

Introduction: Water data, Python and You

Waterhackweek 2020

Bart Nijssen

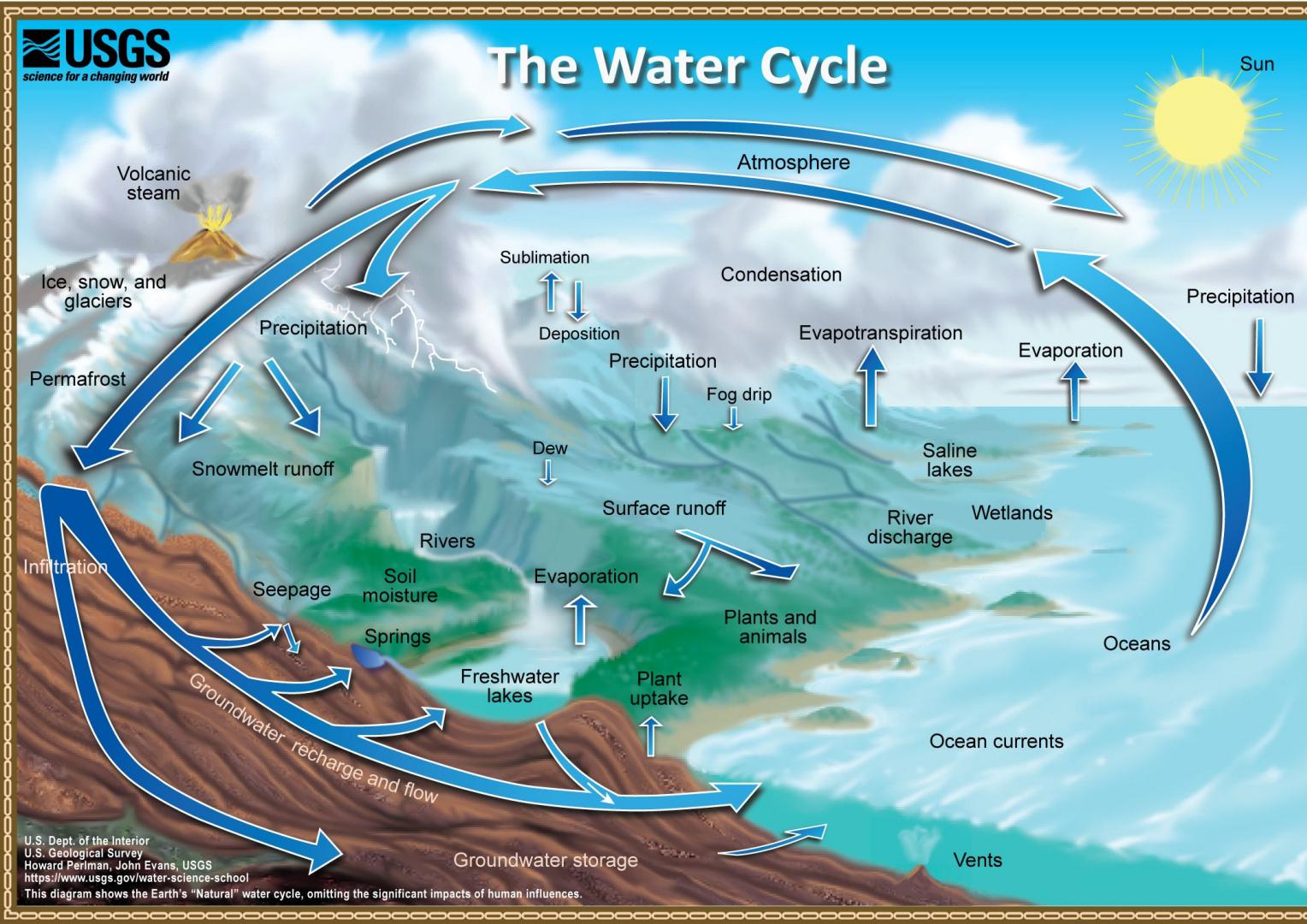
Presentation: 30 minutes

Tuesday Schedule

Post questions on slack in #tutorials

08:00 PDT	Project Updates
08:30 PDT	Introduction: Water data, Python and You Bart Nijssen
09:00 PDT	Break
09:15 PDT	Access and analyze point time series data Yifan Cheng
10:00 PDT	Break
10:15 PDT	Access and analyze raster and multi-dimensional gridded data Steven Pestana
11:00 PDT	Break
11:15 PDT	Water data mash up Emilio Mayorga
