.org/10.7910/dvn/ichcbh
	41	Goats	2010	https://doi.org/10.7910/dvn/ocph42
	42	Horses	2010	https://doi.org/10.7910/dvn/7q52mv
	43	Pigs	2010	https://doi.org/10.7910/dvn/33n0jg
	44	Sheep	2010	https://doi.org/10.7910/dvn/blwpzn

Table 2. Potential driving factors. *N/A* not available, *"avg"* average (because only average data were released).

The conversion resistance of each land system, $P\_res_{T(c,t,0)}$, was determined using historical land system changes with Eqs. (4)–(6), where $h1 = 2010$ and $h2 = 2020$. The value of $Con(T(c,t,i),k)$ was set by checking whether these are conversions from the land system type of $T(c,t,i)$ in 2010 to the k-th land system type in 2020. To avoid noise, we introduced a threshold of 1% in the check. Only when the area of conversion is larger than
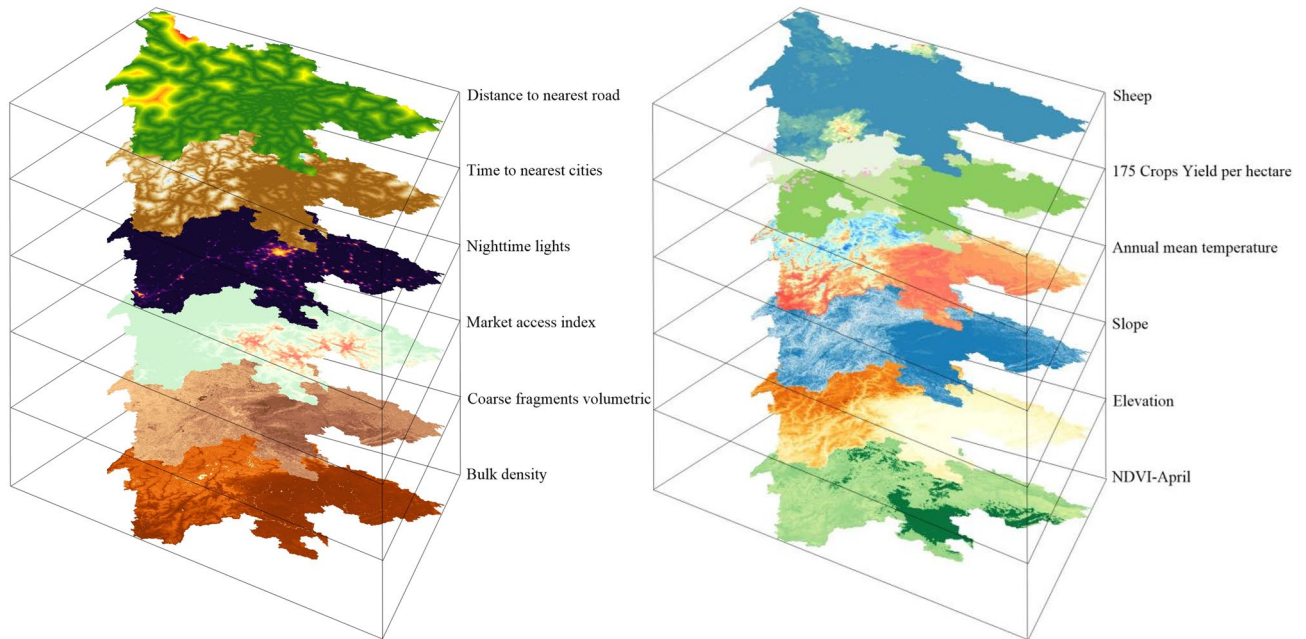
Figure 4. Some potential driving factors prepared for the study area of Sichuan. The data source can be found in Table 2. The figure is generated using Surfer 11 (https://www.goldensoftware.com/products/surfer).

1% of the study area, the value $\text{Con}(\text{T}(c, t, i), k)$ was set as 1; otherwise, it was set as 0. In the experimental evaluation, we do not employ spatial and temporal restrictions, which are optional in Eq. (1).

Benchmarks and evaluation metrics. We have both comparative and benchmark experiments. These two categories of experiments shared the same experimental settings for a study area but different determinations of conversion orders. In the comparative experiments, conversion orders were determined using the automatic method proposed in this study. By contrast, in the benchmark experiments, conversion orders were determined objectively according to the capabilities of different land systems to supply services and also by ensuring their reflection of "the relative order of the land systems contribution to fulfilling a specific demand type"[7], as follows:

$$L_{j,d} = \begin{cases} -1 & S_{j,d} = 0 \\ Rank(S_{j,d}) & S_{j,d} \neq 0 \end{cases} \tag{14}$$

where $Rank(S_{j,d})$ returns the order (starting from 1) of $S_{j,d}$ in the ascending sequence of $\{S_{j,d} \neq 0\}_d$.

