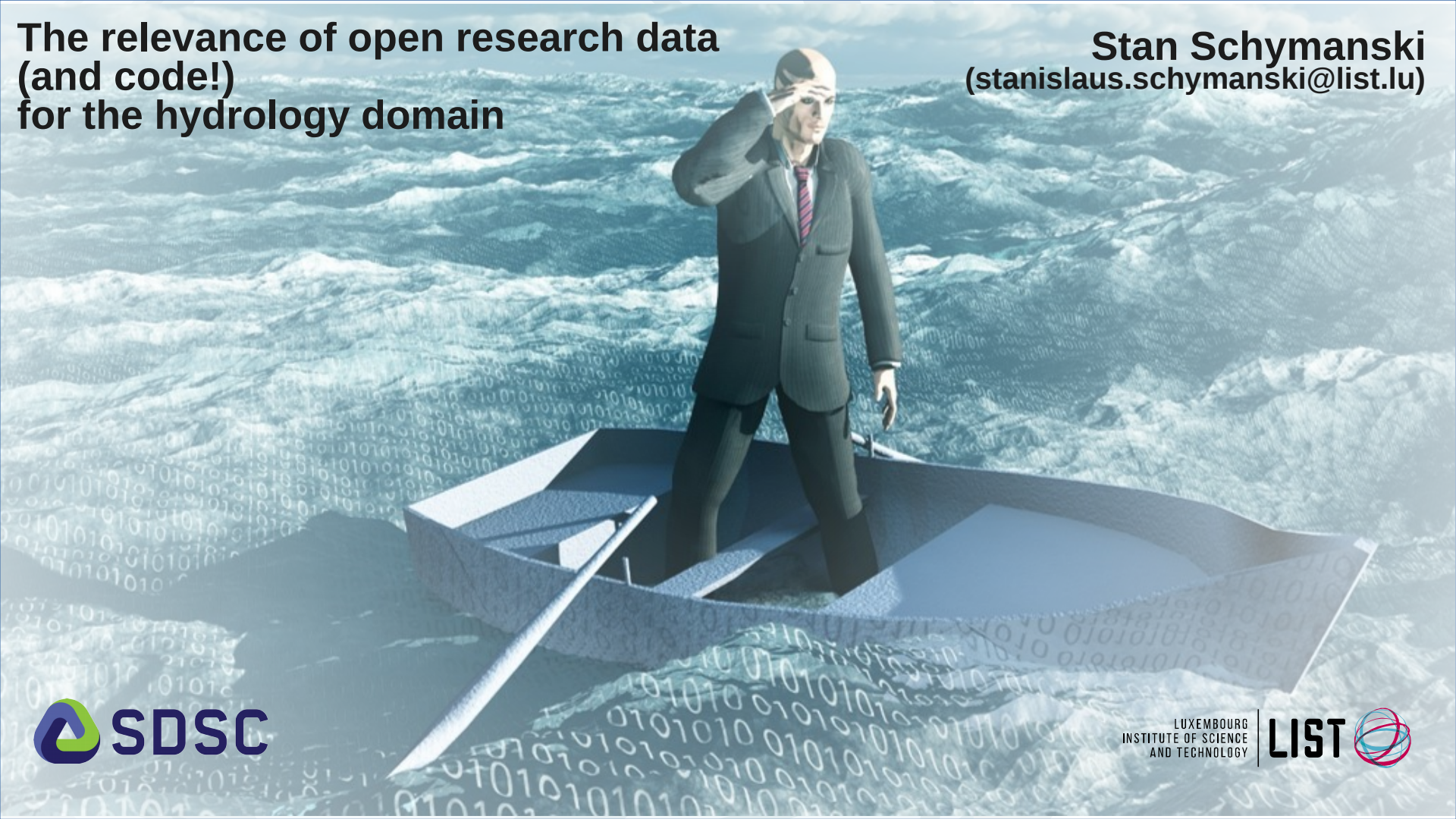


# The relevance of open research data (and code!) for the hydrology domain

**Stan Schymanski**  
([stanislaus.schymanski@list.lu](mailto:stanislaus.schymanski@list.lu))



# THEORY: SCIENTIFIC USE OF DATA AND MODELS

Data and models

Reality I

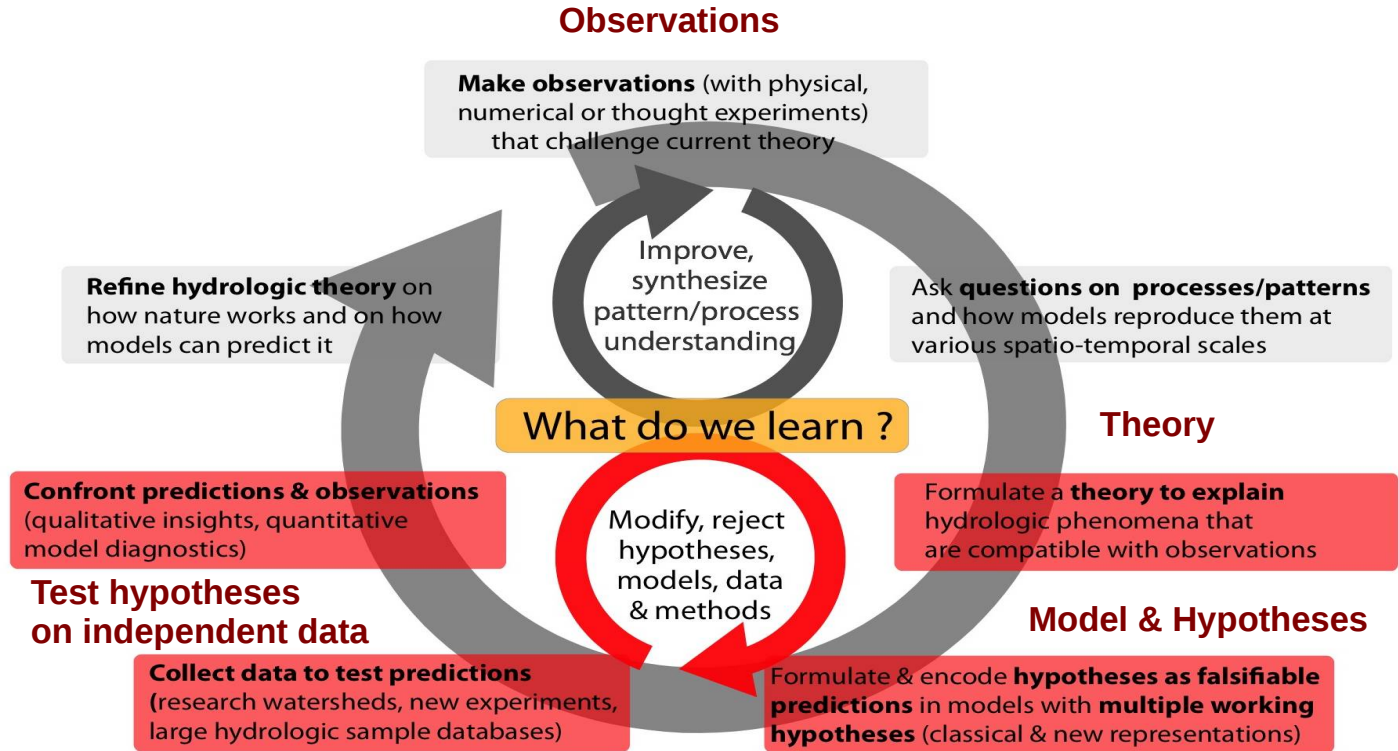
Community

Multi-source

Open data?