12:00 PDT	Break: Qiqochat Zoom rooms & YoTribe Open Mary Dumas

The Water Cycle



Data comes in many forms from many sources ...



Some examples:

Point time series of environmental observations

Advantage: Measurement in real world

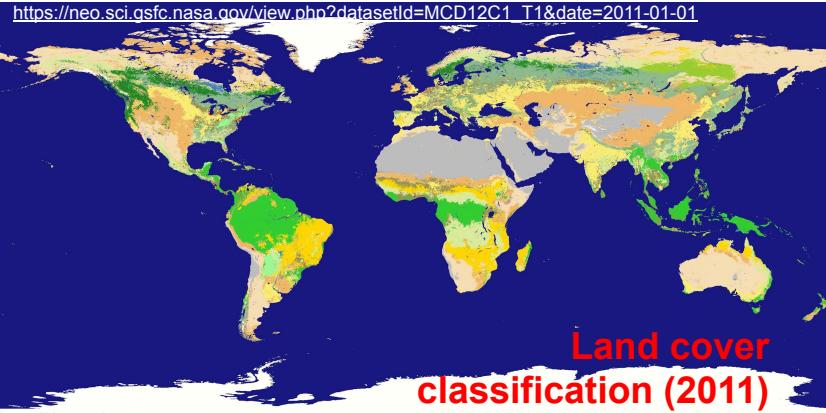


Challenges:

- Representativeness
- Measurement error
- Missing data
- Changes in technique and location
- Record length
- Data format
- etc.



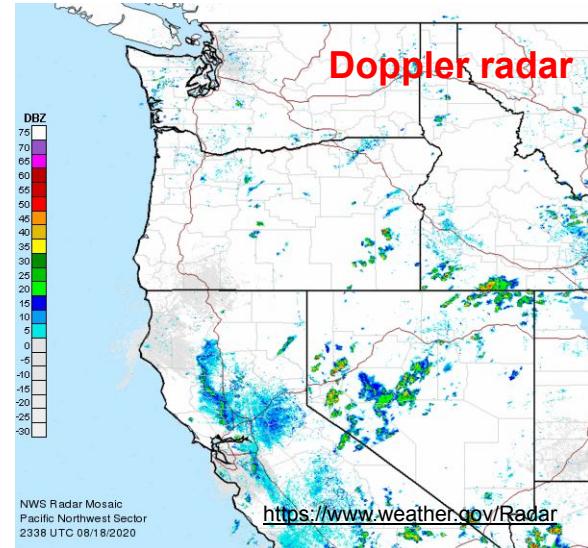
Data comes in many forms from many sources ...



