

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}

¹Utah State University

²DePaul University

³University 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

The slide is titled "Integration of Reproducible Methods into Community Cyberinfrastructure" and lists authors: David G Tarboton (1), Tanu Malik (2), Jonathan L Goodall (3), and Young-Don Choi (3). It is divided into several sections:

- Problem:** Discusses reproducibility issues in research, noting that while reproducibility is important, it is often not achieved. It mentions that reproducibility is particularly difficult for large-scale data and complex workflows.
- Computational Reproducibility:** A pyramid diagram with three levels: **Reproducibility** (top), **Repeatability** (middle), and **Reliability** (bottom). The pyramid is labeled with "Data" on the left and "Code" on the right. Below the pyramid, it states: "The reproducibility taxonomy for complex computational studies" (C. Wang et al., 2016).
- Repository and Compute:** Shows a screenshot of a repository interface with a "Reproducible" label and a "Compute" button.
- Automatic Containerization of Execution Dependencies:** A diagram showing a workflow: "Application" -> "Scholar Client" -> "Container Engine" -> "Container" -> "Scholar Server".
- Goals of ReproBench EarthCube Project:** Lists goals such as advancing computational reproducibility, making it easier to use, and ensuring that research results are reproducible.
- An Actionable Approach to Reproducible Research:** A diagram showing a cycle: "Data" -> "Code" -> "Data".

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:



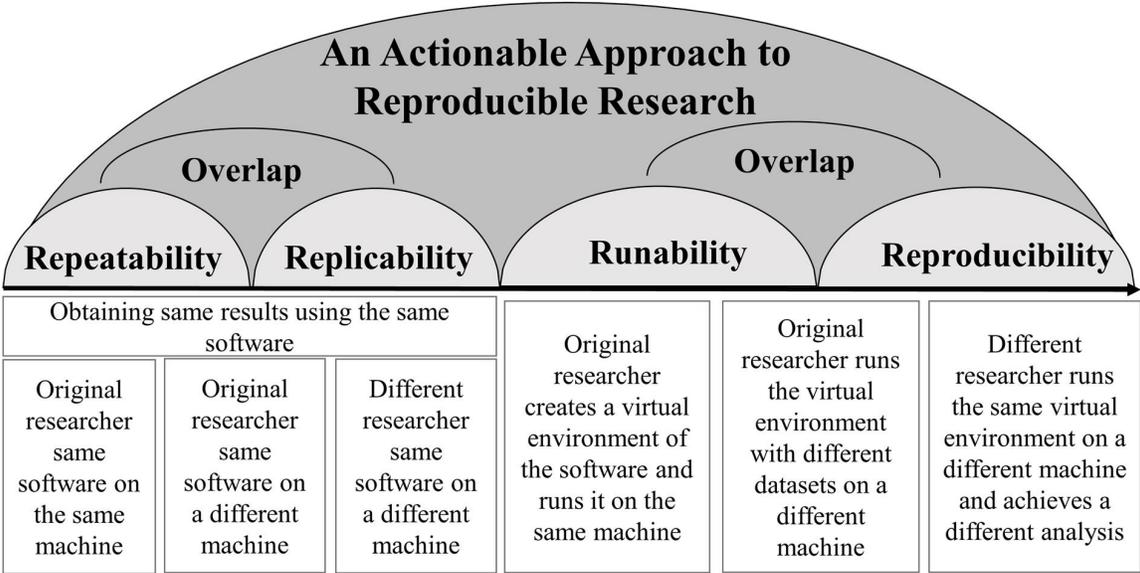
2020 EarthCube Annual Meeting
Virtual – June 18, 2020