Reality II

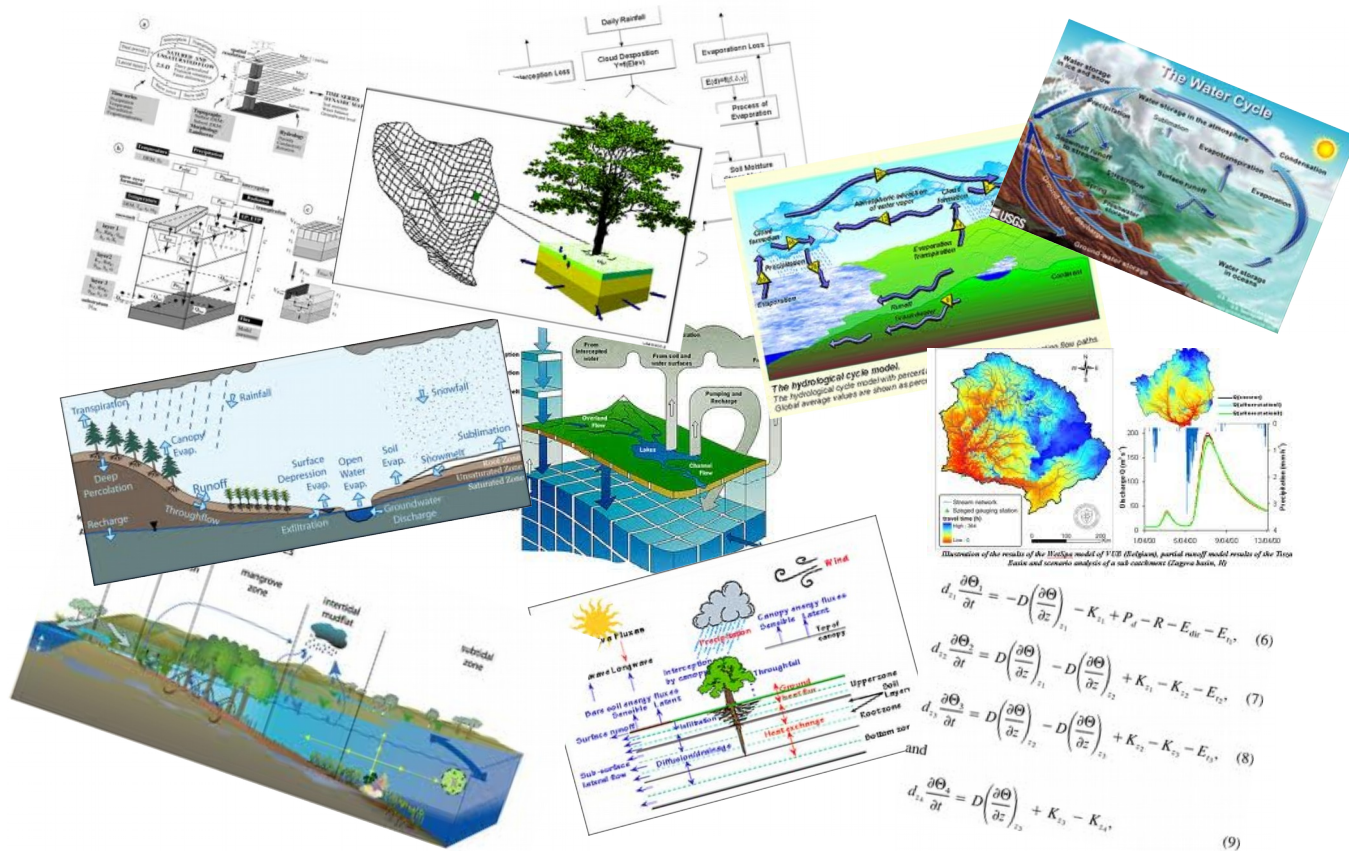
Open hydrology



# REALITY: A MODEL FOR EACH DATA SET

Data and models

- Reality I
- Community
- Multi-source
- Open data?
- Reality II
- Open hydrology



# WHAT DO WE LEARN?

Data and models

Reality I

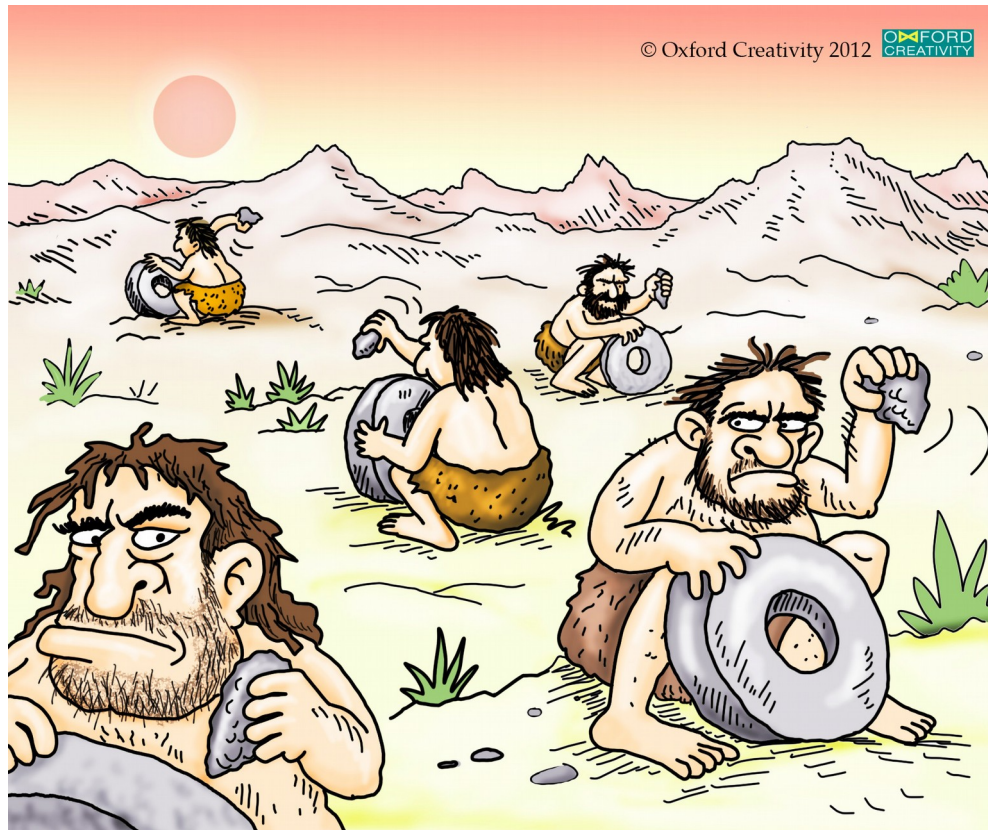
Community

Multi-source

Open data?

Reality II

Open hydrology



Credit: Oxford Creativity, [www.triz.co.uk](http://www.triz.co.uk)

# THEORY → MODEL WORKSHOP 2016

Need a way to follow up published analysis and reproduce it easily

Need to assess hydrological modelling choices/components

Want to re-use building blocks of old models to build new structure

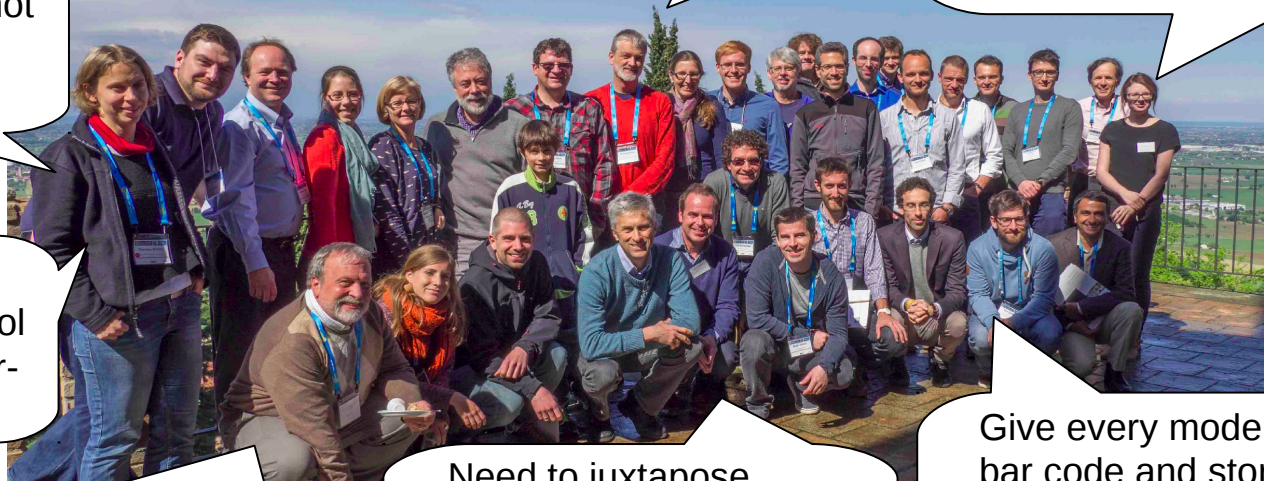
Community is not using available data effectively

Need model testing protocol similar to peer-review.

Need benchmark data sets to test desired model capability.

Need to juxtapose complexity of input variables to complexity of model

Give every model a bar code and store in data base including parameterisations etc.



Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrolo

# Community Hypothesis Evaluation Framework (CHEF)

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

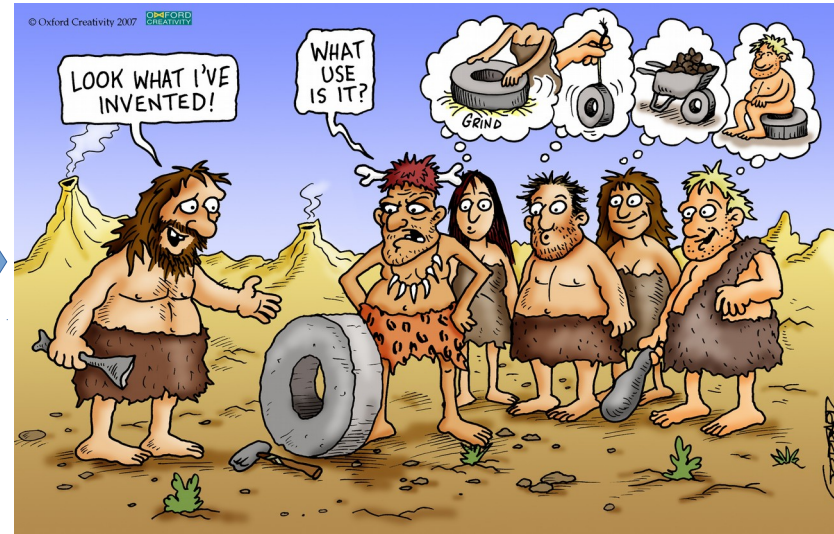
Open hydrology

- Modular models allowing evaluation of one component at a time
- Integration of multiple benchmark data sets (for testing generality of hypotheses)
- Computation of meaningful performance indicators to discriminate between alternative hypotheses (accounting for information content and uncertainty in input, output and reference data)

