Journal of
open research software

Calvin, K, et al. 2019 gcamland v1.0 – An R Package for Modelling Land Use and Land Cover Change. *Journal of Open Research Software,* 7: 31. DOI: https://doi.org/10.5334/jors.233

SOFTWARE METAPAPER

gcamland v1.0 – An R Package for Modelling Land Use and Land Cover Change

Katherine Calvin, Robert Link and Marshall Wise

Joint Global Change Research Institute, Pacific Northwest National laboratory, College Park, Maryland, US

Corresponding author: Katherine Calvin (Katherine.calvin@pnnl.gov)

gcamland v1.0 is an open source R package that was built to allocate land across a variety of uses based on changes in agricultural yield and commodity price. The land allocation algorithm is based on the one included in the Global Change Assessment Model (GCAM). *gcamland v1.0* includes the ability to run in a historical mode, enabling model validation and parameter estimation, or in a future mode, simulating changes in land use/land cover in the future. For both modes, *gcamland v1.0* can run a single simulation or a large ensemble of simulations with different parameters. When ensembles are generated in the historical mode, *gcamland v1.0* calculates the likelihood of a given parameter set by comparing to observational data. *gcamland v1.0* is publicly available via GitHub and has can be adjusted to represent alternative scenarios or configured to different regions and land types.

Keywords: Global Change Assessment Model (GCAM); Land-use land-cover change (LULCC); integrated assessment modelling; R

Funding statement: This research was supported by the Office of Science of the U.S. Department of Energy as part of the Multi-Sector Dynamics Program. Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

(1) Overview

Introduction

Integrated Assessment Models (IAMs) link together representations of multiple sectors, including energy, water, and land, examining the interactions between these systems [1]. These models typically initialize to a specific historic year and then project key variables into the future in annual to decadal time steps. For example, GCAM [2, 3] calculates information on socioeconomics (gross domestic product, population), energy (production, consumption, price, and emissions), agriculture (production, consumption, price, and emissions), land (land use, land cover, emissions), climate (concentrations, forcing, global mean temperature rise), and interactions between these systems. For example, GCAM can examine the effect of changes in the energy system on land allocation [4] or the implications of changes in water availability on crop production [5]. While the fully coupled model can help address many science questions, separating the individual component models (1) reduces computational expense, (2) enables their use in other modelling systems, (3) facilitates ensemble calculations, and (4) facilitates hindcast and parameter estimation efforts. For these reasons, *gcamland v1.0* isolates the land allocation mechanism from GCAM [6].

Users can run two scenario types: "Reference" and "Hindcast". In Reference mode, the model is initialized to a recent year (2010) and estimates future land allocation through 2100 as economic conditions evolve. In this mode, users can examine how land use and land cover might evolve in the future as agricultural yields or commodity prices change. In Hindcast mode, *gcamland* is initialized and run over a historical period. Such simulations allow the user to evaluate model performance by comparing model results to observational data.

Users can evaluate model uncertainty by including the ability to build random parameter ensembles. The ensemble function in *gcamland v1.0* randomly samples from uniform distributions of key parameters, generating an ensemble with a user-specified number of simulations. Ensembles can be run in either Hindcast mode (e.g., to help with parameter estimation) or Reference mode (e.g., to help understand the implications of uncertainty in parameters on future land use). These simulations can be run in parallel using the doParallel package from R, spreading the simulations over the user specified number of cores.

Implementation and architecture

The *gcamland* package is written in R, using a suite of readily available libraries. R was chosen to encourage broad use and community involvement as the language is freely available and widely used. R is free, open-source and broadly used by scientists [7].

Scenario Types

The *gcamland* package has two different scenario types: *Reference* and *Hindcast*. The default scenario type is a Reference type, but this can be changed by adjusting the scenario info object. The Reference type generates a future simulation, where land allocation is calculated in the future, with changes induced by changes in price or agricultural productivity. The Hindcast type simulates land allocation in the past, and it can be used to evaluate the model.

