The Journal of Open Source Software

HySetter: A Python Package for Reproducible Hydroclimate Data Subsetting over CONUS

Taher Chegini [1]

1 Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, US

DOI: 10.21105/joss.08301

Software

- Review
- Repository
- Archive

Editor: James Gaboardi

Reviewers:

- @mhweber
- @Zeitsperre

Submitted: 31 March 2025
Published: 18 July 2025

Summary

Hydrological modeling requires integrating diverse hydroclimate datasets, which often creates a significant technical burden for researchers. HySetter addresses this challenge by providing a command-line interface for automated subsetting of hydroclimate data across the conterminous United States (CONUS). Built upon the HyRiver software stack, HySetter streamlines data acquisition through a unified configuration file that eliminates writing complex scripts for different data sources. By centralizing the workflow, HySetter enables researchers to focus on scientific questions rather than data preparation, while enhancing the reproducibility of hydrological research.

Statement of Need

Hydrological research demands integration of diverse datasets spanning multiple scales. The acquisition and processing of these datasets presents several challenges: (1) Traditional approaches require complex scripts for different data sources, leading to fragmented workflows; (2) Researchers often spend more time on data collection than scientific analysis; (3) Technical overhead impacts novice researchers with limited programming skills; (4) Experienced researchers develop redundant code for data acquisition across projects.

These challenges undermine the adoption of FAIR principles (Wilkinson et al., 2016) in hydrological research. While tools like HydroMT (Eilander et al., 2023) address similar challenges for model building, a comparable solution for hydroclimate data acquisition for CONUS is lacking.

HySetter fills this gap by providing a user-friendly framework requiring minimal technical expertise while supporting advanced applications. Through a YAML configuration file and command-line interface, HySetter democratizes access to hydroclimate data for researchers of varying technical backgrounds.

Functionality

HySetter employs a streamlined workflow in four steps: (1) The user defines the area of interest in a YAML configuration file using HUCs, GAGES-II basins, NHDPlusV2 catchments, or user-defined geometries; (2) The user specifies desired datasets and parameters in the same file; (3) HySetter processes this configuration to access data through HyRiver; (4) Data is standardized and saved in analysis-ready formats.

This approach accommodates different technical backgrounds: novice users can create simple configuration files without coding, while experienced users can programmatically generate configurations for complex analyses.

Figure 1 shows a typical configuration file, and Figure 2 demonstrates HySetter's command-line interface with progress indicators.

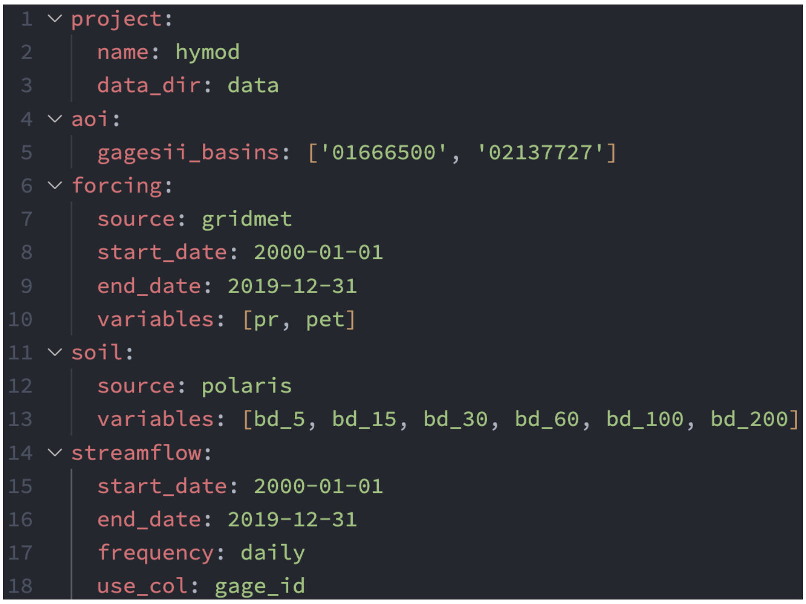

Figure 1: Example YAML configuration file for HySetter specifying project parameters, area of interest, and data sources.

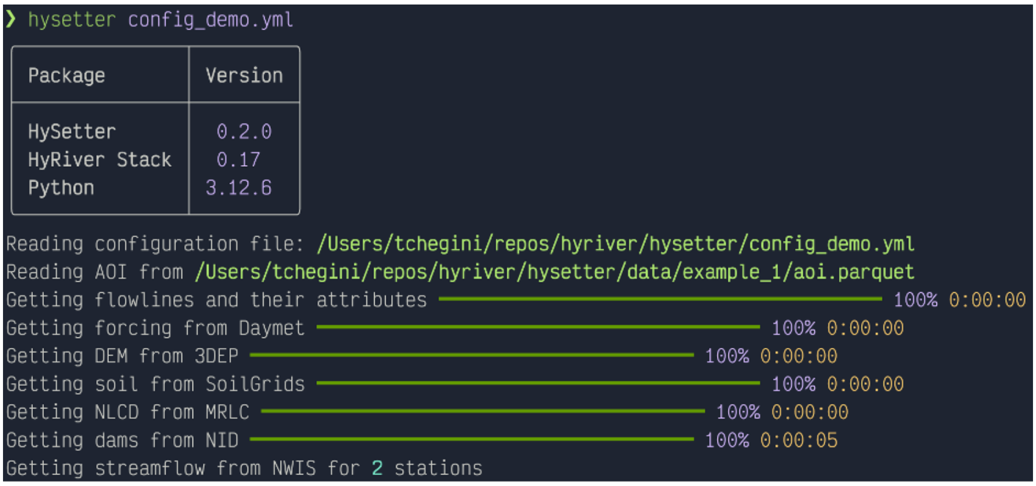

Figure 2: HySetter command-line interface showing real-time progress during data acquisition processes.

