https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development Discussions
Open Access

CLUMondo v2.0: Improved model by adaptive determination of conversion orders for simulating land system changes with many-to-many demand-supply relationships

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

[1]Key Laboratory of Environmental Change and Natural Disaster, Beijing Normal University, Beijing, 100875, China
[2]Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, Beijing, 100875, China

Correspondence to: Changqing Song (songcq@bnu.edu.cn)

Abstract. 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. Its central mechanism is complex and has not been fully revealed or clearly explained, thus preventing further improvement. In this study, we first investigated the source code of CLUMondo, providing for the first time the complete, detailed mechanism of this model. 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. Still, the setting of this parameter should be improved because it is a manual process according to the characteristics of each study area and based on 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. We revised the source code of CLUMondo to incorporate the proposed automated method, resulting CLUMondo Version 2.0. Comparative experiments demonstrated the proposed automated method's validity, high effectiveness, and universal applicability. They showed that the new version of CLUMondo is more effective and easier to use than the existing version. A case study showed that the simulation performance has improved as high as 103.36%. This study facilitates future improvement on CLUMondo and deep coupling with other earth system models, clearly describing its mechanism. It also helps to exploit the full potential of CLUMondo with a new version.

1 Introduction

The sustainable management and conservation of land resources have been central to human society, 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" (Young, 2000). A critical focus of the management and conservation is on land-use and land-cover change, or land change for short (e.g., Song et al., 2018; Wang et al., 2018; Kong et al., 2021). 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 (Lambin et al. 2000; Li et al. 2018; Feng et al. 2020), which in turn affects land resources (Van Asselen and Verburg, 2013; Escobar and Britz, 2021). 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., Borrelli et al., 2017; Bai et al., 2018).

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 (Grundy et al., 2016). A typical example is the Global Change Assessment Model (GCAM, Calvin et al., 2019; Zhang and Hanaoka, 2021), a maker model for the famous Shared Socioeconomic Pathways (O'neill et al., 2014; Schandl et al., 2020). 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 (Liu et al., 2017a) and Land Use Scenario Dynamics-urban (LUSD-urban) model (He et al., 2017)—and suitability-based models, e.g., Conversion of Land Use and its Effects at Small regional extent (CLUE-S) model (Verburg et al., 2002) and its latest version CLUMondo (Van Vliet and Verburg, 2018). 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 have been 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., Jiang et al., 2015; Arunyawat and Shrestha, 2018; Mei et al., 2018). 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., Dong et al., 2018; Nie et al., 2020). 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., Jin et al., 2019; Wang et al., 2019). The demand for

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development
Discussions
Open Access

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.

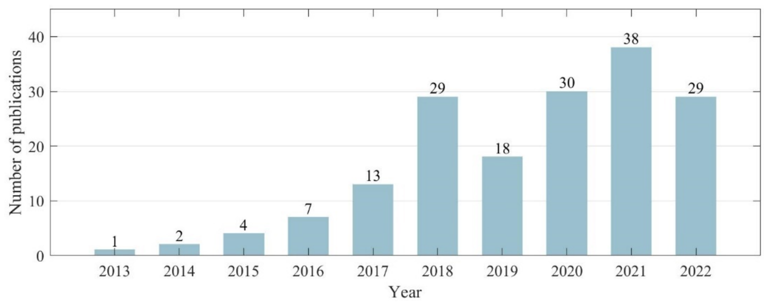


Figure 1. Increasing number of publications retrieved from Google Scholar with the keyword "CLUMondo."

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. However, as noted in Section 2, such a central mechanism has not been fully revealed and clearly explained, especially how CLUMondo balances demands and supplies in a many-to-many mode. The key question is how the so-called competitive advantage is determined and used. Therefore, in this study, we first investigated the detailed mechanism of competitive advantage, making it possible for the future improvement of CLUMondo. The investigation found that the function of competitive advantage depends on an input of conversion orders, which should be manually set by users and require both expert knowledge and fine-tuning. More importantly, the values of conversion orders should vary with studies areas and land system characteristics; thus, the determination of conversion orders is rather sophisticated. To address this problem, we developed an automatic method for adaptively determining conversion orders. Users with this method no longer require expert knowledge and fine-tuning. We updated CLUMondo by incorporating this proposed method, resulting in a new version.

2 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., Van Asselen and Verburg, 2012; Liu et al., 2017b; Wang et al., 2021b), namely land system and land system services. In the

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

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" (Van Vliet and Verburg, 2018). For example, in the case study by Jin et al. (2019), types of land systems include low/medium/high-covered natural grassland, low/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 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 (Gao et al., 2022), 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:

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

where c denotes the c-th cell. k and j denote the k-th and j-th land system type, respectively ($1 \leq k, j \leq n$). $\mathrm{T}(c,t,i)$ 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. $\xi\big(\mathrm{T}(c,t,0)\big)$ calculates how many time steps (usually years) c has been maintained at the initial land system type of this time step, namely $\mathrm{T}(c,t,0)$. $\tau\big(\mathrm{T}(c,t,0)\big)$ is a non-negative integer representing the minimum time steps that the land system type $\mathrm{T}(c,t,0)$ 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. $\mathrm{Con}\big(\mathrm{T}(c,t,i),k\big)$ indicates whether the conversion from $\mathrm{T}(c,t,i)$ to k is allowed according to the settings of users, 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 (2013), the standard determination of $P_{c,j}$ is a linear combination process involving three basic factors:

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development Discussions

$$P_{c,j} = \begin{cases} P\_loc_{c,j} + P\_res_{\mathrm{T}(c,t,0)} + P\_cmp_{i,j} & if\ \mathrm{T}(c,t,0) = j \\ P\_loc_{c,j} + P\_cmp_{i,j} & if\ \mathrm{T}(c,t,0) \neq j \end{cases}, \quad (2)$$

where $P\_loc_{c,j}$, P_res_j, 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):

$$\ln\left(\frac{P\_loc_{c,j}}{1-P\_loc_{c,j}}\right) = \beta_{0,j} + \beta_{1,j}X_{1,c} + \beta_{2,j}X_{2,c} + \cdots + \beta_{m,j}X_{m,c}, \quad (3)$$

  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.