Spatial data from remote sensing platforms

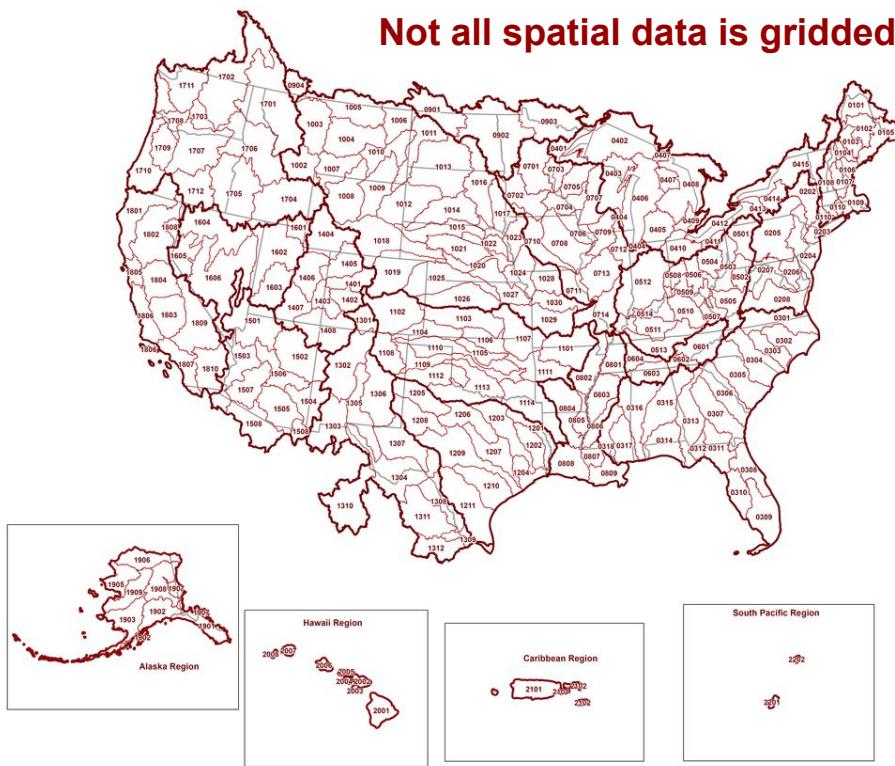
Challenges:

- Measurement error
- Missing data
- Changes in technique
- Record length
- Data format
- Data size
- Projection
- Many different spatial arrangements
- etc.



Data comes in many forms from many sources ...

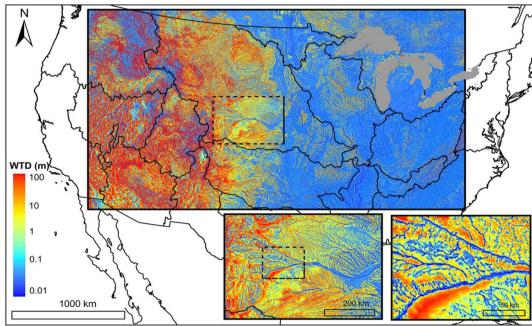
Not all spatial data is gridded



USGS Watershed Boundary Dataset

<https://www.usgs.gov/core-science-systems/ngp/national-hydrography/watershed-boundary-dataset>

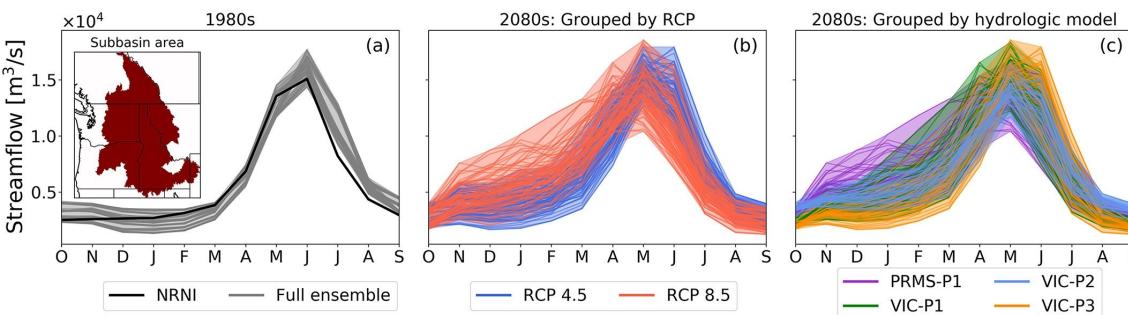
Data comes in many forms from many sources ...