To evaluate the performance of our logistic regressions, we drew receiver operating characteristic (ROC)[51] curves to assess the fit of the logistic regression established for each land system. We employed a measure developed with the ROC curves to quantify each regression's goodness of fit: Area Under the Curve (AUC, sometimes referred to as the ROC value)[52]. The theoretical value of AUC ranges from 0.5 to 1, where a higher value indicates a better fit. According to expert experience e.g.,[22,27,53] an AUC value of 0.7 or above means good fit, and that of 0.9 or above demonstrates excellent fit.

We utilized two popular metrics to evaluate the performance of land change simulation: the standard Kappa index of agreement and the total disagreement. The former metric is also called the Kappa statistic[54] or the Kappa index[23]. It is an improved measure compared with fraction correct (also called proportion correct or proportion agreement), which is biased in most cases when applied to land system maps with unevenly distributed categories of cells. Its calculation incorporates the expected proportion of agreement due to chance, as follows:

$$Kappa = (P_0 - P_e)/(100\% - P_e) \tag{14}$$

where P_0 is the proportion of agreement calculated between the simulated and the actual land systems in 2020, and P_e is the expected proportion of agreement due to chance. The Kappa statistic is a positive metric: The greater the Kappa statistic, the better the performance of a land change simulation.

The latter metric (D) was proposed by Pontius Jr and Millones[55] as an alternative to the Kappa statistic, as follows:

$$\begin{cases} D = 0.5 * \left(\sum_{g=1}^{J} q_g + \sum_{g=1}^{J} a_g \right) \\ q_g = \left| \sum_{i=1}^{J} p_{ig} - \sum_{j=1}^{J} p_{gj} \right| \\ a_g = 2 * min\left[\left(\sum_{i=1}^{J} p_{ig} \right) - p_{gg}, \left(\sum_{j=1}^{J} p_{gj} \right) - p_{gg} \right] \end{cases} \tag{15}$$
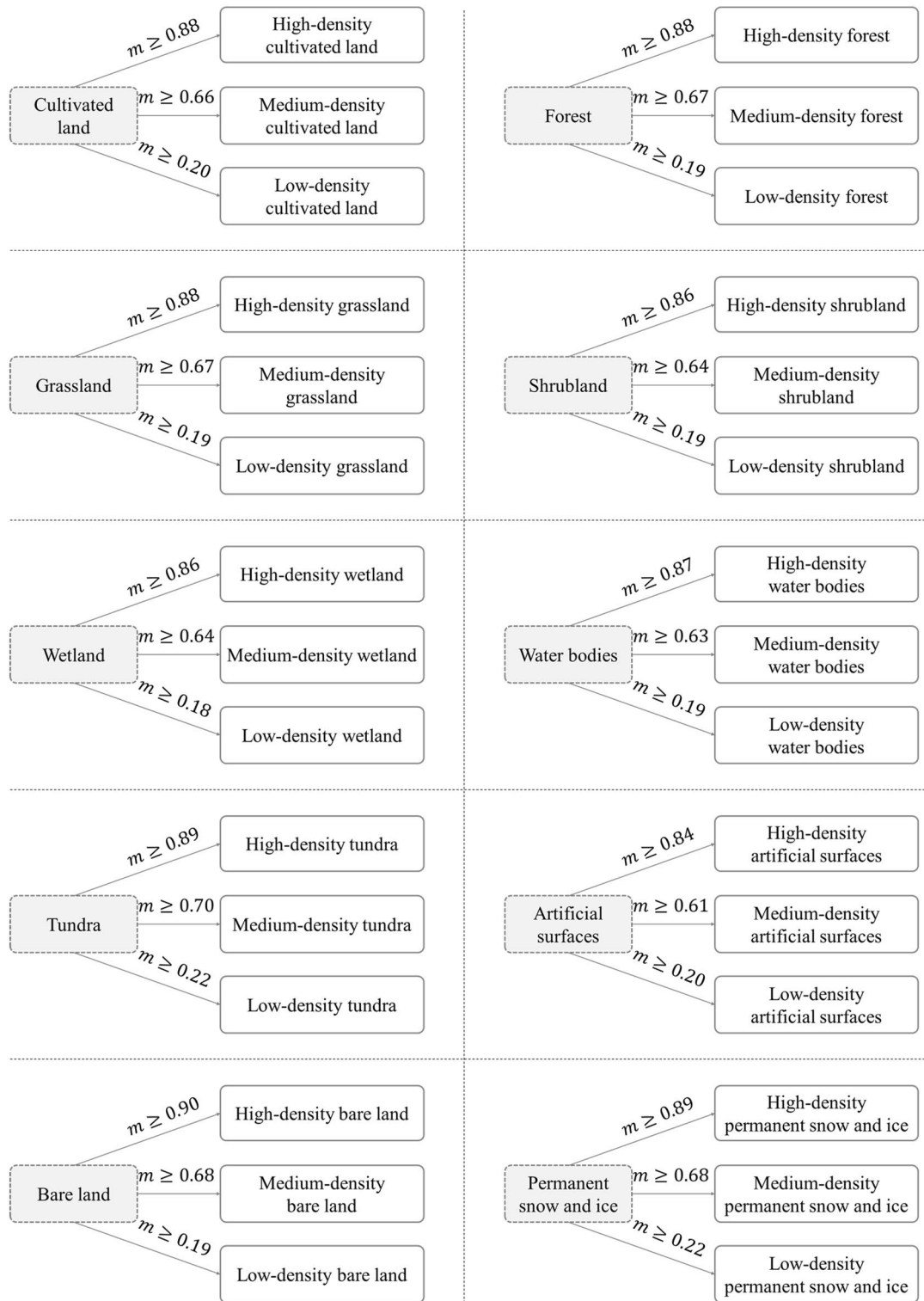
Figure 5. Taxonomy of land systems for our study areas (*m* denotes the threshold).