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

$$P\_res_{\mathrm{T}(c,t,0)} = \sum_{c'} y^{h1,h2}_{c',\mathrm{T}(c,t,0)} / \sum_{c'} y^{h1}_{c',\mathrm{T}(c,t,0)}, \quad (4)$$

$$y^{h1,h2}_{c',\mathrm{T}(c,t,0)} = \begin{cases} 1 & if\ \mathrm{T}(c',h2,0) = \mathrm{T}(c',h1,0) = \mathrm{T}(c,t,0) \\ 0 & else \end{cases}, \quad (5)$$

$$y^{h1}_{c',\mathrm{T}(c,t,0)} = \begin{cases} 1 & if\ \mathrm{T}(c',h1,0) = \mathrm{T}(c,t,0) \\ 0 & else \end{cases}, \quad (6)$$

  where c' denotes the c'-th cell. $h1$ and $h2$ denote two historical years, and $h1 < h2$.

- 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 (2018), $P\_cmp_{i,j}$ has the following properties:

$$\begin{cases} P\_cmp_{i,j} = \sum_g P\_cmp_{i,j,d} \\ P\_cmp_{i,j,d} \propto S_{j,d}, d_{i,d} \end{cases}, \quad (7)$$

  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. It should be noted that only these properties of $P\_cmp_{i,j}$ have been disclosed to users and readers. The detailed mechanism of $P\_cmp_{i,j}$ is absent from both the literature and user manuals.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development
Discussions

3 Investigated mechanism and novel method

3.1 Detailed mechanism of the competitive advantage and problems with its conversion order

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). As a result, this detailed mechanism was discovered for the first time and mathematically expressed in this study as Eqs. (8–10).

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

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).

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

$$Speed_i = \begin{cases} 0.05 & i = 1 \\ Speed_{iter-1} + 0.0002 & i \geq 2 \end{cases}, \quad (10)$$

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:

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

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 in the case of small changes to land systems are needed.

Although the detailed mechanism of the competitive advantage is clear now, the practical problem with exploiting it lies in the determination of the conversion orders in Eq. (8). Currently, no automatic method has been developed for this determination. Only Van Asselen and Verburg (2013) illustrated a determination result, as shown in Table 1. On the explanation of this table, Van Asselen and Verburg (2013) 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.

Table 1 Capability and conversion orders determined by Van Asselen and Verburg (2013) 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 (km^2).

	Crop production (tons)		Land-based livestock (number)		Landless Livestock (number)		Built-up area (km^2)	
	Capability	Order	Capability	Order	Capability	Order	Capability	Order
Cropland extensive, few livestock (ls.)	8,977	4	4,658	2	9,620	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	3,317	2	81,968	3	0.11	1
Cropland med. intensive, few ls.	11,695	5	3,704	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	4,387	2	102,894	4	0.43	1
Cropland intensive, few ls.	24,076	6	2,076	2	7,934	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	4,264	2	172,779	5	0.67	1
Mosaic cropland & grassland, land-based ls.	13,563	4	13,843	4	132,327	4	0.52	1
Mosaic cropland & grassland, landless ls.	16,080	4	4,005	2	122,532	4	0.46	1
Mosaic cropland (ext.) & grassland, few ls.	3,871	2	4,736	2	8,152	2	0.09	1
Mosaic cropland (med.) & grassland, few ls.	6,504	3	4,403	2	11,890	2	0.25	1
Mosaic cropland (int.) & grassland, few ls.	10,984	4	3,374	2	8,815	2	0.47	1
Mosaic cropland & forest, landless ls.	14,548	3	3,815	2	112,431	4	0.33	1
Mosaic cropland (ext.) & open forest, few ls.	6,104	2	3,754	2	15,727	2	0.09	1
Mosaic cropland (med. int.) & forest, few ls.	6,752	3	3,511	2	13,386	2	0.17	1
Mosaic cropland (intensive) & forest, few ls.	9,774	4	3,127	2	12,680	2	0.32	1

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Dense forest	1,478	1	2,368	−1	28,849	−1	0.07	1
Open forest, few ls.	1,459	1	2,302	1	8,976	1	0.09	1
Open forest, landless ls.	4,576	−1	3,073	−1	89,483	3	0.13	1
Mosaic grassland & forest	3,043	1	3,441	1	38,943	1	0.14	1
Mosaic grassland & bare	381	1	2,824	1	3,294	1	0.08	1
Natural grassland	749	1	0	1	0	1	0	1
Grassland, few ls.	1,610	1	2,720	1	18,250	1	0.13	1
Grassland, land-based ls.	2,059	−1	14,159	4	37,991	−1	0.23	1
Bare	18	−1	4	−1	4	−1	0	1
Bare, few ls.	430	−1	2,928	1	2,948	1	0.04	1
Peri-urban & villages	22,056	−1	9,110	−1	184,526	−1	8.97	2
Urban	17,796	−1	5,010	−1	193,283	−1	37.60	3

3.2 Automatic determination of conversion orders using an adaptive data classification scheme

This study presents a simple but powerful 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 universally applicable.

Before developing the method, we rethink the functionality of the conversion order as a parameter of the competitive advantage. As noted in Section 2.1, 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 2.2, 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 capability 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 spoorer 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 Liu et al., 2008 for the nominal, ordinal, interval, and ratio scales of measurement).

From this deep 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 scheme for univariate data, namely Natural Breaks (Jenks, 1967; Jenks and Caspall, 1971; Cheng et al., 2019).

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

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. 2), from which the one with the largest goodness of variance fit (GVF, Eq. 12) is selected.

