The Journal of Open Source Software

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 A. Wise** [1], **Katherine V. Calvin** [1], **Kanishka Narayan** [1], **and Christopher R. 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.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Hugo Ledoux

Reviewers:

- @klau506
- @HenriKajasilta

Submitted: 10 November 2023
Published: unpublished

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., 2020; Zhao, Calvin, Wise, Patel, et al., 2021). 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 (Chepeliev, 2022). It is noteworthy that each agroeconomic modeling team typically develops its own assumptions and methods to prepare and process FAOSTAT data (Bond-Lamberty et al., 2019). While largely overlooked, the uncertainty in the base data calibration approach likely contribute to the disparities in model outcomes (Lampe et al., 2014; Zhao, Calvin, Wise, & Iyer, 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. `gcamfaostat` bridges a crucial gap in the literature by offering several key features and capabilities.

1. **Transparency and Reproducibility**: **`gcamfaostat`** 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. We have documented and demonstrated the use of the package in generating and updating agroeconomic data needed for GCAM v7 (Bond-Lamberty et al., 2023).

2. **Expandability and Consistency**: **`gcamfaostat`** 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, **`gcamfaostat`** 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

Figure 1 shows a standard framework of using FAOSTAT data in GCAM. GCAM is a widely recognized global economic and multisector dynamic model complemented by the `gcamdata` R package, which serves as its data processing system. Particularly, `gcamdata` includes modules (data processing chunks) and functions to convert raw data inputs into hundreds of XML input files used by GCAM (Bond-Lamberty et al., 2019). As an illustration, in the latest GCAM version, GCAM v7 (Bond-Lamberty et al., 2023), 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 `gcamdata` has traditionally involved manual downloads and may have undergone preprocessing. In light of the increasing data needs, maintaining the FAOSTAT data processing tasks in `gcamdata` has become increasingly challenging. In addition, the processing of FAOSTAT data in the AgLU modules of `gcamdata` 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 **`gcamfaostat`** package aims to address these limitations (Figure 2). The targeted approach incorporates data preparation, processing, and synthesis capabilities within a dedicated package, **`gcamfaostat`**, while regional and sectoral aggregation functions in the model data system are implemented using standalone routines within the `gcamdata` package. This strategy not only ensures the streamlined operation of **`gcamfaostat`** but also contributes to keeping model data system lightweight and

more straightforward to maintain.

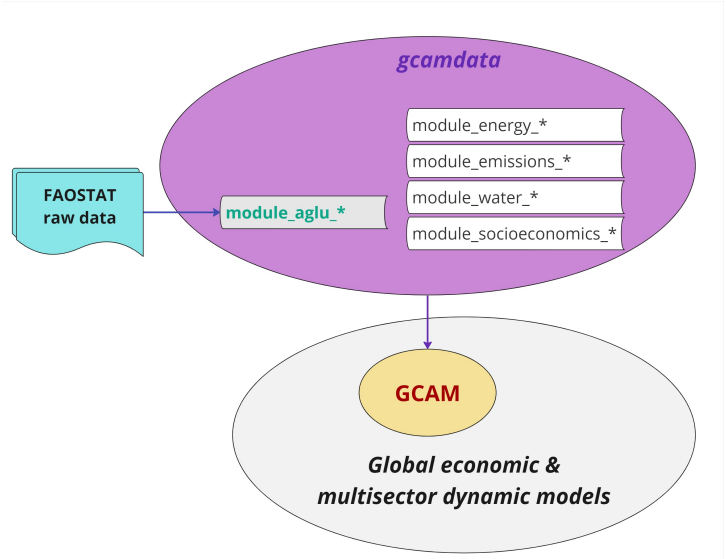

Figure 1: Original framework of utilizing FAOSTAT data in GCAM and similar large-scale models. Note that FAOSTAT data is mainly processed in the AgLU modules in gcamdata while there could be interdependency across data processing modules.

