Integration of Reproducible Methods into Community Cyberinfrastructure

David Tarboton[1,1], Tanu Malik[2,2], Jonathan Goodall[3,3], and Young-Don Choi[3,3]

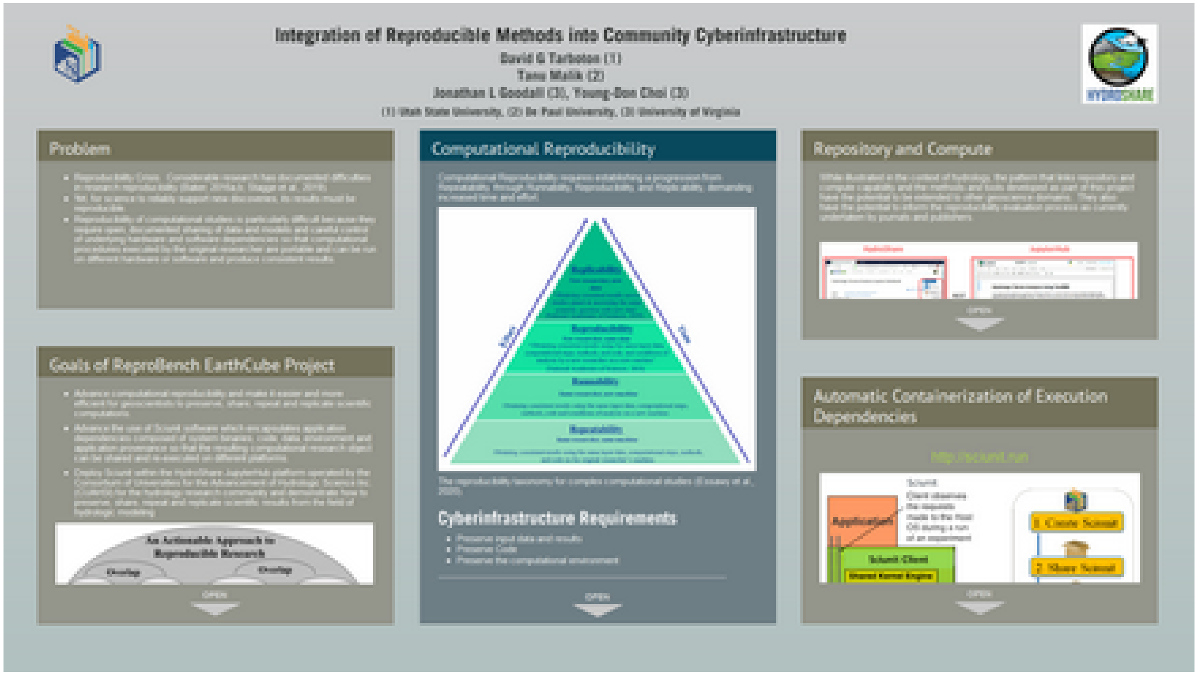
[1]Utah State University
[2]DePaul University
[3]University of Virginia

November 30, 2022

Abstract

For science to reliably support new discoveries, its results must be reproducible. This has proven to be a challenge in many fields including fields that rely on computational methods as a means for supporting new discoveries. Reproducibility in these studies is particularly difficult because they require open, documented sharing of data and models and careful control of underlying hardware and software dependencies so that computational procedures executed by the original researcher are portable and can be run on different hardware or software and produce consistent results. Despite recent advances in making scientific work more findable, accessible, interoperable and reusable (FAIR), fundamental questions in the conduct of reproducible computational studies remain: Can published results be repeated in different computing environments? If yes, how similar are they to previous results? Can we further verify and build on the results by using additional data or changing computational methods? Can these changes be automatically and systematically tracked? This presentation will describe our EarthCube project to advance computational reproducibility and make it easier and more efficient for geoscientists to preserve, share, repeat and replicate scientific computations. Our approach is based on Sciunit software developed by prior EarthCube projects which encapsulates application dependencies composed of system binaries, code, data, environment and application provenance so that the resulting computational research object can be shared and re-executed on different platforms. We have deployed Sciunit within the HydroShare JupyterHub platform operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) for the hydrology research community and will present use cases that demonstrate how to preserve, share, repeat and replicate scientific results from the field of hydrologic modeling. While illustrated in the context of hydrology, the methods and tools developed as part of this project have the potential to be extended to other geoscience domains. They also have the potential to inform the reproducibility evaluation process as currently undertaken by journals and publishers.