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

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. 2(a), Natural Breaks works with a user-specified number of classes. In the case of CLUMondo, this number should not be static and can vary 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, which is to select the classification with the largest GVF by enumerating all possible classifications under all possible numbers of classes. But such an 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 $\mathrm{SDCM} = 0$) 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 algorithm of Natural Breaks. Our core idea is to set a threshold of GVF to stop the complete enumeration. Specifically, the modified algorithm enumerates all possible numbers of classes from the smallest (i.e., 2) to the largest (i.e., the total number of values). Under each number of classes, the modified algorithm further enumerates all possible classifications and finds the one with the largest GVF. At the end of the further enumeration under each number of classes, a comparison will be made between the obtained largest GVF and the threshold. If that GVF is greater than the threshold, then the enumeration will be stopped, and the classification corresponding to that GVF will be finally adopted. The workflow of the modified algorithm is summarized in Fig. 2(b). It can be seen the modified algorithm is adaptive in that the algorithm does not require a user-specified number of classes.

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., Long et al., 2022) 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 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 $\vartheta 1$-th class is smaller than that of the $\vartheta 2$-th class. This resultant classification is translated into conversion orders according to the following mapping: $L_{j,d} = \kappa - 1$ where $S_{j,d} \in \Phi_\kappa$.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

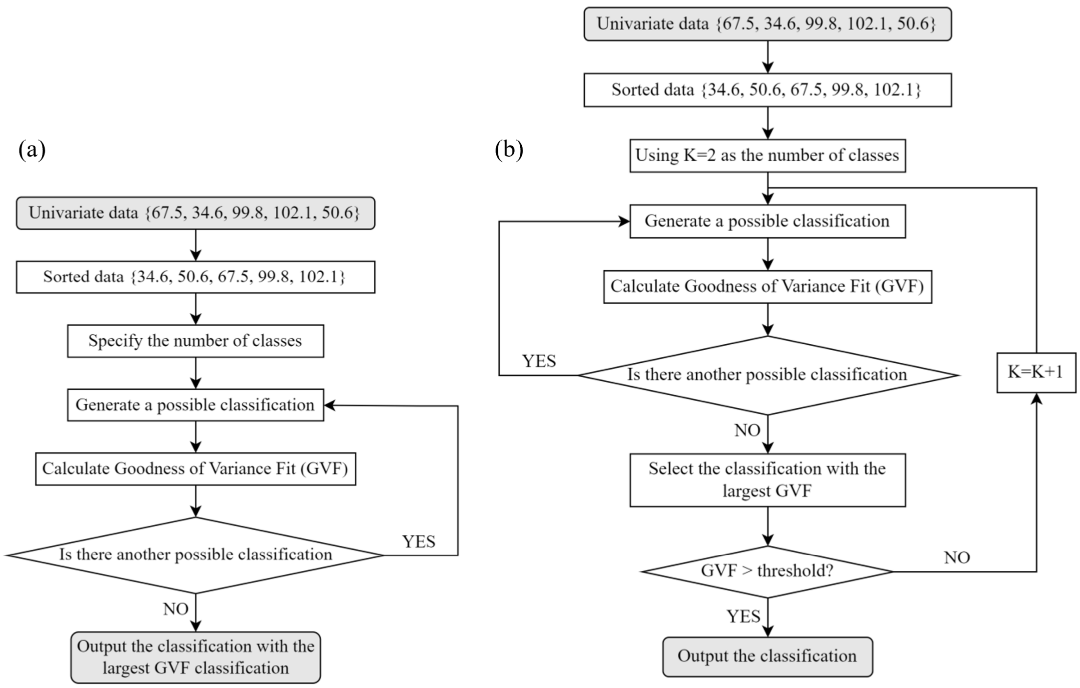


Figure 2 General algorithm of Natural Breaks (a) and the modified version proposed in this study (b)

4 Experimental evaluation

4.1 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. 4. Sichuan has a large area of 486,000 km^2, ranking the fifth among the 34 Chinese

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

provinces (or equivalent administrative units). Its area is larger than most countries all over the world, such as Japan (377,975 km^2), Germany (357,022 km^2), and the United Kingdom (244,820 km^2). 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. 4(a). 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 (Chen et al., 2015).

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 the second out of the 34 Chinese provinces (or equivalent administrative units, National Bureau of Statistics of China, 2021a). The population of Henan is 99.3 million (National Bureau of Statistics of China, 2021b), which ranks the third in China and is great than 94% of countries (or dependent territories) according to the data by United Nations (2019). It has an area of 167,000 km^2, which is approximately twice that of Austria (83,879 km^2). In comparison to Sichuan, the topography of Henan is dominated by a flat plain with a few highlands, as shown in Fig. 4(b). 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 the following two criteria in 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 (Chen et al., 2015), 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 7 categories as shown in Table 2. Some of the potential driving factors are visualized in Fig. 5.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development
Discussions
Open Access
EGU

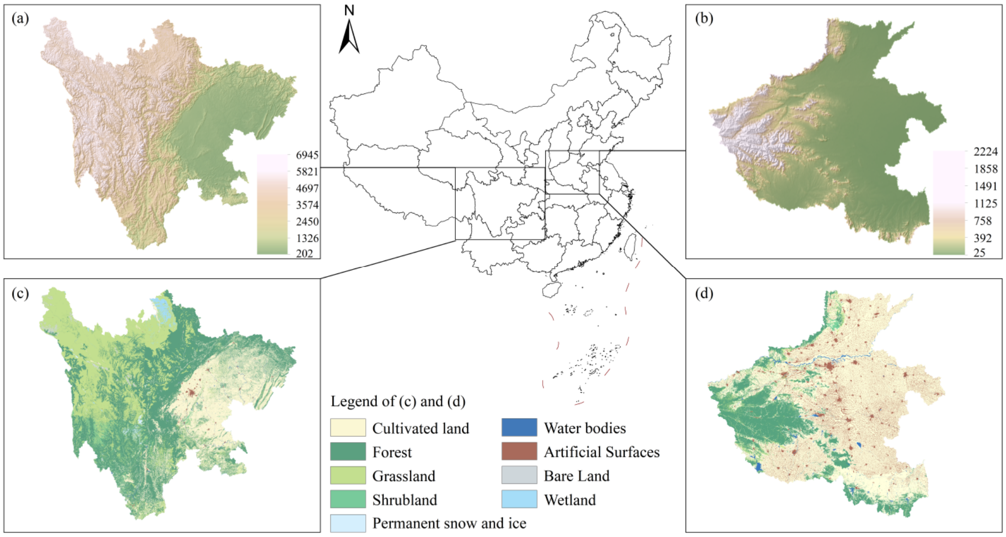


Figure 3 Study areas. (a) Topography of Sichuan; (b) GlobeLand30 land use/cover of Sichuan; (c) Topography of Henan; and (b) GlobeLand30 land use/cover of Henan