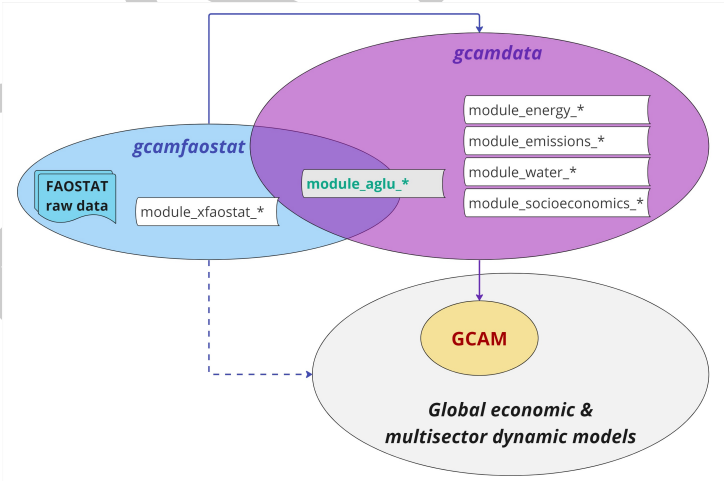

Figure 2: New framework of utilizing FAOSTAT data in GCAM and similar large-scale models through gcamfaostat. Modules with identifier "*xfaostat*" only exist in gcamfaostat. The AgLU-related modules ("*aglu*") that rely on outputs from gcamfaostat can run in both packages. Other gcamdata modules that process data in such areas as energy, emissions, water, and socioeconomics only exist in gcamdata.

Key functions

In this section we describe key functions included in **`gcamfaostat (v1.0.0)`**. More details about the functions and documentations can be found in the online **User Guide**.

Data preparation

`gcamfaostat` includes functions to generate metadata (`gcamfaostat_metadata`) and download FAOSTAT raw data from either a remote archive (`FF_download_RemoteArchive`) or directly

from FAOSTAT (`FF_download_FAOSTAT`).

`gcamfaostat_metadata()`

- The function accesses both the latest FAOSTAT metadata and local data information and returns a summary table including the dataset information needed for **gcamfaostat** (see Table 1 below).
- The function will save the latest FAOSTAT metadata to the metadata_log
- The dataset code needed were specified in the function to get a subset of the FAOSTAT metadata. The function will return only dataset code required when setting `OnlyReturnDatasetCodeRequired = FALSE`.
- The function will check whether FAOSTAT raw data exists locally (`Exist_Local`) and in Prebuilt Data (`Exist_Prebuilt`). If `Exist_Prebuilt` is `TRUE` for all dataset, the package is ready to be built based on the Prebuilt package data.
- `FAO update data` and `FAO size` indicate the information based on the latest FAOSTAT metadata.

- Users can use `FF_rawdata_info()` function to download nonexist raw data from a remote archive or FAOSTAT.

Table 1. FAOSTAT dataset processed in **gcamfaostat v1.0.0**.

Dataset Code	Dataset Name	Exist_Local	Exist_Prebuilt	FAO update date	FAO size
CB	Food Balances: Commodity Balances (non-food) (2010-)	TRUE	TRUE	8/25/2022	1MB
FBSH	Food Balances: Food Balances (-2013, old methodology and population)	TRUE	TRUE	3/10/2023	69MB
TM	Trade: Detailed trade matrix	TRUE	TRUE	2/14/2022	454MB
OA	Population and Employment: Annual population	TRUE	TRUE	10/24/2022	2MB
FO	Forestry: Forestry Production and Trade	TRUE	TRUE	9/5/2023	15MB
QCL	Production: Crops and livestock products	TRUE	TRUE	3/22/2023	29MB
PD	Prices: Deflators	TRUE	TRUE	8/16/2023	1MB
TCL	Trade: Crops and livestock products	TRUE	TRUE	8/14/2023	229MB
FBS	Food Balances: Food Balances (2010-)	TRUE	TRUE	5/4/2023	50MB
RFN	Land, Inputs and Sustainability: Fertilizers by Nutrient	TRUE	TRUE	7/5/2023	2MB
RL	Land, Inputs and Sustainability: Land Use	TRUE	TRUE	7/10/2023	2MB
PP	Prices: Producer Prices	TRUE	TRUE	2/23/2023	10MB
SCL	Food Balances: Supply Utilization Accounts (2010-)	TRUE	TRUE	4/26/2023	59MB