Integration of Reproducible Methods into Community Cyberinfrastructure



David G Tarboton (1)
Tanu Malik (2)
Jonathan L Goodall (3), Young-Don Choi (3)

(1) Utah State University, (2) De Paul University, (3) University of Virginia

HYDROSHARE

PRESENTED AT:

EarthCube

2020 EarthCube Annual Meeting
Virtual – June 18, 2020

PROBLEM

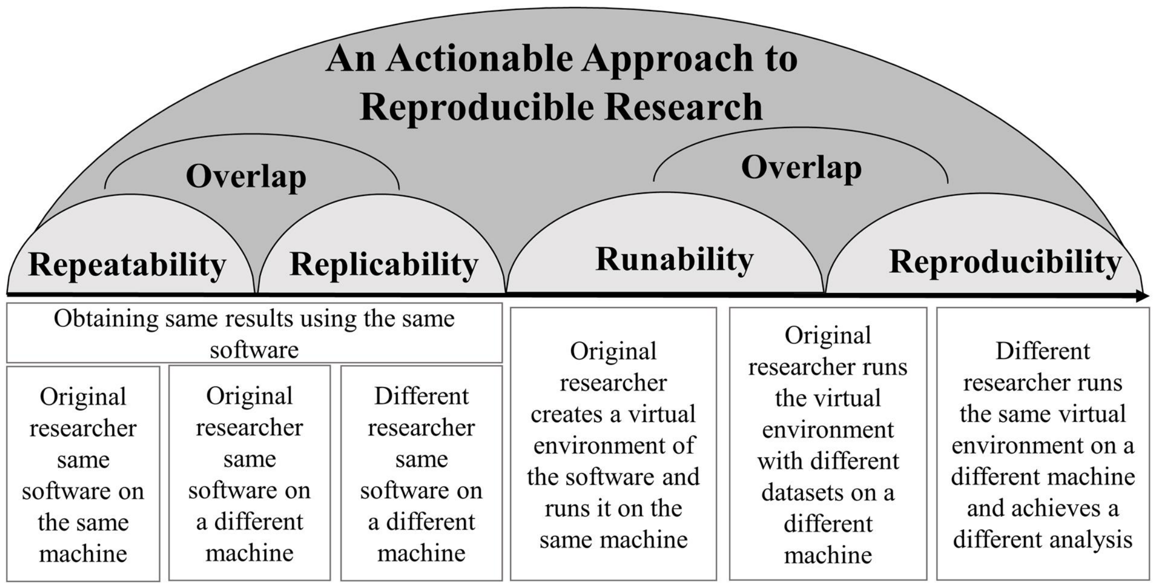
- Reproducibility Crisis. Considerable research has documented difficulties in research reproducibility (Baker, 2016a,b; Stagge et al., 2019).
- Yet, for science to reliably support new discoveries, its results must be reproducible.
- Reproducibility of computational studies is particularly difficult because they require open, documented sharing of data and models and careful control of underlying hardware and software dependencies so that computational procedures executed by the original researcher are portable and can be run on different hardware or software and produce consistent results.

GOALS OF REPROBENCH EARTHCUBE PROJECT

- Advance computational reproducibility and make it easier and more efficient for geoscientists to preserve, share, repeat and replicate scientific computations.

- Advance the use of Sciunit software which encapsulates application dependencies composed of system binaries, code, data, environment and application provenance so that the resulting computational research object can be shared and re-executed on different platforms.