Table 2 Potential driving factors

Category	No.	Data	Year	Source
Soil	1	Bulk density	2017	(Hengl et al., 2017)
	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	(Verburg et al., 2011)
	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	(Kummu et al., 2018)

Geoscientific Model Development Discussions
Open Access
EGU

	16	Gridded population of the world	2010	(Doxsey-Whitfield et al., 2015)
Accessibility	17	Time to nearest cities	2015	(Weiss et al., 2018)
	18	Distance to nearest river	N/A	Calculated in this study
	19	Distance to nearest road	N/A	
	20	Distance to nearest railway	N/A	
	21	Travel time one meter (motorized)	2019	(Weiss et al., 2020)
	22	Travel time one meter (walking-only)	2019	
	23	Time to nearest healthcare facility (motorized)	2019	
	24	Time to nearest healthcare facility (walking-only)	2019	
Agriculture & Vegetation	25	175 crops yield per hectare	2000	(Monfreda et al., 2008)
	26	Gross primary production-March	2010	(Wang et al., 2021a)
	27	Gross primary production-June	2010	
	28	Gross primary production-September	2010	
	29	Gross primary production-December	2010	
	30	Normalized Difference Vegetation Index (NDVI)-March	2010	DOI: 10.5067/modis/myd13a2.006
	31	NDVI-June	2010	
	32	NDVI-September	2010	
	33	NDVI-December	2010	
Terrain	34	Elevation	2000	(Fick and Hijmans, 2017)
	35	Variance of elevation	N/A	Calculated in this study
	36	Slope	N/A	
	37	Aspect	N/A	
Climate	38	Annual mean precipitation	2007-2018 avg	DOI: 10.5281/zenodo.3256275
	39	Mean precipitation-March	2007-2018 avg	
	40	Mean precipitation-June	2007-2018 avg	
	41	Mean precipitation-September	2007-2018 avg	
	42	Mean precipitation-December	2007-2018 avg	
	43	Annual mean temperature	2000-2017 avg	DOI: 10.5281/zenodo.1435938
	44	Mean temperature-March	2000-2017 avg	
	45	Mean temperature-June	2000-2017 avg	
	46	Mean temperature-September	2000-2017 avg	
	47	Mean temperature-December	2000-2017 avg	
Livestock	48	Buffaloes	2010	DOI: 10.7910/dvn/5u8mwi
	49	Cattle	2010	DOI: 10.7910/dvn/givq75
	50	Chickens	2010	DOI: 10.7910/dvn/sufasb
	51	Ducks	2010	DOI: 10.7910/dvn/ichcbh
	52	Goats	2010	DOI: 10.7910/dvn/ocph42

Geoscientific Model Development
Discussions
Open Access

53	Horses	2010	DOI: 10.7910/dvn/7q52mv
54	Pigs	2010	DOI: 10.7910/dvn/33n0jg
55	Sheep	2010	DOI: 10.7910/dvn/blwpzn

Note: N/A = not available; "avg" = average (because only average data were released)

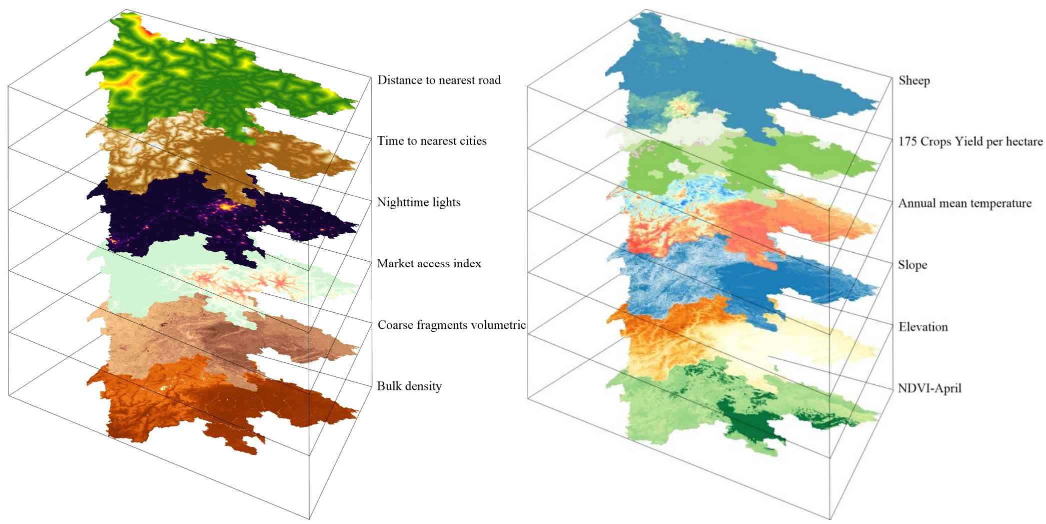


Figure 4 Some potential driving factors prepared for the study area of Sichuan

4.2 Establishment of many-to-many demand-supply relationships

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 many-to-many demand-supply relationships.