Data processing

Module structure

The architecture of **gcamfaostat** processing modules is depicted in Figure 3. This framework currently comprises eight preprocessing modules and nine processing and synthesizing modules, generating twelve output files tailored for GCAM v7. Each module is essentially an R function

with well-defined inputs and outputs. 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. Moreover, 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.

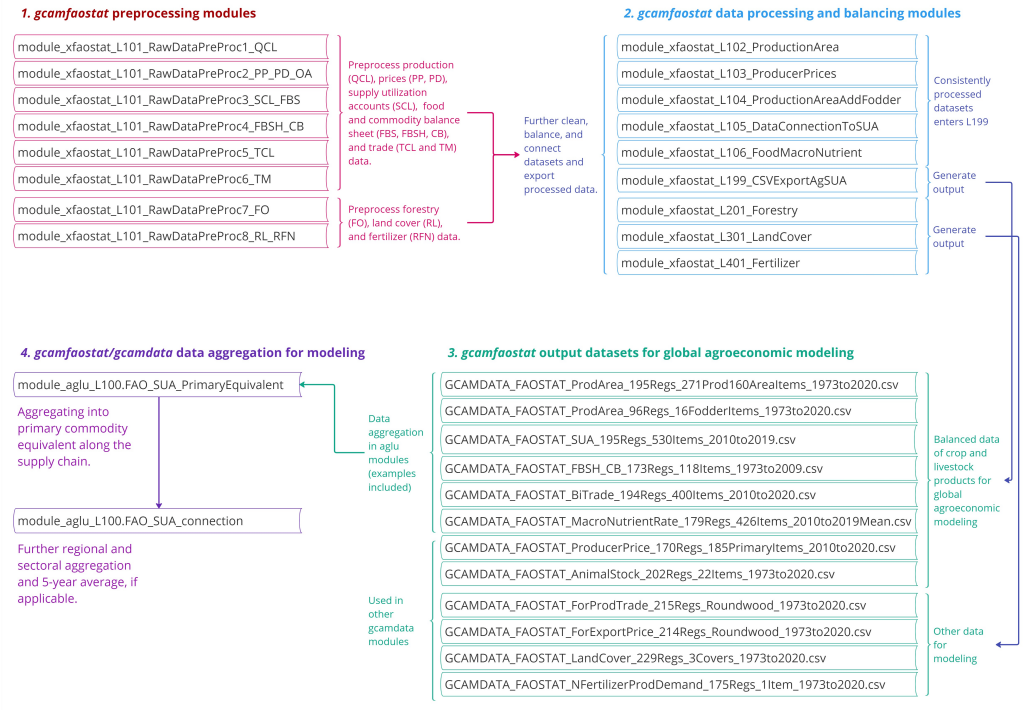

Figure 3: Data processing architecture in gcamfaostat.

Drive the modules

`driver_drake()`

- The function runs data processing modules sequentially to generate intermediate data outputs and final output (e.g., csv or other files) for GCAM (`gcamdata`) or other models.
- The function is inherited from `gcamdata` and it uses the drake (Landau, 2018) pipeline framework, which simplifies module updates, data tracing, and results visualization process.
- It stores the outputs in a drake cache so that when the function is run again, it skips the steps that are up-to-date.
- In `constants.R`, users can set `OUTPUT_Export_CSV = TRUE` and specify the output directory (`DIR_OUTPUT_CSV`) to export and store the output csv files (currently the default option for GCAM v7).

Data tracing

As **`gcamfaostat`** is built upon the foundation of `gcamdata` and leverages the powerful drake framework, it inherits functions designed for tracking data flows. Here we describe several key functions.