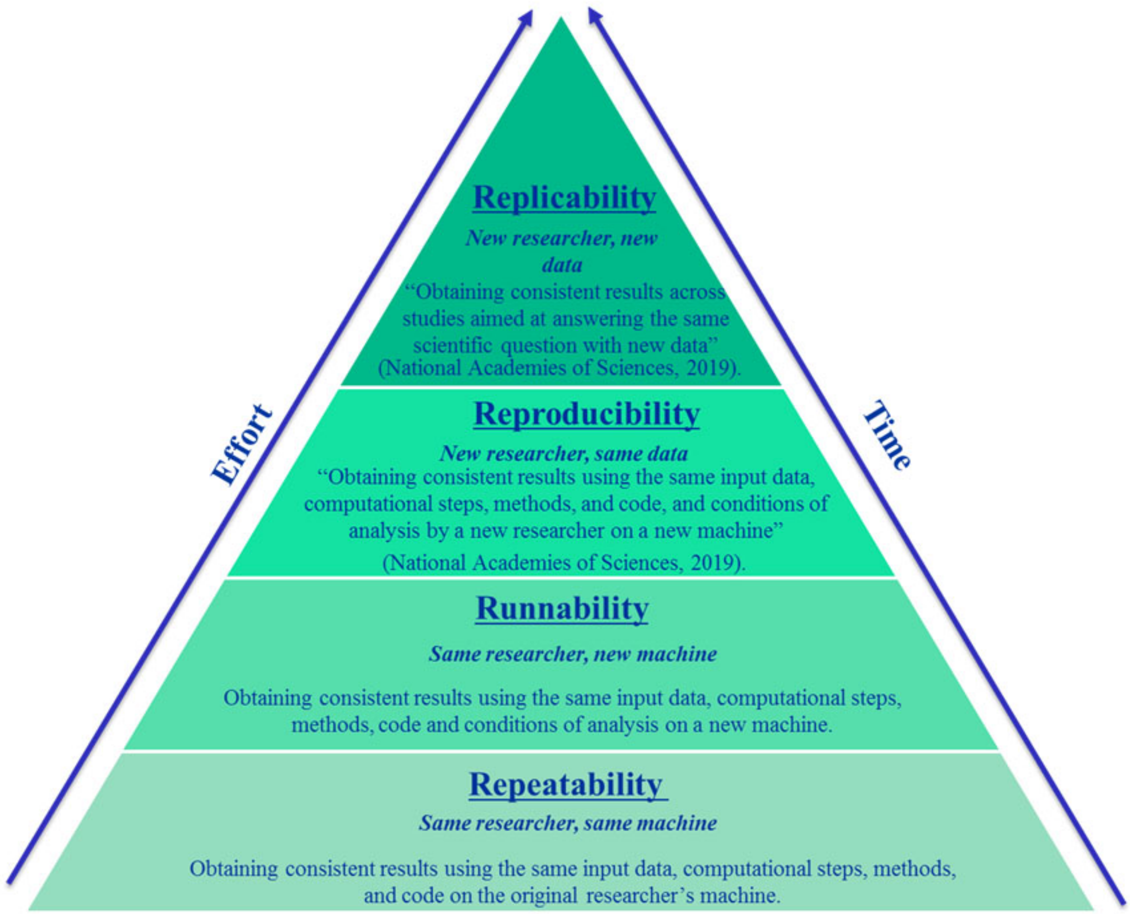
- Deploy Sciunit within the HydroShare JupyterHub platform operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) for the hydrology research community and demonstrate how to preserve, share, repeat and replicate scientific results from the field of hydrologic modeling



Actionable approach for moving geoscience workflows from the Runability to Reproducibility level.

COMPUTATIONAL REPRODUCIBILITY

Computational Reproducibility requires establishing a progression from Repeatability, through Runnability, Reproducibility, and Replicability, demanding increased time and effort.



The reproducibility taxonomy for complex computational studies (Essawy et al., 2020).

Cyberinfrastructure Requirements

- Preserve input data and results
- Preserve Code
- Preserve the computational environment