where J is the total number of land system types, and p_{ij} is the proportion of the study area that is of i-th land system type in the simulation result and the j-th land system type in the reference result. The total disagreement is a negative metric: The smaller the total disagreement, the better the performance of a land change simulation.

Results and analysis. The evaluation results of our logistic regressions are shown in Table 3, which consists of the AUCs of all logistic regressions established for each land system of the two study areas. For the study area of Sichuan, we can see from this table that all AUCs are greater than 0.700 and averaged at 0.913. The proportion

ID	Land system type	AUC
0	Low-density cultivated land	0.856
1	Medium-density cultivated land	0.891
2	High-density cultivated land	0.964
3	Low-density forest	0.730
4	Medium-density forest	0.756
5	High-density forest	0.908
6	Low-density grassland	0.730
7	Medium-density grassland	0.802
8	High-density grassland	0.945
9	Low-density shrubland	0.907
10	Medium-density shrubland	0.924
11	High-density shrubland	0.911
12	Low-density wetland	0.964
13	Medium-density wetland	0.975
14	High-density wetland	0.999
15	Low-density water surfaces	0.889
16	Medium-density water surfaces	0.822
17	High-density water surfaces	0.864
18	Low-density artificial surfaces	0.964
19	Medium-density artificial surfaces	0.981
20	High-density artificial surfaces	0.999
21	Low-density bare land	0.918
22	Medium-density bare land	0.977
23	High-density bare land	0.994
24	Low-density permanent snow and ice	0.987
25	Medium-density permanent snow and ice	0.992
26	High-density permanent snow and ice	0.999

Table 3. Evaluation results (AUCs) of logistic regressions in the study area of Sichuan.

of AUCs greater than 0.900 is 66% (18 out of 27), and that of AUCs greater than 0.800 is as high as 89% (24 out of 27). These results demonstrate that our incorporation of a large number of diverse, potential driving factors into logistic regression is valid and highly effective. These results also demonstrate the excellent fit of the vast majority of the established logistic regressions. We also noticed from Table 3 the following pattern: the AUC generally decreases from a high-density land system to the corresponding medium-density land system and then the low-density one. This pattern makes sense because the cell-level heterogeneity reduces from a high-density land system to the corresponding medium-density and low-density ones. The higher cell-level heterogeneity a land system has, the more significant relationship can be established between the land system and its driving factors.

Similar findings can be made for the study area of Henan. As shown in Table 4, more than half (61%, 14 out of 23) of Henan's AUCs have a value greater than 0.950. The proportion of AUCs greater than 0.900 reached 78% (18 out of 23), and that of AUCs greater than 0.800 was 87% (20 out of 23). The average of all AUCs is 0.928, which is even greater than that of Sichuan's AUCs. Therefore, our logistic regressions for Henan are also valid and highly effective.

