The Journal of Open Source Software

SubsetTools: A Python package to subset data to build and run ParFlow hydrologic models

Amanda K. Triplett [3*¶], **Georgios Artavanis** [1,2*¶], **William M. Hasling**[1,2], **Reed M. Maxwell** [2,5,6¶], **Amy Defnet** [1,2], **Amy M. Johnson**[4], **William Lytle**[3], **Andrew Bennett**[3], **Elena Leonarduzzi**[2], **Lisa K. Gallagher**[2,5], and **Laura E. Condon** [3¶]

1 Research Software Engineering, Princeton University, USA **2** Integrated GroundWater Modeling Center, Princeton University, USA **3** Department of Hydrology and Atmospheric Sciences, University of Arizona, USA **4** CyVerse, USA **5** Department of Civil and Environmental Engineering, Princeton University, USA **6** High Meadows Environmental Institute, Princeton University, USA ¶ Corresponding author * These authors contributed equally.

DOI: 10.21105/joss.06752

Software

- Review
- Repository
- Archive

Editor: Taher Chegini

Reviewers:

- @JannisHoch
- @dvalters

Submitted: 15 March 2024
Published: 17 July 2024

Summary

Hydrologic models are an integral part of understanding and managing water supply. There are countless hydrologic models available that differ in their complexity, scale and focus on different parts of the hydrologic cycle. ParFlow is a fully integrated, physics-based model that simulates surface and subsurface flow simultaneously (Ashby & Falgout, 1996; Jones & Woodward, 2001; Kollet & Maxwell, 2006; Maxwell, 2013). ParFlow is also coupled with a land surface model which allows it to simulate the full terrestrial hydrologic cycle from bedrock to treetops (Kollet & Maxwell, 2008; Maxwell & Miller, 2005). It has been applied to a myriad of watersheds across the US and around the world to answer questions of water supply and groundwater–surface water interactions.

ParFlow is a scientifically rigorous hydrologic model; however, its application by the broader community has been limited to a degree by its technical complexity which creates a high barrier to entry for new users. Intensive training and hydrologic expertise is required to appropriately build a ParFlow model from scratch.

`SubsetTools` is a Python package that seeks to lower the barrier to entry by allowing a user to subset published and verified ParFlow inputs and model configurations to build their own watershed models. These tools allow a user to set up and run a model in a matter of minutes, rather than weeks or months. `SubsetTools` is designed to interface with two domains covering the contiguous United States (CONUS), CONUS1 (Maxwell et al., 2015, 2015; O'Neill et al., 2021) and CONUS2 (Yang et al., 2023). These domains determine the structure and attributes of the hydrogeologic inputs used to build the ParFlow model. `SubsetTools` is the first package of its kind to fetch and process all necessary inputs and create a functional ParFlow model, all in a single workflow.

Statement of need

There are three primary barriers to building a hydrologic model from scratch. `SubsetTools` helps to resolve them in the following ways:

Barrier one: Finding quality data and then using it within a model framework is challenging. It requires significant time and expertise to assemble and process all of the input datasets that a model requires.

Solution: Our team has spent years developing a national geofabric for the ParFlow CONUS simulations (Maxwell et al., 2015; Yang et al., 2023). We conducted large data assembly and analysis projects to develop hydrologically consistent topographic datasets (Zhang et al., 2021) and spatially consistent and continuous hydrostratigraphy (Tijerina-Kreuzer et al., 2024). Rather than repeating this effort, `SubsetTools` users can start from all of the input datasets that have already been developed and tested for hydrologic consistency. This assures that model inputs have the correct format, units, spatial resolution, and orientation to run a new subset model.

Barrier two: It requires modeling expertise to set up a ParFlow run script. A run script often includes more than a hundred input keys and parameters that need to be configured for a simulation to run smoothly.

Solution: We have multiple working model configurations already developed for our national platform and can easily adapt these scripts for watershed simulation. Subset tools users are handed a working script and several tutorials on how to modify this script for a range of modeling scenarios.

Barrier 3: Groundwater models require a very long initialization known as 'spinup' to develop a steady state groundwater configuration. This has to be completed before any transient simulations are run and can require significant computational resources.

Solution: Because we have already developed steady state conditions at the national level for CONUS1 and CONUS2 (Maxwell et al., 2015; Yang et al., 2023), users of `SubsetTools` can start from a pre-initialized groundwater configuration. Thus they can skip the spin up step and directly run their model.