For the Reference type, the years included in the historical period are 1975, 1990, 2005, and 2010; the future period spans from 2015 to 2100 in five-year increments. In Hindcast mode, the historical period only includes the year 1975; the future period spans from 1976 to 2010 in annual increments. These years can be modified in the constants file; the only requirement is that calibration data is needed for the "history years" and prices are needed for the "future years".

For Reference simulations, the user can adjust agricultural prices and productivity growth to craft their own scenario (see below). For Hindcast simulations, price and productivity estimates are from the Food and Agriculture Organization.

Required Input

The *gcamland* package has three types of input: initialization data, hindcast data, and scenario data. There are defaults provided for all types of data in the 'inst/extdata' directory. Initialization data defines the structure of the land nest (see **Figure 1**) and provides calibration data for the historical period. That calibration data includes land allocation, the value of unmanaged land, agricultural production for managed land types, non-land cost of production, and logit exponents that parameterize the degree of competition among land types [6]. Hindcast data includes prices and yields for all agricultural commodities from 1975 to 2010. Scenario data includes prices for all agricultural commodities for years simulated in the Reference scenario type, and productivity growth (i.e., adjustments to future yields) for managed land types in the future period for the Reference scenario type. We anticipate that users will adjust the scenario data to craft a new scenario (e.g., examine alternative future price or yield trajectories). Users can also adjust the logit exponents for certain levels of the nest by adjusting arguments in the scenario info object.

Calculation Steps

Each *gcamland* simulation has four steps: setup, initial calculation, final calculation, and reporting. These steps are called from the 'run_model' method. The setup step is called once per simulation. This step initializes the land allocator, reading in calibration data (described above) and calculating the required parameters. The calibration portion of the setup step calculates expected profit and land shares in the historical period and then uses that information to calculate a share weight that is used to ensure that model-calculated land allocations match read-in land allocations in the historical period.

The initial and final calculation steps are run for each model year (both history and future). The initial calculation step ensures that expected price, cost, expected yield, and expected profit are initialized for the future period. The final calculation step calculates land shares for each type within each nest based on the expected profit, logit exponent, and share weight. Using these land shares, the amount of land by type or use is also calculated.

The reporting step gathers all outputs and prints this to a file if requested (described below).

Model Output

Model output includes yield, expected yield, price, expected price, expected profit, land shares, and land allocation for all land types. The default file format for most outputs is an rds file (a compressed, serialized R object), but the