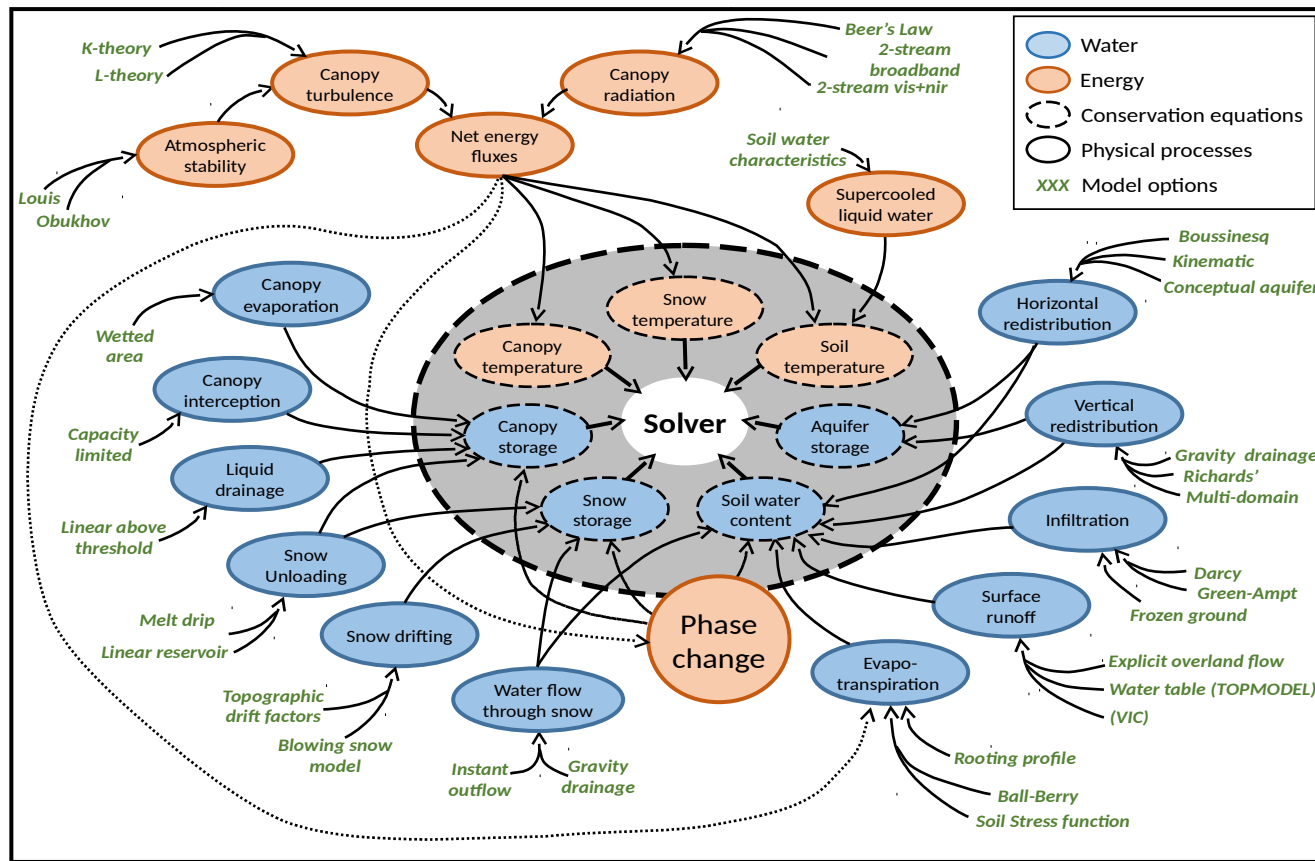
→ **Need community involvement to develop common Hypothesis Evaluation Framework allowing to build on each others modelling advances and compare against increasing body of observational evidence.**

# SHARING → COMMUNITY → PROGRESS

Data and models  
Reality I  
**Community**  
Multi-source  
Open data?  
Reality II  
Open hydrology



# Structure for Unifying Multiple Modeling Alternatives (SUMMA)



Clark et al. (2015) WRR

Data and models  
Reality I

Community

Multi-source

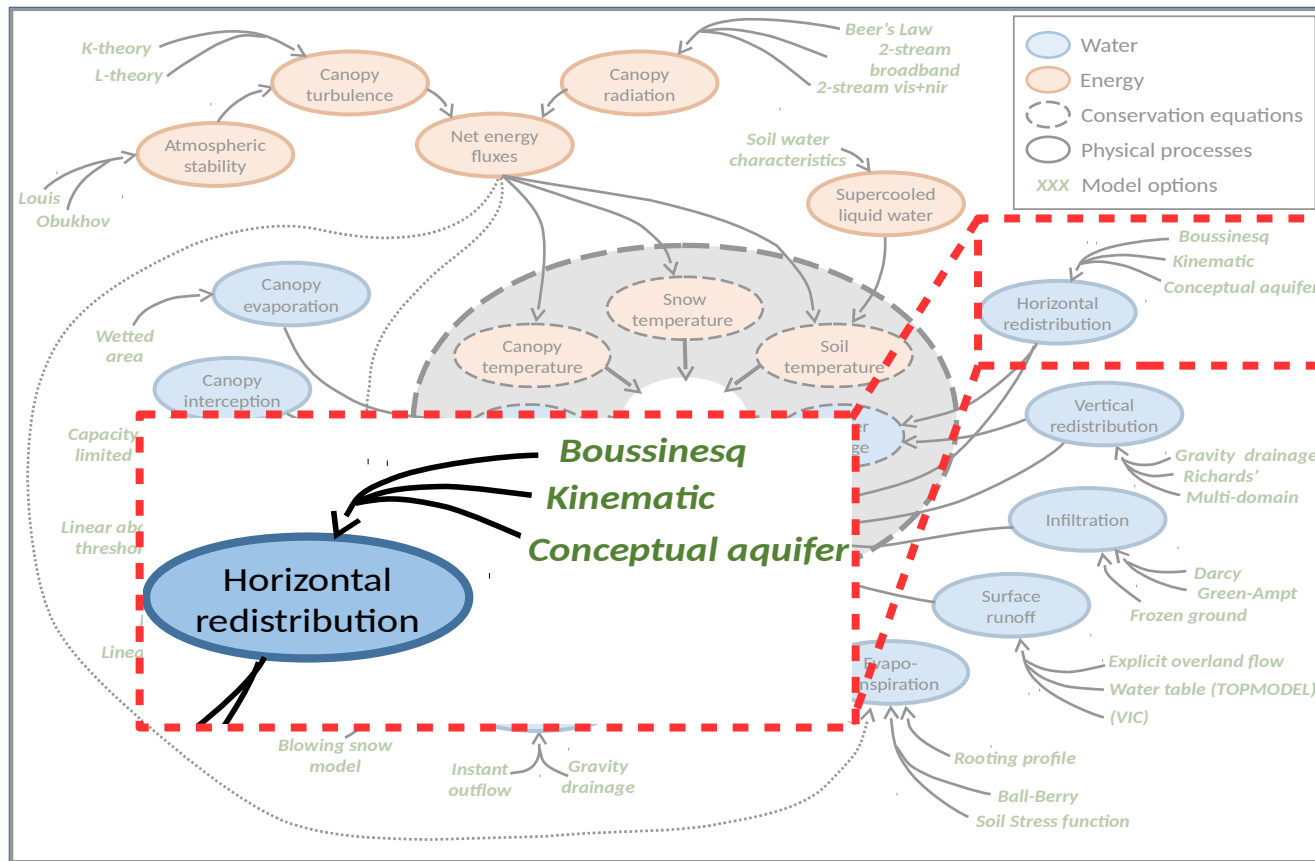
Open data?

Reality II

Open hydrology



# Structure for Unifying Multiple Modeling Alternatives (SUMMA)



Clark et al. (2015) WRR

Data and models  
Reality I  
**Community**  
Multi-source  
Open data?  
Reality II  
Open hydrology

# NEED FOR MULTI-SOURCE DATA

- Meteorological forcing
- Soils, Geology
- Land use/vegetation
- Topography
- Streamflow
- Evapotranspiration
- Tracers
- Soil moisture

GOSS Portal



**HYDROSHARE**

<http://apps.ecmwf.int/archive-catalogue/>

<http://www.hydroterre.psu.edu/>

<https://ldas.gsfc.nasa.gov/>

<https://search.earthdata.nasa.gov/>

<http://waterportal.ceos.org/>

<http://www.bom.gov.au/waterdata/>

<http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases>

<http://www.cimis.water.ca.gov>

<https://nex.nasa.gov/nex/>

## AGU journals

AGU encourages authors to identify and archive their data in **approved data centers**. If there is no relevant public repository available, and the data are such that they cannot easily be included in a supplement, authors are expected to **curate the above data for at least 5 years after publication** and provide a transparent process to make the data available to anyone upon request.

<http://publications.agu.org/author-resource-center/publication-policies/data-policy>

## EGU journals

Copernicus Publications recommends depositing data that correspond to journal articles in **reliable (public) data repositories, assigning digital object identifiers, and properly citing data sets as individual contributions**. Please find your appropriate data repository in the registry for research data repositories [re3data.org](https://re3data.org). A **data citation in a publication should resemble a bibliographic citation** and be located in the publication's reference list. (...)

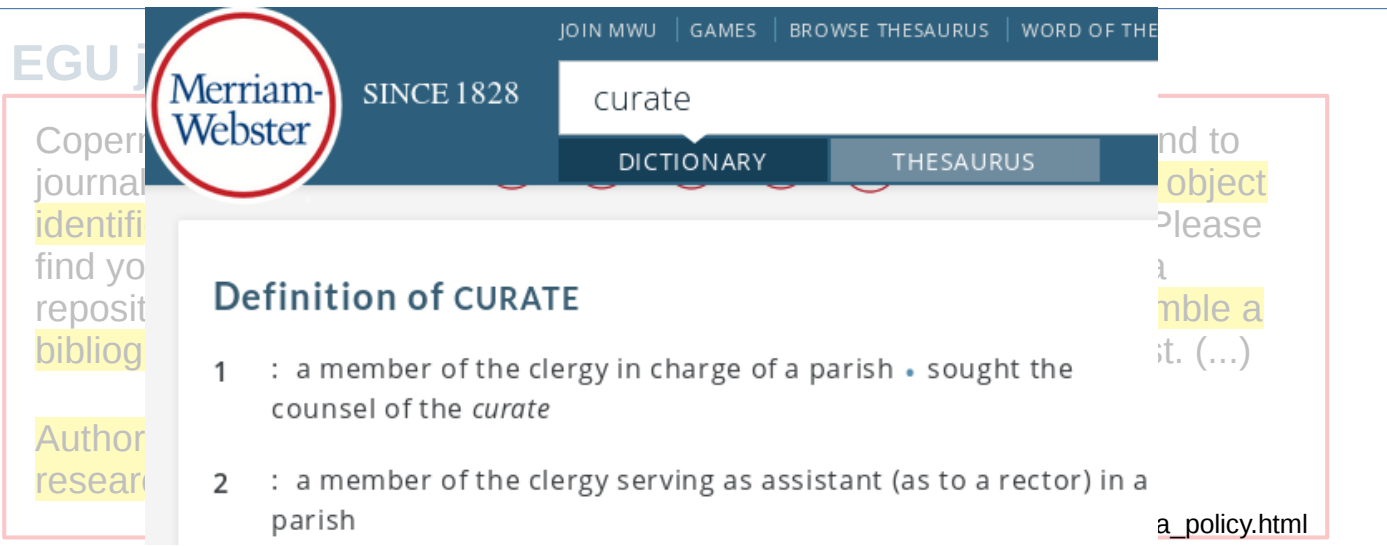
**Authors are required to provide a statement on how their underlying research data can be accessed.**

[https://www.hydrology-and-earth-system-sciences.net/about/data\\_policy.html](https://www.hydrology-and-earth-system-sciences.net/about/data_policy.html)

## AGU journals

AGU encourages authors to identify and archive their data in **approved data centers**. If there is no relevant public repository available, and the data are such that they cannot easily be included in a supplement, authors are expected to **curate** the above data for at least 5 years after publication and provide a transparent process to make the data available to anyone upon request.

