Data Access and Time-series Statistics

- **Emilio Mayorga**, University of Washington - APL
- **Yifan Cheng**, University of Washington - CEE

WaterHackWeek (https://waterhackweek.github.io) Cyberseminar. February 7, 2019

Data types and this seminar

- Spatial representation? Point, gridded, etc.
- Long time series, near-real-time data, high-frequency sensor data?
- USA, International, Global?
- Academic research or government monitoring and research?
- Atmospheric, surface water, groundwater? Variables of interest?

- **Seminar focused on time series data from sites commonly represented as fixed geographical points.** And we'll emphasize surface waters.

Where do I find the data I need? Google?

- **Google Search.**
  - Overwhelming. Many irrelevant results.
- **Google Earth Engine.** https://earthengine.google.com/ (https://earthengine.google.com/)
  - Fantastic for remote sensing data and processing. Not yet there for site time series, site data
- **Google Dataset Search.** https://toolbox.google.com/datasetsearch (https://toolbox.google.com/datasetsearch)
  - New, promising. Probably a good place to start.

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

- But some systems are good, broad starting points.
- Some make it easier to identify, ingest and use the *"granular"* data consistently across sources, while some only target the *"dataset"* and metadata level.
- See this nice (but a bit dated) **GLEON** page comparing DataONE and CUAHSI HIS-HydroClient (http://gleon.org/data/repositories). Valuable comparison beyond those two systems.
- All have pros and cons.

Catalogs: User Interface vs APIs

- Most catalog systems provide both a user interface (UI) for interactive browsing, and an *Application Programming Interface (API)* for remote programmatic access through the web. The UI is typically easier to use, but the API facilitates repeated tasks, large queries, reuse, and reproducibility.

Data Search Strategies, Considerations

- Clarify the type of data your research question requires
- Ask your colleagues about sources of relevant data!
- Do try Google. But also use specialized (Earth Sciences, hydrology, etc) but still broad catalogs.
- Does the catalog make it easier to access data in consistent data formats across data sources? Or is it focused on "datasets"?
- Go to each individual data provider's web site, or try to find a system or tool that spans multiple data sets?

... continued

- Manual file downloading, or programmatic access ("web services"; say, from Python)?
- What options does the data provider have? Manual file downloads, custom or standards-based web service API, custom or standards-based data formats?
- Are there multi-dataset code packages that can handle the dataset you want? Is the package fairly easy to use, well documented, with plenty of examples, and does it have an active community of users and developers (say, on GitHub)?
- Are parameters ("water temperature", "stream discharge", "rainfall", etc) well described? Do different data sources use different parameter names, meanings, and units?

Searching for Data: Useful systems/catalogs we'll explore

1. **CUAHSI HIS-HydroClient.** https://www.cuahsi.org/data-models (https://www.cuahsi.org/data-models). HydroClient, http://data.cuahsi.org (http://data.cuahsi.org), is a web application providing discovery, visualization and download capability for site time series from many different sources accessible through the CUAHSI "HIS Central" catalog http://his.cuahsi.org/ (http://his.cuahsi.org/)
2. **CUAHSI HydroShare.** https://www.hydroshare.org/search/ (https://www.hydroshare.org/search/) (already discussed in previous seminars)
3. **USGS NWIS.** National Water Information System - USGS Water Data for the Nation: https://waterdata.usgs.gov/nwis (https://waterdata.usgs.gov/nwis). Water quantity and quality from surface waters and groundwater.
4. **Observatories, Field site networks.** NEON (National Ecological Observatory Network) (https://www.neonscience.org), LTER (https://lternet.edu/), CZO (http://criticalzone.org/national/), GLEON (http://gleon.org/), USDA watersheds.
5. **MonitorMyWatershed / ModelMyWatershed web application.** https://modelmywatershed.org (https://modelmywatershed.org). Provides user-friendly search capabilities across 4 catalogs: HydroShare, CUAHSI HIS-HydroClient, CINERGI, and Water Quality Portal; plus integration with HydroShare.