We present in this section a subtle and feasible method for the establishment, first qualitatively and then quantitatively. The qualitative establishment involves two steps: generating land systems based on scale transformations and defining the services of different land systems. The quantitative establishment is to quantify the land system-dependent supply (in 2010) of and the aggregated demand (in 2020) for each land system service.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

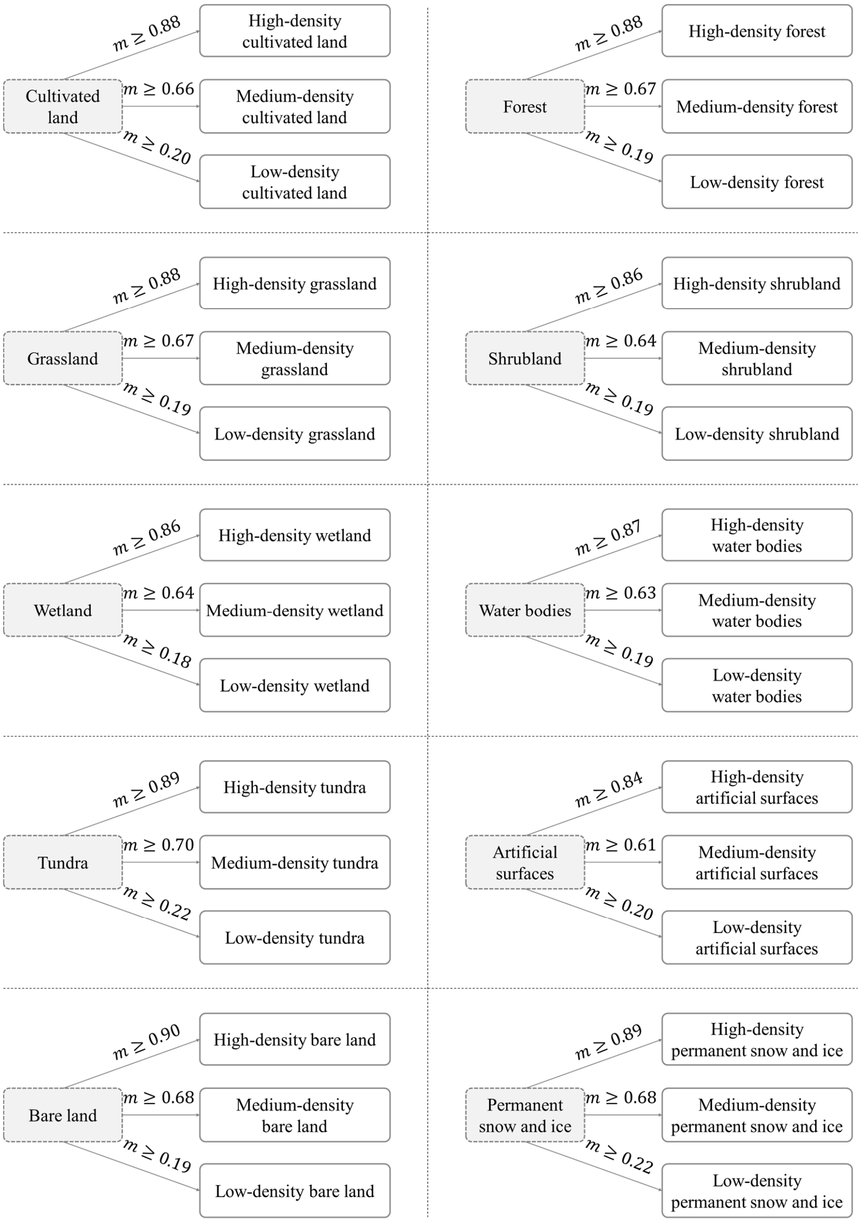
A well-designed taxonomy of land systems is the basis of many-to-many demand-supply relationships. To generate such land systems, we first upscaled the raw land use/cover data from the initial spatial resolution of 30 m to a coarser resolution of 990 m, by aggregating every 33 × 33 pixels (referred to as micro-pixels) of the raw data into new ones (referred to as macro-pixels). 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. 6. 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. 6. Finally, the land systems were slightly resampled to the spatial resolution of 1 km to match the resolution of potential driving factors.

Services of different land systems are difficult to define, especially for a comprehensive evaluation. However, our generation of land systems allows a subtle but easy definition as well as the quantification of supply and demand. Because our land systems were generated by transforming the spatial scale of GlobeLand30, a pixel of any land system (a macro-pixel) corresponds to many GlobeLand30 pixels (i.e., micro-pixels) with usually diverse types. Accordingly, our land system services were defined as the area of each of the ten types of GlobeLand30 land use/cover. Under this definition, each land system would become multifunctional to supply all ten services potentially.

To quantify the capability of each land system in supplying every service 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 data 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 2020 was calculated as the total area of the d-th type of pixel (i.e., micro-pixels) in the 2020 data of GlobeLand30.



335 **Figure 5 Taxonomy of land systems for our study areas (m denotes the threshold)**

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

4.3 Settings of other simulation parameters

In addition to the establishment of demand-supply relationships, there are some other parameters to 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 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 a 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 one is more correlated with others, we calculated the sum of SRCCs between the one potential driving factor and each of the others. 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 improved the independence of our samples.

The conversion resistance of each land system, $P\_res_{\mathrm{T}(c,t,0)}$, was determined using historical land system changes with Eqs. (4–6), where $h1 = 2010$ and $h2 = 2020$. The value of $\mathrm{Con}(\mathrm{T}(c,t,i),k)$ was set by checking whether these are conversions from the land system type of $\mathrm{T}(c,t,i)$ 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 1% of the study area, the value $\mathrm{Con}(\mathrm{T}(c,t,i),k)$ 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).

4.4 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" (Van Asselen and Verburg, 2013), 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; Chang et al., 2013) curves to access 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, namely Area Under the Curve (AUC, sometimes also referred to as the ROC value; Lin et al., 2011). 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., Hu and Lo, 2007; Mei et al., 2018; Jin et al., 2019), an AUC value of 0.7 or above means good fit, and that of 0.9 or above demonstrates excellent fit.