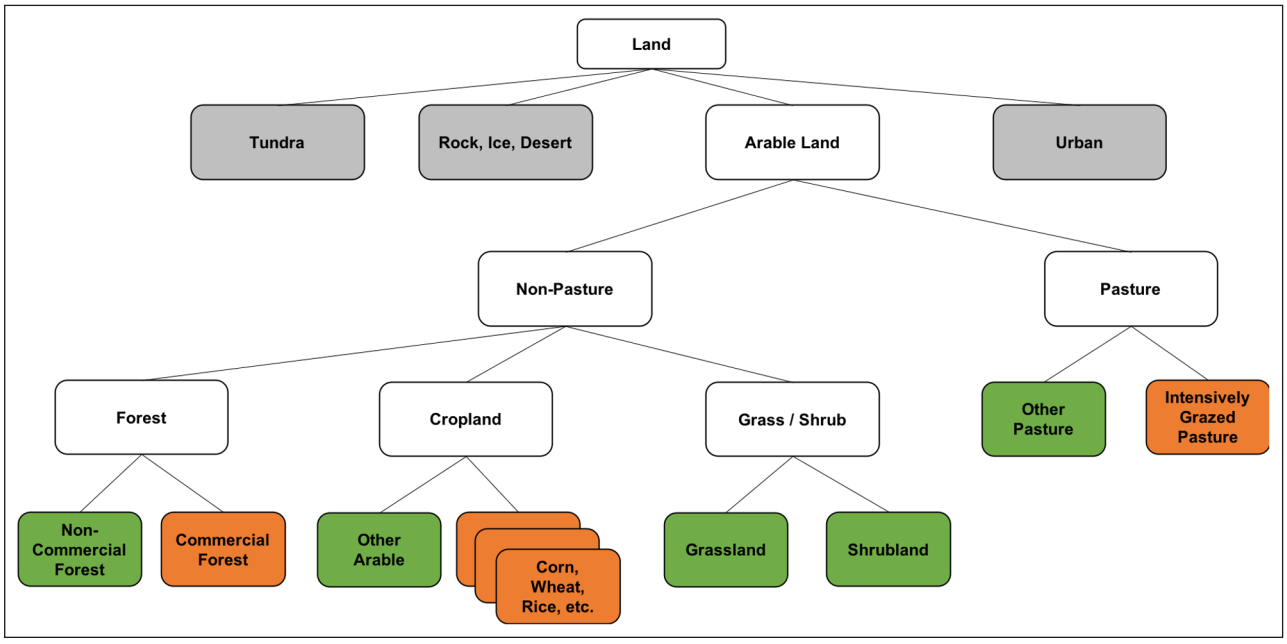

Figure 1: Default nesting structure used in *gcamland*.

'export_data' function enables the output from a single scenario to be written as a csv file. There are also optional plotting functions that plot land area by type and region/subregion (**Figure 2**).

Ensemble Mode

The ensemble mode, invoked through the 'run_ensemble' function, allows users to generate an ensemble of parameter sets. Two types of parameters are varied in the ensemble mode: the logit exponents used at different points in the land nest (**Figure 1**) and the algorithm for forming expectations about yield and price. The parameters used in the samples are drawn quasi-randomly from a uniform distribution. These are used to generate input values for the 'run_model' function. When ensembles are run in Hindcast mode, *gcamland* will calculate the likelihood of a given set of parameters by comparing calculated land allocation to observed land allocation. Such information can be used to estimate a probability density function for each parameter.

Analysis Functions

The package includes a suite of Bayesian analysis functions. These functions are used in conjunction with the ensemble mode discussed above to estimate characteristics of the posterior probability density function (PDF) for the model parameters. A user can specify both the Bayesian prior for the model parameters and the likelihood function used to compare model outputs to historical data. Functions are supplied to compute the posterior probability density (PPD) of the parameters sampled in the ensemble, as well as statistics summarizing the PDF. Currently supported statistics include maximum *a posteriori* parameters (i.e., the set of parameters with the highest PPD), marginal expectation values, and highest-posterior-density intervals.

The package also includes a function to compute the Widely Applicable Information Criterion (WAIC) [8, 9]. This statistic allows a user to estimate the out of sample performance of a model fit to a particular data set. The WAIC is also useful for model comparison, allowing users to estimate a Bayesian odds ratio for model families, based on their expected out of sample performance.

Installing and Running the Package

The package can be installed directly from its github repository using the R *devtools* package. From an R prompt, run the command:

```
devtools::install_github('jgcri/gcamland')
```

The steps to run an individual *gcamland* simulation are:

1. (Optional) Attach the *gcamland* namespace: `library('gcamland')`
2. (Optional) Adjust the scenario information object
3. Run a simulation `run_model(SCENARIO.INFO, aVerbose=TRUE)`
4. (Optional) Export the output to csv `export_results(SCENARIO.INFO)`
5. (Optional) Plot the results `print(plotRegionalLandAllocation (SCENARIO.INFO))`

The steps to run an ensemble of *gcamland* simulations are:

1. (Optional) Attach the *gcamland* namespace: `library('gcamland')`
2. Run an ensemble of simulations `run_ensemble (N, aOutputDir = MY.DIRECTORY, atype= "Hindcast")`

A list of the main methods is provided in **Table 1**, including the inputs and outputs of each method.