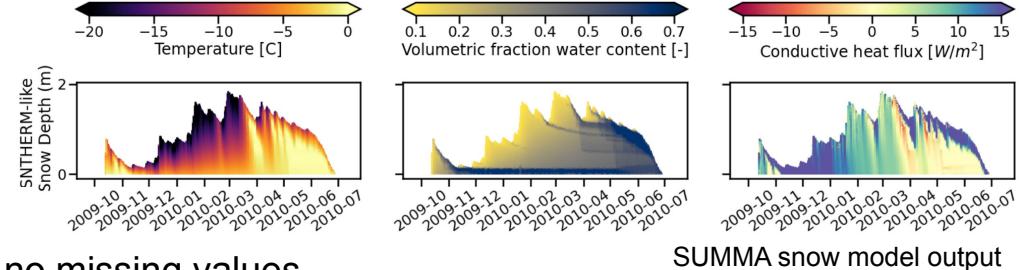
ParFlow: Modeled water table depth
Kuffour et al., 2020:
<https://doi.org/10.5194/gmd-13-1373-2020>

Climate change projections for streamflow at the Dalles along the Columbia River

Chegwidden et al., 2019: <https://doi.org/10.1029/2018EF001047>



Model output



Advantages:

- Typically no missing values
- Can provide insight into variables that cannot be measured easily (or at all)
- Can provide insight into conditions that do not exist: scenarios, forecasting, climate change, land use change, etc.: **what-if**

SUMMA snow model output

Challenges:

- Model error
- Model uncertainty
- Data format
- Data size
- Many different spatial arrangements
- etc.

Data discovery

Data tends to be scattered across many sites, repositories, etc. and may be difficult to locate at times.

Data is also made available under many different licenses and with many different restrictions and is easier to access in some geographic regions than in others.

A number of initiatives exist to make data discovery easier, but it often remains a significant effort.



Where do I find the data I need?

No single catalog will meet all your needs for search, discovery, and convenient access to data, equally well, in all domains of water research!

Possible starting points:

- **Google Search**
Overwhelming. Many irrelevant results (but often the first place to start)
- **Google Earth Engine:** <https://earthengine.google.com/>
Fantastic for remote sensing data and processing. Also holds some gridded model products. Not yet there for site time series, site data
- **Google Dataset Search:** <https://toolbox.google.com/datasetsearch>
New, promising, but still rather messy

Where do I find the data I need?

Many topical or organizational repos, e.g. in U.S.:

- USGS: <https://waterdata.usgs.gov/nwis>

The screenshot shows the USGS National Water Information System (NWIS) homepage. At the top, there's a banner with five images: a flooded area, a forest, a fire, a damaged vehicle, and two people working. To the right of the banner are links to 'USGS Home', 'Contact USGS', and 'Search USGS'. Below the banner, the title 'National Water Information System: Web Interface' is displayed, along with 'USGS Water Resources' and dropdown menus for 'Data Category' (set to 'Home') and 'Geographic Area' (set to 'United States'). A 'GO' button is also present. A yellow callout box contains a link to 'Click to hide News Bulletins' and two news items: 'Introducing The Next Generation of USGS Water Data for the Nation' and 'Full News' with a RSS icon. The main content area features a large blue header 'USGS Water Data for the Nation'. Below it are sections for 'Search for Sites With Data' (with 'Current Conditions' and 'Site Information' buttons) and 'Introduction'.

USGS Water Data for the Nation

Search for Sites With Data

Current Conditions Sites with real-time or recent surface-water, groundwater, or water-quality data.

Site Information Descriptive site information for all sites with links to all available water data for individual sites.