To evaluate the performance of land change simulation, we utilized one of the most popular and widely accepted measure called the standard Kappa index of agreement, also called Kappa statistic (Hagen, 2002) or Kappa index (Jiang et al., 2015). 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.

4.5 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 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.

Geoscientific Model Development
Discussions
Open Access

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

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

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' 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.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

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

ID	Land system type	AUC
0	Low-density cultivated land	0.664
1	Medium-density cultivated land	0.712
2	High-density cultivated land	0.782
3	Low-density forest	0.881
4	Medium-density forest	0.902
5	High-density forest	0.983
6	Low-density grassland	0.936
7	Medium-density grassland	0.956
8	High-density grassland	0.975
9	Low-density shrubland	0.998
10	Medium-density shrubland	0.999
11	High-density shrubland	0.999
12	Low-density wetland	0.948
13	Medium-density wetland	0.984
14	High-density wetland	1
15	Low-density water surfaces	0.963
16	Medium-density water surfaces	0.982
17	High-density water surfaces	0.998
18	Low-density artificial surfaces	0.831
19	Medium-density artificial surfaces	0.908
20	High-density artificial surfaces	0.986
21	Low-density bare land	0.992
22	Medium-density bare land	0.974

The evaluation results of our land change simulations in two study areas are shown in 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 0.6618 and 0.6812, respectively. Thus, the Kappa statistic obtained in the comparative experiment increased 2.93% in comparison with that in the benchmark experiment. For the study area of Henan, the Kappa statistic obtained in the benchmark experiment is only 0.4287, implying that the CLUMondo model in this case (without the proposed method for adaptively determining conversion orders) is much ineffective and intensive calibration is needed. By contrast, the Kappa statistic obtained in the comparative experiment witnessed a substantial increase of 103.36%, reaching 0.8718. This considerable increase demonstrates the effectiveness of the CLUMondo model in this case and the usefulness of the proposed method for adaptively determining conversion orders.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Table 5 Evaluation results of our land change simulations

Study area	Experiment	Kappa
Sichuan	Benchmark	0.6618
	Comparative	0.6812
Henan	Benchmark	0.4287
	Comparative	0.8718

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 universal applicability. The method is especially of use if the simulation performance of CLUMondo is poor.

5 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 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 25 with the study area of Sichuan (as Sichuan has 27 land systems), and it went from 1 to 21 with the study area of Henan (as Henan has 23 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 $S_{j,d} \in \Phi_{\kappa}$. With each set of conversion orders, we performed independent experiments and calculated the Kappa statistic.

The results are shown in Fig. 7. For both study areas, no simulation results (and Kappa statistics) 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 when the Kappa statistic was unavailable, e.g., if the number of classes was 6 for Sichuan or 10 for Henan. 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 Kappa statistic was available, we have the following two findings:

- For the study area of Sichuan, the Kappa statistic researched its highest value (i.e., 0.8771) with only two classes of conversion orders. Then, the Kappa statistic underwent a decreasing trend with more classes until the number of classes grow to 14. After that, the Kappa statistic has a sharp increase from 0.6867 to 0.8545. The increased value

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development Discussions
Open Access

remained even though the number of classes grew further. In comparison with all these values, the Kappa statistic obtained using our proposed method (i.e., 0.8718 as noted in the previous section) is among the highest ones.

- For the study area of Henan, most values of the Kappa statistic fell into the range of 0.6600 to 0.6700. Only two experiments produced a Kappa statistic greater than 0.6700, namely the experiment with three classes (Kappa = 0.6812) and that with four classes (Kappa = 0.6927). By contrast, the Kappa statistic obtained using our proposed method (i.e., 0.6812) is still among the highest ones.

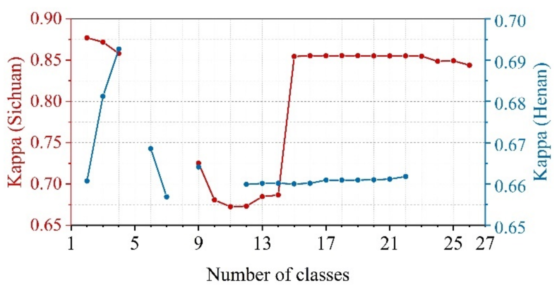


Figure 6 Kappa indices with every possible number of classes.

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). If a greater (or smaller) empirical value of GVF was adopted, the quality of simulation result will become worse (or better) with the study area of Sichuan but will get better (or worse) with the study area of Henan.

6 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 changes of land cover types but also land-use intensification (e.g., Van Asselen and Verburg, 2013; Jin et al., 2019). Therefore, it has found an increasing number of applications, where the simulation results serve as the basis of kinds of analysis. In this case, the effectiveness of CLUMondo is crucial and should be improved, if possible, but no improvement has been reported. The reason behind this is that although CLUMondo was initially developed in 2013 and open-sourced in 2016 (https://github.com/vueg/clumondo), its

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

central mechanism is complex and has not been fully revealed or clearly explained. All studies were focused on its applications; only Van Vliet and Verburg (2018) provided a very brief explanation of the central mechanism, some key details were not revealed, especially how CLUMondo balances demands and supplies in a many-to-many mode.

In this study, we first investigated the source code of CLUMondo, providing for the first time the complete, detailed mechanism of this model. 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, but the setting of this parameter should be improved. Specifically, 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, resulting CLUMondo Version 2.0. To demonstrate the validity and effectiveness of CLUMondo v2.0, 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. The results showed that Sichuan's simulation performance (AUC) was improved by 2.93% from 0.6618 to 0.6812, and that of Henan was improved substantially by 103.36% from 0.4287 to 0.8718. Further analysis showed that these improvements had achieved the highest level that can be expected by revising conversion orders. From these results, we made the following four conclusions:

- CLUMondo is valid and effective in simulating land system changes with many-to-many demand-supply relationships. Although more than 20 land systems were involved, the AUCs by CLUMondo (both the original and here improved versions) were high. Note that AUC decreases when the number of land systems increases;
- 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 (including the one by this study);
- 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;
- Our proposed automatic method for adaptively determining conversion orders is valid, highly effective, and universally applicable. The resultant new version of CLUMondo is more effective and easier to use than the existing version.