gcamland v1.0 Example

By following the steps to run an individual simulation listed above, *gcamland v1.0* will calculate land use and land cover in the USA through 2100 assuming commodity

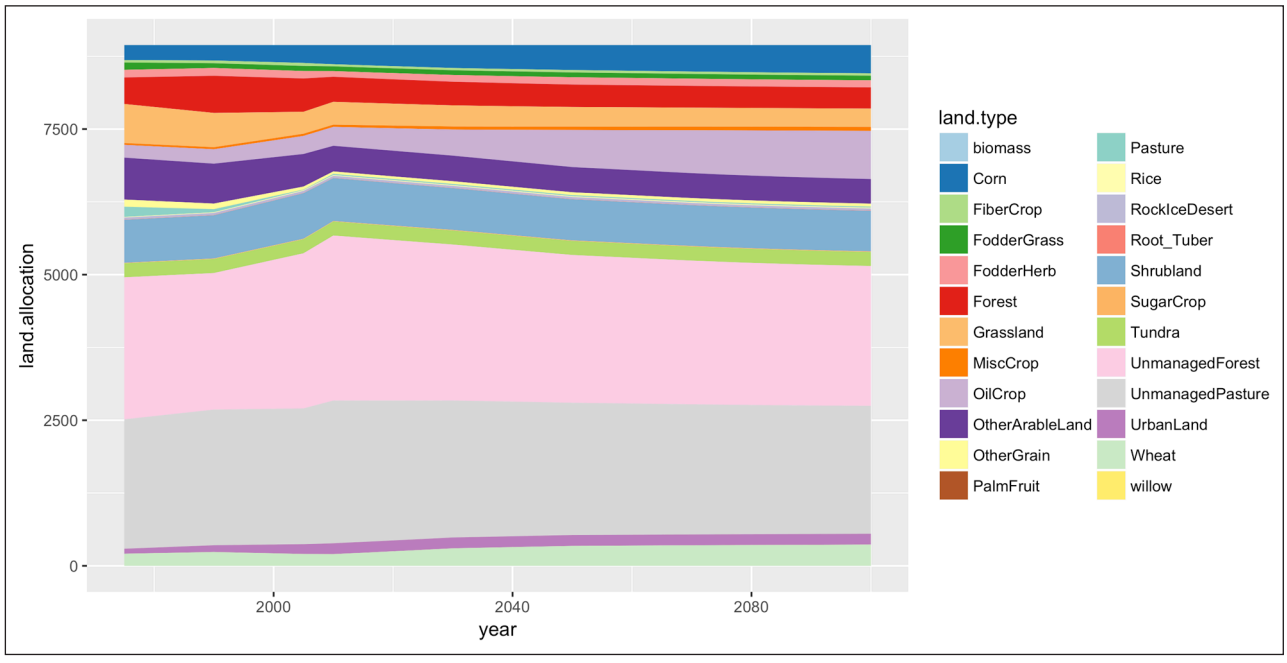

Figure 2: Sample land allocation by use and type in the USA for a Reference scenario.

Table 1: Table of the main methods in the gcamland v1.0 package, including their inputs and outputs. Note that only methods that can be called directly are listed in the table.