`info()`

- The function returns information of an object, including name, metadata information, precursors and dependents.

`load_from_cache()`

- If a drake cache is available, e.g., when `driver_drake()` had been run, this function, if given a list of object names, loads the objects from the cache into a list of data frames.
- The function `get_data_list` can be used to assign each object in the list to a data frame.

Visualization and Other capabilities

In addition to generating data for modeling purposes, we also provide illustrative examples for visualizing the key data elements. Other functions and capabilities including raw data updates and generating new outputs are discussed in Use Cases.

Future work

Data development is never a once and for all task, and continued efforts are needed to sustain and improve the processing procedures. Further improvements might include:

1. **Sustain processing functions for updated raw data**: ensuring that our processing functions remain up-to-date when raw data undergoes revisions is imperative.

2. **Evaluate and enhance assumptions**: a critical examination of the assumptions utilized in processes like interpolation, extrapolation, aggregation, disaggregation, and mapping is essential and should be an ongoing endeavor.

3. **Revise assumptions in low-quality data zones**: regions and sectors with little or low-quality data require careful consideration. We will need to adjust our assumptions when improved data becomes available.

4. **Promoting broader applications**: leveraging data processed by **`gcamfaostat`** can significantly contribute to harmonizing input data in global agroeconomic modeling. Encouraging the utilization of this data and fostering collaboration to enhance data processing is crucial.

5. **Assess sensitivity in downstream applications**: understanding the sensitivity of downstream data applications, e.g., global agroeconomic projections, to upstream data processing assumptions is crucial. This awareness empowers us to make informed decisions and refinements.

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

Bond-Lamberty, B., Dorheim, K., Cui, R., Horowitz, R., Snyder, A., Calvin, K., Feng, L., Hoesly, R., Horing, J., Kyle, G. P., & others. (2019). Gcamdata: An r package for preparation, synthesis, and tracking of input data for the GCAM integrated human-earth systems model. *Journal of Open Research Software*, *7*(1).

Bond-Lamberty, B., Patel, P., Lurz, J., kyle, P., Calvin, K., Smith, S., Snyder, A., Dorheim, K. R., Binsted, M., Link, R., Kim, S., Graham, N., Narayan, K., S., A., Feng, L., Lochner, E., Roney, C., Lynch, C., Horing, J., … Weber, M. (2023). *JGCRI/gcam-core: GCAM 7.0* (gcam-v7.0). Zenodo. https://doi.org/10.5281/zenodo.8010145

Bruckner, M., Wood, R., Moran, D., Kuschnig, N., Wieland, H., Maus, V., & Börner, J. (2019). FABIO—the construction of the food and agriculture biomass input–output model. *Environmental Science & Technology*, *53*(19), 11302–11312. https://doi.org/10.1021/acs.est.9b03554

Calvin, K. V., Snyder, A., Zhao, X., & Wise, M. (2022). Modeling land use and land cover change: Using a hindcast to estimate economic parameters in gcamland v2.0. *Geoscientific Model Development*, *15*(2), 429–447. https://doi.org/10.5194/gmd-15-429-2022

Chepeliev, M. (2022). Incorporating nutritional accounts to the GTAP data base. *Journal of Global Economic Analysis*, *7*(1), 1–43. https://doi.org/10.21642/JGEA.070101AF

Doelman, J. C., Beier, F. D., Stehfest, E., Bodirsky, B. L., Beusen, A. H. W., Humpenöder, F., Mishra, A., Popp, A., Vuuren, van D. P., Vos, de L., Weindl, I., Zeist, van W.-J., & Kram, T. (2022). Quantifying synergies and trade-offs in the global water-land-food-climate nexus using a multi-model scenario approach. *Environmental Research Letters*, *17*(4), 045004. https://doi.org/10.1088/1748-9326/ac5766

