Di Vittorio, A V, et al. 2020 Moirai Version 3: A Data Processing System to Generate Recent Historical Land Inputs for Global Modeling Applications at Various Scales. *Journal of Open Research Software,* 8: 1. DOI: https://doi.org/10.5334/jors.266

SOFTWARE METAPAPER

Moirai Version 3: A Data Processing System to Generate Recent Historical Land Inputs for Global Modeling Applications at Various Scales

Alan V. Di Vittorio[1], Chris R. Vernon[2] and Shijie Shu[3]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, US

[2] Pacific Northwest National Laboratory, Richland, WA, US

[3] University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, US

Corresponding author: Alan V. Di Vittorio (avdivittorio@lbl.gov)

The Moirai land data system, written in C and R, is designed to produce recent historical land inputs for an integrated human-Earth systems model. The primary function of Moirai is to combine spatially explicit input data (e.g., raster images) with tabular input data (e.g., crop price table) to generate spatially-referenced tabular data of crop production, crop harvested area, land value, irrigated and rainfed crop area, water footprint, soil and vegetation carbon density of unmanaged land, and historical land use/cover. These data are aggregated to user-defined geographic boundaries within 231 countries and the default boundaries are defined globally by 235 watersheds. The production, harvested area, and land value outputs reconstruct those available from the Global Trade Analysis Project, while the other outputs provide additional information for various applications, such as initializing or evaluating a land use change model or an economic general/partial equilibrium model. Furthermore, Moirai is a modular system that can be updated and customized through replacement and addition of source data.

Keywords: moirai; land cover; land use; model; data; system

(1) Overview

Introduction

Economic models of agriculture and land use require a wide variety of data for initialization and calibration, but the availability, format, relevance, and spatial delineation are limited. As a result, for the past 17 years most general and partial equilibrium models have relied on Global Trade Analysis Project (GTAP) land data [e.g., 1], which provides tabular data of crop production, harvested area, and land value for 226 countries, delineated by up to 18 Agro-Ecological Zones (AEZs). The GTAP land use data cover 175 crops and have been updated three times since 2001 to reference four different years of Food and Agricultural Organization (FAO) data (2001, 2004, 2007, and 2011), but with the exception of year 2001 [1–4] these data must be purchased. Model development during this time has had to either align with this particular spatial configuration of the data or implement complex ways of mapping these data to native grids. Recently, Di Vittorio et al. [5] have shown that AEZs do not necessarily confer the desired spatial benefits of homogeneity and are not compatible with watershed-based modules that are being used to provide water runoff and accessibility to multi-sectoral models.

The Moirai land data system (Moirai) is designed to provide open-source land data corresponding and additional to GTAP data, with the flexibility and modularity to customize spatial delineation and update data sources commonly used by the Global Change Assessment Model (GCAM) [6]. As a result, Moirai can provide customizable outputs for a variety of models and applications. Moirai is written in C for speed and efficiency. A single run with 235 water basins, 231 countries, and no output-year recalibration takes under one hour for most modern computing platforms. With input data on the order of 100GB, the maximum runtime memory usage is on the order of 2GB. The diagnostic scripts are written in R for ease of use.

Implementation and architecture

Moirai is a land data integration system, so we first present the output products to facilitate understanding of the subsequent sections on data sources and processing. There is a single input file to Moirai that includes:

- A diagnostic output flag
- An output recalibration year for crop production, harvested area, and land rent
- An output USD recalibration year for land rent

- The year associated with the detailed input crop data
- The USD year associated with the input land rent data
- Data paths and file names for all input and output data, including a runtime log file.

Each of the inputs and outputs described below and in **Table 1** (raster inputs) and **Table 2** (text inputs) have a corresponding name or data path in the Moirai input file (see functioning examples in the "input_files" folder). The other four entries in the input file are 1) general input data path for data that does not have a specified input path, 2) general output data path for all outputs, 3) path to which the eight primary data output files will be copied, and 4) path to which the two primary mapping output files will be copied. Four R diagnostics scripts are located in a "diagnostics" folder and the path to their outputs is specified in each script, with the default being a specific folder in the diagnostics folder.

Table 1: Raster inputs (these have been accessed in June 2018 unless noted otherwise).

Data	Details	Source
Crop yield and harvested area	5 arcmin, 175 crops-same as GTAP, circa 2000, area provided as fraction of land area in grid cell	Monfreda et al., 2008; http://www.earthstat.org/data-download/
Cropland physical extent	5 arcmin, circa 2000, provided as fraction of land area in grid cell	Ramankutty et al., 2008; http://www.earthstat.org/data-download/
Irrigated and rainfed crop harvested area	5 arcmin, hectares, 26 crop classes, circa 2000	Portmann et al., 2010; https://www.uni-frankfurt.de/45218031/data_download/
Crop water footprint data	5 arcmin, mm/yr, 18 crop types, 3 water types, circa 2000	Mekonnen and Hoekstra, 2011; http://waterfootprint.org/en/resources/waterstat/product-water-footprint-statistics/
Fraction of land area in grid cell for crop and water footprint data above	5 arcmin, spherical earth with WGS84 mean radius	D. Plouff and N. Ramankutty provided these data corresponding to the above cropland data (in late 2013). Note that these are the same data used to provide the area values in the current crop yield and harvested area data above
Potential vegetation	5 arcmin, thematic, 15 vegetation types, circa 2000 if no historical land use had occurred	Ramankutty and Foley, 1999; http://www.earthstat.org/data-download/
Land use area	5 arcmin, km^2, 1700-2016 (decadal up to 2000), HYDE 3.2.000 baseline, 12 land use types	Klein Goldewijk et al., 2017; ftp://ftp.pbl.nl/hyde/hyde3.2/2017_beta_release/
Land area in grid cell	5 arcmin, km^2, circa 2000, spherical earth with WGS84 mean radius, with Greenland and several arctic islands added based on fraction of land area in grid cell for crop area and potential vegetation; this is the working grid	Klein Goldewijk et al., 2017; ftp://ftp.pbl.nl/hyde/hyde3.2/2017_beta_release/
Total grid cell area	5 arcmin, km^2, spherical earth with WGS84 mean radius, with Greenland and several arctic islands added based on fraction of land area in grid cell for crop area and potential vegetation; this is the working grid	Klein Goldewijk et al., 2017; ftp://ftp.pbl.nl/hyde/hyde3.2/2017_beta_release/
234 Country boundaries	5 arcmin, from VMAP0 vector data (the source of FAO country boundaries), added East Timor based on a map, and merged some countries to reflect FAO data	VMAP0: http://gis.ess.washington.edu/data/raster/GlobalData/ (last accessed in 2013, now restricted to UW, but these data are currently available in four parts at http://gis-lab.info/qa/vmap0-eng.html);

(Contd.)