<http://publications.agu.org/author-resource-center/publication-policies/data-policy>



The screenshot shows the Merriam-Webster website with the search term "curate" entered. The word "curate" is circled in red. The definition of CURATE is displayed below the search bar. The Merriam-Webster logo is circled in red. The text "EGU" is visible in the top left corner of the screenshot. The text "Copern journal" is visible in the top left corner of the screenshot. The text "identifi" is visible in the top left corner of the screenshot. The text "find yo" is visible in the top left corner of the screenshot. The text "reposit" is visible in the top left corner of the screenshot. The text "bibliog" is visible in the top left corner of the screenshot. The text "Author" is visible in the top left corner of the screenshot. The text "resear" is visible in the top left corner of the screenshot. The text "nd to" is visible in the top right corner of the screenshot. The text "object" is visible in the top right corner of the screenshot. The text "Please" is visible in the top right corner of the screenshot. The text "l" is visible in the top right corner of the screenshot. The text "mble a" is visible in the top right corner of the screenshot. The text "it. (...)" is visible in the top right corner of the screenshot. The text "a\_policy.html" is visible in the bottom right corner of the screenshot.

Merriam-Webster SINCE 1828

curate

DICTIONARY THESAURUS

**Definition of CURATE**

- 1 : a member of the clergy in charge of a parish • sought the counsel of the *curate*
- 2 : a member of the clergy serving as assistant (as to a rector) in a parish

a\_policy.html

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrology

## AGU journals

AGU encourages authors to identify and archive their data in **approved data centers**. If there is no relevant public repository available, and the data are such that they cannot easily be included in a supplement, authors are expected to **curate the above data for at least 5 years after publication** and provide a transparent process to make the data available to anyone upon request.

<http://publications.agu.org/author-resource-center/publication-policies/data-policy>

## EGU journals

Copernicus Publications recommends depositing data that correspond to journal articles in **reliable (public) data repositories, assigning digital object identifiers, and properly citing data sets as individual contributions**. Please find your appropriate data repository in the registry for research data repositories **re3data.org**. **A data citation in a publication should resemble a bibliographic citation** and be located in the publication's reference list. (...)

**Authors are required to provide a statement on how their underlying research data can be accessed.**

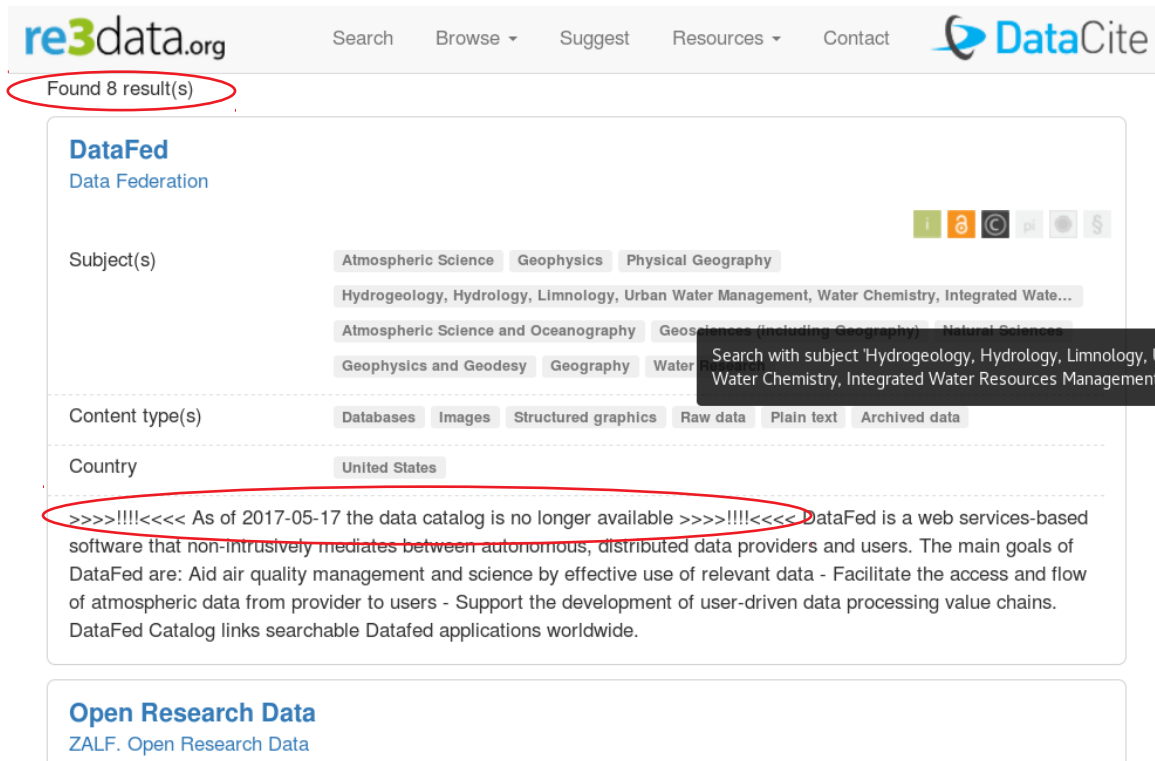
[https://www.hydrology-and-earth-system-sciences.net/about/data\\_policy.html](https://www.hydrology-and-earth-system-sciences.net/about/data_policy.html)

# HOW TO FIND “APPROPRIATE” REPO?

Data and models  
Reality I  
Community  
Multi-source  
**Open data?**  
Reality II  
Open hydrology

8 results for  
“Water Research”.

1<sup>st</sup> result: “no  
longer available”?!?



The screenshot shows the re3data.org search results page. At the top, the re3data.org logo is on the left, and navigation links for Search, Browse, Suggest, Resources, and Contact are on the right. The DataCite logo is also present. Below the navigation, a red circle highlights the text "Found 8 result(s)".

The first result is for "DataFed" (Data Federation). It includes subject tags such as Atmospheric Science, Geophysics, Physical Geography, Hydrogeology, Hydrology, Limnology, Urban Water Management, Water Chemistry, Integrated Water Resources Management, Atmospheric Science and Oceanography, Geosciences (including Geography), Natural Sciences, Geophysics and Geodesy, Geography, and Water. Content type tags include Databases, Images, Structured graphics, Raw data, Plain text, and Archived data. The country is listed as United States.

The description for DataFed is circled in red and contains the text: ">>>>!!!!<<<< As of 2017-05-17 the data catalog is no longer available >>>>!!!!<<<<". A black tooltip box is overlaid on the right side of the description, containing the text: "Search with subject 'Hydrogeology, Hydrology, Limnology, Urban Water Chemistry, Integrated Water Resources Management'".

The second result is "Open Research Data" from ZALF.

# REALITY: IT AIN'T THAT EASY...

- Am I allowed to deposit all data?
- Where and how to deposit?

Maybe I'd better just describe where I got the data from...

Example: HESSD,  
<https://doi.org/10.5194/hess-2017-473>

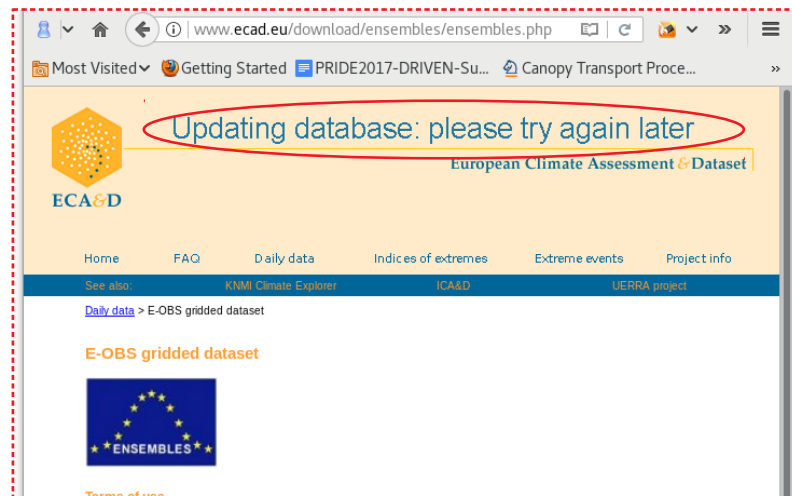


# DATA AVAILABILITY STATEMENT IN HESS

## 8 Data availability

Both the E-OBS and CPC precipitation datasets are freely available (from respectively <http://www.ecad.eu/download/ensembles/ensembles.php> and [ftp://ftp.cdc.noaa.gov/Datasets/cpc\\_us\\_precip/](ftp://ftp.cdc.noaa.gov/Datasets/cpc_us_precip/)). The ERA-Interim, ERA-Interim LAND and ERA20C re-analysis data are freely available from the ECMWF. The GLEAM dataset can be downloaded via <https://www.gleam.eu/>. The discharge data is obtained from the Global Runoff Data Centre and is freely available. The EC-Earth model output and the W3RA model output is available upon request by the author.