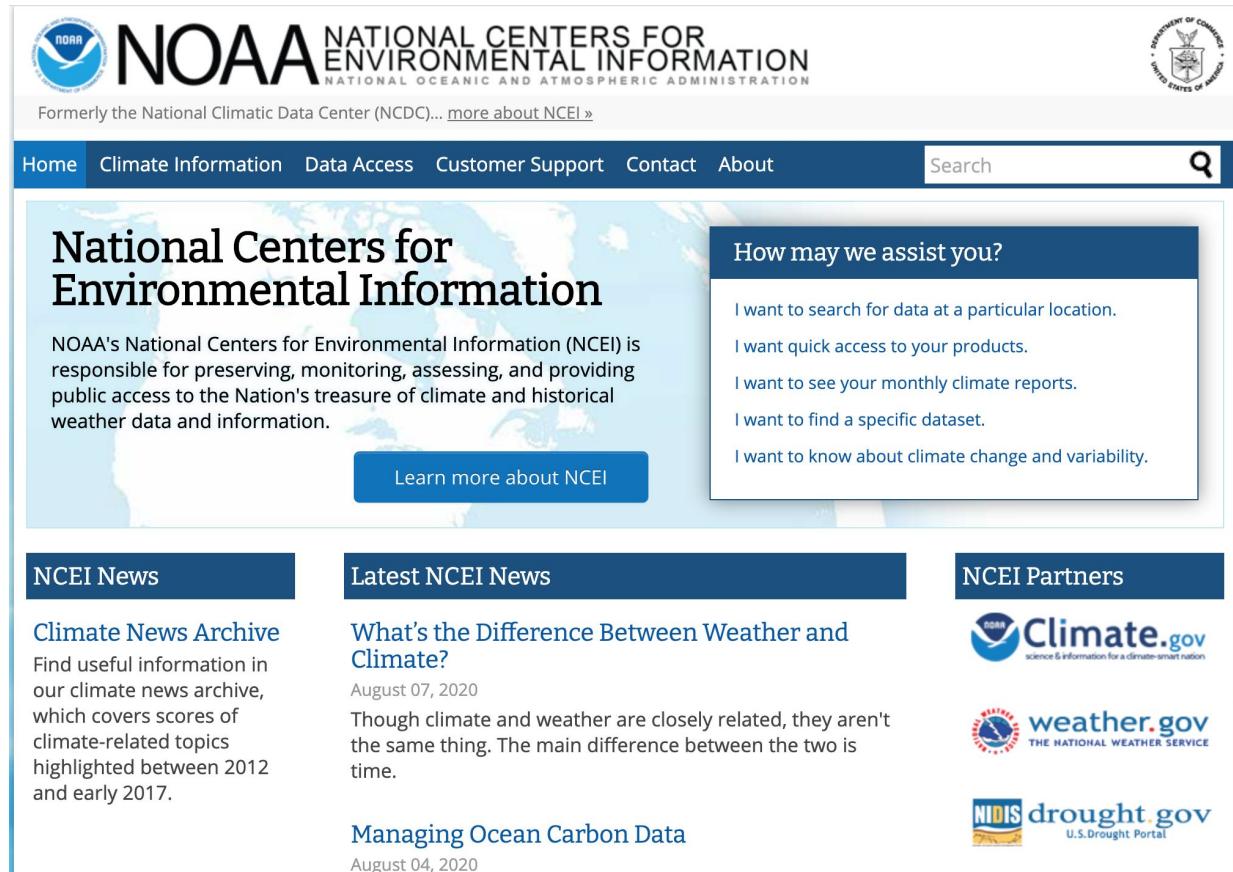
Introduction

These pages provide access to water-resources data collected at approximately 1.9 million sites in all 50 States, the District of Columbia, Puerto Rico, the Virgin Islands, Guam, American Samoa and the Commonwealth of the Northern Mariana Islands. Online access to this data is organized around the categories listed to the left.

Where do I find the data I need?

Many topical or organizational repos, e.g. in U.S.:

- NOAA NCDC:
<https://www.ncdc.noaa.gov/>



The screenshot shows the homepage of the NOAA National Centers for Environmental Information (NCEI). The header features the NOAA logo and the text "NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". A search bar is at the top right. Below the header, a main banner reads "National Centers for Environmental Information" and describes NCEI's mission to preserve, monitor, assess, and provide public access to climate and historical weather data. A "Learn more about NCEI" button is in the center of the banner. To the right, a sidebar titled "How may we assist you?" lists five options: "I want to search for data at a particular location.", "I want quick access to your products.", "I want to see your monthly climate reports.", "I want to find a specific dataset.", and "I want to know about climate change and variability.". At the bottom, there are three sections: "NCEI News", "Latest NCEI News", and "NCEI Partners".