Data	Details	Source
Original AEZ boundaries	5 arcmin, 1961–1990 data, 160 country boundaries, GTAP Land Use Database, Release 2.1	Lee et al., 2005; Lee et al., 2009; Monfreda et al., 2009; https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900
Output Geographic Land Unit (GLU) boundaries	5 arcmin, thematic, 235 water basins	Developed for the water module of the Global Change Assessment Model, aggregated from a 1/8-degree global watershed data set

Land cover area data	half-degree,
	1800–2016 (decadal up to 2000),
	23 land cover types,
	fraction of grid cell and grid cell area
		Produced specifically for Moirai using HYDE 3.2.000 data;
		http://climate.atmos.uiuc.edu/atuljain/availabledata.html;
		Previous public version available here: https://www.atmos.illinois.edu/~meiyapp2/datasets.htm (Meiyappan and Jain, 2012)

Protected land area	5 arcmin, thematic, protected or not-protected	Derived from a previous version of the World Database on Protected Areas; (current version available at: https://www.iucn.org/theme/protected-areas/our-work/world-database-protected-areas)

Outputs

Moirai takes a given set of Geographic Land Unit (GLU) boundaries and generates 10 production output files for use by a global land use model, plus a runtime log file. Further processing of these tabular text files may be necessary for a given application, such as performed by the GCAM data system, but since the information has historically been shared in text tables, many users may already have a basis for processing Moirai outputs. Eight of the production output files contain data and two contain mapping values between countries or land types. Moirai also outputs five raster files containing the final country, region, and GLU boundaries as determined by the intersection of the input data. If specified in the input file, Moirai generates many diagnostic files including raster outputs of land use and land cover, some of which are needed by the R diagnostic scripts.

The three primary outputs are 1) crop harvested area and 2) production for 175 crops at the level of each GLU within each country, and 3) land rent value for 12 land types at the level of each GLU within each economic region. The content of these files reconstructs that of the GTAP land use database, and the output file names are specified in the input file. The default output year for crop data is circa 2000, with land rent in year-2001 US dollars. The user can independently specify the output years for crop (1995–2014) and USD-year recalibration (1970–2017). The default spatial boundaries are 235 water basins as GLUs, 231 countries, and 87 economic regions defined by GTAP, with each of these specified as an input to Moirai.

The five remaining output data files provide (at the level of each GLU within each country) 1) irrigated and 2) rainfed harvested area for 26 crop classes (circa 2000), 3) land type area for 47 years from 1700 to 2016, including land use and the land type prior to conversion, 4) soil and vegetation carbon density for unmanaged land types, and 5) water use footprint of 18 crops for three water types (circa 2000).

The two mapping files are cross-reference tables that 1) map spatial country codes to iso3 codes and FAO country names for each GLU within each country, and 2) map land type codes present in the area and carbon density outputs to reference vegetation, land use, and protected status.

Source data

The source data are all publicly available and are included with the Moirai distribution (see section (2) for distribution details). Raster and text input data names, details, and sources can be viewed in the docs section of the Moirai GitHub repository. Some of these data have undergone pre-processing to add relevant information, (dis)aggregate data, or change the data format (e.g., from vector to raster data). Each data set is a specified input to Moirai and as such can be updated by the user, although this may require modification of the read-in code if the data format is different.

In most cases the primary customization will be GLU boundaries, which are defined by one raster file and one text file and do not require code modification if generated in the following format. The GLU data are integers that thematically assign each pixel to a single GLU, with ocean assigned the no-data value of –9999. There are separate land area input data that determine which pixels are land pixels, but the substitute GLU data can also include other water bodies with no-data values. The raster file is a single-band binary file with 4-byte signed integer values, no header, and 5 arcmin resolution. It uses the WGS84 geographic datum with no projection, and the first value in the file is the pixel with upper left corner at –180 degrees longitude and 90 degrees latitude. The values are stored in order of ascending longitude in each descending latitude row, with longitude varying faster. The mapping of the thematic values to names is defined in the GLU text file. This is a comma-separated-value file with two columns and one header line. The first column contains the GLU integer, the second column contains the GLU name, and the header text is not used.

Table 2: Text inputs as comma separated value files (these have been accessed in June 2018 unless noted otherwise).

Data	Details	Source[a]
87 economic regions	Tabular, GTAP Land Use Database, Release 2.1	Lee et al., 2005; Lee et al., 2009; Monfreda et al., 2009; https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900
Land rent for 13 use sectors	Tabular, 87 regions by 18 AEZs, GTAP Land Use Database, Release 2.1	Lee et al., 2005; Lee et al., 2009; Monfreda et al., 2009; https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900
FAO 235 countries	Tabular, with some added countries to match the VMAP0 data, there are two input files containing these data: one maps the countries to the economic regions and the other maps the countries to the raster country data	http://faostat.fao.org/; accessed Aug 2013
Geographic Land Unit (GLU) list	Thematic codes and names for the GLU raster data	Developed for the water module of the Global Change Assessment Model, aggregated from a 1/8-degree global watershed data set
GCAM region list	Names and integer codes for GCAM regions, used in some diagnostics	GCAM
Country to GCAM region mapping	Cross-reference table mapping FAO countries to GCAM regions	FAO country and GCAM region data
HYDE3.2.000 list	Thematic codes and names for HYDE3.2.000 raster data	Klein Goldewijk et al., 2017; ftp://ftp.pbl.nl/hyde/hyde3.2/2017_beta_release/
Land cover to land use mapping	Cross-reference table mapping land cover to potential vegetation and land use	Land cover, land use, and potential vegetation data
GTAP product use list	Tabular, Codes and names for 13 GTAP use sectors	Lee et al., 2005; Lee et al., 2009; Monfreda et al., 2009; https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900
175 crop to FAO crop and GTAP use mapping	Cross-reference table mapping 175 crops to FAO crops to GTAP use	175 Crop, FAO crop, and GTAP use data
FAO production data	Tabular, up to 169 crops, 1997–2007	http://faostat.fao.org/faostat/; accessed July 2018
FAO yield data	Tabular, up to 160 crops, 1997–2007, for diagnostics only	http://faostat.fao.org/faostat/; accessed July 2018
FAO harvested area data	Tabular, up to 161 crops, 1997–2007, for year recalibration only	http://faostat.fao.org/faostat/; accessed July 2018
FAO Crop producer prices data	Tabular, up to 205 crops, 1997–2007	http://faostat.fao.org/faostat/; accessed July 2018
USD-year conversion list	Factors to convert input FAO 2005 USD to output 2001 USD	Derived from consumer price index centered at 1982–1984 (https://www.bls.gov/cpi/data.htm)
Potential vegetation list	Thematic codes and names for potential vegetation raster data	Ramankutty and Foley, 1999; http://www.earthstat.org/data-download/
Vegetation carbon density for potential vegetation	Average vegetation carbon densities for the potential vegetation types	Literature review
Soil carbon density for potential vegetation	Average soil carbon densities for the potential vegetation types	Literature review