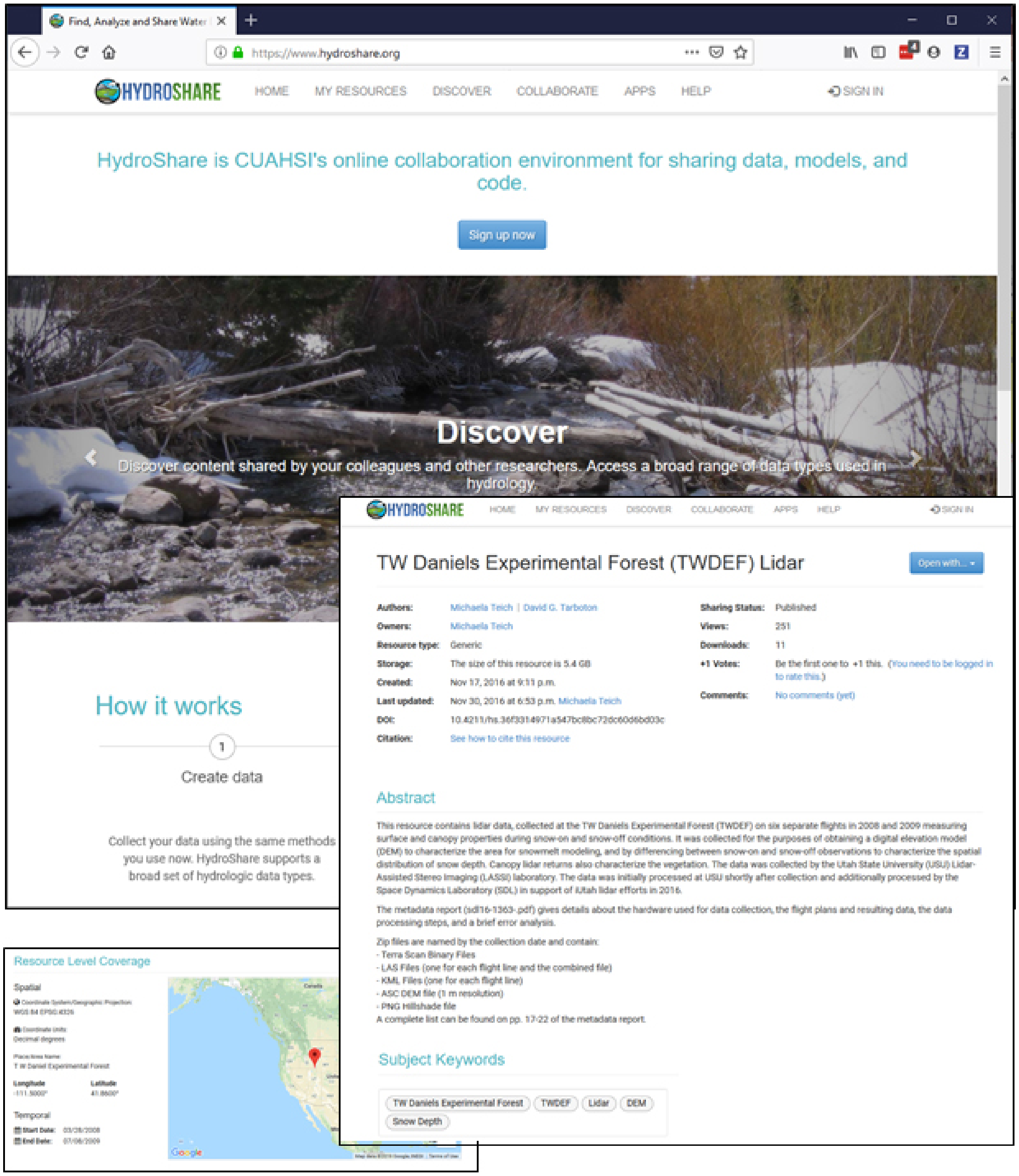
Solution

- HydroShare platform for sharing and archiving data and models
- JupyterHub compute platforms linked to HydroShare for model execution
- Sciunit Software for encapsulating computational dependencies

HydroShare Data and Model Repository

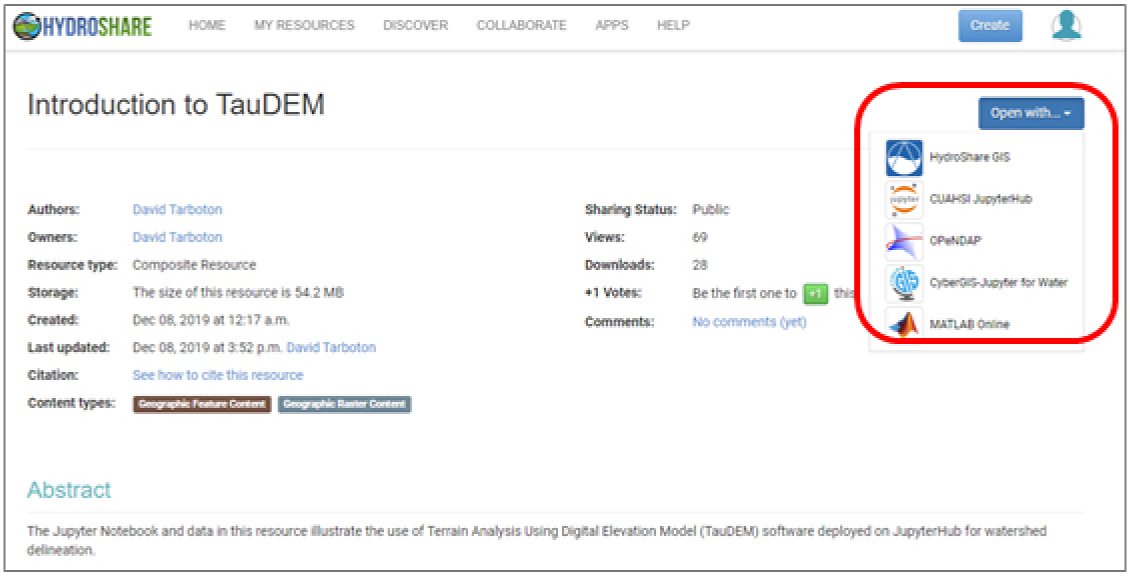
- Manage data (and models and workflows) throughout research life cycle
- Share data, models, and other research products
- Permanent publication of data and models with citable digital object identifiers (DOIs)
- Fulfill Findable, Accessible, Interoperable, Reusable (FAIR) open data mandate