Since our proposed method has exploited the full potential of conversion orders, they would no longer be a reason for the possible poor performance of CLUMondo. Accordingly, future research is recommended to test the effects of other

500 parameters involved in the mechanism of CLUMondo disclosed in this study. Also, future applications are recommended to employ this new version of CLUMondo to improve simulation performance and reduce subjectivity in parameter settings.

Code and data availability. The source code for the model (i.e., CLUMondo v2.0) and all data used to produce the Sichuan 505 results present in this paper are archived on Zenodo (10.5281/zenodo.6594722), and that used to produce the Henan results present in this paper are also archived on Zenodo (10.5281/zenodo.6594815).

Author contributions. PG and CS designed the study. PG led the analysis of results and wrote the paper. YG developed the source code and performed the experiments. YG and SY analyzed the experimental results. XZ visualized some 510 experimental results.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Financial support. This research has been supported by the Second Tibetan Plateau Scientific Expedition and Research 515 Program of China (Grant No. 2019QZKK0608), Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA23100303), and National Natural Science Foundation of China (Grant No. 42171088).

References

Arunyawat, S. and Shrestha, R. P.: Simulating future land use and ecosystem services in Northern Thailand, J. Land Use Sci., 13, 146-165, 10.1080/1747423X.2018.1496157, 2018.

Bai, Y., Wong, C. P., Jiang, B., Hughes, A. C., Wang, M., and Wang, Q.: Developing China's Ecological Redline Policy using ecosystem services assessments for land use planning, Nature Communications, 9, 3034, 2018.

Borrelli, P., Robinson, D. A., Fleischer, L. R., Lugato, E., Ballabio, C., Alewell, C., Meusburger, K., Modugno, S., Schütt, B., and Ferro, V.: An assessment of the global impact of 21st century land use change on soil erosion, Nature Communications, 8, 1-13, 2017.

Calvin, K., Patel, P., Clarke, L., Asrar, G., Bond-Lamberty, B., Cui, R. Y. Y., Vittorio, A. D., Dorheim, K., Edmonds, J., and Hartin, C.: GCAM v5.1: representing the linkages between energy, water, land, climate, and economic systems, Geoscientific Model Development, 12, 677-698, 2019.

Chang, Y., Zhu, Z. L., Bu, R. C., Chen, H. W., Feng, Y. T., Li, Y. H., Hu, Y. M., and Wang, Z. C.: Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China, Landscape Ecology, 28, 1989-2004, 10.1007/s10980-013-9935-4, 2013.

Chen, J., Chen, J., Liao, A. P., Cao, X., Chen, L. J., Chen, X. H., He, C. Y., Han, G., Peng, S., and Lu, M.: Global land cover mapping at 30m resolution: A POK-based operational approach, Int. J. Photogramm. Remote Sens., 103, 7-27, 2015.

Cheng, C. X., Zhang, T., Su, K., Gao, P. C., and Shen, S.: Assessing the intensity of the population affected by a complex natural disaster using social media data, ISPRS International Journal of Geo-Information, 8, 358, 2019.

Dong, N., You, L., Cai, W. J., Li, G., and Lin, H.: Land use projections in China under global socioeconomic and emission scenarios: Utilizing a scenario-based land-use change assessment framework, Global Environmental Change, 50, 164-177, 2018.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., and Baptista, S. R.: Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4, Papers in Applied Geography, 1, 226-234, 2015.

Escobar, N. and Britz, W.: Metrics on the sustainability of region-specific bioplastics production, considering global land use change effects, Resources, Conservation and Recycling, 167, 105345, 2021.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol., 37, 4302-4315, 2017.

Gao, P. C., Xie, Y. R., Song, C. Q., Cheng, C. X., and 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. Geog. Sci., in press, 2022.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Grundy, M. J., Bryan, B. A., Nolan, M., Battaglia, M., Hatfield-Dodds, S., Connor, J. D., and Keating, B. A.: Scenarios for Australian agricultural production and land use to 2050, Agricultural Systems, 142, 70-83, 2016.

Hagen, A.: Multi-method assessment of map similarity, Proceedings of the 5th AGILE Conference on Geographic Information Science, 171-182,

He, C. Y., Li, J. W., Zhang, X. L., Liu, Z. F., and 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, Journal of Cleaner Production, 165, 57-69, 2017.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., and Bauer-Marschallinger, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS One, 12, e0169748, 2017.

Hu, Z. Y. and Lo, C. P.: Modeling urban growth in Atlanta using logistic regression, Computers, Environment and Urban Systems, 31, 667-688, https://doi.org/10.1016/j.compenvurbsys.2006.11.001, 2007.

Jenks, G. F.: The data model concept in statistical mapping, International Yearbook of Cartography, 7, 186-190, 1967.

Jenks, G. F. and Caspall, F. C.: Error on choroplethic maps: definition, measurement, reduction, Annals of the Association of American Geographers, 61, 217-244, 1971.

Jiang, W. G., Chen, Z., Lei, X., Jia, K., and Wu, Y. F.: Simulating urban land use change by incorporating an autologistic regression model into a CLUE-S model, J. Geog. Sci., 25, 836-850, 10.1007/s11442-015-1205-8, 2015.

Jin, X. L., Jiang, P. H., Ma, D. X., and Li, M. C.: Land system evolution of Qinghai-Tibetan Plateau under various development strategies, Appl. Geog., 104, 1-9, 2019.

Kong, X. S., Zhou, Z. Z., and Jiao, L. M.: Hotspots of land-use change in global biodiversity hotspots, Resources, Conservation and Recycling, 174, 105770, 2021.

Kummu, M., Taka, M., and Guillaume, J. H. A.: Gridded global datasets for Gross Domestic Product and Human Development Index over 1990–2015, Scientific Data, 5, 180004, 10.1038/sdata.2018.4, 2018.