Method	Description	Inputs		Output
		Name	Description	
run_model	Calculates land use and land cover for a given region and set of time periods	aScenarioInfo	Scenario-related information, including names, logits, expectations.	Table of model results
		aPeriods	Integer vector of periods to run. Default is all periods defined for the scenario type.	
		aVerbose	If TRUE, output additional debugging information.	
run_ensemble	Generates a suite of different scenarios and calculates land use and land cover for each (by calling run_model)	N	Number of parameter sets to select	List of ScenarioInfo objects for the ensemble members
		aOutputDir	Directory where outputs are saved	
		skip	Number of iterations to skip (i.e., if building on another run.)	
		atype	Scenario type: either "Reference" or "Hindcast"	
		logparallel	Name of directory to use for parallel workers' log files. If NULL, then don't write log files.	
export_results	Saves results from a specified scenario as a csv file	aScenarioInfo	Scenario-related information, including names, logits, expectations.	csv file with model results
plotLandAllocation	Plots allocation over time by land type, with subregional detail	aScenarioInfo	Scenario-related information, including names, logits, expectations.	ggplot plot
plotRegionalLand Allocation	Plots allocation over time by land type, aggregated to region (as in Figure 2)	aScenarioInfo	Scenario-related information, including names, logits, expectations.	ggplot plot
ScenarioInfo	Creates a scenario information object that can be used in the functions above	aExpectationType	String indicating whether to use "Perfect", "Lagged", or "Linear" expectations	ScenarioInfo object
		aLaggedShareOld	Weight to put on older information if "Lagged" expectations are used	
		aLinearYears	Number of years to use in extrapolation if "Linear" expectations are used	
		aLogitUseDefault	Boolean indicating whether to use default logits	
		aLogitAgroForest	AgroForest logit exponent (assuming mLogitUseDefault == FALSE)	
		aLogitAgroForest_ NonPasture	AgroForest_NonPasture logit exponent (assuming mLogitUseDefault == FALSE)	
		aLogitCropland	Cropland logit exponent (assuming mLogitUseDefault == FALSE)	
		aScenarioType	Type of scenario to run: either "Reference" or "Hindcast"	
		aScenarioName	Complete scenario name, with expectations and logit information	
		aFileName	File name	
		aOutputDir	Output directory	
		aSerialNum	Serial number for a run that is part of a series.	
		aRegion	Region to use in the calculation. Right now we only run a single region at a time.	

prices are constant and crop yields increase as specified in [10]. This results in an increase in land allocated to crops (e.g., corn, rice, wheat, etc.), as yield increases drive an increase in profits, and a decline in unmanaged forest and other natural lands (**Figure 2**). *gcamland v1.0* can examine the effects of other yield and price trajectories on land use and land cover. For example, by altering the future prices specified in 'inst/extdata/scenario-data/AgPrices_Reference.csv', one can explore the effect of changes in commodity prices on future land use and land cover. *gcamland v1.0* can also be used to calculate land use and land cover in other regions. For example, by changing the 'DEFAULT.REGION' in the 'constants.R' file, the model can estimate land use and land cover for 31 pre-specified regions. Prices and yields can then be adjusted for each of these regions, as described above.

Quality control

The *gcamland* package includes automated functional and unit testing. Unit testing is run using the R *testthat* package and currently covers just over 80% of the code by line count. The ensemble capability described above is not amenable to automated testing, due to the size of the minimum meaningful run. This functionality is hand-tested whenever changes to that part of the code base occur. The Bayesian inference functions have an additional set of validation tests that are run on a mock dataset generated from a linear model. This model was also analyzed using the *rethinking* package [11], and the unit tests require results of the *gcamland* analysis functions to be consistent with the results from the previously published package.

Functional testing is provided by the R built-in package checker, which verifies that packages can be installed, loaded, and unloaded cleanly and that the package code and structure conform to established best practices. The *gcamland* repository uses continuous integration, provided by Travis CI, ensuring that the test suite is run for every pull request. Tests must pass for the pull request to be merged into the repository. Any pull request that modifies production code or data must also be peer reviewed by another member of the development team.

(2) Availability

Operating system

Mac OS Sierra; Windows 7; Unix

Programming language

R (version 3.3 or greater)

Additional system requirements

N/A

Dependencies

- assertthat (>=0.2),
- dplyr (>=0.4.3),
- tibble (>=1.1),
- tidyr (>=0.6.0),
- readr (>=1.0.0),
- ggplot2 (>=2.2.1),
- igraph (>=1.0.1),
- randtoolbox (>=1.17),
- doParallel (>=1.0),
- foreach (>=1.4)

Software location

Archive

Name: Zenodo
Persistent identifier: http://doi.org/10.5281/zenodo.1264706
Licence: GPL2
Publisher: Katherine V. Calvin
Version published: v1.0.0
Date published: 04/06/18

