gcamfaostat: An R package to prepare, process, and synthesize FAOSTAT data for global agroeconomic and multisector dynamic modeling

Xin Zhao 1, Maksym Chepeliev 2, Pralit Patel 1, Marshall Wise 1, Katherine Calvin 1, Kanishka Narayan 1, and Chris Vernon 1

1 Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA 2 Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University, West Lafayette, IN, USA

DOI: 10.21105/joss.06388

Software
- Review
- Repository
- Archive

Summary

The gcamfaostat R package is designed for the preparation, processing, and synthesis of the Food and Agriculture Organization (FAO) Statistics (FAOSTAT) agroeconomic data. The primary purpose is to facilitate FAOSTAT data use in global economic and multisector dynamic models while ensuring transparency, traceability, and reproducibility. Here, we provide an overview of the development of gcamfaostat v1.0.0 and demonstrate its capabilities in generating and maintaining agroeconomic data required for the Global Change Analysis Model (GCAM). Our initiative seeks to enhance the quality and accessibility of data for the global agroeconomic modeling community, with the aim of fostering more robust and harmonized outcomes in a collaborative, efficient, and open-source framework. The processed data and visualizations offered by gcamfaostat can also be valuable to a broader audience interested in gaining insights into the intricacies of global agriculture.

Statement of need

Global economic and multisector dynamic models have become pivotal tools for investigating complex interactions between human activities and the environment, as evident in recent research (Doelman et al., 2022; Fujimori et al., 2022; Ven et al., 2023). Agriculture and land use (AgLU) plays a critical role in these models, particularly when used to address key agroeconomic questions (Graham et al., 2023; Yarlagadda et al., 2023; Zhang et al., 2023; Zhao et al., 2024). Sound economic modeling hinges significantly upon the accessibility and quality of data (Bruckner et al., 2019; Calvin et al., 2022; Chepeliev, 2022). The FAOSTAT serves as one of the key global data sources, offering open-access data on country-level agricultural production, land use, trade, food consumption, nutrient content, prices, and more (FAO, 2023). However, the raw data from FAOSTAT requires cleaning, balancing, and synthesis, involving assumptions such as interpolation and mapping, which can introduce uncertainties. In addition, some of the core datasets reported by FAOSTAT, such as FAO’s Food Balance Sheets (FBS), are compiled at a specific level of aggregation, combining together primary and processed commodities (e.g., wheat and flour), which creates additional data processing challenges for the agroeconomic modeling community (Ceploiev, 2022). It is noteworthy that each agroeconomic modeling team typically develops its own assumptions and methods to prepare and process FAOSTAT data. While largely overlooked, the uncertainty in the base data calibration approach likely contribute to the disparities in model outcomes (Lampe et al., 2014; Zhao et al., 2021). Hence, our motivation is to create an open-source tool (gcamfaostat) for the preparation, processing, and synthesis of FAOSTAT data for global agroeconomic modeling.

To the best of our knowledge, such a tool has not been developed yet. gcmafaoostat bridges a crucial gap in the literature by offering several key features and capabilities.

1. **Transparency and Reproducibility**: gcmafaoostat incorporates functions for downloading, cleaning, synthesizing, and balancing agroeconomic datasets in a traceable, transparent, and reproducible manner (Wilkinson et al., 2016). This enhances the credibility of the processing and allows for better scrutiny of the methods. Here we document and demonstrate the use of the package in generating and updating agroeconomic data needed for GCAM v7.0 (Bond-Lamberty et al., 2023).

2. **Expandability and Consistency**: gcmafaoostat can be used to flexibly process and update agroeconomic data for any agroeconomic model. The package framework can be also easily expanded to include new modules for consistently processing new data.

3. **Community Collaboration and Efficiency**: The package provides an open-source platform for researchers to continually enhance the processing methods. This collaborative approach, which establishes a standardized and streamlined process for data preparation and processing, carries benefits that extend to all modeling groups. By reducing the effort required for data processing and fostering harmonized base data calibration, it contributes to a reduction in modeling uncertainty and enhances the overall research efficiency.

4. **User Accessibility**: Where applicable, the processed data can be mapped and aggregated to user-specified regions and sectors for agroeconomic modeling. However, beyond the modeling community, gcmafaoostat can be valuable to a broader range of users interested in understanding global agriculture trends and dynamics, as it provides user-friendly data processing and visualization tools.