The evaluation results of our land change simulations in the two study areas are shown in Fig. 6 and Table 5. It can be seen from this table that for the study area of Sichuan, the Kappa statistics of its benchmark and comparative experiments are 0.4287 and 0.8656, respectively. Thus, the Kappa statistic obtained in the comparative experiment increased by 101.91% compared to that in the benchmark experiment. This considerable increase demonstrates that the proposed method is highly effective. Similar conclusions can be drawn from the total disagreement. The total disagreement of the benchmark experiment is 0.5047, whereas that of the comparative experiment was reduced substantially to 0.1169 (with a reduction rate of 76.84%). For the study area of Henan, both the Kappa statistic and the total disagreement demonstrated the effectiveness of the proposed method. Specifically, the Kappa statistic was increased from 0.6475 in the benchmark experiment to 0.6823 in the comparative experiment, and the total disagreement was decreased from 0.2824 to 0.2535.

Overall, our evaluation results with the two study areas demonstrate not only the effectiveness of the proposed method for adaptively determining conversion orders but also the method's high applicability. The method is especially of use if the simulation performance of CLUMondo is poor (e.g., with the study area of Sichuan).

ID	Land system type	AUC
0	Low-density cultivated land	0.658
1	Medium-density cultivated land	0.715
2	High-density cultivated land	0.776
3	Low-density forest	0.877
4	Medium-density forest	0.903
5	High-density forest	0.982
6	Low-density grassland	0.905
7	Medium-density grassland	0.918
8	High-density grassland	0.965
9	Low-density shrubland	0.998
10	Medium-density shrubland	0.997
11	High-density shrubland	0.995
12	Low-density wetland	0.994
13	Medium-density wetland	0.937
14	High-density wetland	0.963
15	Low-density water surfaces	0.955
16	Medium-density water surfaces	0.980
17	High-density water surfaces	0.997
18	Low-density artificial surfaces	0.846
19	Medium-density artificial surfaces	0.918
20	High-density artificial surfaces	0.991
21	Low-density bare land	0.992
22	Medium-density bare land	0.989
23	High-density bare land	–

Table 4. Evaluation results (AUCs) of logistic regressions in the study area of Henan.

Discussion

In the proposed method for adaptively determining conversion orders, we adopted an empirical value of GVF (= 0.8) in adopting our modified algorithm of Natural Breaks. To test the effectiveness of this empirical value, we performed further experiments in the two study areas. In these experiments, we first employed the general algorithm of Natural Breaks instead of our modified algorithm. In utilizing the general algorithm, we enumerated and tried every possible number of classes. This number ranged from 1 to 27 with the study area of Sichuan (as Sichuan has 27 land systems), and it went from 1 to 24 with the study area of Henan (as Henan has 24 land systems). Each possible number of classes results in a unique classification of the capabilities of different land systems (i.e., $L_{j,d}$). Then, each classification was translated into a unique set of conversion orders using the same method of the modified algorithm, namely $L_{j,d} = \kappa - 1$ where $S_{j,d} \in \Phi_\kappa$. We performed independent experiments with each set of conversion orders (i.e., each possible number of classes).

The evaluation results of each experiment are shown in Fig. 7. For both study areas, no simulation results (and thus no evaluation results) were obtained when the number of classes equaled one (i.e., when all land systems have the same conversion orders). This fact justified the importance of conversion orders and the necessity of studying how to determine them. There are also some other cases where the evaluation results were not obtained. These cases were due to the failure of CLUMondo to produce a simulation result and thus excluded from our analysis. For the experiments where the evaluation results were available, we have the following two findings:

- For the study area of Sichuan, the Kappa statistic researched its highest level (i.e., greater than 86%) when the number of classes is small (i.e., 2–4). Then, the Kappa statistic underwent a decreasing trend with more classes, and a sharp decrease can be observed when the number of classes was increased from 23 to 24. When the number of classes equalled or exceeded 24, the Kappa statistic would become smaller than 45%. The trend shown by the values of total disagreement is opposite to that of the Kappa statistic in this case. We further calculated the correlation between these two sets of values, finding that their SRCC is as high as – 0.999.
- For the study area of Henan, the Kappa statistic and the total disagreement witnessed a general downward trend and a general upward trend along with the increase of the number of classes, respectively. The SRCC between the two metrics is – 0.992 in this case. Our proposed method corresponds to the experiment in which the number of classes is three. It can be seen the proposed method led to the best performance.