In summary, the `SubsetTools` package provides functions that simplify the process of setting up and running a ParFlow model within the Continental US. It allows the user to subset all required hydrogeologic and climate forcing datasets from `HydroData`. It also provides template model runscripts which are designed to link seamlessly with functions that edit the model keys corresponding to the domain and model configuration specified by the user. These features enable a more rapid and replicable application of ParFlow for hydrologic simulations.

`SubsetTools` is designed to be used by both hydrology students, researchers and practitioners. For students, the functions and examples provided in the package can be run with little programming or hydrologic knowledge to start teaching concepts. However, the functions have been thoughtfully designed to be flexible and transparent so that more advanced users can develop customized workflows that meet their modeling needs. `SubsetTools` has already been used at a workshop for high school students at the Watershed Institute and as part of a graduate course in hydrology at Princeton University.

Functionality

The source code for `SubsetTools` is available on GitHub and the entire package is covered under the MIT License. The documentation for the package is hosted on ReadTheDocs and includes installation instructions, short tutorials, example notebooks, a complete API reference, as well as contributing guidelines. In the section below we will go through an abbreviated outline of how a user can interact with `SubsetTools` to build and run their own ParFlow model.

First, the user should supply information about what geographic area they want to model. This may be given as a hydrologic unit code (HUC) or list of HUCs, a bounding box or a single point of latitude and longitude. The user should also specify timing information such as a date range for their simulation. Finally, they should choose what dataset they would like to use in their simulation, for example CONUS1 or CONUS2.

With the information provided above, users can subset all required model input files. An example is shown below for subsetting static and climate forcing data on the CONUS2 grid for

the Upper Verde region.

```
import subsettools as st

static_paths = st.subset_static(
    ij_bounds=[815, 1200, 924, 1290], # CONUS2 grid bounds for the Upper Verde region
    dataset="conus2_domain",
    write_dir="/path/to/your/output/directory",
)

forcing_paths = st.subset_forcing(
    ij_bounds=[815, 1200, 924, 1290],
    grid="conus2",
    start="2012-10-11",
    end="2013-10-11",
    dataset="CW3E",
    write_dir="/path/to/your/output/directory",
    forcing_vars=('precipitation', 'air_temp',),
)
```

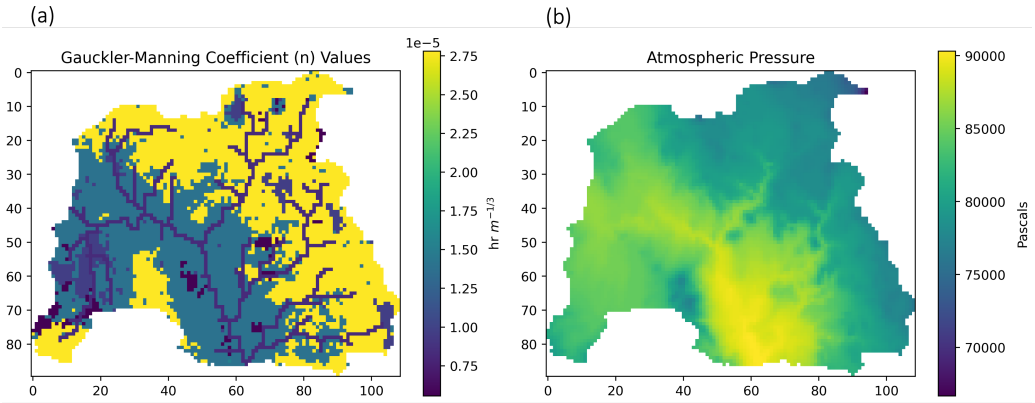

Figure 1: Two example subset outputs for HUC 15060202, the Upper Verde Watershed in Arizona. (a) shows the subset Gauckler-Manning friction coefficient, *n*, for this domain as a result of the function st.subset_static(). (b) shows atmospheric pressure, one of the forcing variables output by the function st.subset_forcing().