FAO. (2023). *FAOSTAT database*. https://www.fao.org/faostat/en/#data

Fujimori, S., Wu, W., Doelman, J., Frank, S., Hristov, J., Kyle, P., Sands, R., Zeist, W.-J. van, Havlik, P., Domínguez, I. P., Sahoo, A., Stehfest, E., Tabeau, A., Valin, H., Meijl, H. van, Hasegawa, T., & Takahashi, K. (2022). Land-based climate change mitigation measures can affect agricultural markets and food security. *Nature Food*, *3*(2), 110–121. https://doi.org/10.1038/s43016-022-00464-4

Graham, N. T., Iyer, G., Wild, T. B., Dolan, F., Lamontagne, J., & Calvin, K. (2023). Agricultural market integration preserves future global water resources. *One Earth*, *6*(9), 1235–1245. https://doi.org/10.1016/j.oneear.2023.08.003

Lampe, M. von, Willenbockel, D., Ahammad, H., Blanc, E., Cai, Y., Calvin, K., Fujimori, S., Hasegawa, T., Havlik, P., Heyhoe, E., Kyle, P., Lotze-Campen, H., Mason d'Croz, D., Nelson, G. C., Sands, R. D., Schmitz, C., Tabeau, A., Valin, H., Mensbrugghe, D. van der, & Meijl, H. van. (2014). Why do global long-term scenarios for agriculture differ? An overview of the AgMIP global economic model intercomparison. *Agricultural Economics*, *45*(1), 3–20. https://doi.org/10.1111/agec.12086

Landau, W. M. (2018). The drake r package: A pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *3*(21), 550. https://doi.org/10.21105/joss.00550

Ven, D.-J. van de, Mittal, S., Gambhir, A., Lamboll, R. D., Doukas, H., Giarola, S., Hawkes, A., Koasidis, K., Köberle, A. C., McJeon, H., Perdana, S., Peters, G. P., Rogelj, J., Sognnaes, I., Vielle, M., & Nikas, A. (2023). A multimodel analysis of post-glasgow climate targets and feasibility challenges. *Nature Climate Change*, *13*(6), 570–578. https://doi.org/10.1038/s41558-023-01661-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes,

A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yarlagadda, B., Wild, T., Zhao, X., Clarke, L., Cui, R., Khan, Z., Birnbaum, A., & Lamontagne, J. (2023). Trade and climate mitigation interactions create agro-economic opportunities with social and environmental trade-offs in latin america and the caribbean. *Earth's Future*, *11*(4), e2022EF003063. https://doi.org/10.1029/2022EF003063

Zhang, Y., Waldhoff, S., Wise, M., Edmonds, J., & Patel, P. (2023). Agriculture, bioenergy, and water implications of constrained cereal trade and climate change impacts. *PLOS ONE*, *18*(9), e0291577. https://doi.org/10.1371/journal.pone.0291577

Zhao, X., Calvin, K. V., & Wise, M. A. (2020). The critical role of conversion cost and comparative advantage in modeling agricultural land use change. *Climate Change Economics*, *11*(01), 2050004. https://doi.org/10.1142/s2010007820500049

Zhao, X., Calvin, K. V., Wise, M. A., & Iyer, G. (2021). The role of global agricultural market integration in multiregional economic modeling: Using hindcast experiments to validate an armington model. *Economic Analysis and Policy*, *72*, 1–17. https://doi.org/10.1016/j.eap.2021.07.007

Zhao, X., Calvin, K. V., Wise, M. A., Patel, P. L., Snyder, A. C., Waldhoff, S. T., Hejazi, M. I., & Edmonds, J. A. (2021). Global agricultural responses to interannual climate and biophysical variability. *Environmental Research Letters*, *16*(10), 104037. https://doi.org/10.1088/1748-9326/ac2965

Zhao, X., & Wise, M. (2023). *Core model proposal# 360: GCAM agriculture and land use (AgLU) data and method updates: Connecting land hectares to food calories. PNNL-34313.* https://jgcri.github.io/gcam-doc/cmp/360_AgLU_data_and_methods.pdf