Based on these two findings, we concluded that the empirical value of GVF (= 0.8) is effective and an excellent choice. It is effective because its resultant values of the Kappa statistic are among the highest ones for both study areas. In addition, it is an excellent choice as it should no longer be increased (or decreased), especially with the study area of Henan.

Figure 6. Land change simulation results of Henan (**a**) and Sichuan (**b**). The data were generated in this study. The figure is generated using ArcGIS 10.7 (http://www.esri.com/software/arcgis/arcgis-for-desktop).

Study area	Experiment	Kappa	Total disagreement
Sichuan	Benchmark	0.4287	0.5047
	Comparative	0.8656	0.1169
Henan	Benchmark	0.6475	0.2824
	Comparative	0.6823	0.2535

Table 5. Evaluation results of our land change simulations.

Our proposed method facilitates applications of CLUMondo with complex land systems (in terms of not only the number of land use/cover types but also the consideration of land use intensity). As noted in the introduction section, CLUMondo is becoming increasingly popular. But recent applications of CLUMondo still rely on simple land systems, which include only several land use/cover types. Let us take some recent studies as examples. Wang et al.[56] performed a cost–benefit analysis of China's forest landscape restoration policy with CLUMondo simulations involving six land use/cover types. While assessing the impact of global initiatives on land restoration scenarios in India, Edrisi et al. simulated the changes of eight land use/cover types using CLUMondo[57]. By contrast, Zhao et al. simulated the changes of only five land use/cover types using CLUMondo when assessing the effects of land use policies on ecosystem services[58]. Our proposed method facilitates the setting of conversion orders even if the number of land system types is larger, e.g., more than 20 in this study, and when demand–supply relationships are many-to-many, like in this study.

Conclusions

CLUMondo is the only model that simulates land changes by incorporating the multifunctionality of a land system. This incorporation enables CLUMondo to support kinds of demands, both area and non-area, and to establish many-to-many relationships between diverse demands and different types of land systems, thus allowing a more realistic and useful simulation of land changes. For example, it has been used to explore not only the
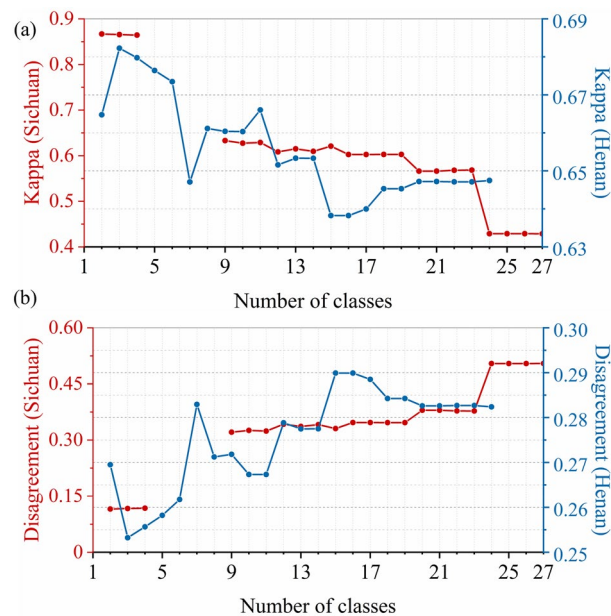
Figure 7. Evaluation results with each possible number of classes: (**a**) the Kappa statistic and (**b**) the total disagreement.

changes of land cover types but also land-use intensification e.g.,[7,27]. Therefore, it has found an increasing number of applications, where the simulation results serve as the basis of diverse analysis.

In this study, we first investigated the source code of CLUMondo, providing for this model's complete, detailed mechanism. By doing so, we facilitate future improvement on CLUMondo and its deep coupling with other earth system models. More importantly, we found that the featured function of CLUMondo—balancing demands and supplies in a many-to-many mode—relies on a parameter called conversion order. This parameter should be set manually according to the characteristics of each study area and based on expert knowledge, which is not feasible for users without understanding the whole, detailed mechanism. Therefore, the second contribution of this study is the development of an automatic method for adaptively determining conversion orders. Users with the method no longer require expert knowledge and fine-tuning for any study area. We revised the source code of CLUMondo to incorporate the proposed automatic method. To demonstrate its validity and effectiveness, we performed comparative experiments using two representative case studies, i.e., Sichuan and Henan. To ensure the experiments involved the featured function of CLUMondo, we established land systems and many-to-many demand–supply relationships (10 demands met by the supply by more than 20–30 land systems) for simulation in both case studies. From these results, we made the following three conclusions:

- Our investigation into the complete, detailed mechanism of CLUMondo is successful in that it allows the identification of core parameters of the model and future improvements;
- Conversion order is a core parameter that affects the simulation performance of CLUMondo; the performance might be unacceptably poor if conversion orders are not well specified; and
- Our proposed automatic method for adaptively determining conversion orders is valid and highly effective.