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

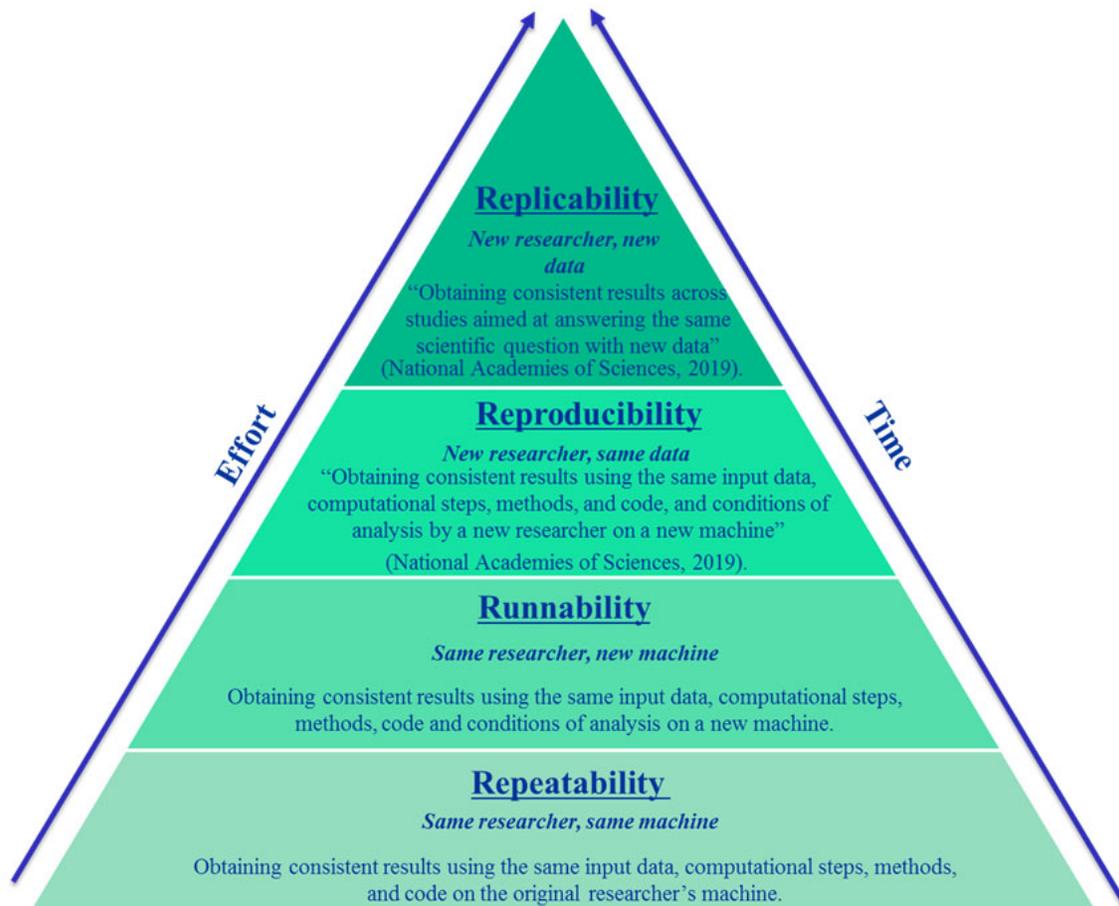
- Advance computational reproducibility and make it easier and more efficient for geoscientists to preserve, share, repeat and replicate scientific computations.
- Advance the use of Sciunit software which encapsulates application dependencies composed of system binaries, code, data, environment and 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

The screenshot displays the HydroShare website interface. At the top, the navigation bar includes 'HOME', 'MY RESOURCES', 'DISCOVER', 'COLLABORATE', 'APPS', and 'HELP', along with a 'SIGN IN' button. The main header states: 'HydroShare is CUAHSI's online collaboration environment for sharing data, models, and code.' Below this is a 'Sign up now' button. A large banner image shows a snowy landscape with the text 'Discover' and 'Discover content shared by your colleagues and other researchers. Access a broad range of data types used in hydrology.' Below the banner, there are two main sections: 'How it works' and 'Resource Level Coverage'. The 'How it works' section shows a step '1 Create data' with the text: 'Collect your data using the same methods you use now. HydroShare supports a broad set of hydrologic data types.' The 'Resource Level Coverage' section includes a map of the TW Daniels Experimental Forest and metadata: 'Spatial: Coordinate System/Geographic Projection: WGS 84 EPSG:4326; Coordinate Units: Decimal degrees; Place/area Name: TW Daniels Experimental Forest; Longitude: -111.5000°; Latitude: 43.8600°; Temporal: Start Date: 03/25/2008; End Date: 07/06/2009'. The right side of the screenshot shows a detailed view of the resource 'TW Daniels Experimental Forest (TWDEF) Lidar'. It includes metadata such as Authors (Michaela Teich | David G. Tarboton), Owners (Michaela Teich), Resource type (Generic), Storage (5.4 GB), Created (Nov 17, 2016), Last updated (Nov 30, 2016), DOI (10.4211/hs.36f3314971a547bc8bc72dc60d9bd03c), and Citation. It also shows Sharing Status (Published), Views (251), Downloads (11), and a +1 Vote. The abstract describes the lidar data collection at TWDEF in 2008 and 2009 for snowmelt modeling and vegetation characterization. The subject keywords are 'TW Daniels Experimental Forest', 'TWDEF', 'Lidar', and 'DEM', with 'Snow Depth' also listed.