The 1800 to 2016 land cover area data have been generated specifically for Moirai. They are produced by combining the crop and pasture area from the History Database of the Global Environment (HYDE 3.2) [7] with a map of potential vegetation representing preindustrial land cover in the absence of human activities [8]. The HYDE data are linearly interpolated to annual values. The land cover at year 1800 is used as the starting point and the changed land area of each land cover type are calculated and then attributed to 92 possible land use transitional types based on the algorithm described by Meiyappan and Jain [8], with recent updates to incorporate more information. We then overlap the land use transition onto the land cover data to generate the land cover fraction for the next year and continue the process through 2016. The final land cover product contains the 28 different land cover types in **Table 1** of Meiyaappan and Jain [9], plus 6 managed land cover types (rice, crop without rice, intensively managed pasture, rangeland, irrigated crop and irrigated rice) that provide additional detail for the crop and pasture land cover types. Moirai uses 32 of these input land types, decadal data prior to 2000, and annual data from 2000 forward.

Data processing

While the development of Moirai's core has been presented by [5], Moirai itself is an updated version that provides additional outputs, an updated spatial foundation, a new land cover basis, and more flexibility in terms of desired GLUs and output years. Therefore, we provide an updated description of Moirai's implementation here.

The bulk of the data processing reconciles inconsistencies among all the input data sets (**Figure 1**). The final outputs are based on data that are in a valid land cell, as determined

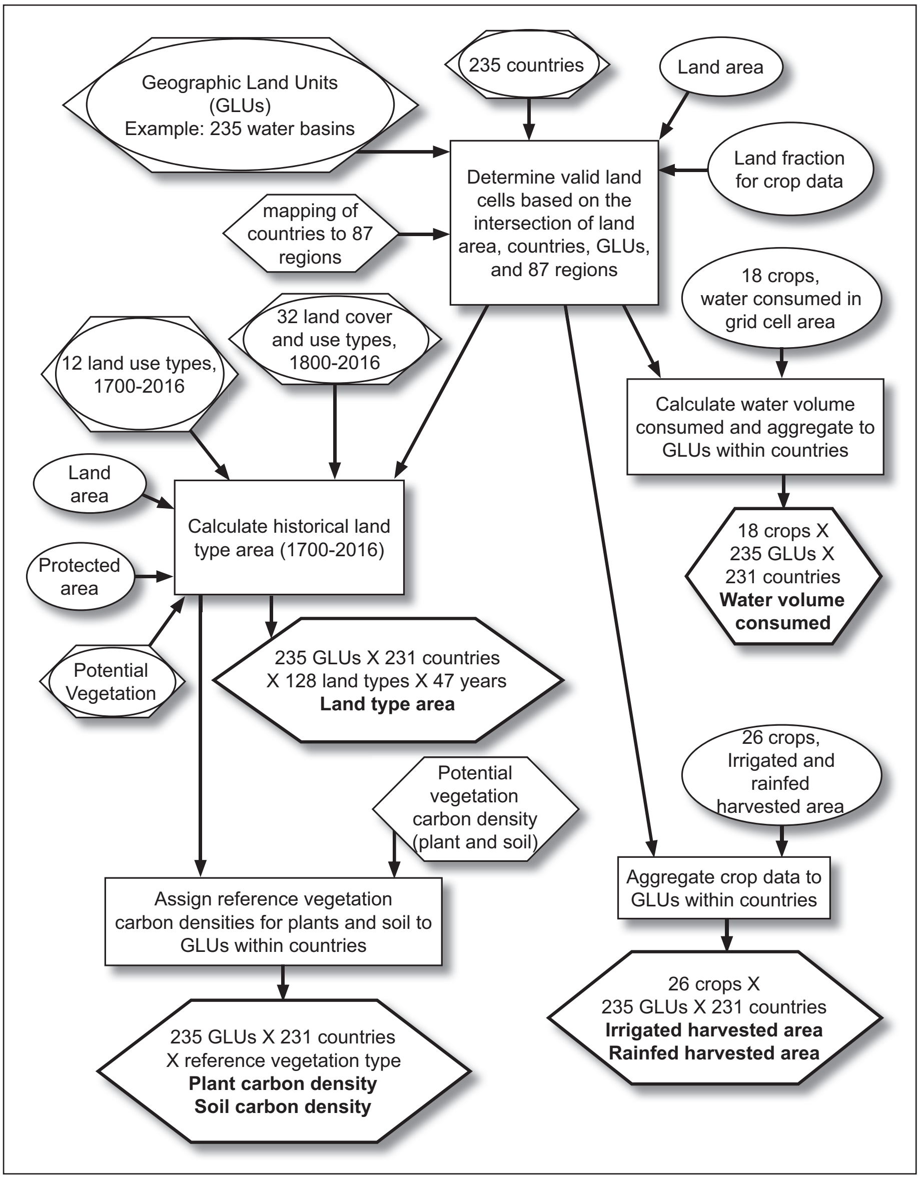
Geographic Land Units (GLUs) Example: 235 water basins
235 countries
Land area
Land fraction for crop data
Determine valid land cells based on the intersection of land area, countries, GLUs, and 87 regions
mapping of countries to 87 regions
18 crops, water consumed in grid cell area
32 land cover and use types, 1800-2016
12 land use types, 1700-2016
Calculate water volume consumed and aggregate to GLUs within countries
Land area
Calculate historical land type area (1700-2016)
18 crops X 235 GLUs X 231 countries Water volume consumed
Protected area
Potential Vegetation
235 GLUs X 231 countries X 128 land types X 47 years Land type area
26 crops, Irrigated and rainfed harvested area
Potential vegetation carbon density (plant and soil)
Aggregate crop data to GLUs within countries
Assign reference vegetation carbon densities for plants and soil to GLUs within countries
26 crops X 235 GLUs X 231 countries Irrigated harvested area Rainfed harvested area
235 GLUs X 231 countries X reference vegetation type Plant carbon density Soil carbon density

Figure 1: Spatial basis of the Moirai land data system and pathways to five outputs (the irrigated and rainfed data are output as separate files). The Geographic Land Units (GLUs) are defined by the user. Ovals are raster data (all at 5 arcmin except for the land cover/use data, which are at half-degree), hexagons are tabular data, boxes are processes, and the outputs are in underlined, bold lettering within bold hexagons.