http://www.hydroshare.org (http://www.hydroshare.org)

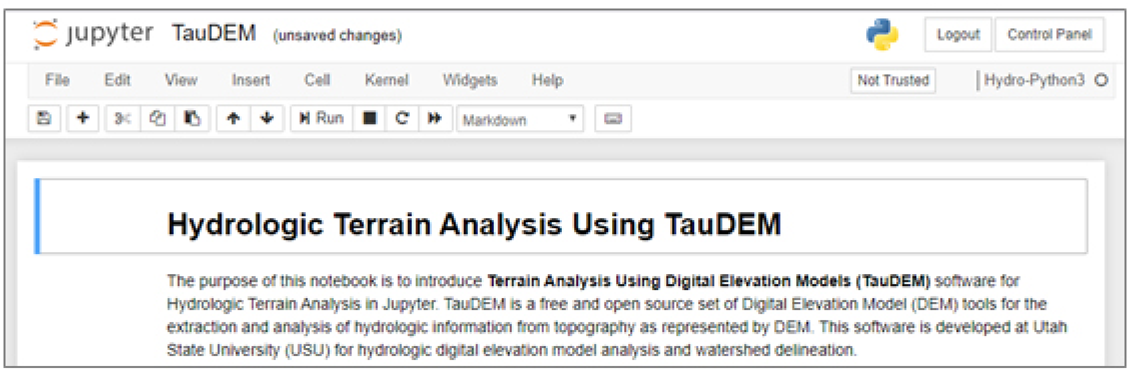


CUAHSI and CyberGIS Jupyter for Water Gateways to computing

- Provide immediate value
  - What can I do now that I may not be able to easily do on my PC
- Model input data preparation
- Model execution
- Visualization and analysis (best of practice tools)
- Reduced needs for software installation and configuration (platform independence)
- Write and execute code in a Jupyter Notebook, acting on content of HydroShare resources and saving results back to HydroShare Repository
  - Collaboration
  - Access to enhanced computation (HPC, Big data)
- Enhanced trust in research through transparency, replicability and reproducibility