Code repository

Name: GitHub
Identifier: https://github.com/JGCRI/gcamland
Licence: GPL2 (https://github.com/JGCRI/gcamland/blob/master/LICENSE)
Date published: dd/mm/yy

Language

English

(3) Reuse potential

gcamland methods and functions use Roxygen to generate documentation via R's built-in help function. The code base is also thoroughly documented to encourage community development and use. Other regional aggregations, land types, and initialization data can be used through modest modifications to the initialization scripts. gcamland can also be used in other coupled model studies since it can be called from other models.

A user guide on how to use and modify gcamland is available at: https://github.com/JGCRI/gcamland/wiki.

An issue tracker is available at: https://github.com/jgcri/gcamland/issues.

Competing Interests

The authors have no competing interests to declare.

References

1. **Weyant, J** 2017 "Some Contributions of Integrated Assessment Models of Global Climate Change." *Review of Environmental Economics and Policy*, 11(1): 115–137. DOI: https://doi.org/10.1093/reep/rew018
2. **Calvin, K, Bond-Lamberty, B, Clarke, L, Edmonds, J, Eom, J, Hartin, C, Kim, S, Kyle, P, Link, R, Moss, R, McJeon, H, Patel, P, Smith, S, Waldhoff, S** and **Wise, M** 2017 "The SSP4: A world of deepening inequality." *Global Environmental Change*, 42: 284–296. DOI: https://doi.org/10.1016/j.gloenvcha.2016.06.010
3. **Calvin, K V** and Coauthors 2019 GCAM v5.1: Representing the linkages between energy, water, land, climate, and economic systems. *Geosci. Model Dev.*, 12: 1–22. DOI: https://doi.org/10.5194/gmd-2018-214
4. **Calvin, K, Wise, M, Luckow, P, Kyle, P, Clarke, L** and **Edmonds, J** 2016 Implications of uncertain future fossil energy resources on bioenergy use and terrestrial

carbon emissions. *Clim. Change*, 136. DOI: https://doi.org/10.1007/s10584-013-0923-0

5. **Cui, R Y** and Coauthors 2018 Regional responses to future, demand-driven water scarcity. *Environ. Res. Lett.*, 13: 94006. http://stacks.iop.org/1748-9326/13/i=9/a=094006. DOI: https://doi.org/10.1088/1748-9326/aad8f7
6. **Wise, M, Calvin, K, Kyle, P, Luckow, P** and **Edmonds, J** 2014 "Economic and Physical Modeling of Land Use in GCAM 3.0 and an Application to Agricultural Productivity, Land, and Terrestrial Carbon." *Climate Change Economics*, 5(2). DOI: https://doi.org/10.1142/S2010007814500031
7. **Tippmann, S** 2014 Programming tools: Adventures with R. *Nature*, 517: 109–110. DOI: https://doi.org/10.1038/517109a
8. **Gelman, A, Hwang, J** and **Vehtari, A** 2014 Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24: 997–1016. DOI: https://doi.org/10.1007/s11222-013-9416-2
9. **Watanabe, S** 2010 Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11: 3571–3594.
10. **Briunsma, J** 2009 The Resource Outlook to 2050: By How Much Do Land, Water, and Crop Yields Need to Increase by 2050? *Expert Meeting on How to Feed the World in 2050, Food and Agriculture Organization of the United Nations.*
11. **McElreath, R** 2016 Rethinking: Statistical Rethinking book package. *R package version 1.59.1.*

How to cite this article: Calvin, K, Link, R and Wise, M 2019 gcamland v1.0 – An R Package for Modelling Land Use and Land Cover Change. *Journal of Open Research Software,* 7: 31. DOI: https://doi.org/10.5334/jors.233

Submitted: 04 June 2018 **Accepted:** 07 October 2019 **Published:** 22 October 2019

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

OPEN ACCESS