Core Model Proposal #325: gcamdata chunk re-name

Product: Global Change Analysis Model (GCAM)

Institution: Joint Global Change Research Institute (JGCRI)

Authors: G. Page Kyle, Ellie Lochner

Reviewers: Pralit Patel, Maridee A Weber

Date committed: 04 Jun 2023

IR document number: PNNL-34385

Related sector: gcamdata

Type of development: bugfix

Purpose: The R/ folder of the gcamdata package is difficult to work with because it contains a large number of "chunk" files (i.e., files prefixed with "zchunk_"), from eight different modules (aglu, climate, emissions, energy, modeltime, socioeconomics, water, and gcamusa), with no module-specific sorting or sub-folders that would help to distinguish chunks from the different modules. This proposal addresses this issue by re-naming all chunk files, and re-naming functions where necessary, so that alpha-numeric sorting appropriately differentiates the eight modules. It also fixes a few bugs to ensure the gcamdata tests pass.

Description of Changes

This proposal modifies the filenames of the "chunks" (zchunk_*.R) in the /R/ folder of the gcamdata package, so that each chunk's module names are readily apparent, and the alpha-numeric sorting results in chunk files being grouped by module. It also updates function names within the chunks to be consistent with the file names. This is done through a function that we've added to admin.R, called "rename_gcamdata_chunks". This function is split into two parts. The first part renames the file based on criteria described below. Before it actually renames anything, it checks for any duplicate file names and will error if so. If all checks pass, it renames the files. The second part of the function renames the module functions within the file to be consistent with the new file name. Renames and function updates will only happen for R scripts that begin with "zchunk" as this string indicates the file and content within the file has not been updated.

Changes that this function implements include:

1. Removing "zchunk" and replacing it with z<module-name>
    1. We do this for clarity and for grouping by module. The module name is obtained by reading the 3rd line of each file, which should specify the module name in a comment. If the module isn't specified on the third line, which is the case in a number of chunks, the code will error with a message telling the user which file needs to be manually fixed. These can usually be fixed with slight reorganization of the comments at the top of files.
2. Renaming "socioeconomics" to "socio"
    1. We have been inconsistent whether we refer to socioeconomics as "socio" or "socioeconomics". For consistency and conciseness, we've chosen "socio" and thus rename all files and functions accordingly.
3. Removing LA/LB prefixes
    1. Additional letters (LA or LB) were used in some cases for prescribing a specific file run order in the old gcam-data-system. This is no longer relevant, so we remove them. We retain the numerical strings prefixed by "L". (e.g., LA101 is re-set to L101).
4. Cleaning up GCAM-USA modules
    1. We've included a "gcamusa" module option (i.e. zgcamusa_*), so for all GCAM-USA files, we remove the appended "_USA" in the file and function names, which is now redundant.
5. Rename "batch" to "XML"
    1. The word "batch" is removed from file and function names, and replaced with "_xml", for clarify and to continue differentiating between chunks that do and don't directly create XMLs.

Other changes from this proposal:

One function that was unintentionally getting "exported" in the NAMESPACE (module_aglu_LB151.ag_MIRCA_ctry_C_GLU_irr) was modified, and a function

(chunk_readylist) in admin.R was removed, as well as its corresponding test (test_chunk_readylist), as it is no longer necessary. The function relied on a hard-wired (and outdated) list of filenames in the /R folder, and was made in order to track progress during the course of the data system re-write.

This proposal also includes an update to the gcamdata documentation and re-builds the GCAM data map, in R/sysdata.rda.

Finally, this proposal addresses recent bugs within the data system that cause testing failures, including fixing capitalization of constants, removing successive mutates, and fixing timeshift failures.

FOR OTHER BRANCHES:

For assistance with other branches that have old file names, we've added a gcamdata function, "rename_gcamdata_chunks", in admin.R that automatically renames chunks for a user. This proposal only updates files in the master branch, so all other files on other branches have not been updated. *All branches that intend to be merged into master will have to run this function before being merged*. Since this method is now a function in the data system, it should be fairly easy for users to update their own branches. We've also added a test in tests/testthat/test_chunks that will tell a user if they have any chunks that have to be renamed. The test will pass if all chunks are named correctly, and will fail if there are chunks not yet renamed. (i.e., there are any "zchunks" in /R/,

Note, this function only renames chunks that begin with "zchunk", as this indicates that the chunks have not been updated yet. Changes to file content (part 2 of the function) have to be run at the same time as the file rename (part 1) for function renaming to happen correctly. This would only become a problem if there was a break/error after part 1, and before part 2 completed, and the current R session was aborted. When you'd try to rerun the function, it would no longer recognize any files or file content needing to be updated as they would no longer contain the "zchunk" prefix. If this does happen, the data system should still run as normal and the function names can be manually adjusted another time.

Validation

No model runs are necessary for validation as the contents of the /xml folder are unaffected by the changes in this proposal. Enclosed is a screenshot of the XML folder before and after the changes in this proposal.