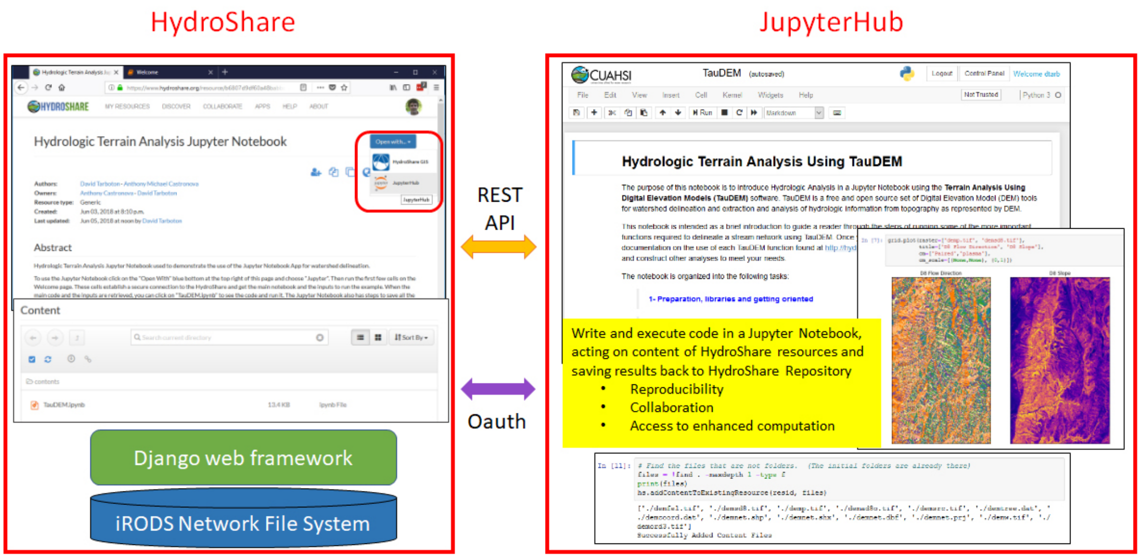

HYDROSHARE
HOME
MY RESOURCES
DISCOVER
COLLABORATE
APPS
HELP
Create
Introduction to TauDEM
Open with...
HydroShare GIS
CUAHSI JupyterHub
OPeNDAP
CyberGIS-Jupyter for Water
MATLAB Online
Authors: David Tarboton
Owners: David Tarboton
Resource type: Composite Resource
Storage: The size of this resource is 54.2 MB
Created: Dec 08, 2019 at 12:17 a.m.
Last updated: Dec 08, 2019 at 3:52 p.m. David Tarboton
Citation: See how to cite this resource
Content types: Geographic Feature Content Geographic Raster Content
Sharing Status: Public
Views: 69
Downloads: 28
+1 Votes: Be the first one to +1 this
Comments: No comments (yet)
Abstract
The Jupyter Notebook and data in this resource illustrate the use of Terrain Analysis Using Digital Elevation Model (TauDEM) software deployed on JupyterHub for watershed delineation.


js-168-155.jetstream-cloud.org/user/demo/tree/Downloads/18984997bf8f44dd99a246d4fbece903/18984997bf8f...
jupyter
Logout
Control Panel
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Downloads / 18984997bf8f44dd99a246d4fbece903 / 18984997bf8f44dd99a246d4fbece903 / data / contents
Name
Last Modified
File size
seconds ago
TauDEM.ipynb
Running
13 minutes ago
17.8 kB
logan.tif
13 minutes ago
56.8 MB
logan.vrt
13 minutes ago
1.73 kB


jupyter TauDEM (unsaved changes)
Logout
Control Panel
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Hydro-Python3
Run
Markdown
Hydrologic Terrain Analysis Using TauDEM
The purpose of this notebook is to introduce Terrain Analysis Using Digital Elevation Models (TauDEM) software for Hydrologic Terrain Analysis in Jupyter. TauDEM is a free and open source set of Digital Elevation Model (DEM) tools for the extraction and analysis of hydrologic information from topography as represented by DEM. This software is developed at Utah State University (USU) for hydrologic digital elevation model analysis and watershed delineation.

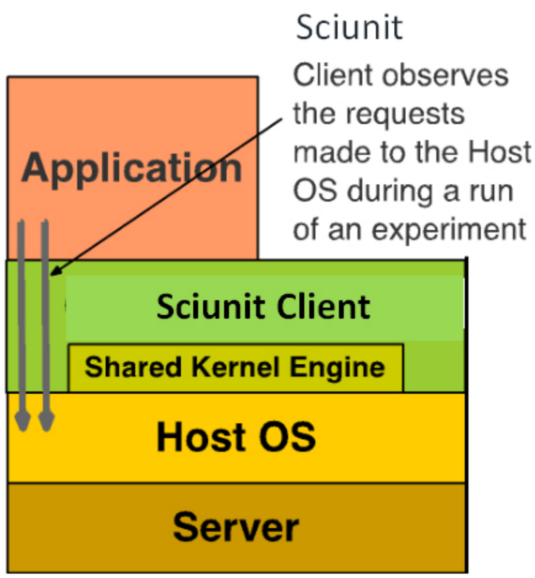
REPOSITORY AND COMPUTE

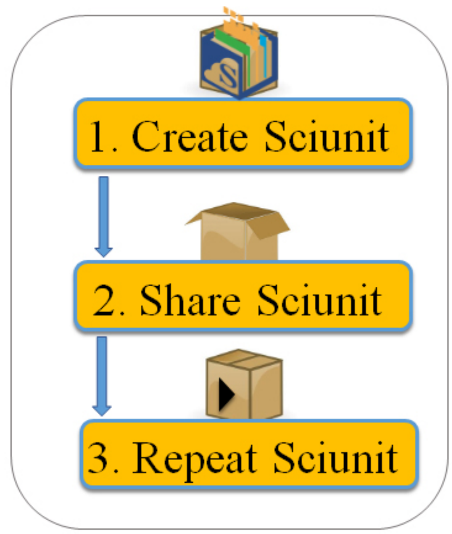
While illustrated in the context of hydrology, the pattern that links repository and compute capability and the methods and tools developed as part of this project have the potential to be extended to other geoscience domains. They also have the potential to inform the reproducibility evaluation process as currently undertaken by journals and publishers.