by the intersection of the HYDE-based land area, the FAO country boundaries, the GLU boundaries, and whether a country is assigned to one of the 87 GTAP economic regions [2, 3]. Input data that cannot be associated with a valid land cell are not used. The historical land cover data have been generated using the same version of HYDE data that is currently input to Moirai [7]. These data provide a coarse-resolution merging of land cover and land use every 10 years from 1800 to 2000 and every year from 2001 to 2016. These data are spatially disaggregated and are used to determine the reference, or non-managed, vegetation in each land cell, with gaps filled by the potential vegetation data [8]. The land cover types are aggregated to the potential vegetation types, which are used as the reference vegetation types. The proportions of reference vegetation and various land uses within each land cell at each of 47 years from 1700–2016 are then reconciled with the finer-resolution HYDE land area and use data. The year 1800 land cover data are used with the HYDE data for the 10 decadal years from 1700–1790. If the land area is less than the total crop, pasture, and urban area, then first urban, then pasture, and then crop area are reduced until their total matches the land area. These data for the year 2000 are used to reconcile the data that share the spatial and temporal bases of the cropland extent data [10], including the 175 crop data [4, 11], irrigated/rainfed crops [12], and water footprint data [13], through normalization by cropland extent and conversion to the spatial basis defined above. The raster data are mapped to the tabular countries and regions for further processing of the data, causing the raster data to be aggregated in some cases (e.g., Serbia and Montenegro are merged to match historical data) and omitted in others (e.g., Antarctica), resulting in 231 output countries.

The five remaining outputs are generated directly by aggregating finer-resolution data to the specified GLUs within each country, based on the pixel-level spatial reconciliation described above (**Figure 1**). The output land type area data quantify the spatial basis for 47 years from 1700–2016 and represent the completely reconciled land surface as it changes over time. The output vegetation and soil carbon densities are area-weighted averages for each reference vegetation type within each GLU in each country. There are 26 crop classes in the input rainfed/irrigated data and their respective annual harvested areas circa 2000 are summed to the intersection of GLUs and countries. The annual water footprint data include 18 specific crops that used at least 67% of the annual crop water use circa 2000, and these data are output as water use volume for each GLU within each country.

Once the common spatial basis has been established, the crop production and harvested area for each crop in each GLU in each country are determined from the gridded 175-crop yield and harvested area data (**Figure 2**). Note that not every crop or GLU are present in each country. Harvested area and production (harvested area multiplied by yield) are simply summed across all grid cells within a given boundary because the source data have already been normalized to 1997–2003 average annual FAO data to obtain the circa 2000 output year [4, 11].

However, the user can specify recalibration of these input data to a different average FAO data year. This recalibration affects only the crop production and harvested area and non-forestry land rent outputs. The recalibration is based on a 5-year average of FAO country-level data centered on the specified year and is performed for each land cell and crop. The harvested area is recalibrated first as:

$$H_{p,i} = H'_{p,i}\left[H^{FAO}_{l,i,t} \Big/ \sum_{p\in l} H'_{p,i}\right], \tag{1}$$

where $H_{p,i}$ is the new harvested area for pixel p and crop i, $H'_{p,i}$ is the input harvested area for pixel p and crop i, and $H^{FAO}_{l,i,t}$ is the average FAO harvested area for country l (that includes pixel p) and crop i centered at year t. The yield is then recalibrated using the new harvested area ($H_{p,i}$) from equation (1) and the same production data used for land rent calculations below:

$$Y_{p,i} = Y'_{p,i}\left[Q^{FAO}_{l,i,t} \Big/ \sum_{p\in l} Y'_{p,i} H_{p,i}\right], \tag{2}$$

where $Y_{p,i}$ is the new harvested area for pixel p and crop i, $Y'_{p,i}$ is the input harvested area for pixel p and crop i, and $Q^{FAO}_{l,i,t}$ is the average FAO production for country l (that includes pixel p) and crop i centered at year t.

The agricultural and forestry land rent data are generated separately at the level of 87 regions (**Figure 2**). There are 12 agricultural sectors, including three livestock sectors for meat, dairy, and fiber production, and a non-ruminant livestock sector that has zero land rent value because it is assumed to not use substantial area in production. For each region, the eight crop sector land rents are distributed among the GLUs following [1] and [2]:

$$L_{c,g} = L_c\left[\sum_{i\in SECTOR=c} P^{FAO}_i Q_{i,g} \Big/ \sum_{g\in GLU}\ \sum_{i\in SECTOR=c} P^{FAO}_i Q_{i,g}\right], \tag{3}$$

where $L_{c,g}$ is the land rent for sector c in GLU g (USD), L_c is the original total land rent of sector c within a region, P^{FAO}_i is the FAO production-weighted annual average price per metric ton of crop i, $Q_{i,a}$ is the output production for crop i in GLU g (metric tons, from above), *SECTOR* is the set of GTAP sectors within a region, and *GLU* is the set of GLUs within a region. The default USD-year is 2001 and the default averaging years for price and production are 1997–2003, but the user can change these two defaults independently, with the price and production averages corresponding to the recalibration years as described above.

The livestock sector land rent calculation for each region uses the average cereal grain sector price and yield along with the pasture area, and the proportion assigned to each GLU is the same for each livestock sector. Using the production-weighted price and the area-weighted yield (from output production divided by output harvested area above) essentially scales the value term in eq. (3) by the ratio of pasture area to total harvested grain area within a GLU:

$$L_{l,g} = L_l\left[\frac{A_{pasture,g}}{\sum\limits_{i\in SECTOR=grain} A_{i,g}} \sum_{i\in SECTOR=grain} P_i^{FAO} Q_{i,g} \Bigg/ \sum_{g\in GLU}\left(\frac{A_{pasture,g}}{\sum\limits_{i\in SECTOR=grain} A_{i,g}} \sum_{i\in SECTOR=grain} P_i^{FAO} Q_{i,g}\right)\right], \qquad (4)$$

where $L_{l,g}$ is the land rent for livestock sector l in GLU g (USD), L_l is the original total land rent of livestock sector l, $A_{pasture,g}$ is the pasture area in GLU g, $A_{i,g}$ is the harvested area of crop i in GLU g, *grain* denotes the grain sector, and P_i^{FAO}, $Q_{i,g}$, *SECTOR*, and *GLU* are the same as in eq. (3).