Oops! Is it going to be the same data after the update?



The screenshot shows a web browser window with the URL [www.ecad.eu/download/ensembles/ensembles.php](http://www.ecad.eu/download/ensembles/ensembles.php). The page header includes the ECA&D logo and the text "European Climate Assessment Dataset". A red oval highlights a message that says "Updating database: please try again later". Below the message, there are navigation links for Home, FAQ, Daily data, Indices of extremes, Extreme events, and Project info. A blue bar contains links for "See also: KNMI Climate Explorer, ICA&D, UERRA project". The main content area shows a breadcrumb trail "Daily data > E-OBS gridded dataset" and the title "E-OBS gridded dataset" in orange. At the bottom, there is a logo for "ENSEMBLES" featuring the European Union flag.

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrology

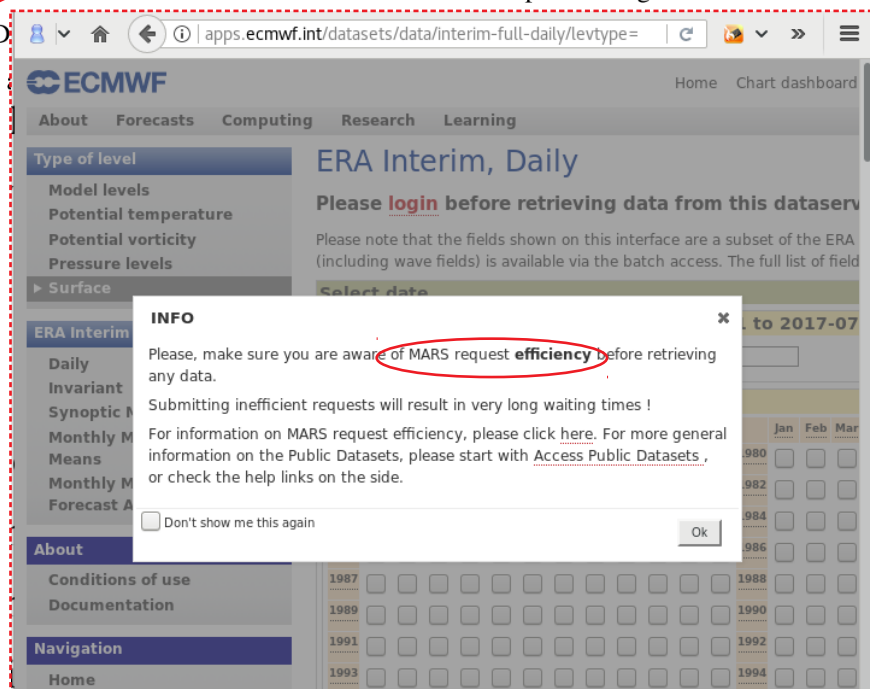


# DATA AVAILABILITY STATEMENT IN HESS

## 8 Data availability

Both the E-OBS and CPC precipitation datasets are freely available (from respectively <http://www.ecad.eu/download/ensembles/ensembles.php> and [ftp://ftp.cdc.noaa.gov/Datasets/cpc\\_us\\_precip/](ftp://ftp.cdc.noaa.gov/Datasets/cpc_us_precip/)). The ERA-Interim, ERA-Interim LAND and ERA20C re-analysis data are freely available from the ECMWF. The GLEAM dataset can be downloaded via <https://www.gleam.eu/>. The discharge data is obtained from the Global Runoff Database. The W3RA model output is available upon request by the

“Request efficiency”? How do I request the same data as in the paper?



The screenshot shows the ECMWF website interface for the ERA Interim, Daily dataset. The browser address bar displays `apps.ecmwf.int/datasets/data/interim-full-daily/levtype=`. The page title is "ERA Interim, Daily" and it includes a "Please login before retrieving data from this dataset" warning. An "INFO" pop-up dialog box is overlaid on the page, containing the following text: "Please, make sure you are aware of MARS request efficiency before retrieving any data. Submitting inefficient requests will result in very long waiting times ! For information on MARS request efficiency, please click [here](#). For more general information on the Public Datasets, please start with [Access Public Datasets](#), or check the help links on the side." The dialog box has an "Ok" button and a checkbox labeled "Don't show me this again".

Data and models

Reality I

Community

Multi-source

Open data?

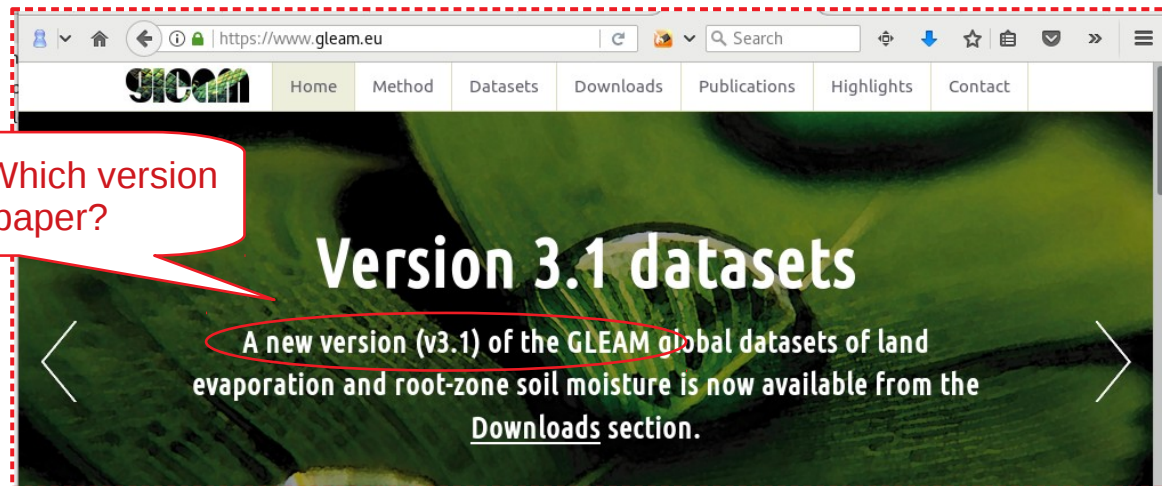
Reality II

Open hydrology

# DATA AVAILABILITY STATEMENT IN HESS

## 8 Data availability

Both the E-OBS and CPC precipitation datasets are freely available (from respectively <http://www.ecad.eu/download/ensembles/ensembles.php> and [ftp://ftp.cdc.noaa.gov/Datasets/cpc\\_us\\_precip/](ftp://ftp.cdc.noaa.gov/Datasets/cpc_us_precip/)). The ERA-Interim, ERA-Interim LAND and ERA20C re-analysis data are freely available from the ECMWF. The GLEAM dataset can be downloaded via <https://www.gleam.eu/>. The discharge data is obtained from the Global Runoff Data Centre and is freely available. The EC-Earth model output and the W3RA model output is available upon request by the author.



“New version”? Which version was used in the paper?

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrology

# DATA AVAILABILITY STATEMENT IN HESS

## 8 Data availability

Both the E-OBS and CPC precipitation datasets are freely available (from respectively <http://www.ecad.eu/download/ensembles/ensembles.php> and [ftp://ftp.cdc.noaa.gov/Datasets/cpc\\_us\\_precip/](ftp://ftp.cdc.noaa.gov/Datasets/cpc_us_precip/)). The ERA-Interim, ERA-Interim LAND and ERA20C re-analysis data are freely available from the ECMWF. The GLEAM dataset can be downloaded via <https://www.gleam.eu/>. The discharge data is obtained from the Global Runoff Data Centre and is freely available. The EC-Earth model output and the W3RA model output is available upon request by the author.

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrology

# DATA AVAILABILITY

## 8 Data availability

Both the E-OBS and CPC reanalysis ensembles.php and ftp://ftp.cgd.cmc.gov.cn/ftp/analysis data are freely available. discharge data is obtained from W3RA model output is available

The screenshot shows a web browser window at www.bafg.de/GRDC/EN/02\_srvcs/21\_tmsr: with a search bar and navigation icons. The breadcrumb trail reads: You are here: GRDC > Standard Services > River Discharge Data. The left sidebar contains a menu with the following items: River Discharge Data (dropdown), Station Catalogues, Stations in Google Earth, Geospatial Data Products (arrow), GRDC Report Series (arrow), and GRDC Hydro Terms (arrow). Below this is a 'Services' section with links to Global Runoff Database, River Discharge Data, and Geospatial Data. The main content area is titled 'Steps to Order River Discharge Time Series' and contains a 7-step list. A red circle highlights the first three steps. A red speech bubble points to the list with the text 'Aaaargh! 7 steps to get the data, including explanation of research topic!'. At the bottom right, there is a world map showing GRDC stations with monthly data, color-coded by time series length. The legend indicates: 1919-1979 (blue), 1980-1989 (green), 1990-1999 (yellow), 2000-2009 (orange), and 2010-2019 (red). The map is titled '6431 GRDC stations with monthly data, incl. data derived from daily data (Status: 28 Sep 2017)' and includes the GRDC logo.

Steps to Order River Discharge Time Series

1. Read the Policy Guidelines and agree to the GRDC User Declaration.
2. Examine the GRDC station maps (see right margin) to see whether GRDC data may be useful for your research project.
3. Download the GRDC Catalogue (XLS) from the catalogue menu item, or the KMZ files for use with Google Earth, and select your stations of interest.
4. Prepare a list of selected stations and indicate the time period of interest, ideally in standard text (DOS ASCII) or MS-Excel format (XLS). Alternatively, you can use the GRDC order form (see right margin) for your data request.
5. Write an explanatory summary of your research project (one page).
6. Send Order Form, Station List, and Project Summary to the GRDC, preferably via e-mail (mailto:grdc@bafg.de).
7. Please do not forget to send the signed *User Declaration*. Send it to the GRDC via fax (+49 261 13065722). Alternatively to fax letter, electronic formats like PDF or a graphic format will be accepted.