HySetter currently supports subsetting from the following data sources:

- **Area of Interest**: HUC watersheds, GAGES-II basins, NHDPlusV2 catchments and their attributes (StreamCat and NLDI), or user-defined geometries
- **Drainage Network**: NHDPlusV2
- **Climate Forcing**: Daymet, GridMET, NLDAS2
- **Streamflow**: National Water Information System (NWIS)
- **Soil**: gNATSGO, SoilGrids
- **Topography**: 3D Elevation Program (3DEP)
- **Dams**: National Inventory of Dams (NID)

- **Land Use/Land Cover**: Multi-Resolution Land Characteristics (MRLC) products including land cover, canopy, imperviousness, and urban descriptors

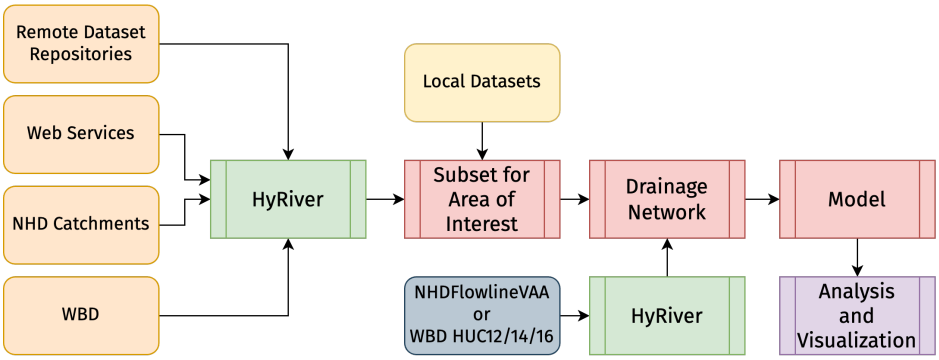

Figure 3: HySetter workflow showing data flow from remote dataset repositories and web services through HyRiver to create area-specific datasets for hydrological modeling.

Implementation

HySetter's workflow is illustrated in Figure 3, showing data flow from various sources through HyRiver to create subsets for specific areas. HySetter builds upon the HyRiver stack (Chegini et al., 2021), which provides access to numerous hydroclimate web services. While HyRiver requires Python programming knowledge, HySetter adds an abstraction layer that translates user-friendly configuration files into appropriate function calls.

The implementation follows modular design with three components: (1) **Configuration Parser** processes YAML files to extract user requirements; (2) **Data Acquisition Engine** interfaces with HyRiver to retrieve data; (3) **Output Manager** standardizes and exports processed data for analysis.

HySetter provides detailed feedback through progress bars and messages during data acquisition, helping users track operations and diagnose issues. The software includes extensive error handling to manage challenges in accessing remote services.

Advanced users can utilize HySetter as a Python library with a user-friendly API for further post-processing operations and integration with other libraries.

Research Applications

HySetter has demonstrated utility in: (1) **Hydrological Modeling**: Implementation and calibration of models across watersheds, leveraging consistent input datasets for parameterization; (2) **Large-Scale Studies**: Analysis of regional patterns requiring consistent data inputs across numerous basins; (3) **Educational Contexts**: Supporting classroom exercises where students focus on hydrological processes rather than data acquisition techniques.

The standardized, reproducible nature of HySetter workflows ensures research outputs can be easily validated and extended, promoting transparency and collaboration. By abstracting data retrieval complexities, HySetter enables researchers to focus on answering critical hydrological questions rather than data preparation.

Comparison with Existing Tools

Tools like HydroMT (Eilander et al., 2023) provide frameworks for model building and analysis, while SWAT+AW (Chawanda et al., 2020) offers automated workflows for the SWAT model.

However, these tools primarily target model building rather than standardized data acquisition.

HySetter differentiates itself by focusing specifically on the data acquisition stage, offering a streamlined solution for accessing hydroclimate data across CONUS. HySetter is tailored to datasets commonly used in hydrological studies over CONUS, providing optimized workflows. Its YAML-based configuration approach requires minimal technical expertise, making it accessible to users with limited programming experience.

Conclusions and Future Work

HySetter addresses a critical gap in hydrological modeling by providing a reproducible framework for hydroclimate data acquisition across CONUS. By simplifying access to diverse data sources through a unified configuration approach, it enables researchers to focus on scientific questions rather than technical challenges.

Future development plans include: (1) Support for additional data sources based on community needs; (2) Enhanced tools for large-scale studies requiring data acquisition over extensive areas; (3) Additional documentation examples demonstrating HySetter's application in diverse hydrological contexts.

Acknowledgements

This work was supported by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) through the Hydroinformatics Innovation Fellowship program.

References

Chawanda, C. J., George, C., Thiery, W., Griensven, A. van, Tech, J., Arnold, J., & Srinivasan, R. (2020). User-friendly workflows for catchment modelling: Towards reproducible SWAT+ model studies. *Environmental Modelling & Software*, *134*, 104812. https://doi.org/10.1016/j.envsoft.2020.104812

Chegini, T., Li, H.-Y., & Leung, L. R. (2021). HyRiver: Hydroclimate Data Retriever. *Journal of Open Source Software*, *6*(66), 1–3. https://doi.org/10.21105/joss.03175

Eilander, D., Boisgontier, H., Bouaziz, L. J. e., Buitink, J., Couasnon, A., Dalmijn, B., Hegnauer, M., Jong, T. de, Loos, S., Marth, I., & Verseveld, W. van. (2023). HydroMT: Automated and reproducible model building and analysis. *Journal of Open Source Software*, *8*(83), 4897. https://doi.org/10.21105/joss.04897

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., & others. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18