Lin, Y. P., Chu, H. J., Wu, C. F., and Verburg, P. H.: Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling – a case study, International Journal of Geographical Information Science, 25, 65-87, 10.1080/13658811003752332, 2011.

Liu, X. P., Liang, X., Li, X., Xu, X. C., Ou, J. P., Chen, Y. M., Li, S. Y., Wang, S. J., and Pei, F. S.: A future land use simulation model (FLUS) for simulating multiple land use scenarios by coupling human and natural effects, Landscape and Urban Planning, 168, 94-116, 2017a.

Liu, Y., Goodchild, M. F., Guo, Q., Tian, Y., and Wu, L.: Towards a general field model and its order in GIS, International Journal of Geographical Information Science, 22, 623-643, 2008.

Liu, Z. F., Verburg, P. H., Wu, J. G., and He, C. Y.: Understanding Land System Change Through Scenario-Based Simulations: A Case Study from the Drylands in Northern China, Environmental Management, 59, 440-454, 10.1007/s00267-016-0802-3, 2017b.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Geoscientific Model Development
Discussions

Long, Y., Song, Y. M., and Chen, L.: Identifying subcenters with a nonparametric method and ubiquitous point-of-interest data: A case study of 284 Chinese cities, Environment and Planning B: Urban Analytics and City Science, 49, 58-75, 10.1177/2399808321996705, 2022.

Mei, Z. X., Wu, H., and 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, Frontiers of Earth Science, 12, 299-310, 10.1007/s11707-017-0639-y, 2018.

Monfreda, C., Ramankutty, N., and Foley, J. A.: Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, Global Biogeochem. Cycles, 22, 2008.

Announcement of Statistics on Grain Production: http://www.stats.gov.cn/xxgk/sjfb/zxfb2020/202112/t20211206_1825071.html, last access: 13 April.

National Bureau of Statistics of China: China Statistical Yearbook, China Statistics Press, Beijing, China2021b.

Nie, X., Lu, B., Chen, Z. P., Yang, Y. W., Chen, S., Chen, Z. H., and Wang, H.: 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, Ecological Indicators, 118, 106708, 2020.

O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., Mathur, R., and van Vuuren, D. P.: A new scenario framework for climate change research: the concept of shared socioeconomic pathways, Clim. Change, 122, 387-400, 2014.

Schandl, H., Lu, Y. Y., Che, N., Newth, D., West, J., Frank, S., Obersteiner, M., Rendall, A., and Hatfield-Dodds, S.: Shared socio-economic pathways and their implications for global materials use, Resources, Conservation and Recycling, 160, 104866, 2020.

Song, X. P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F., and Townshend, J. R.: Global land change from 1982 to 2016, Nature, 560, 639-643, 2018.

United Nations: World Population Prospects, 2019.

van Asselen, S. and Verburg, P. H.: A land system representation for global assessments and land-use modeling, Global Change Biology, 18, 3125-3148, 2012.

van Asselen, S. and Verburg, P. H.: Land cover change or land-use intensification: simulating land system change with a global-scale land change model, Global Change Biology, 19, 3648-3667, https://doi.org/10.1111/gcb.12331, 2013.

van Vliet, J. and Verburg, P. H.: A short presentation of CLUMondo, in: Geomatic Approaches for Modeling Land Change Scenarios, edited by: Camacho Olmedo, M. T., Paegelow, M., Mas, J.-F., and Escobar, F., Springer International Publishing, Cham, 485-492, 10.1007/978-3-319-60801-3_34, 2018.

Verburg, P. H., Ellis, E. C., and Letourneau, A.: A global assessment of market accessibility and market influence for global environmental change studies, Environmental Research Letters, 6, 034019, 2011.

Verburg, P. H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., and Mastura, S. S.: Modeling the spatial dynamics of regional land use: the CLUE-S model, Environmental Management, 30, 391-405, 2002.

https://doi.org/10.5194/gmd-2022-123
Preprint. Discussion started: 2 June 2022

Wang, S. H., Zhang, Y. G., Ju, W. M., Qiu, B., and 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. Tot. Environ., 755, 142569, 2021a.

Wang, S. M., Ma, Q. F., Ding, H. Y., and Liang, H. W.: Detection of urban expansion and land surface temperature change using multi-temporal landsat images, Resources, Conservation and Recycling, 128, 526-534, 2018.

Wang, Y., van Vliet, J., Pu, L. J., and Verburg, P. H.: Modeling different urban change trajectories and their trade-offs with food production in Jiangsu Province, China, Computers, Environment and Urban Systems, 77, 101355, https://doi.org/10.1016/j.compenvurbsys.2019.101355, 2019.

Wang, Y., van Vliet, J., Debonne, N., Pu, L. J., and Verburg, P. H.: Settlement changes after peak population: Land system projections for China until 2050, Landscape and Urban Planning, 209, 104045, https://doi.org/10.1016/j.landurbplan.2021.104045, 2021b.

Weiss, D., Nelson, A., Vargas-Ruiz, C., Gligorić, K., Bavadekar, S., Gabrilovich, E., Bertozzi-Villa, A., Rozier, J., Gibson, H., and Shekel, T.: Global maps of travel time to healthcare facilities, Nature Medicine, 26, 1835-1838, 2020.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S., Kang, S. Y., Cameron, E., Bisanzio, D., Battle, K. E., Bhatt, S., and Gething, P. W.: A global map of travel time to cities to assess inequalities in accessibility in 2015, Nature, 553, 333-336, 10.1038/nature25181, 2018.

Young, A.: Land Resources: Now and for the Future, Cambridge University Press, Cambridge, UK2000.

Zhang, R. S. and Hanaoka, T.: Deployment of electric vehicles in China to meet the carbon neutral target by 2060: Provincial disparities in energy systems, CO2 emissions, and cost effectiveness, Resources, Conservation and Recycling, 170, 105622, 2021.