Aaaargh!  
7 steps to get the data, including explanation of research topic!

GRDC Stations Time Series Length (years)

- 1919 - 1979
- 1980 - 1989
- 1990 - 1999
- 2000 - 2009
- 2010 - 2019

6431 GRDC stations with monthly data, incl. data derived from daily data (Status: 28 Sep 2017)  
Koblenz: Global Runoff Data Centre, 2017. GRDC

# REPRODUCIBILITY FROM THE START

*“Opening science once the science has been done is too late” (Olivier Verscheure)*

Data and models

Reality I

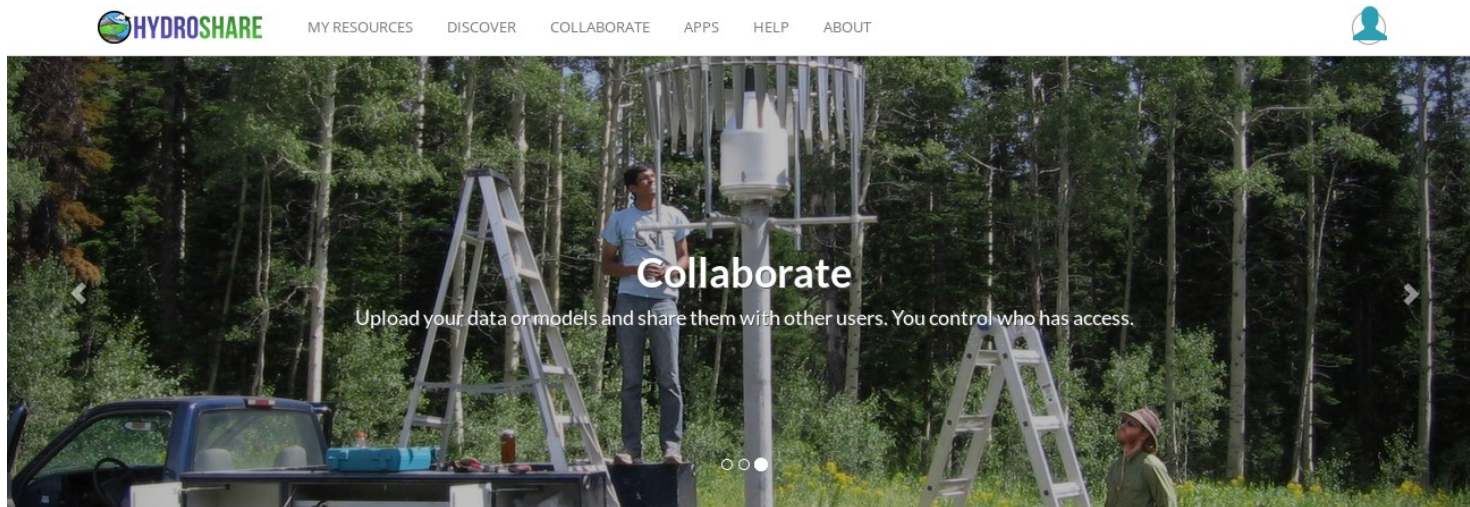
Community

Multi-source

Open data?

Reality II

Open hydrology



## How it works

1

Create data

2

Upload to HydroShare

3

Describe with metadata

4

Share with colleagues

# REPRODUCIBILITY FROM THE START

- Data and models
- Reality I
- Community
- Multi-source
- Open data?
- Reality II
- Open hydrology**

**HYDROSHARE**

MY RESOURCES DISCOVER COLLABORATE APPS HELP ABOUT

## 1 Hydrologic Terrain Analysis of Logan Watershed Using Jupyter Notebook TauDEM

Authors: [Maya Scott](#)

Owners: [Maya Scott](#)

Resource type: Generic







Created: May 17, 2017, 2:11 p.m.

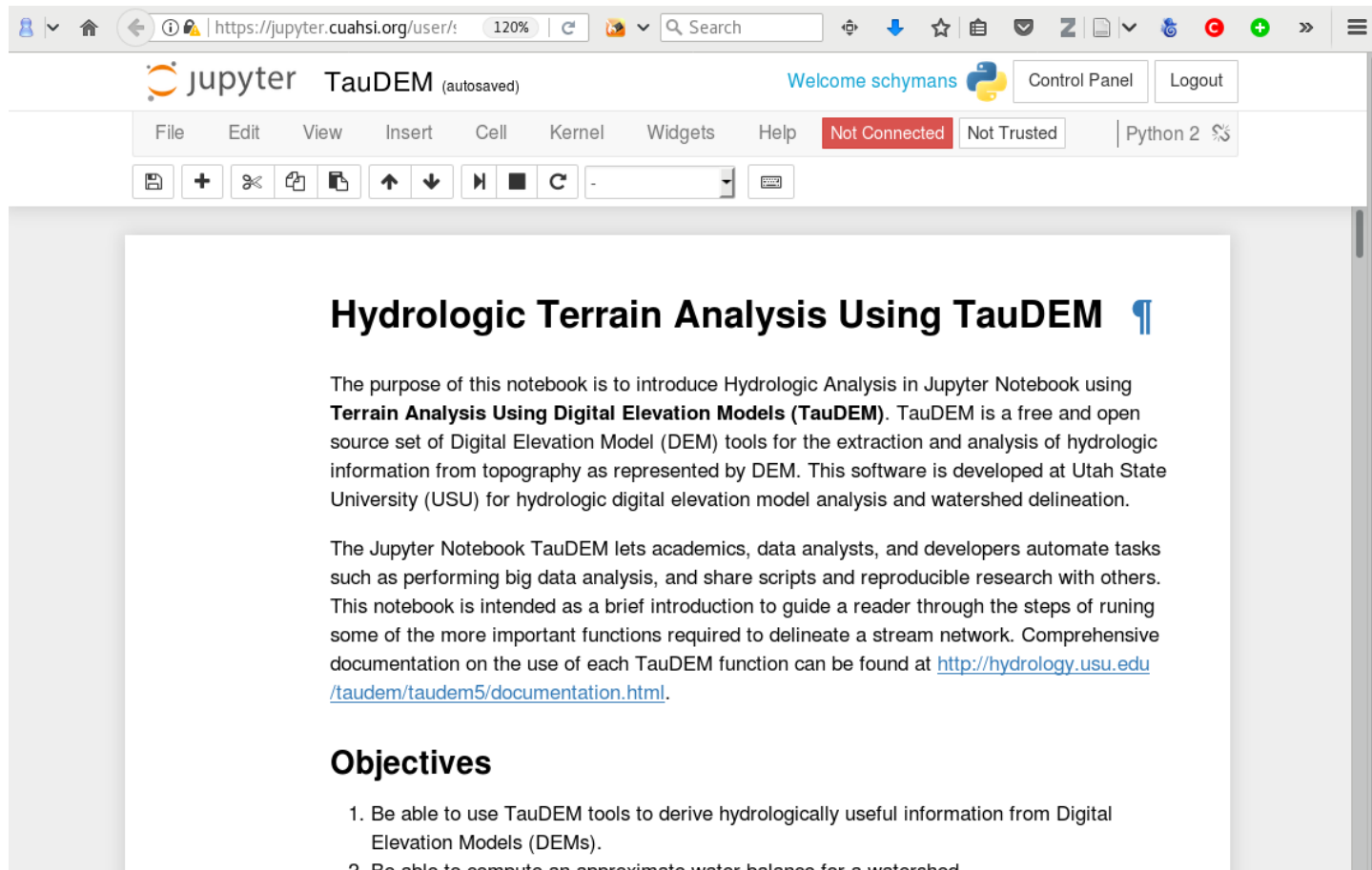
Last updated: May 17, 2017, 2:35 p.m. by [Maya Scott](#)

### Abstract

Jupyter Notebook TauDEM was used to define streamflow and subwatersheds in the Logan River Utah. To start, "logan.tif" Digital Elevation Model (DEM) data and "LoganOutlet.shp" Logan Outlet the main inputs. The final results were "loganpw.tif" subwatershed and "logannet.shp" stream networks. This

Open with... ▾

-  HydroShare GIS
-  JupyterHub NCSA
-  GeoTrust Server
-  TauDEM CyberGIS Hydro...
-  GeoTrust
-  Landlab Jupyter Termi...



The screenshot shows a web browser window displaying a Jupyter Notebook. The browser address bar shows the URL `https://jupyter.cuahsi.org/user/!`. The Jupyter interface includes a header with the Jupyter logo, the notebook name "TauDEM (autosaved)", a welcome message "Welcome schymans", and buttons for "Control Panel" and "Logout". Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, Help, and a status indicator "Not Connected". The notebook content area displays a document with the following text:

## Hydrologic Terrain Analysis Using TauDEM ¶

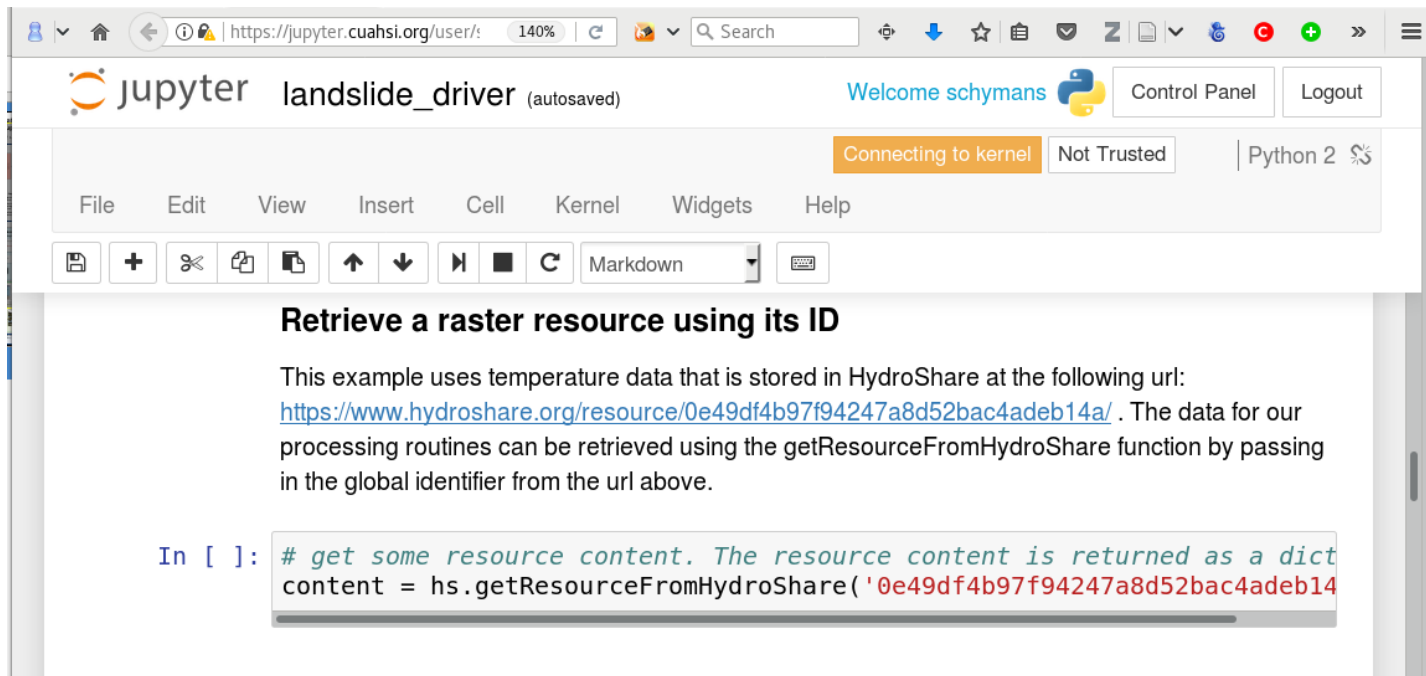
The purpose of this notebook is to introduce Hydrologic Analysis in Jupyter Notebook using **Terrain Analysis Using Digital Elevation Models (TauDEM)**. TauDEM is a free and open source set of Digital Elevation Model (DEM) tools for the extraction and analysis of hydrologic information from topography as represented by DEM. This software is developed at Utah State University (USU) for hydrologic digital elevation model analysis and watershed delineation.

The Jupyter Notebook TauDEM lets academics, data analysts, and developers automate tasks such as performing big data analysis, and share scripts and reproducible research with others. This notebook is intended as a brief introduction to guide a reader through the steps of running some of the more important functions required to delineate a stream network. Comprehensive documentation on the use of each TauDEM function can be found at <http://hydrology.usu.edu/taudem/taudem5/documentation.html>.

### Objectives

1. Be able to use TauDEM tools to derive hydrologically useful information from Digital Elevation Models (DEMs).
2. Be able to compute an approximate water balance for a watershed.

# RECORD SOURCES OF DATA



The screenshot shows a JupyterLab interface in a browser window. The address bar shows the URL `https://jupyter.cuahsi.org/user/!`. The page title is "jupyter landslide\_driver (autosaved)". The user is identified as "schymans". The interface includes a "Control Panel" and "Logout" button. A status bar indicates "Connecting to kernel" and "Not Trusted". The menu bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. The toolbar contains icons for saving, adding, undo, redo, and other actions. The main content area displays a notebook cell with the following text:

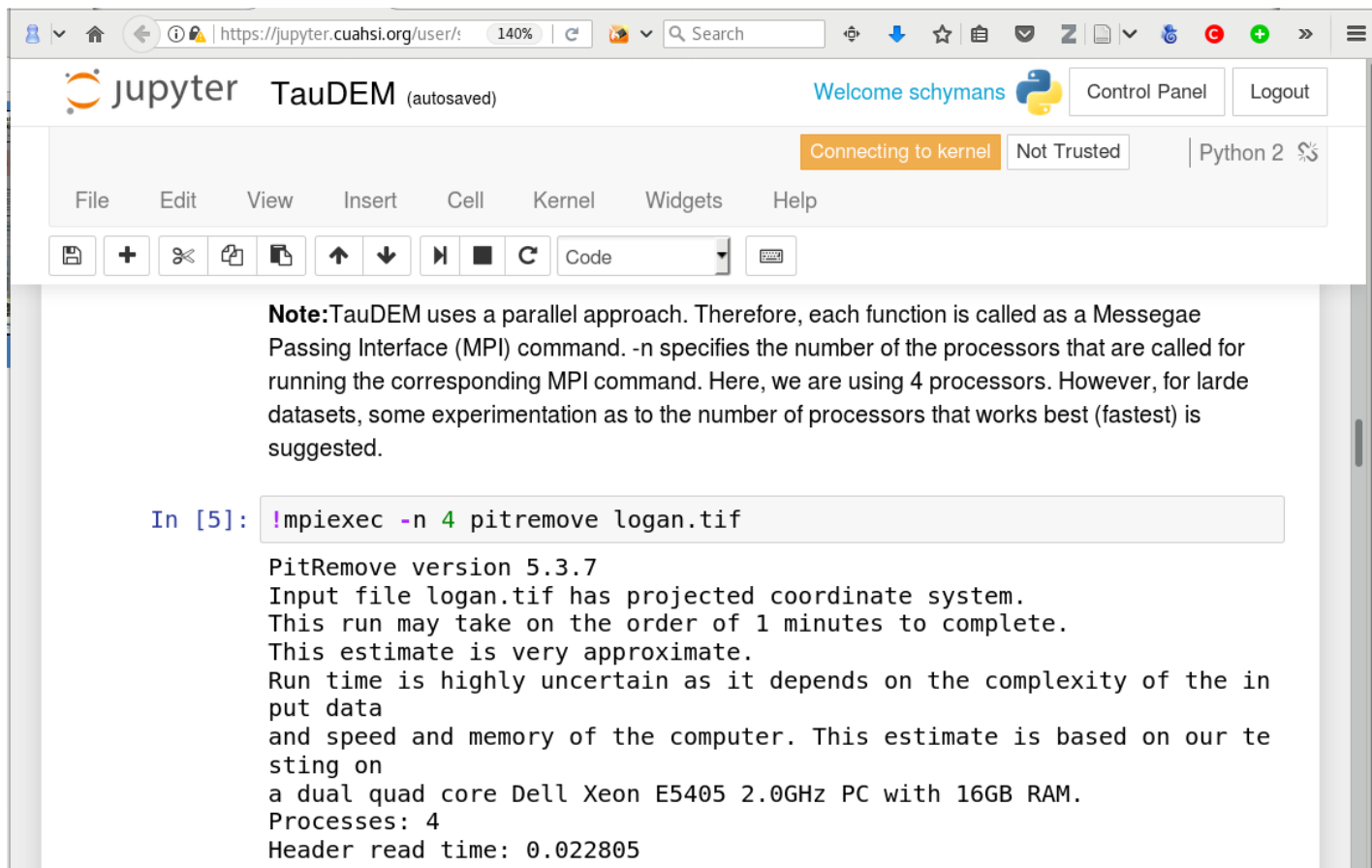
**Retrieve a raster resource using its ID**

This example uses temperature data that is stored in HydroShare at the following url: <https://www.hydroshare.org/resource/0e49df4b97f94247a8d52bac4adeb14a/> . The data for our processing routines can be retrieved using the `getResourceFromHydroShare` function by passing in the global identifier from the url above.

```
In [ ]: # get some resource content. The resource content is returned as a dict
content = hs.getResourceFromHydroShare('0e49df4b97f94247a8d52bac4adeb14
```



# FAST MULTI-CORE EXECUTION



The screenshot shows a JupyterLab interface in a web browser. The browser address bar shows the URL `https://jupyter.cuahsi.org/user/`. The JupyterLab header includes the logo, the name "TauDEM (autosaved)", a "Welcome schymans" message, and buttons for "Control Panel" and "Logout". Below the header, there is a status bar indicating "Connecting to kernel", "Not Trusted", and "Python 2". The main interface has a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". Below the menu bar is a toolbar with icons for saving, adding, undo, redo, copy, paste, and running code. The code cell contains the following text:

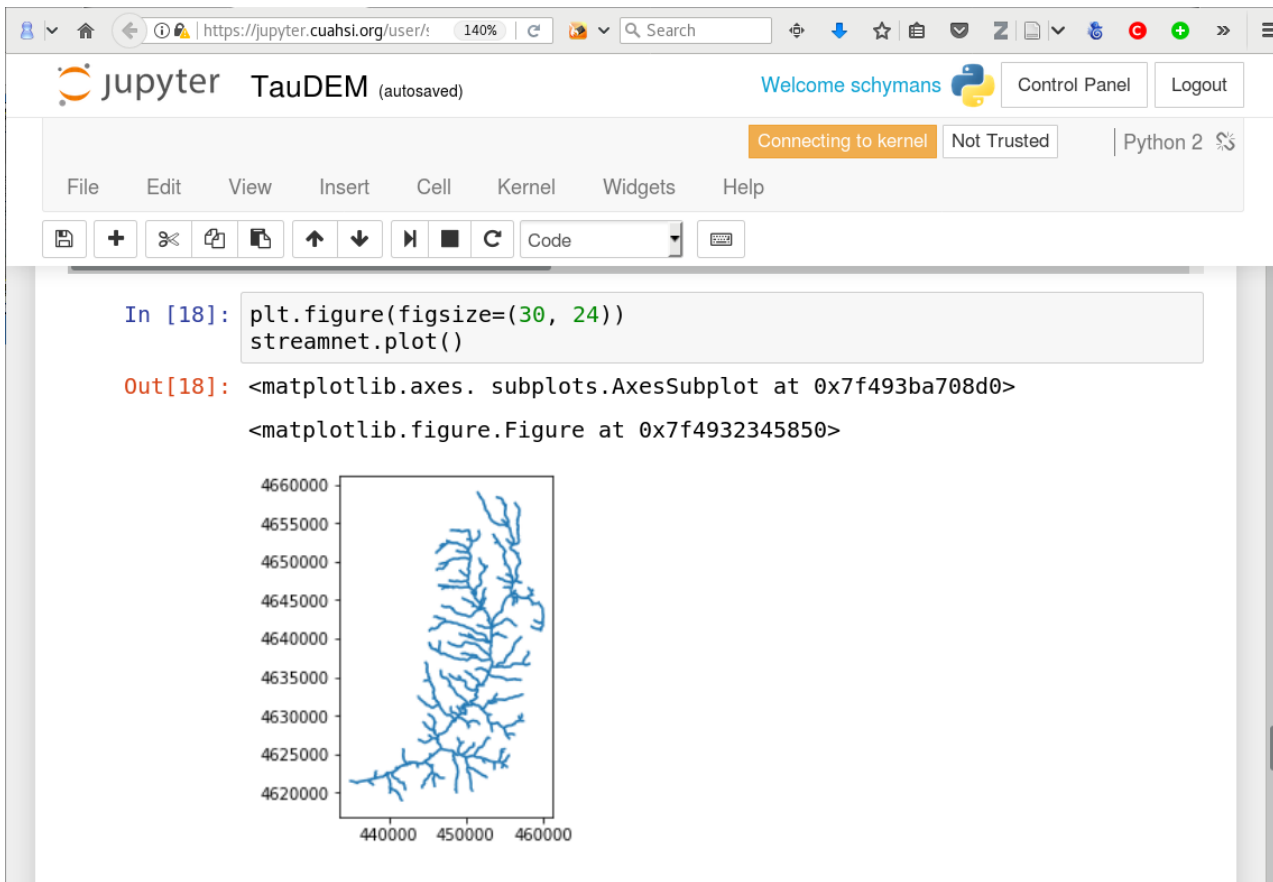
```
In [5]: !mpiexec -n 4 pitremove logan.tif
```

The output of the code cell is as follows:

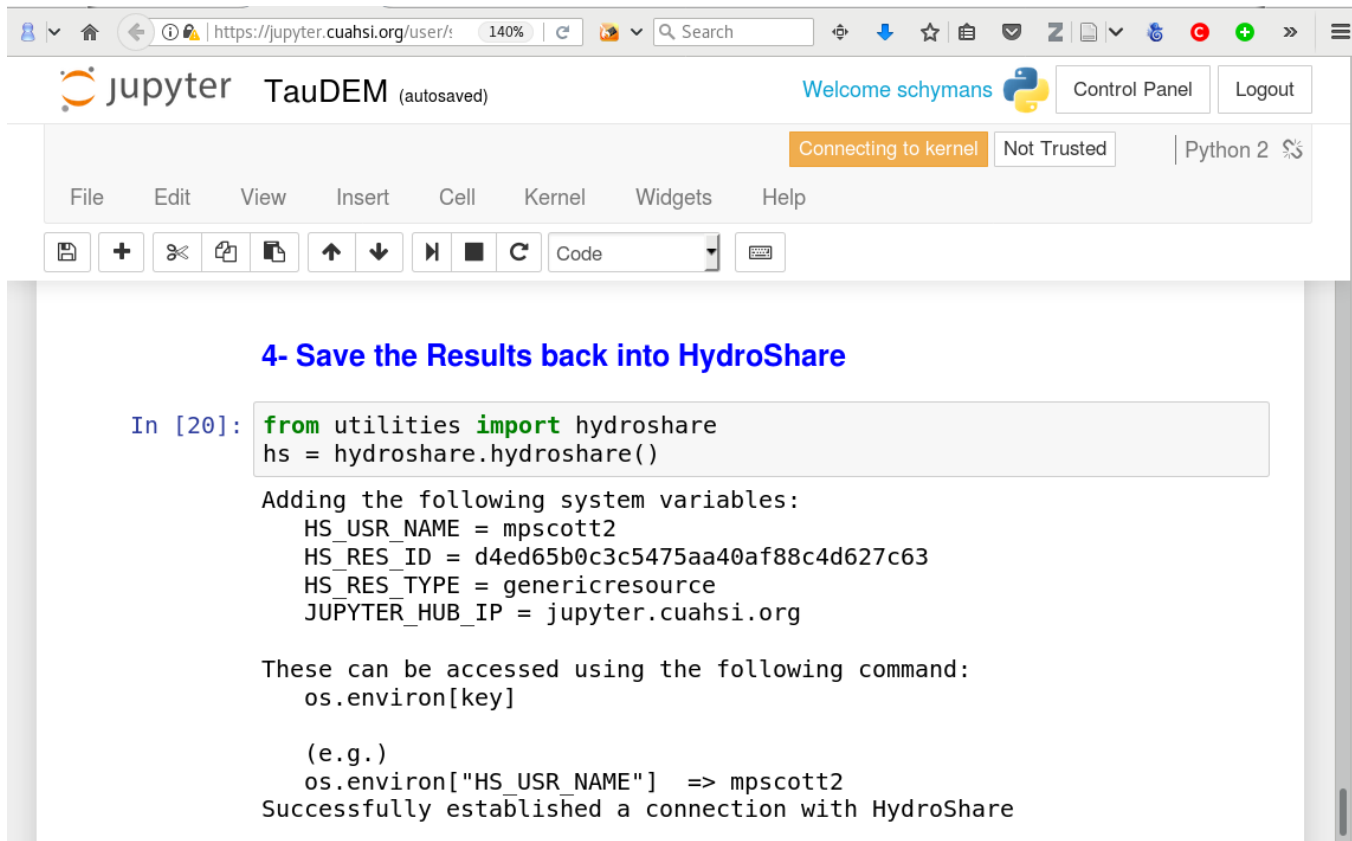
```
PitRemove version 5.3.7  
Input file logan.tif has projected coordinate system.  
This run may take on the order of 1 minutes to complete.  
This estimate is very approximate.  
Run time is highly uncertain as it depends on the complexity of the in  
put data  
and speed and memory of the computer. This estimate is based on our te  
sting on  
a dual quad core Dell Xeon E5405 2.0GHz PC with 16GB RAM.  
Processes: 4  
Header read time: 0.022805
```

# PUBLICATION-READY PLOTS

Data and models  
Reality I  
Community  
Multi-source  
Open data?  
Reality II  
**Open hydrology**



# DEPOSIT EVERYTHING INCL. WORKFLOW BACK ON HYDROSHARE



The screenshot shows a JupyterLab interface in a web browser. The browser address bar shows the URL `https://jupyter.cuahsi.org/user/`. The JupyterLab header includes the logo, the name "TauDEM (autosaved)", and a "Welcome schymans" message. There are buttons for "Control Panel" and "Logout". Below the header is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". A toolbar contains icons for file operations and a "Code" dropdown menu. The main content area displays a code cell with the following text:

```
In [20]: from utilities import hydroshare
hs = hydroshare.hydroshare()

Adding the following system variables:
HS_USER_NAME = mpscott2
HS_RES_ID = d4ed65b0c3c5475aa40af88c4d627c63
HS_RES_TYPE = genericresource
JUPYTER_HUB_IP = jupyter.cuahsi.org

These can be accessed using the following command:
os.environ[key]

(e.g.)
os.environ["HS_USER_NAME"] => mpscott2
Successfully established a connection with HydroShare
```

# HOW ABOUT RENGA ?

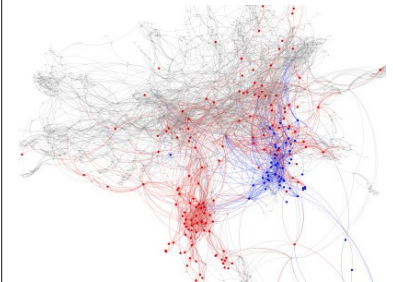
Data and models  
Reality I  
Community  
Multi-source  
Open data?  
Reality II  
**Open hydrology**



GitHub

zenodo

- (1) Download and convert data from external resource (logger, repository, etc.): `renga.load(Data1)`
- (2) Import any third-party code:  
`renga.import(Mod1)`
- (3) Interactive development, analysis and documentation in jupyter notebook, using your preferred analysis tool (e.g. python, R, matlab...)
- (4) Create publication-quality figures
- (5) Create docker containers for new components:  
`renga.dockerize(components)`
- (6) Submit to renga: `renga.submit(project1)`
- (7) Control access rights:  
`renga.share(project1, reviewers)`
- (8) Submit paper to scientific journal and link to project: `renga.addlink(project1, paper1)`
- (9) Publish project components:  
`renga.publish(project1, target=zenodo.org)`



## ENVIRONMENTAL Science & Technology

Viewpoint

[pubs.acs.org/est](https://pubs.acs.org/est)

### Open Science for Identifying “Known Unknown” Chemicals

Emma L. Schymanski<sup>\*,†</sup> and Antony J. Williams<sup>\*,‡</sup>

<sup>†</sup>Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

<sup>‡</sup>National Center for Computational Toxicology, U.S. EPA, Research Triangle Park, Durham, North Carolina 27711, United States

SCIENTIFIC  
OPINION  
NON PEER



quality and correctness is common, yet detrimental to many studies requiring quality data sources,<sup>3</sup> leading to unintentional errors.<sup>1</sup> Challenges for chemical databases include encouraging scientists to submit *high-quality* data to online platforms and, in return, providing access to curated chemical structures and related data for suspect prioritization such as (1) experimental/

## SUMMARY

- Need open data and open models
- Need to involve hydrological community
- Need an operational collaborative platform
- Restrictive data licenses are in the way of reproducible research
- Would data providers remove restrictions if lineage was recorded and credit attributed correctly?

THANK YOU FOR YOUR QUESTIONS/COMMENTS!  
stanislaus.schymanski@list.lu

Data and models

Reality I

Community

Multi-source

Open data?

Reality II

Open hydrology