**Design and Functionality**

**Bridging the gap between FAOSTAT and global economic modeling**

GCAM is a widely recognized global economic and multisector dynamic model complemented by the gcmdat R package, which serves as its data processing system (Bond-Lamberty et al., 2019). Particularly, gcmdat includes modules (data processing chunks) and functions to convert raw data inputs into hundreds of XML input files used by GCAM. As an illustration, in the latest GCAM version, GCAM v7.0, about 280 XML files, with a combined size of 4.1 GB, are generated. Although AgLU-related XMLs represent only about 10% of the total number of files, they contribute over 50% in size (~2.1 GB). The majority of AgLU-related data, whether directly or indirectly, rely on raw data sourced from FAOSTAT.

Nonetheless, the FAOSTAT data employed within gcmdat has traditionally involved manual downloads and may have undergone preprocessing. In light of the increasing data needs, maintaining the FAOSTAT data processing tasks in gcmdat has become increasingly challenging. In addition, the processing of FAOSTAT data in the AgLU modules of gcmdat is tailored specifically for GCAM. Consequently, the integration of FAOSTAT data updates has proven to be a non-trivial task, and the data processed by the AgLU module has limited applicability in other modeling contexts (Zhao & Wise, 2023). The gcmafaoostat package aims to address these limitations (Figure 1). The targeted approach incorporates data preparation, processing, and synthesis capabilities within a dedicated package, gcmafaoostat, while regional and sectoral aggregation functions in the model data system are implemented using standalone routines within the gcmdat package. This strategy not only ensures the streamlined operation of gcmafaoostat but also contributes to keeping data systems lightweight and more straightforward to maintain.

Figure 1: New framework of utilizing FAOSTAT data in GCAM and similar large-scale models through gcamfaostat. Modules with identifier "xfaostat" only exist in gcamfaostat. The AgLU-related modules ("aglu") that rely on outputs from gcamfaostat can run in both packages. Other gcamdata modules that process data in such areas as energy, emissions, water, and socioeconomics only exist in gcamdata.

Key functions

Here we describe key functions included in gcamfaostat (v1.0.0) focusing on the data preparing and processing. More details about functions and examples for data tracing, visualization and other capabilities are illustrated in the online User Guide.

The architecture of gcamfaostat processing modules is depicted in Figure 2. This framework currently comprises eight preprocessing modules and nine processing and synthesizing modules, generating twelve output files tailored for GCAM v7.0. Each module is essentially an R function with well-defined inputs and outputs.

Note that by default, the preprocessed FAOSTAT data, i.e., outputs of the xfaostat_L101_* modules, have been stored in the Prebuilt Data of the package. gcamfaostat includes a function to generate metadata (gcamfaostat_metadata). It accesses both the latest FAOSTAT metadata and local data information and returns a summary table including the dataset information needed for gcamfaostat. There are also functions to download FAOSTAT raw data from either a remote archive (FF_download_RemoteArchive) or directly from FAOSTAT (FF_download_FAOSTAT).

To showcase the flexibility and expandability of our package, we also incorporated two AgLU modules (from gcamdata) that exemplify the data aggregation processes, e.g., across regions, sectors, and time. More importantly, the driver_drake() function plays a pivotal role by executing all available data processing modules, thereby generating both intermediate and final outputs, which are vital components of our comprehensive data processing pipeline. The function was inherited from gcamdata and it uses the drake (Landau, 2018) pipeline framework, which simplifies module updates, data tracing, and results visualization process.

Finally, data development is never a once and for all task, and continued efforts are needed to sustain and improve the processing procedures. Future work and community contribution are also detailed in the online User Guide.

Acknowledgements

This research was supported by the US Department of Energy, Office of Science, as part of research in MultiSector Dynamics, Earth and Environmental System Modeling Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. Dr. Calvin is currently detailed to the National Aeronautics and Space Administration. Dr. Calvin’s contributions to this article occurred prior to her detail. The views expressed are her own and do not necessarily represent the views of the National Aeronautics and Space Administration or the United States Government. We extend our sincere appreciation to Matthew Binsted and Page Kyle for their invaluable contributions.

References