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Formerly the National Climatic Data Center (NCDC)... [more about NCEI](#)

Home Climate Information Data Access Customer Support Contact About Search

National Centers for Environmental Information

NOAA's National Centers for Environmental Information (NCEI) is responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of climate and historical weather data and information.

Learn more about NCEI

How may we assist you?

I want to search for data at a particular location.
I want quick access to your products.
I want to see your monthly climate reports.
I want to find a specific dataset.
I want to know about climate change and variability.

NCEI News

Climate News Archive
Find useful information in our climate news archive, which covers scores of climate-related topics highlighted between 2012 and early 2017.

Latest NCEI News

What's the Difference Between Weather and Climate?
August 07, 2020
Though climate and weather are closely related, they aren't the same thing. The main difference between the two is time.

NCEI Partners

 Climate.gov
science & information for a climate-smart nation

 weather.gov
THE NATIONAL WEATHER SERVICE

 NIDIS drought.gov
U.S.Drought Portal

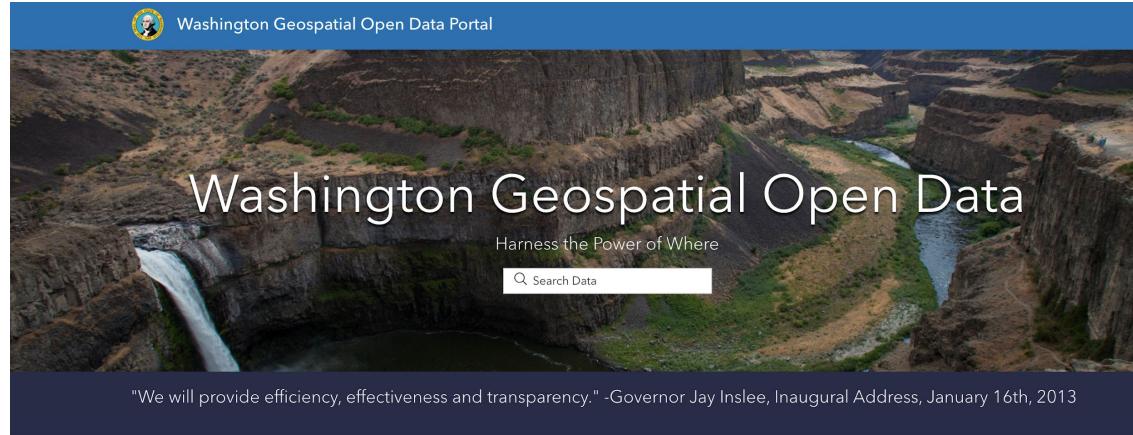
Managing Ocean Carbon Data
August 04, 2020

Where do I find the data I need?

Many topical or organizational repos, e.g. in U.S.:

- Washington State:

<https://geo.wa.gov>



Highlights

Explore Washington Applications

This section displays three cards related to Washington State emergency operations. The first card on the left is a map of Washington state's county boundaries, with each county labeled with its name. The second card in the center features the logo of the Washington State Emergency Operations Center (WISE) and the text "Washington State Emergency Operations Center Dashboard". The third card on the right is a map showing a network of red and blue lines and dots, likely representing emergency response routes or incident locations.

Where do I find the data I need?

CUAHSI also provides data discovery tools that are relevant for hydrology:

- CUAHSI: <https://www.cuahsi.org/data-models/discovery-and-analysis>

The screenshot shows the CUAHSI Discovery and Analysis tool. At the top, there is a navigation bar with links for News & Events, Contact Us, Quick Links, ABOUT, EDUCATION, DATA & MODELS (which is highlighted in blue), PROJECTS, FUNDING OPPORTUNITIES, COMMUNITY, LIBRARY, and a search icon. The CUAHSI logo is prominently displayed, along with a note that it is sponsored by the National Science Foundation.