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

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  this

Comments: No comments (yet)

Open with... ▾

-  HydroShare GIS
-  CUMSI JupyterHub
-  CPeNDAP
-  CyberGIS-Jupyter for Water
-  MATLAB Online

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-168-155jetstream-cloud.org/user/demo/tree/Downloads/18984997bf8f44dd99a246d4f4bece903/18984997bf8f... ☆  

 Logout Control Panel

Files Running Clusters

Select items to perform actions on them. Upload New ▾ 

0 ▾ / Downloads / 18984997bf8f44dd99a246d4f4bece903 / 18984997bf8f44dd99a246d4f4bece903 / data / contents

	Name ▾	Last Modified	File size
	TauDEM.ipynb	seconds ago	
	logan.tif	13 minutes ago	56.8 MB
	logan.vrt	13 minutes ago	1.73 kB

 **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.

HydroShare

The screenshot shows the HydroShare website interface. At the top, there's a navigation bar with 'HYDROSHARE' and links for 'MY RESOURCES', 'DISCOVER', 'COLLABORATE', 'APPS', and 'HELP'. The main content area displays 'Hydrologic Terrain Analysis Jupyter Notebook' with metadata including authors (David Tarboton, Anthony Michael Catherine), creation date (June 03, 2018), and last updated date (June 05, 2018). Below the metadata is an 'Abstract' section and a 'Content' section showing a file named 'TauDEM.ipynb' (12.4 KB, Jupyter File). At the bottom, there are two stacked boxes: a green one labeled 'Django web framework' and a blue one labeled 'iRODS Network File System'.

JupyterHub

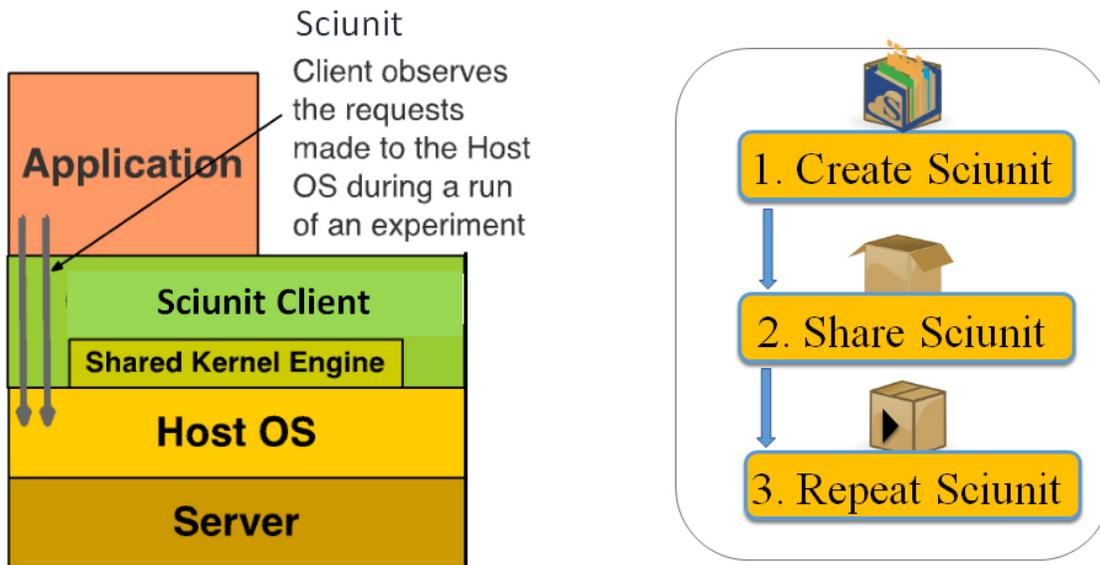
The screenshot shows the JupyterHub interface for 'TauDEM (autores)'. The top navigation bar includes 'CUAHSI', 'TauDEM (autores)', 'Logout', 'Control Panel', and 'Welcome dashboard'. The main content area displays 'Hydrologic Terrain Analysis Using TauDEM' with an introductory text and a list of tasks: '1- Preparation, libraries and getting oriented'. A yellow callout box contains the text: 'Write and execute code in a Jupyter Notebook, acting on content of HydroShare resources and saving results back to HydroShare Repository', followed by a bulleted list: '• Reproducibility', '• Collaboration', and '• Access to enhanced computation'. Below the callout is a code cell with Python code for finding files and adding content to a resource. The code includes comments and function calls like 'addContentToExistingResource' and 'Resourcefully Add Content Files'. To the right of the code, there are two small maps showing 'DR Flow Direction' and 'DR Slope'.

REST API

Oauth

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>.