We modified the source code of CLUMondo to integrate the proposed method as an option for users (who can still set conversion orders manually). To distinguish the modified CLUMondo from the official version, we referred to this modified one as CLUMondo-BNU v1.0 and also released it for public use. It is important to note that both the original and improved models rely on logistic regressions, and the focus of this study is not to improve the regression module. However, as one of the core modules of CLUMondo, regression is recommended to improve by future studies using advance techniques such as auto-models for spatially autocorrelated occupancy and abundance data[59], the geographically weighted temporally correlated logistic regression model[60], the Maximum Entropy Model (i.e., Maxent)[61,62], artificial intelligence[63].

Code availability

The source code for the model (i.e., CLUMondo-BNU v1.0) and its manual are archived on Zenodo (https://doi.org/10.5281/zenodo.7051199). All data used to produce the Sichuan results present in this paper are archived on Zenodo (https://doi.org/10.5281/zenodo.6594722), and that used to produce the Henan results present in this paper are also archived on Zenodo (https://doi.org/10.5281/zenodo.6594815).

References

1. Wang, Y. H. *et al.* Modelling and evaluating the economy-resource-ecological environment system of a third-polar city using system dynamics and ranked weights-based coupling coordination degree model. *Cities* **133**, 104151 (2023).
2. Gao, P. C. *et al.* Sustainable land-use optimization using NSGA-II: Theoretical and experimental comparisons of improved algorithms. *Landsc. Ecol.* **36**, 1877–1892 (2021).
3. Young, A. *Land Resources: Now and for the Future* (Cambridge University Press, 2000).
4. Kong, X. S., Zhou, Z. Z. & Jiao, L. M. Hotspots of land-use change in global biodiversity hotspots. *Resour. Conserv. Recycl.* **174**, 105770 (2021).
5. Wang, S. M., Ma, Q. F., Ding, H. Y. & Liang, H. W. Detection of urban expansion and land surface temperature change using multi-temporal landsat images. *Resour. Conserv. Recycl.* **128**, 526–534 (2018).
6. Song, X. P. *et al.* Global land change from 1982 to 2016. *Nature* **560**, 639–643 (2018).
7. van Asselen, S. & Verburg, P. H. Land cover change or land-use intensification: simulating land system change with a global-scale land change model. *Glob. Change Biol.* **19**, 3648–3667 (2013).
8. Escobar, N. & Britz, W. Metrics on the sustainability of region-specific bioplastics production, considering global land use change effects. *Resour. Conserv. Recycl.* **167**, 105345 (2021).
9. Borrelli, P. *et al.* An assessment of the global impact of 21st century land use change on soil erosion. *Nat. Commun.* **8**, 1–13 (2017).
10. Bai, Y. *et al.* Developing China's ecological redline policy using ecosystem services assessments for land use planning. *Nat. Commun.* **9**, 3034 (2018).
11. Couto, E. V. D., Oliveira, P. B., Vieira, L. M., Schmitz, M. H. & Ferreira, J. H. D. Integrating environmental, geographical and social data to assess sustainability in hydrographic basins: The ESI approach. *Sustainability* **12**, 3057 (2020).
12. Grundy, M. J. *et al.* Scenarios for Australian agricultural production and land use to 2050. *Agric. Syst.* **142**, 70–83 (2016).
13. Zhang, R. S. & Hanaoka, T. Deployment of electric vehicles in China to meet the carbon neutral target by 2060: Provincial disparities in energy systems, CO_2 emissions, and cost effectiveness. *Resour. Conserv. Recycl.* **170**, 105622 (2021).
14. Calvin, K. *et al.* GCAM v5.1: Representing the linkages between energy, water, land, climate, and economic systems. *Geosci. Model Dev.* **12**, 677–698 (2019).
15. O'Neill, B. C. *et al.* A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Clim. Change* **122**, 387–400 (2014).
16. Schandl, H. *et al.* Shared socio-economic pathways and their implications for global materials use. *Resour. Conserv. Recycl.* **160**, 104866 (2020).
17. Liu, X. P. *et al.* A future land use simulation model (FLUS) for simulating multiple land use scenarios by coupling human and natural effects. *Landsc. Urban Plan.* **168**, 94–116 (2017).
18. He, C. Y., Li, J. W., Zhang, X. L., Liu, Z. F. & Zhang, D. Will rapid urban expansion in the drylands of northern China continue: A scenario analysis based on the land use scenario dynamics-urban model and the shared socioeconomic pathways. *J. Clean. Prod.* **165**, 57–69 (2017).