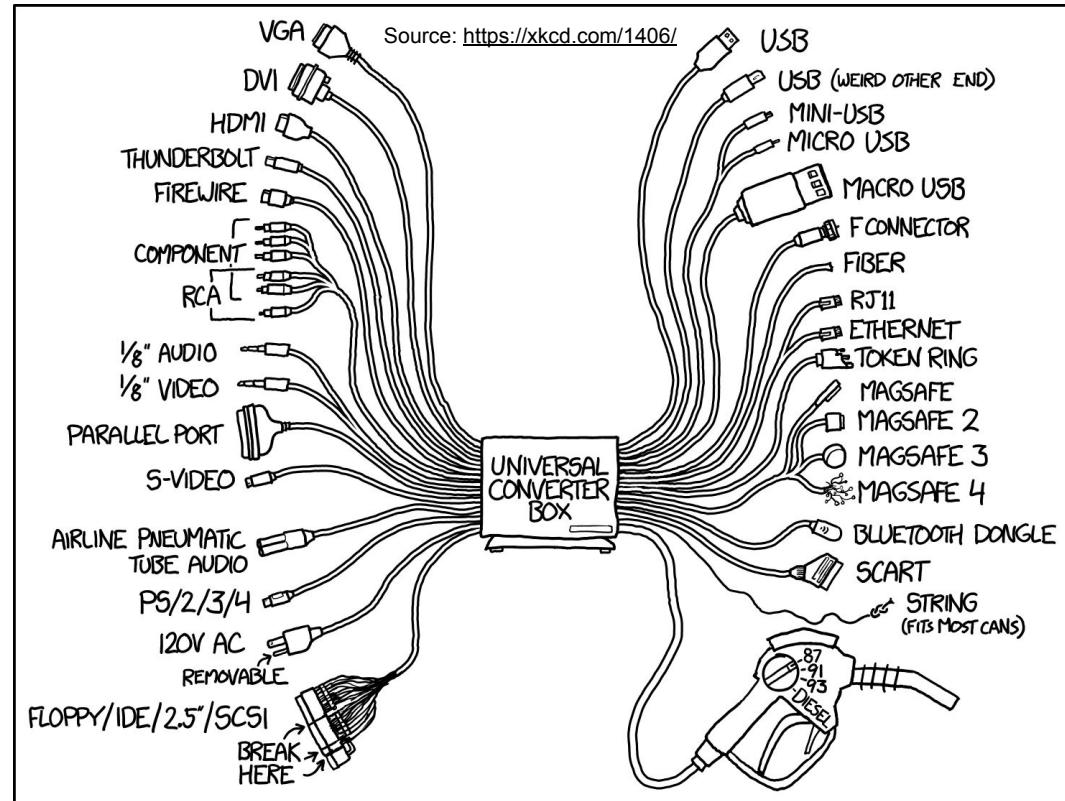
The main content area has a title "Discovery and Analysis". Below it, a sub-section titled "DATA & MODELS" lists various tools: Hydroinformatics Innovation Fellowship, Discovery and Analysis (which is selected and highlighted in orange), Publish Your Data, Data Management Plans, Data Portals, Community Models, For Developers, Legacy Tools, and FAQs.

The central part of the page features a map of the New England region, specifically showing Massachusetts and parts of Connecticut and Rhode Island. The map displays numerous blue circular markers representing data sources or monitoring stations, with labels like "Mashpee", "Nauset", "Dracut", "Lowell", "Lawrence", "Beverly", "Gloucester", "Rowley", "Plum Island", "Haverhill", "Salisbury", "Westford", "Acton", "Concord", "Burlington", "Woburn", "Danvers", "Salem", "Manchester-by-the-Sea", "Beverly", "Lynn", "Cambridge", "Waltham", and "Winthrop". A "Layer Control" dropdown menu is open, showing options like "Select Date Range" (From: 06/01/1900, To: 06/01/2017), "Select Keyword(s)" (All), and "Select Data Service(s)" (Selected 87 of 94). Buttons for "Search Map" and "Reset" are also visible.