... Other Systems (explore on your own)

- EarthCube Data Discovery Studio (formerly "CINERGI") (http://DataDiscoveryStudio.org). Broad "meta" catalog that integrates and harmonizes many different catalogs across the Earth Sciences. Operates at the dataset level.
- DataONE (https://search.dataone.org). Broad catalog operating at the dataset level, with a diverse range of "Earth observational" data, including water data. US based and strongest for US data, but integrates datasets from other regions.
- PANGEA (https://www.pangaea.de). "archiving, publishing and distributing georeferenced data from earth system research". Based in Germany, PANGEA operates at the dataset level and archives both large and small datasets, often via agreements with journals.
- Water Quality Portal (https://www.waterqualitydata.us). "A cooperative service sponsored by USGS, EPA and the National Water Quality Monitoring Council (NWQMC) that integrates publicly available water quality data from USGS NWIS, the EPA STORET Data Warehouse, and USDA STEWARDS."
- **NOAA NCDC.** Climate data -- next slide.
- US agencies
  - Other federal agencies (Army Corps. Eng., DOE, USDA)
  - State agencies (and local agencies!)

Climate Data

- **NOAA NCDC** Land-based station data (US and Global) (https://www.ncdc.noaa.gov/data-access/land-based-station-data)
  - Multiple datasets by a single agency provider
  - Multiple access mechanisms (custom REST API; standards-based APIs; etc)
  - Access for a subset of datasets via independent, specialized tool: CUAHSI HIS/HydroClient, `ulmo`
- **NASA ORNL Daymet** (US) (https://daymet.ornl.gov)
  - 1km x 1km gridded daily data, including query access via lat-lon points
  - Multiple access mechanisms
  - Access via independent, specialized tools: `ulmo`, daymetpy (http://khufkens.github.io/daymetpy/)

Brief NCDC drill in online:

- NCDC Land-Based Station Data pages (Datasets, CDO (https://www.ncdc.noaa.gov/cdo-web/), web services (https://www.ncdc.noaa.gov/cdo-web/webservices/ncdcwebservices), CDO REST API (https://www.ncdc.noaa.gov/cdo-web/webservices/v2)): https://www.ncdc.noaa.gov/data-access/land-based-station-data (https://www.ncdc.noaa.gov/data-access/land-based-station-data)
- `ulmo` NCDC Global Historical Climate Network Daily (GHCN) plugin (https://ulmo.readthedocs.io/en/latest/api.html#module-ulmo.ncdc.ghcn_daily). Also NCDC GSOD, CRIS plugins.

Python

- **Pandas** is the backbone. "DataFrame" tabular data structure. Incorporates lots of functionality and core Python tools: read/write, data organization, data exploration, cleaning, and summarizing; Numpy, matplotlib plotting
- **GeoPandas.** Geospatially enabled Pandas, incorporating several useful geospatial tools.
- Matplotlib plotting.
- **ulmo** data access package. https://ulmo.readthedocs.io (https://ulmo.readthedocs.io)
- Python datetime handling.
  - Beware of different Python datetime types: Python standard `datetime` type; Numpy datetime; Pandas `Timestamp`
  - Timezone handling; datetime utilities, conversion
  - See this (https://medium.com/jbennetcodes/dealing-with-datetimes-like-a-pro-in-pandas-b80d3d808a7f) and that (https://medium.com/jbennetcodes/dealing-with-datetimes-like-a-pro-in-python-fb3ac0feb94b) blog posts

Data access and ingest. Common approaches, tools

- Manual browsing, donwloads, and reading local files (but issues of reproducibility, efficiency, thorougness)
- `requests` Python package (https://stackabuse.com/the-python-requests-module/) (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 (http://his.cuahsi.org/wofws.html)
  - 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.

... Other Python packages

- We won't discuss these, but you can explore them
- `climata` (https://github.com/heigeo/climata)
- `MetPy` (https://github.com/Unidata/MetP)
- `obsio` (https://github.com/jaredwo/obsio)
- provider-specific packages, such as `daymetpy`

Next: Use cases with Python examples

- Southeast US (Yifan's research area)
- Surface water temperature, discharge and water quality
- Running Python code: conda (https://geohackweek.github.io/datasharing/01-conda-tutorial/), conda environment, and Jupyter notebooks
- All materials available on GitHub repository, https://github.com/waterhackweek/tsdata_access (https://github.com/waterhackweek/tsdata_access)
- Use cases
    1. River and reservoir water temperature. (https://nbviewer.jupyter.org/github/waterhackweek/tsdata_access/blob/master/Part1-ulmo_nwis_reservoirtemperature.ipynb) Yifan
    2. Search, access and initial look. Notebook 1 (https://nbviewer.jupyter.org/github/waterhackweek/tsdata_access/blob/master/Part2-ulmo_nwis_and_cuahsi.ipynb), and notebook 2 (https://nbviewer.jupyter.org/github/waterhackweek/tsdata_access/blob/master/Part2-NEON_and_Observatories_RESTAPI.ipynb). Emilio