19. Verburg, P. H. *et al.* Modeling the spatial dynamics of regional land use: The CLUE-S model. *Environ. Manage.* **30**, 391–405 (2002).
20. van Vliet, J. & Verburg, P. H. A short presentation of CLUMondo. In *Geomatic Approaches for Modeling Land Change Scenarios* (eds Camacho Olmedo, M. T. *et al.*) (Springer, 2018).
21. Arunyawat, S. & Shrestha, R. P. Simulating future land use and ecosystem services in Northern Thailand. *J. Land Use Sci.* **13**, 146–165 (2018).
22. Mei, Z. X., Wu, H. & Li, S. Y. Simulating land-use changes by incorporating spatial autocorrelation and self-organization in CLUE-S modeling: A case study in Zengcheng District, Guangzhou, China. *Front. Earth Sci.* **12**, 299–310 (2018).
23. Jiang, W. G., Chen, Z., Lei, X., Jia, K. & Wu, Y. F. Simulating urban land use change by incorporating an autologistic regression model into a CLUE-S model. *J. Geogr. Sci.* **25**, 836–850 (2015).
24. Nie, X. *et al.* Increase or decrease? Integrating the CLUMondo and InVEST models to assess the impact of the implementation of the Major Function Oriented Zone planning on carbon storage. *Ecol. Ind.* **118**, 106708 (2020).
25. Dong, N., You, L., Cai, W. J., Li, G. & Lin, H. Land use projections in China under global socioeconomic and emission scenarios: Utilizing a scenario-based land-use change assessment framework. *Glob. Environ. Chang.* **50**, 164–177 (2018).
26. Wang, Y., van Vliet, J., Pu, L. J. & Verburg, P. H. Modeling different urban change trajectories and their trade-offs with food production in Jiangsu Province, China. *Comput. Environ. Urban Syst.* **77**, 101355 (2019).
27. Jin, X. L., Jiang, P. H., Ma, D. X. & Li, M. C. Land system evolution of Qinghai-Tibetan Plateau under various development strategies. *Appl. Geogr.* **104**, 1–9 (2019).
28. Wang, Y., van Vliet, J., Debonne, N., Pu, L. J. & Verburg, P. H. Settlement changes after peak population: Land system projections for China until 2050. *Landsc. Urban Plan.* **209**, 104045 (2021).
29. Liu, Z. F., Verburg, P. H., Wu, J. G. & He, C. Y. Understanding land system change through scenario-based simulations: A case study from the drylands in Northern China. *Environ. Manage.* **59**, 440–454 (2017).
30. van Asselen, S. & Verburg, P. H. A land system representation for global assessments and land-use modeling. *Glob. Change Biol.* **18**, 3125–3148 (2012).
31. Gao, P. C., Xie, Y. R., Song, C. Q., Cheng, C. X. & Ye, S. J. Exploring detailed urban-rural development under intersecting population growth and food production scenarios: Trajectories for China's most populous agricultural province to 2030. *J. Geogr. Sci.* **33**, 222–244 (2023).
32. Liu, Y., Goodchild, M. F., Guo, Q., Tian, Y. & Wu, L. Towards a general field model and its order in GIS. *Int. J. Geogr. Inf. Sci.* **22**, 623–643 (2008).
33. Jenks, G. F. The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **7**, 186–190 (1967).
34. Cheng, C. X., Zhang, T., Su, K., Gao, P. C. & Shen, S. Assessing the intensity of the population affected by a complex natural disaster using social media data. *ISPRS Int. J. Geogr. Inf.* **8**, 358 (2019).
35. Jenks, G. F. & Caspall, F. C. Error on choroplethic maps: Definition, measurement, reduction. *Ann. Assoc. Am. Geogr.* **61**, 217–244 (1971).
36. Long, Y., Song, Y. M. & Chen, L. Identifying subcenters with a nonparametric method and ubiquitous point-of-interest data: A case study of 284 Chinese cities. *Environ. Plan. B* **49**, 58–75 (2022).
37. Chen, J. *et al.* Global land cover mapping at 30m resolution: A POK-based operational approach. *Int. J. Photogramm. Remote Sens.* **103**, 7–27 (2015).
38. National Bureau of Statistics of China. *Announcement of Statistics on Grain Production*. (2021).
39. National Bureau of Statistics of China. *China Statistical Yearbook* (China Statistics Press, 2021).
40. United Nations. *World Population Prospects*. (2019).
41. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 (2017).
42. Verburg, P. H., Ellis, E. C. & Letourneau, A. A global assessment of market accessibility and market influence for global environmental change studies. *Environ. Res. Lett.* **6**, 034019 (2011).
43. Kummu, M., Taka, M. & Guillaume, J. H. A. Gridded global datasets for gross domestic product and human development index over 1990–2015. *Sci. Data* **5**, 180004 (2018).