An appropriate run script must also be selected based on the kind of ParFlow simulation the user wants to perform. The `SubsetTools` package provides eight different templates, which can be used as a starting point for building a ParFlow model. The example function call shown below specifies a transient run using ParFlow-CLM over a solid file domain on the CONUS2 grid.

```
import subsettools as st

reference_run = st.get_template_runscript(
    grid="conus2",
    mode="transient",
    input_file_type="solid",
    write_dir="/path/to/your/output/directory"
)
runscript_path = st.edit_runscript_for_subset(
    ij_bounds=[815, 1200, 924, 1290], # CONUS2 grid bounds for the Upper Verde region
    runscript_path=reference_run,
```

```
    runname="your_runname",
    forcing_dir="/path/to/your/forcing/directory",
)
```

The `SubsetTools` package also provides functions to customize the template runscript, for example by specifying the desired subset domain to match the subset inputs, modifying the file paths of the model input files, and changing the processor topology for the ParFlow run. Once the customized Parflow runscript is ready, the user can launch a ParFlow simulation using the pftools package utilities.

Future Work

The current version of `SubsetTools` is a novel package to create and run ParFlow watershed models. However, this package remains under active development to add new features. For example, we plan to add functions to assist users in restarting a ParFlow run and heuristically estimate processor topology. Further, we plan to expand the methods by which a user can define a domain such as by the area upstream of a point or with a shapefile. Finally, we currenly only create watershed models within CONUS. The package can easily be expanded to cover other areas of the world if the appropriate model input files are provided to be hosted in `HydroData`.

Acknowledgements

This research has been supported by the U.S. Department of Energy Office of Science IDEAS-Watersheds (DE-AC02-05CH11231), the US National Science Foundation Office of Advanced Cyberinfrastructure HydroFrame projects (OAC- 2054506 and OAC-1835855) and the US National Science Foundation HydroGEN project (NSF C-A-2134892).

References

Ashby, S. F., & Falgout, R. D. (1996). A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations. *Nuclear Science and Engineering*, *124*(1), 145–159. https://doi.org/10.13182/NSE96-A24230

Jones, J. E., & Woodward, C. S. (2001). Newton-Krylov-multigrid solvers for large-scale, highly heterogenous, variably saturated flow problems. *Advances in Water Resources*, *24*(7), 763–774. https://doi.org/10.1016/S0309-1708(00)00075-0

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. *Advances in Water Resources*, *29*(7), 945–958. https://doi.org/10.1016/j.advwatres.2005.08.006

Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resources Research*, *44*(2). https://doi.org/10.1029/2007wr006004

Maxwell, R. M. (2013). A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. *Advances in Water Resources*, *53*, 109–117. https://doi.org/10.1016/j.advwatres.2012.10.001

Maxwell, R. M., Condon, L. E., & Kollet, S. J. (2015). A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. *Geoscientific Model Development*, *8*(3), 923–937. https://doi.org/10.5194/gmd-8-923-2015

Maxwell, R. M., & Miller, N. L. (2005). Development of a coupled land surface and groundwater model. *Journal of Hydrometeorology*, *6*(3), 233–247. https://doi.org/10.1175/JHM422.1

O'Neill, M. M. F., Tijerina, D. T., Condon, L. E., & Maxwell, R. M. (2021). Assessment of the ParFlow–CLM CONUS 1.0 integrated hydrologic model: Evaluation of hyper-resolution water balance components across the contiguous United States. *Geoscientific Model Development*, *14*(12), 7223–7254. https://doi.org/10.5194/gmd-14-7223-2021

Tijerina-Kreuzer, D., Swilley, J. S., Tran, H. V., Zhang, J., West, B., Yang, C., Condon, L. E., & Maxwell, R. M. (2024). Continental scale hydrostratigraphy: Basin-scale testing of alternative data-driven approaches. *Ground Water*, *62*(1), 93–110. https://doi.org/10.1111/gwat.13357

Yang, C., Tijerina-Kreuzer, D. T., Tran, H. V., Condon, L. E., & Maxwell, R. M. (2023). A high-resolution, 3D groundwater-surface water simulation of the contiguous US: Advances in the integrated ParFlow CONUS 2.0 modeling platform. *Journal of Hydrology*, *626*. https://doi.org/10.1016/j.jhydrol.2023.130294

Zhang, J., Condon, L. E., Tran, H., & Maxwell, R. M. (2021). A national topographic dataset for hydrological modeling over the contiguous United States. *Earth System Science Data*, *13*(7), 3263–3279. https://doi.org/10.5194/essd-13-3263-2021