The forestry land rent within each region is distributed among new GLUs based on the original land rent per forest area within the original GLUs. First, the original land rent per unit forest area is calculated for each original GLU and region. Next, the new GLU forestry land rent is calculated as the sum of the product of original land rent per area and forest area within the new GLU:

$$L_{f,new} = \sum_{g\in GLU_{Orig}\cap new} L_{f,g}^{Orig} A_{f,g\cap new}, \qquad (5)$$

where $L_{f,new}$ is the forest sector f land rent for GLU *new*, $L_{f,g}^{Orig}$ is the forest land rent per unit area in original GLU g, $A_{f,g\cap new}$ is the forest area of original GLU g within new GLU *new*, and GLU_{orig} is the set of original GLUs within the region. Note that the orginal set of GLUs is the set of original AEZs associated with the GTAP land rent data.

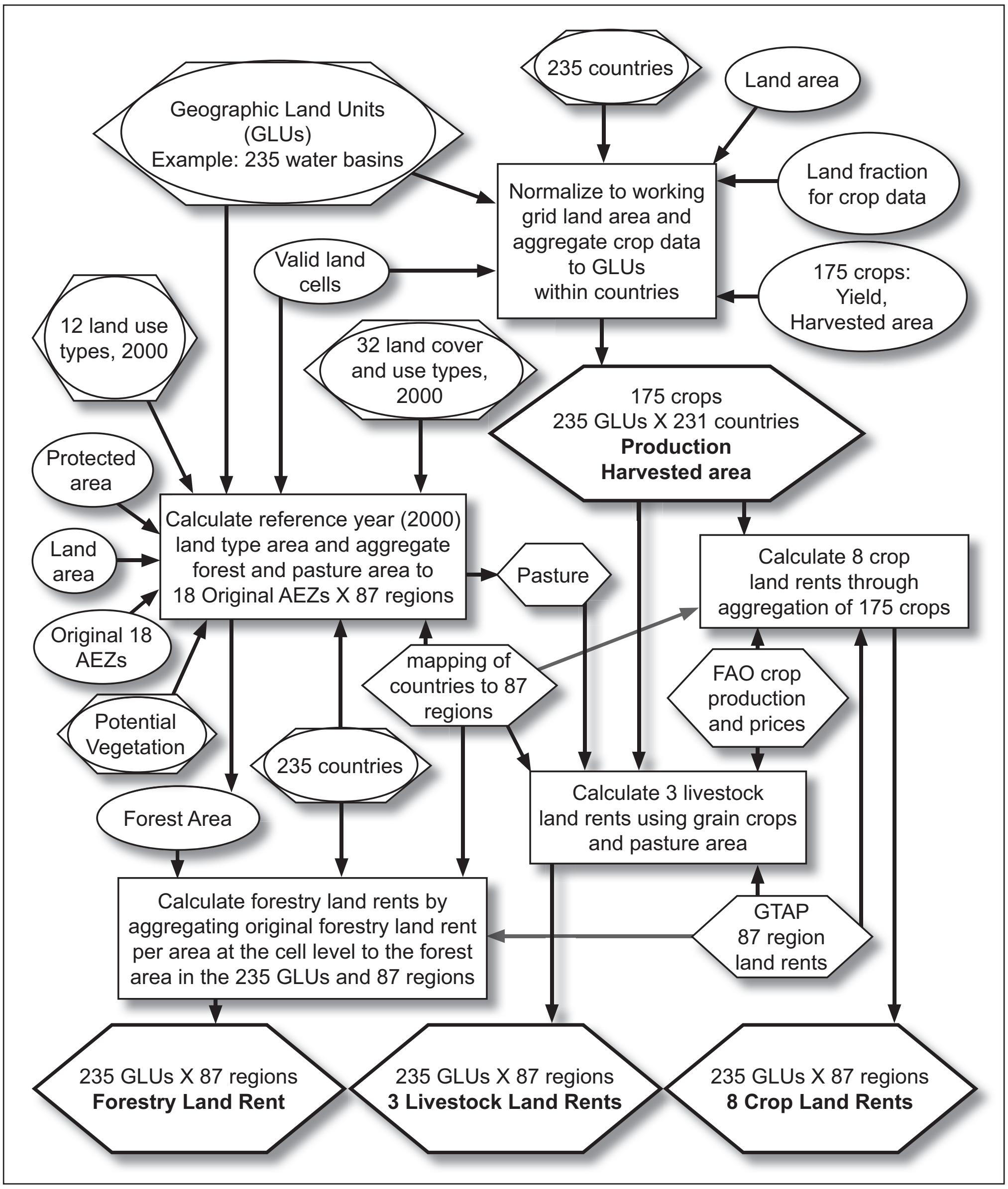
235 countries
Land area
Geographic Land Units (GLUs) Example: 235 water basins
Normalize to working grid land area and aggregate crop data to GLUs within countries
Land fraction for crop data
Valid land cells
175 crops: Yield, Harvested area
12 land use types, 2000
32 land cover and use types, 2000
175 crops 235 GLUs X 231 countries Production Harvested area
Protected area
Calculate reference year (2000) land type area and aggregate forest and pasture area to 18 Original AEZs X 87 regions
Land area
Pasture
Calculate 8 crop land rents through aggregation of 175 crops
Original 18 AEZs
mapping of countries to 87 regions
FAO crop production and prices
Potential Vegetation
235 countries
Forest Area
Calculate 3 livestock land rents using grain crops and pasture area
Calculate forestry land rents by aggregating original forestry land rent per area at the cell level to the forest area in the 235 GLUs and 87 regions
GTAP 87 region land rents
235 GLUs X 87 regions Forestry Land Rent
235 GLUs X 87 regions 3 Livestock Land Rents
235 GLUs X 87 regions 8 Crop Land Rents

Figure 2: Processing three primary Moirai outputs. Recalibration of crop data year occurs before aggregation, and recalibration of land rent price year occurs before land rent calculation. Ovals are raster data (all at 5 arcmin except for the land cover/use data, which are at half-degree), hexagons are tabular data, boxes are processes, and the outputs are in underlined, bold lettering within bold hexagons.

We cannot reproduce the original method because some required source data and metadata are not available, in particular the forest land rent data and associated spatial distribution of forest types.

Diagnostics

Four R diagnostic scripts are provided for generating plots of the outputs, and in general they require the Moirai diagnostic output files. These scripts have been created in the context of GCAM development, so there are some options that are GCAM specific, such as the aggregation to GCAM political regions. Furthermore, while the scripts will generally run as long as the data formats are correct, not all diagnostic comparisons are valid (see scripts for details). In particular, certain comparisons across Moirai outputs and other data sources at the GLU level are not valid unless the GLUs are identical across the compared data.