Data access, analysis, and export

Once you find the right data, accessing it, manipulating it, and exporting it in a usable form can be a significant task.

Fortunately there is a large set of tools that make this process easier. We'll see some of these later today.



Data access and ingest

Common approaches and tools

- Manual browsing, downloads, and reading local files (but issues of reproducibility, efficiency, thoroughness)
- [requests Python package](#) (and wget, curl): generic remote access through the web.
- Pandas read_csv function. Not just local files, but also remote files.
- Custom web APIs (often called "REST" APIs) from the data provider (eg, NEON). Often fairly easy to use, but highly variable across systems.
- Standards-based resources:
 - APIs: OPeNDAP, Open Geospatial Consortium (OGC) Web Services (WFS, SOS, etc), **CUAHSI WaterOneFlow**
 - Formats: WaterML (**CUAHSI WaterML 1.x** vs OGC WaterML 2.0), NetCDF (3 "classic" vs 4), Metadata standards
 - See this [old but still very useful descriptions of CUAHSI "HIS" standards](#)
 - Standards enable reusability across multiple data sources, systems
- ulmo. Water and climate data. Wraps a lot of the underlying complexity into simpler, more user-friendly Python APIs.

Data storage

Data volumes can become very large, very quickly.

Examples:

- Remote sensing time series
- Model output

Storing data on a local hard-drive without backups is (more than) risky.

Downloading data locally can be very challenging when the volumes are large.

For very large data sets: Take computing to the data

- Examples: [Google Earth Engine](#), [Pangeo](#), [Hydroshare](#).

Data publishing - why?

AGU DATA POLICY

AGU affirmed in its 2015 position statement that "*Earth and space science data should be widely accessible in multiple formats and long-term preservation of data is an integral responsibility of scientists and sponsoring institutions.*" Following this statement and to advance scientific exploration and discovery, and allow a full assessment of results presented in AGU's journals, *all data necessary to understand, evaluate, replicate, and build upon the reported research must be made available and accessible whenever possible.*

AGU encourages authors to identify and archive their data in approved data centers.

AGU requires an explicit statement in the "Acknowledgments" section of a paper that clarifies how users can access the data from a paper (via supplements, repositories, other sources, etc.) and states any restrictions on access.

Data publishing - FAIR

FAIR - Findable, Accessible, Interoperable, Reusable

The **FAIR principles** were designed with data-driven and machine-assisted open science in mind. The final aim of following FAIR principles is that machines as well as people can Find, Access, Interoperate and thus Reuse each other's research objects.

FAIR is **not a standard**, although the acronym is frequently used in that context. The GO FAIR view is that standards are needed for the **Internet of FAIR Data and Services** and that ideally, standards, API's and protocols are developed 'following FAIR guiding principles'.

FAIR is **not equivalent to open** (and open is not equivalent to 'free').

FAIR principles do not, in themselves, cover the crucial aspects of **intrinsic data quality or ethics**.

Data publishing - where?

Sites that offer the opportunity to mint a DOI (Digital Object Identifier) for data sets:

- **HydroShare:** [https://help.hydroshare.org/introduction-to-hydroshare/getting-started/permanently-publish-a-resource/Permanently Publish](https://help.hydroshare.org/introduction-to-hydroshare/getting-started/permanently-publish-a-resource/Permanently%20Publish)

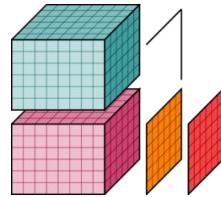
IMPORTANT! You should view publishing a resource the same as publishing a research paper. Once published, you can no longer change a resource's content or metadata description. Please make sure you are ready to publish your resource before you use the following steps.
- **Zenodo:** <https://zenodo.org/>

Python ecosystem

ulmo



Rasterio: access to geospatial raster data



xarray



NumPy 

matplotlib 

More information

- <https://waterhackweek.github.io/learning-resources/>
- <https://www.cuahsi.org/education/cyberseminars/waterhackweek-cyberseminar-series/>
- <https://github.com/waterhackweek>