16

44. Doxsey-Whitfield, E. *et al.* Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Pap. Appl. Geogr.* **1**, 226–234 (2015).
45. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
46. Weiss, D. *et al.* Global maps of travel time to healthcare facilities. *Nat. Med.* **26**, 1835–1838 (2020).
47. Monfreda, C., Ramankutty, N. & Foley, J. A. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Glob. Biogeochem. Cycles* **22**, 1022 (2008).
48. Wang, S. H., Zhang, Y. G., Ju, W. M., Qiu, B. & Zhang, Z. Y. Tracking the seasonal and inter-annual variations of global gross primary production during last four decades using satellite near-infrared reflectance data. *Sci. Total Environ.* **755**, 142569 (2021).
49. Fick, S. E. & Hijmans, R. J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
50. Gao, P. C. & Li, Z. L. Aggregation-based method for computing absolute Boltzmann entropy of landscape gradient with full thermodynamic consistency. *Landsc. Ecol.* **34**, 1837–1847 (2019).
51. Chang, Y. *et al.* Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landsc. Ecol.* **28**, 1989–2004 (2013).
52. Lin, Y. P., Chu, H. J., Wu, C. F. & Verburg, P. H. Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modelling: A case study. *Int. J. Geogr. Inf. Sci.* **25**, 65–87 (2011).
53. Hu, Z. Y. & Lo, C. P. Modeling urban growth in Atlanta using logistic regression. *Comput. Environ. Urban Syst.* **31**, 667–688 (2007).
54. Hagen A. Multi-method assessment of map similarity. In *Proceedings of the 5th AGILE Conference on Geographic Information Science*). (Universitat de les Illes Balears Palma, 2002).
55. Pontius, R. G. Jr. & Millones, M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* **32**, 4407–4429 (2011).
56. Wang, H., Tian, F., Wu, J. X. & Nie, X. Is China forest landscape restoration (FLR) worth it? A cost-benefit analysis and non-equilibrium ecological view. *World Dev.* **161**, 106126 (2023).
57. Edrisi, S. A., Bundela, A. K., Verma, V., Dubey, P. K. & Abhilash, P. C. Assessing the impact of global initiatives on current and future land restoration scenarios in India. *Environ. Res.* **216**, 114413 (2023).
58. Zhao, Y. *et al.* Distinguishing the effects of land use policies on ecosystem services and their trade-offs based on multi-scenario simulations. *Appl. Geogr.* **151**, 102864 (2023).
59. Bardos, D. C., Guillera-Arroita, G. & Wintle, B. A. Valid auto-models for spatially autocorrelated occupancy and abundance data. *Methods Ecol. Evol.* **6**, 1137–1149 (2015).
60. Liu, Y., Lam, K. F., Wu, J. T. & Lam, T.T.-Y. Geographically weighted temporally correlated logistic regression model. *Sci. Rep.* **8**, 1417 (2018).
61. Tan, J. B., Li, A. N., Lei, G. B. & Xie, X. Y. A SD-MaxEnt-CA model for simulating the landscape dynamic of natural ecosystem by considering socio-economic and natural impacts. *Ecol. Model.* **410**, 108783 (2019).
62. Gao, P. C. & Li, Z. L. Computation of the Boltzmann entropy of a landscape: A review and a generalization. *Landsc. Ecol.* **34**, 2183–2196 (2019).
63. Lv, J. J. *et al.* Simulating urban expansion by incorporating an integrated gravitational field model into a demand-driven random forest-cellular automata model. *Cities* **109**, 103044 (2021).

Author contributions
P.G. and C.S. designed the study. P.G. led the analysis of results and wrote the paper. Y.G. developed the source code and performed the experiments. Y.G. and S.Y. analyzed the experimental results. X.Z. visualized some experimental results.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31001-3

Correspondence and requests for materials should be addressed to C.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.