AUTOMATIC CONTAINERIZATION OF EXECUTION DEPENDENCIES

http://sciunit.run (http://sciunit.run)





HydroShare Jupyter Notebook Resources that illustrate the use of Sciunit for reproducibility

- CHOI, Y. (2020). Sciunit SUMMA Result Reproduction Illustration, HydroShare, http://www.hydroshare.org/resource/7d1403636fd3444c87e3c5b40b000b91 (http://www.hydroshare.org/resource/7d1403636fd3444c87e3c5b40b000b91) (This illustrates computational reproducibility using a model and computational environment encapsulated in a Sciunit stored in HydroShare. Details are described in Essawy et al., 2020)
- Choi, Y., J. Goodall, J. Sadler, A. M. Castronova, A. Bennett, T. Malik, B. Nijssen, Z. Li, S. Wang, M. Clark, D. Tarboton, M. Deeds (2020). EarthCube2020: An Approach for Open and Reproducible Environmental Modeling, HydroShare, http://www.hydroshare.org/resource/75f31565dbd24c198450b9d37c6fcf74 (http://www.hydroshare.org/resource/75f31565dbd24c198450b9d37c6fcf74) (This illustrates the cycle involving the creation of a Sciunit container, saving to HydroShare and then re-execution of that container for computational reproducibility).

ABSTRACT

For science to reliably support new discoveries, its results must be reproducible. This has proven to be a challenge in many fields including fields that rely on computational methods as a means for supporting new discoveries. Reproducibility in these studies is particularly difficult because they require open, documented sharing of data and models and careful control of underlying hardware and software dependencies so that computational procedures executed by the original researcher are portable and can be run on different hardware or software and produce consistent results. Despite recent advances in making scientific work more findable, accessible, interoperable and reusable (FAIR), fundamental questions in the conduct of reproducible computational studies remain: Can published results be repeated in different computing environments? If yes, how similar are they to previous results? Can we further verify and build on the results by using additional data or changing computational methods? Can these changes be automatically and systematically tracked? This presentation will describe our EarthCube project to advance computational reproducibility and make it easier and more efficient for geoscientists to preserve, share, repeat and replicate scientific computations. Our approach is based on Sciunit software developed by prior EarthCube projects which encapsulates application dependencies composed of system binaries, code, data, environment and application provenance so that the resulting computational research object can be shared and re-executed on different platforms. We have deployed Sciunit within the HydroShare JupyterHub platform operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) for the hydrology research community and will present use cases that demonstrate how to preserve, share, repeat and replicate scientific results from the field of hydrologic modeling. While illustrated in the context of hydrology, the methods and tools developed as part of this project have the potential to be extended to other geoscience domains. They also have the potential to inform the reproducibility evaluation process as currently undertaken by journals and publishers.

REFERENCES

Baker, M., (2016a), "1500 scientists lift the lid on reproducibility," Nature News, 533(7604): 452-454, https://doi.org/10.1038/533452a.

Baker, M., (2016b), "Muddled meanings hamper efforts to fix reproducibility crisis," Nature News, https://doi.org/10.1038/nature.2016.20076

Essawy, B.T., Goodall, J.L., Voce, D., Morsy, M.M., Sadler, J.M., Choi, Y.D., Tarboton, D.G., Malik, T., 2020. A taxonomy for reproducible and replicable research in environmental modelling. Environ. Model. Softw. 104753. https://doi.org/10.1016/j.envsoft.2020.104753

Stagge, J. H., D. E. Rosenberg, A. M. Abdallah, H. Akbar, N. A. Attallah and R. James, (2019), "Assessing data availability and research reproducibility in hydrology and water resources," Scientific Data, 6: 190030, https://doi.org/10.1038/sdata.2019.30.