Quality control

Moirai has undergone extensive functional testing and has recently been used to provide water-basin level data for a GCAM study on the impacts of potential climate change on U.S. agricultural markets [14]. A previous version has been used to quantify uncertainty in land resource projection due to spatial boundary definition [4]. We have also used extensive intermediate diagnostic outputs to ensure that the input data are correctly read, converted, and reconciled before undergoing final calculations.

Here we present comparisons of Moirai crop harvested area, crop production, and land rent with corresponding FAO and GTAP data. We show that Moirai duplicates well these two data sets at the country level for crop data (**Figure 3**) and at the intersection of GLUs and economic regions for land rent (**Figure 4**). Differences among the data sets are visualized by Cumulative Distribution Functions (CDFs) derived from values associated with each land level (e.g., **Figure 3a**). Based on the Kolmogorov-Smirnov (K-S) test, only one crop production CDF and no crop harvested area CDFs are significantly different ($\alpha = 0.05$) between Moirai and FAO data. In contrast, 16 production and 19 harvested area CDFs are significantly different between GTAP and FAO data. Similarly, 12 production and 20 harvested area CDFs are significantly different between Moirai and GTAP. None of the CDFs for

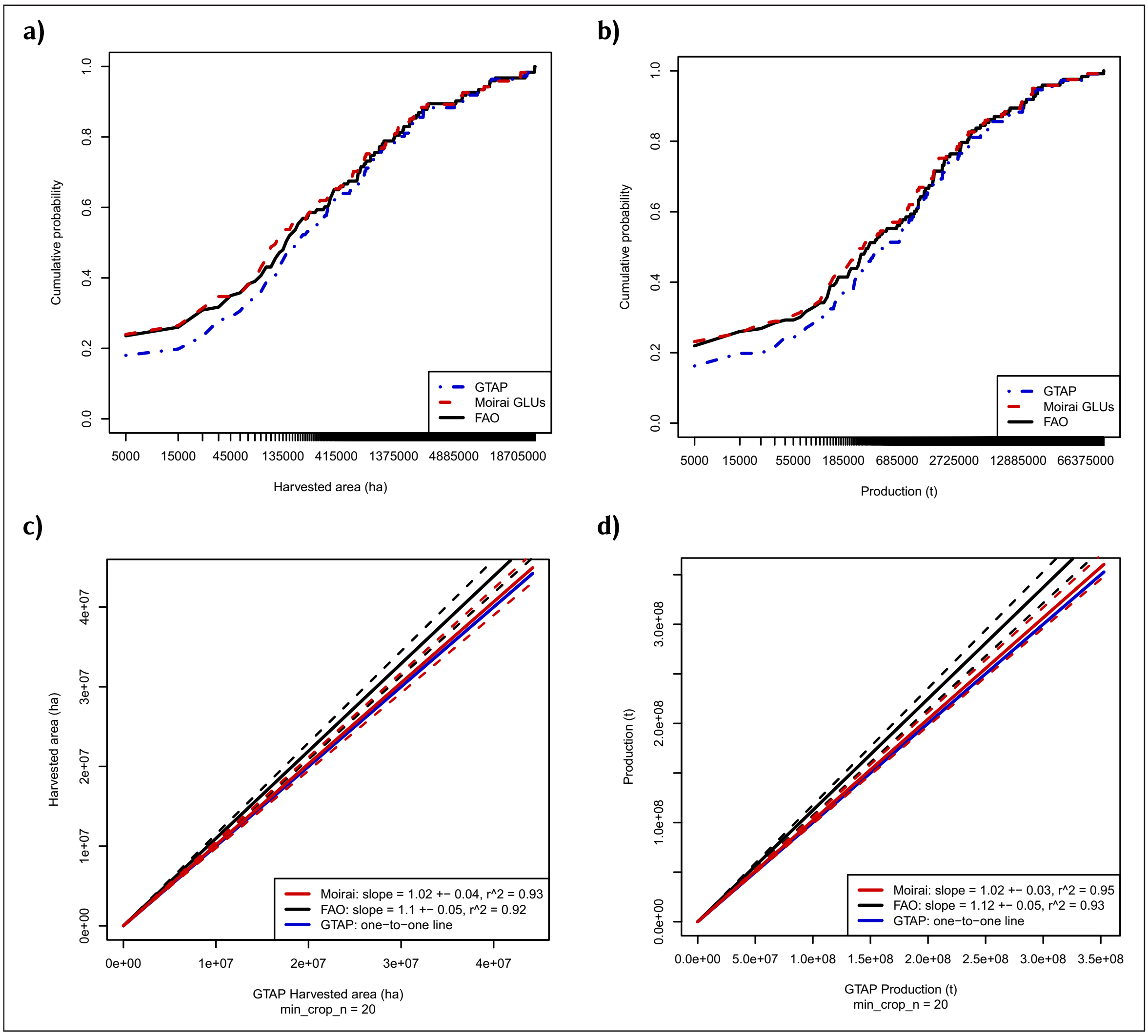

Figure 3: Evaluation of Moirai crop production and harvested area outputs at the country level, using water basins as the Geographic Land Units (GLUs). **a)** Wheat harvested area Cumulative Distribution Function (CDF), **b)** Wheat production CDF, **c)** mean of Moirai versus GTAP and FAO versus GTAP harvested area regressions for 93 and 88 crops, respectively, with n >= 20, **d)** like (c) but for production.

Wheat (**Figure 3a** and **3b**) are significantly different from each other and the K-S statistic is smallest between Moirai and FAO data (0.06), as compared with 0.08 between GTAP and FAO, and 0.10 between Moirai and GTAP. No land rent CDFs are significantly different between Moirai and GTAP, the K-S statistic for Forestry land rent is 0.06 (**Figure 4a**), and the K-S statistic for Wheat land rent is 0.10 (**Figure 4b**). We also calculate the mean of the crop- and sector-specific regression parameters among Moirai, GTAP, and FAO data at the country level for crop data (**Figure 3c** and **3d**) and between Moirai and GTAP at the intersection of GLUs and economic regions for land rent (**Figure 4c**). None of these mean regression slopes are significantly different from each other ($\alpha = 0.05$).

(2) Availability

Operating system

Moirai version 3 has been built and tested on Mac OSX (Xcode and terminal) and Slackware and Ubuntu Linux.

Programming language

Moirai is written in standard ANSII C code, and therefore should compile and run just about anywhere. The diagnostics scripts are written in R and have been run on versions >=3.1.3.

Additional system requirements

The disk space required to install Moirai and store some outputs is 115 GB. The maximum run-time memory usage benchmarks at 1.6 GB.

Dependencies

Moirai requires the C netcdf library (https://www.unidata.ucar.edu/software/netcdf/) to be installed by the user. The other required C libraries are usually included with the C compiler: stdio.h, math.h, stdlib.h, string.h, time.h, ctype.h. The diagnostic R scripts require the stringr and ggplot2 R libraries.

Software location

Archive

Name: Zenodo
Persistent identifier: 10.5281/zenodo.3370875
Licence: Modified BSD-3 license
Publisher: Alan Di Vittorio
Version published: v3.0.1
Date published: 18 August 2019

Code repository

Name: GitHub
Identifier: https://github.com/JGCRI/moirai

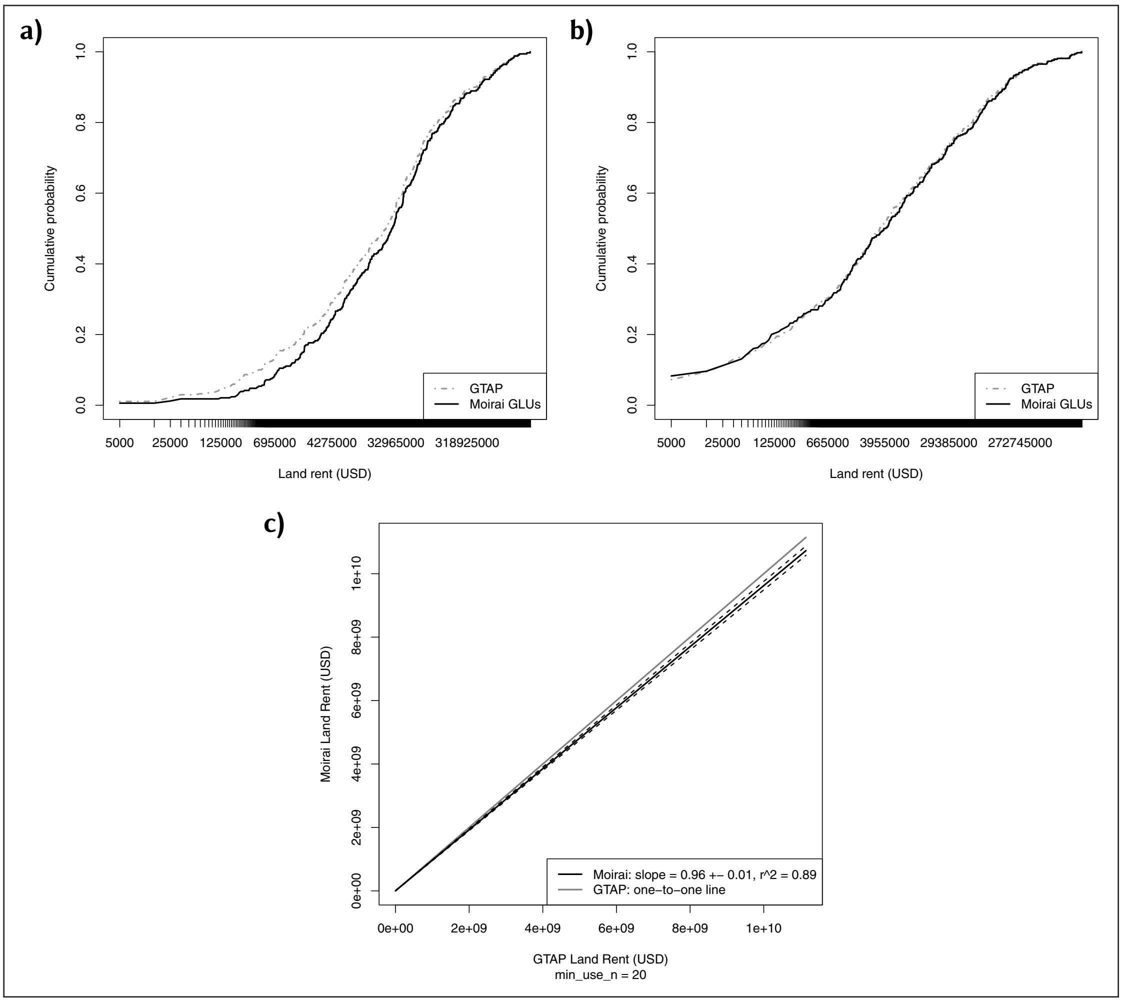

Figure 4: Evaluation of Moirai land rent output0 at the intersection of the original 18 Agro-Ecological Zones (Original and Moirai GLUs) and the 87 GTAP economic regions. **a)** Forestry land rent, **b)** Wheat land rent, **c)** mean of Moirai versus GTAP regressions for 12 land use sectors.

Licence: Modified BSD-3 license
Date published: 18 August 2019

Language

English

Installation

The Moirai LDS can be obtained from the Zenodo digital archive or from GitHub as listed above.

Moirai uses GitLFS (Large File Storage) so users will need download and install Git LFS before cloning the repository from GitHub (see https://github.com/git-lfs/git-lfs/wiki/Installation). Once Git LFS is downloaded, the user will need to install Git LFS by executing:

```
"git lfs install"
```

Moirai can then be cloned from GitHub using:

```
"git clone https://github.com/JGCRI/moirai.git"
```

Downloading the digital archive from Zenodo only requires the user to unzip the file.

Once the moirai directory is expanded on your local machine the 'moirai' command line tool can be compiled using a makefile (on any compatible platform) or Xcode (on a Mac). To compile in Xcode, open the '.../moirai/moirai.xcodeproj' file and set the location of your NetCDF library (see below for NetCDF details) in the Build Settings (click on the top-level 'moirai' project file icon in the navigator window to access these). There are three fields that need to be updated to reflect the location of your NetCDF header file (netcdf.h): 'Search Paths>Header Search Paths', and the Debug and Release fields of 'Search Paths>User Header Search Paths'. The 'Search Paths>Library Search Paths' and 'Linking>Other Linker Flags' fields need to be updated to reflect the location of the actual library file. Once the NetCDF location is set, simply select 'Build' from the 'Product' menu to compile 'moirai'. The current default setting for the location of the executable is '.../moirai/Build/Products/Debug'.

Alternatively, 'moirai' can be compiled using the 'makefile', with which the NetCDF library and header paths are automatically determined. Simply navigate to the '.../moirai' directory on the command line and type "`make`". The 'moirai' executable will be written in '.../moirai/bin', and the objects in '.../moirai/obj'. Note that the executable needs to be called from from the '.../moirai' directory, regardless of how it was compiled, because the example input file path entries are based on this project directory as the working directory (these can be changed by the user, as needed).

Running Moirai LDS

The Moirai LDS is a command line tool that takes the name of the Moirai LDS input file as the only argument (two examples are provided in '.../moirai/input_files' that can be run immediately once the code is built), but it can also be run directly in Xcode. Regardless of how the code is run, the user must also correctly specify the input and output directories (and any other input data files that they may want to substitute) in the Moirai LDS input file (see below) before running the code. To run within Xcode on a Mac, first compile the code with Xcode (see above) and specify the input file in the 'Product>Scheme>Edit schemes...' menu (the default is 'moirai_input_basins235.txt'). Then select 'Product>Run' from the menu, and the outputs will be written as specified in the input file (see below).

Alternatively, 'moirai' can be run directly from the command line by first navigating to the '.../moirai' directory and then typing either:

```
"bin/moirai input_files/moirai_input_basins235.txt"
```

or

```
"Build/debug/moirai input_files/moirai_input_
basins235.txt"
```

depending on where the compiled executable resides (see above). The input file name is the only argument and determines how the outputs are written.

There are two example input files that can be run without modification (see below): 'moirai_input_basins235.txt' and 'moirai_input_aez_orig.txt'. The outputs will be written to '.../moirai/outputs/basins235/' or '.../moirai/outputs/aez_orig/', respectively (the directories will be created automatically). These newly created outputs can be compared with those in '.../moirai/example_outputs/basins235/' or '.../moirai/example_outputs/aez_orig/', respectively.

(3) Reuse potential

Moirai is designed to provide flexibility in defining GLUs and specifying the output data-year and USD-year without making any code modifications. This allows researchers to readily generate land data on spatial grids that match their pre-existing requirements. This is particularly useful for initialization or evaluation of land use change models or economic general/partial equilibrium models that have embedded spatial configurations. However, the software is also designed to allow source data substitution with minimal code modification, especially if the replacement data are in the same format as the original data. Substituting source data entails changing the file name in the input file and, if necessary, modifying or creating the read function for that particular data set and updating constants in the header file. If the replacement data are considerably different, then further modification of the code may be required. In the specific case of land cover data, Moirai can process an identical data set at finer resolution simply by updating the resolution in the header file. In the specific case of HYDE land use data, the most recent version (3.2.1) and the low and high land use scenarios should be directly substitutable without any code modification. Furthermore, each output is generated by its own function, which allows for targeted reformatting of output files. Each function is defined in its own file, and the main function controls the processing flow through function calls and memory management. Please contact the lead author via email if you are considering code modification or further development.

Acknowledgements

This research was supported by the U.S. Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research, as a part of research in the Multi-Sector Dynamics, Earth and Environmental System Modeling Program. Work at Lawrence Berkeley

National Laboratory was performed under contract DE-AC02-05CH11231, and work at Battelle-operated Pacific Northwest National Laboratory was performed under contract DE-AC05-76RL01830. Work performed at University of Illinois, Urbana-Champaign was performed under a DOE award.

Competing Interests

The authors have no competing interests to declare.

References

1. GTAP Land Use Database Release 2.1, 2009. https://www.gtap.agecon.purdue.edu/resources/res_display.asp?RecordID=1900.
2. **Lee, H-L, Hertel, T W, Sohngen, B** and **Ramankutty, N** 2005 Towards an integrated land use database for assessing the potential for greenhouse gas mitigation. Purdue University.
3. **Lee, H-L, Hertel, T W, Rose, S** and **Avetisyan, M** 2009 An integrated global land use data base for CGE analysis of climate policy options. Chapter 4. In: Hertel, T W, Rose, S and Tol, R (eds.), *Economic Analysis of Land Use in Global Climate Change Policy*, 72–88. Abingdon: Routledge. DOI: https://doi.org/10.4324/9780203882962
4. **Monfreda, C, Ramankutty, N** and **Foley, J A** 2008 Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochem. Cycles*, 22: GB1022. DOI: https://doi.org/10.1029/2007GB002947
5. **Di Vittorio, A V, Kyle, P** and **Collins, W D** 2016 What are the effects of Agro-Ecological Zones and land use region boundaries on land resource projection using the Global Change Assessment Model? *Environmental Modelling & Software*, 85: 246–265. DOI: https://doi.org/10.1016/j.envsoft.2016.08.016
6. **Calvin, K, Patel, P, Clarke, L, Asrar, G, Bond-Lamberty, B, Di Vittorio, A, Edmonds, J, Hartin, C, Hejazi, M, Iyer, G, Kyle, P, Kim, S, Link, R, Mcjeon, H, Smith, S J, Waldaff, S** and **Wise, M** 2019 GCAM v5.1: Representing the linkages between energy, water, land, climate, and economic systems. *Geosci. Model Dev.*, 12: 677–698. DOI: https://doi.org/10.5194/gmd-12-677-2019
7. **Klein Goldewijk, K, Beusen, A, Doelman, J** and **Stehfest, E** 2017 Anthropogenic land use estimates for the Holocene – HYDE 3.2. *Earth Syst. Sci. Data*, 9: 927–953. DOI: https://doi.org/10.5194/essd-9-927-2017
8. **Ramankutty, N** and **Foley, J A** 1999 Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochem. Cycles*, 13: 997–1027. DOI: https://doi.org/10.1029/1999GB900046
9. **Meiyappan, P** and **Jain, A** 2012 Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. *Frontiers of Earth Science*, 6: 122–139. DOI: https://doi.org/10.1007/s11707-012-0314-2
10. **Ramankutty, N, Evan, A T, Monfreda, C** and **Foley, J A** 2008 Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochem. Cycles*, 22: GB1003. DOI: https://doi.org/10.1029/2007GB002952
11. **Monfreda, C, Ramankutty, N** and **Hertel, T** 2009 Global agricultural land use data for climate change analysis. Chapter 2. In: Hertel, T W, Rose, S and Tol, R. (eds.), *Economic Analysis of Land Use in Global Climate Change Policy*, 33–48. Abingdon: Routledge.
12. **Portmann, F T, Siebert, S** and **Döll, P** 2010 MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, 24: GB1011. DOI: https://doi.org/10.1029/2008GB003435
13. **Mekonnen, M M** and **Hoekstra, A Y** 2011 The green, blue and grey water footprint of crops and derived crop products. *Hydrol. Earth Syst. Sci.*, 15: 1577–1600. DOI: https://doi.org/10.5194/hess-15-1577-2011
14. **Snyder, A,** et al. in review. The domestic and international implications of future climate for U.S. agriculture.

How to cite this article: Di Vittorio, A V, Vernon, C R and Shu, S 2020 Moirai Version 3: A Data Processing System to Generate Recent Historical Land Inputs for Global Modeling Applications at Various Scales. *Journal of Open Research Software,* 8: 1. DOI: https://doi.org/10.5334/jors.266

Submitted: 05 March 2019 **Accepted:** 20 December 2019 **Published:** 15 January 2020

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS