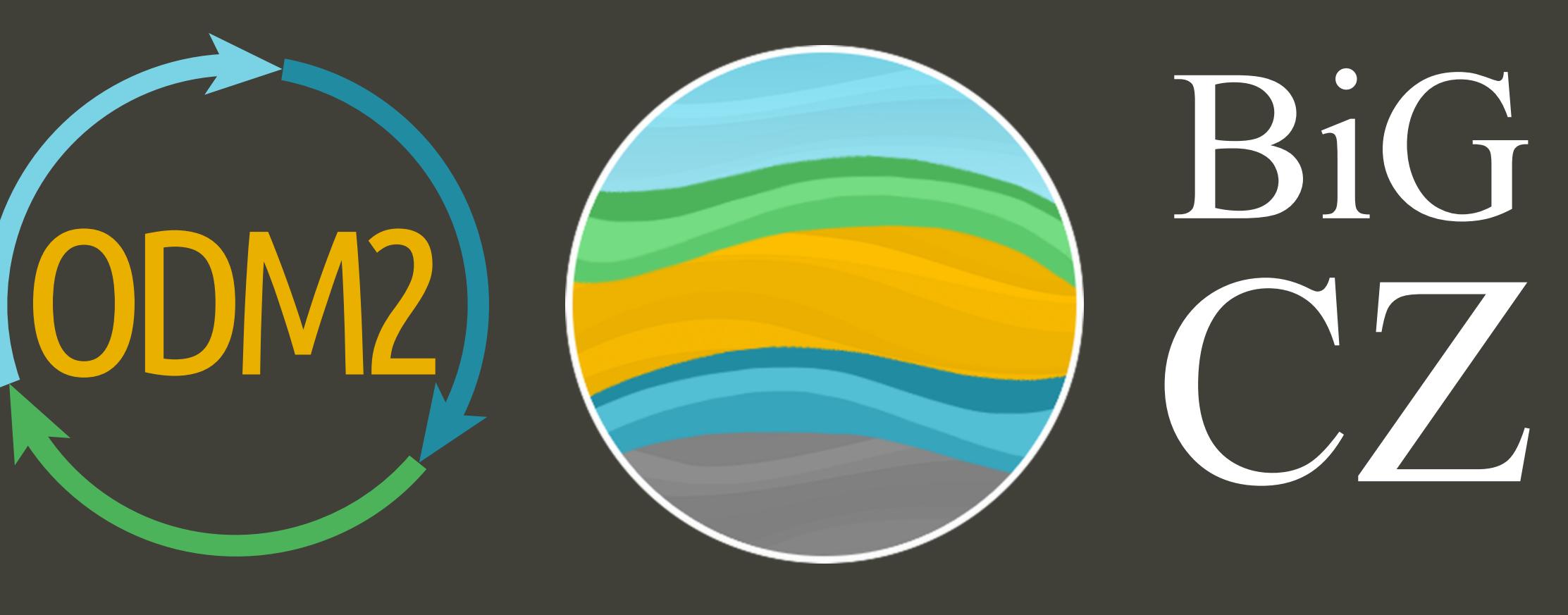


# BiG CZ Data Project: Integration of bio- & geoscience (BiG) data for the Critical Zone (CZ) sciences using the ODM2 standards & software ecosystem

Anthony Aufdenkampe<sup>a</sup>, Emilio Mayorga<sup>b</sup>, Jeffery S. Horsburgh<sup>c</sup>, Kerstin Lehnert<sup>d</sup>, Ilya Zaslavsky<sup>e</sup>, David Valentine<sup>f</sup>, Sara Damiano<sup>m</sup>, Robert Cheetham<sup>f</sup>, Steve Richard<sup>g</sup>, Folker Meyer<sup>h</sup>, Chris Henry<sup>h</sup>, Gary Berg-Cross<sup>i</sup>, Rolof Versteeg<sup>j</sup>, Aaron Packman<sup>k</sup>, Emma Aronson<sup>l</sup>



## Introduction

We have developed the Observations Data Model v2.0 (ODM2) standards and software ecosystem (<http://ODM2.org>) specifically to solve the challenges of managing, sharing and integrating diverse data from of biological and geological (BiG) disciplines for Critical Zone (CZ) science.

ODM2 is at the core of functionality behind the a number of cyber-infrastructure projects, including: Interdisciplinary Earth Data Alliance's (IEDA) EarthChemDB (<http://www.iedadata.org>), EnviroDIY's Water Quality Data Portal (<http://data.envirodiy.org>) and the BiG CZ Data Portal and Toolbox (<http://bigcz.org>).

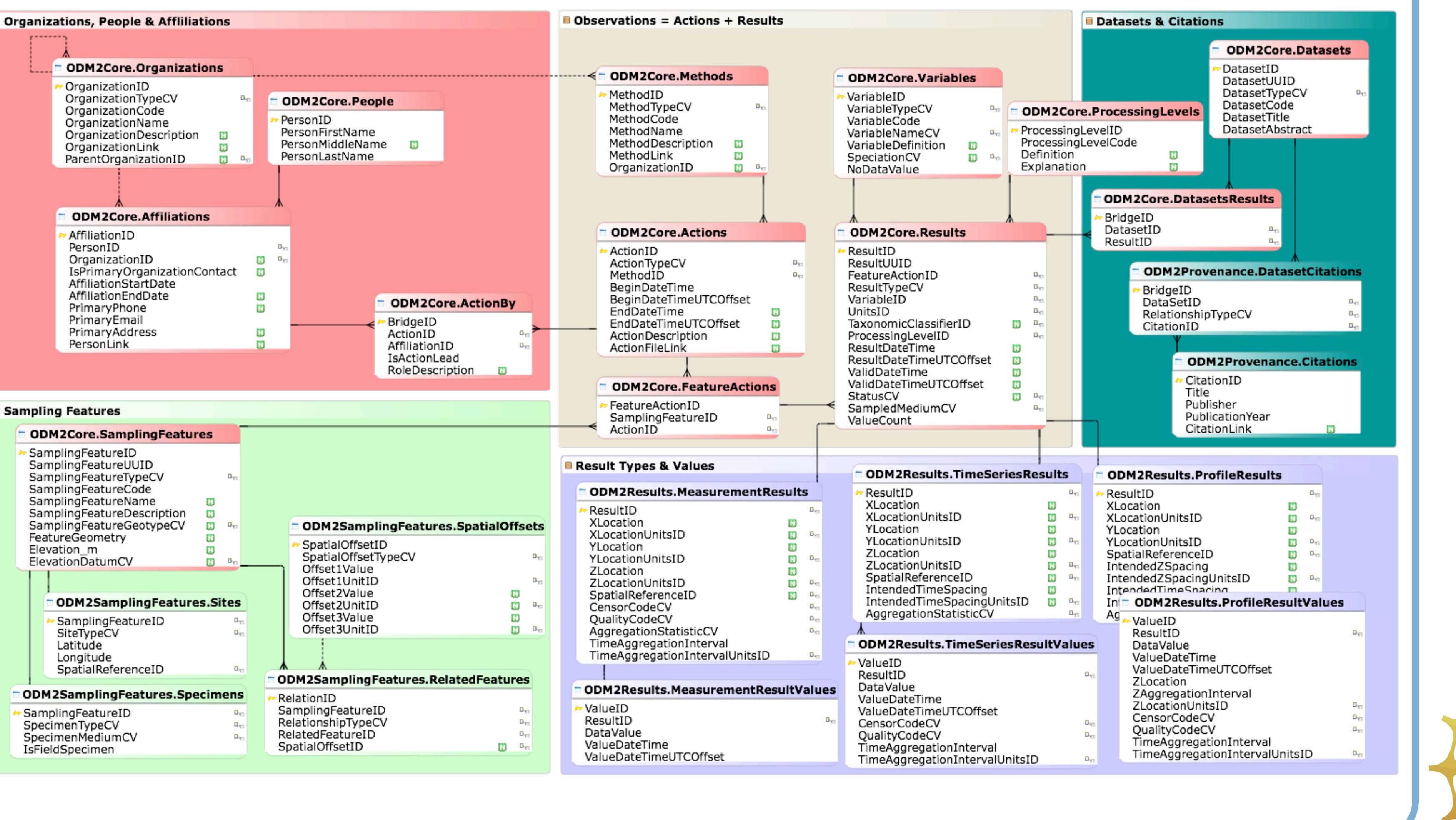
## Observation Data Model v2 (ODM2)

ODM2 is a new information model aimed at facilitating greater interoperability across scientific disciplines and domain cyberinfrastructures, including: hydrological time series, soil and sediment geochemistry, biodiversity surveys, oceanographic sensor profiles, multi-dimensional spectra, and more (Horsburgh et al. 2016).

ODM2 is a profile of Open Geospatial Consortium's (OGC) Observations and Measurements (O&M) standard. ODM2 was developed with community input by integrating and enhancing concepts from ODM1 (from CUAHSI Hydrological Information System), IEDA PetDB & SedDB and other existing information models.

ODM2 metadata structure provides a powerful new way to describe sampling, sensing, and analysis workflows.

- A flexible **Sampling Feature extension** tracks relationships in nested hierarchies of **sampling sites** (weather stations, stream gauges, soil pits, wells, bore holes, etc.), **field specimens** (rocks, cores, soils, water, gas, plants, etc.), and **lab specimens** (i.e. subsamples, fractions, powders, extracts, digests). This extension is fully compatible with OGC O&M Sampling Features and International Geo Sample Numbers (IGSN).
- The ability to record **Actions** (e.g., sensor deployments, specimen collection, lab analyses), relationships between Actions, **People** that perform them, the **Methods** and **Equipment** that are used, and their **Results** and **Values** of many types.
- Object-Relational-Mapping** translates entity types to a class-subclass structure that enables inclusion of domain-specific attributes.

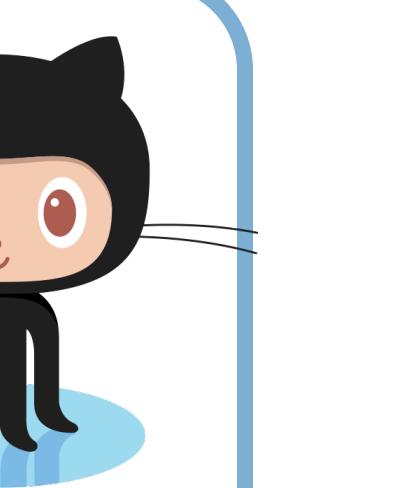


## ODM2 Controlled Vocabularies

ODM2 CVs were developed by combining the best terms from existing CVs (i.e. CUAHSI CVs, IEDA EarthChem CVs) and by developing numerous new "Types" CV lists to facilitate data integration across disciplines.

We developed a new web-based system that combines community moderation with a Simple Knowledge Organization System (SKOS) and associated web services: <http://vocabulary.odm2.org>

## ODM2 Software Ecosystem



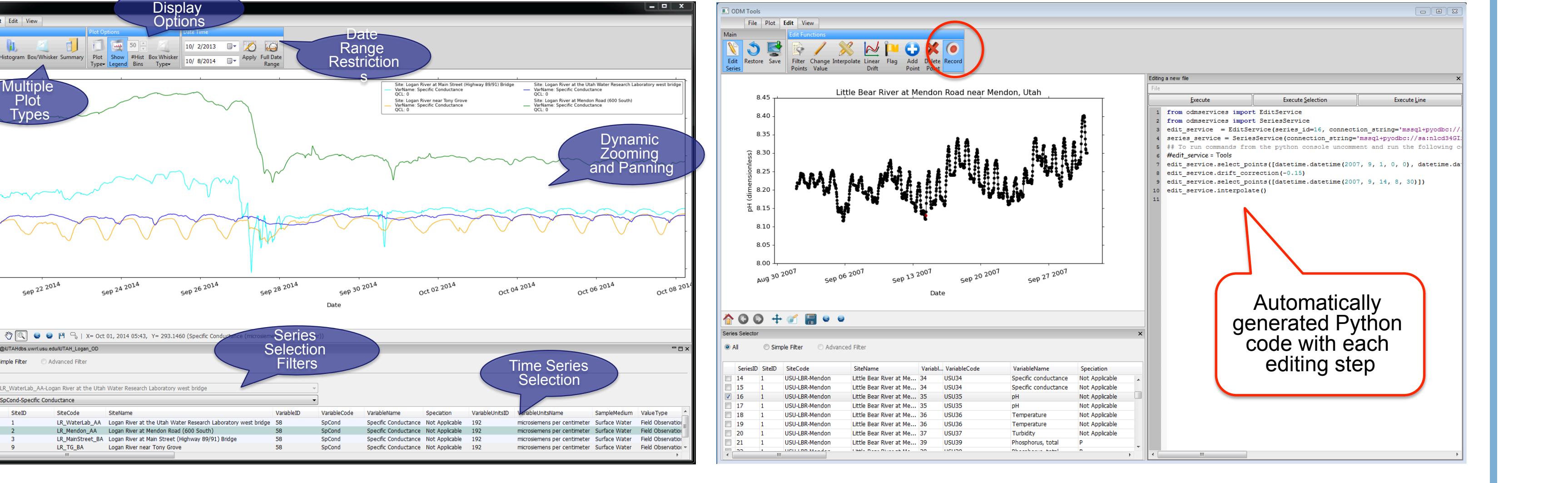
The entire ODM2 standards and software ecosystem is available and documented on GitHub, <https://github.com/ODM2>, and is written in Python for cross-platform (Windows, Mac, Linux) use and to encourage contributions.

### ODM2 Information Model Documentation & Relational Database Schemas:

Documentation and scripts for creating ODM2 relational databases with SQL Server, MySQL, PostgreSQL, and SQLite. <https://github.com/ODM2/ODM2>.

**ODM2 Python Application Programming Interface (API):** A cross platform, Python and SQL Alchemy-based application programming interface (API) for ODM2. <https://github.com/ODM2/ODM2PythonAPI>.

**ODM Tools Python:** A desktop application for visualization and managing sensor datasets in ODM2. <https://github.com/ODM2/ODMToolsPython>.



**ODM2 Streaming Data Loader:** A Python based program for loading streaming sensor data into ODM2. <https://github.com/ODM2/ODM2StreamingDataLoader>.

**ODM2 Web Streaming Data Loader:** A Python/Django-based web application and web service for registering data collection devices and allowing them to stream sensor observations into ODM2. <https://github.com/ODM2/ODM2WebsDL>.

**WaterOneFlow for ODM2:** A web service interface for delivering data from ODM2 via WaterML 1.1 or OGC WaterML 2.0. <https://github.com/ODM2/WOFpy>.

**ODM2 Admin:** A web-based Django/Python admin app for inputting, managing and visualizing data in ODM2. <https://github.com/miguelleon/ODM2-Admin>.

**HydroShare ODM2 Time Series:** An ODM2 in SQLite for storing hydrologic time series in HydroShare. <https://github.com/hydroshare/hydroshare>.

**ODM2 Web Services:** Web services for sharing all data types from ODM2. <https://github.com/ODM2/ODM2RESTfulWebServices> (under development).

**BiG CZ Toolbox:** A single, easy-to-install and easy-to-teach cross-platform package of the ODM2 software ecosystem, using conda packaging system & conda virtual environments. <https://github.com/BiG-CZ/BiG-CZ-Toolbox> (under development).

## YAML Observation Data Archive & Exchange (YODA) File Format & Tools

YODA is a text file representation of an ODM2 dataset, designed to be a self-describing archival and exchange format. <https://github.com/ODM2/YODA-File>.

YODA Profiles describe metadata schemas for the enclosed data array for interchange by ODM2 web services, regardless of syntax (i.e. YAML vs JSON vs XML).

[https://github.com/ODM2/YODA-File/blob/master/doc/YODA\\_profiles.md](https://github.com/ODM2/YODA-File/blob/master/doc/YODA_profiles.md)

YODA Tools is a library to validate data files according to YODA Profiles & ODM2 CVs, and to import/export data w/ an ODM2 database.

<https://github.com/ODM2/YodaTools>.

All under development

## BiG CZ Data Portal

A web application prototype for intuitive, high-performance map-based discovery, visualization, analysis and publication of diverse data, including: points with sensor and sample based observational data, GIS data, and satellite data.

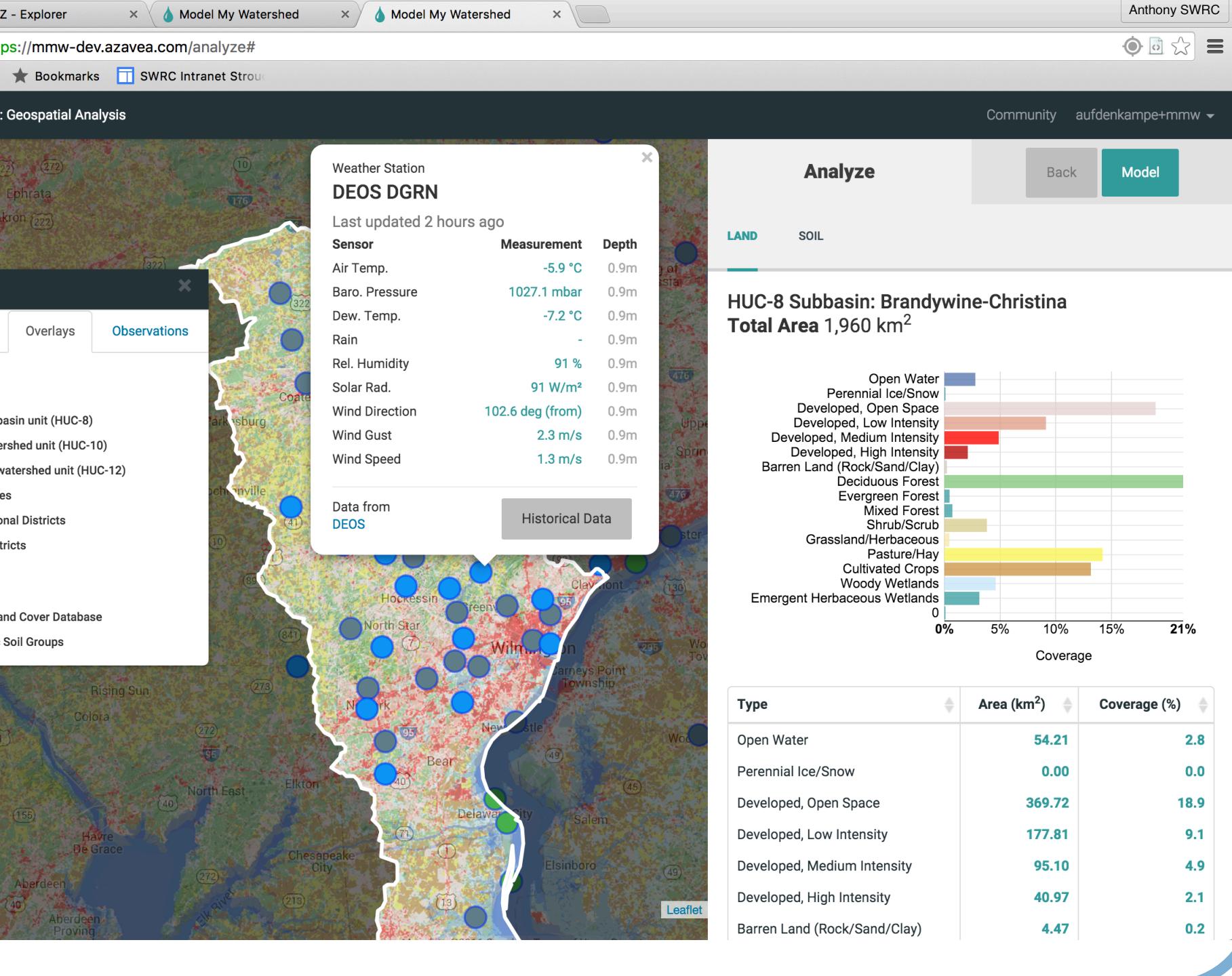
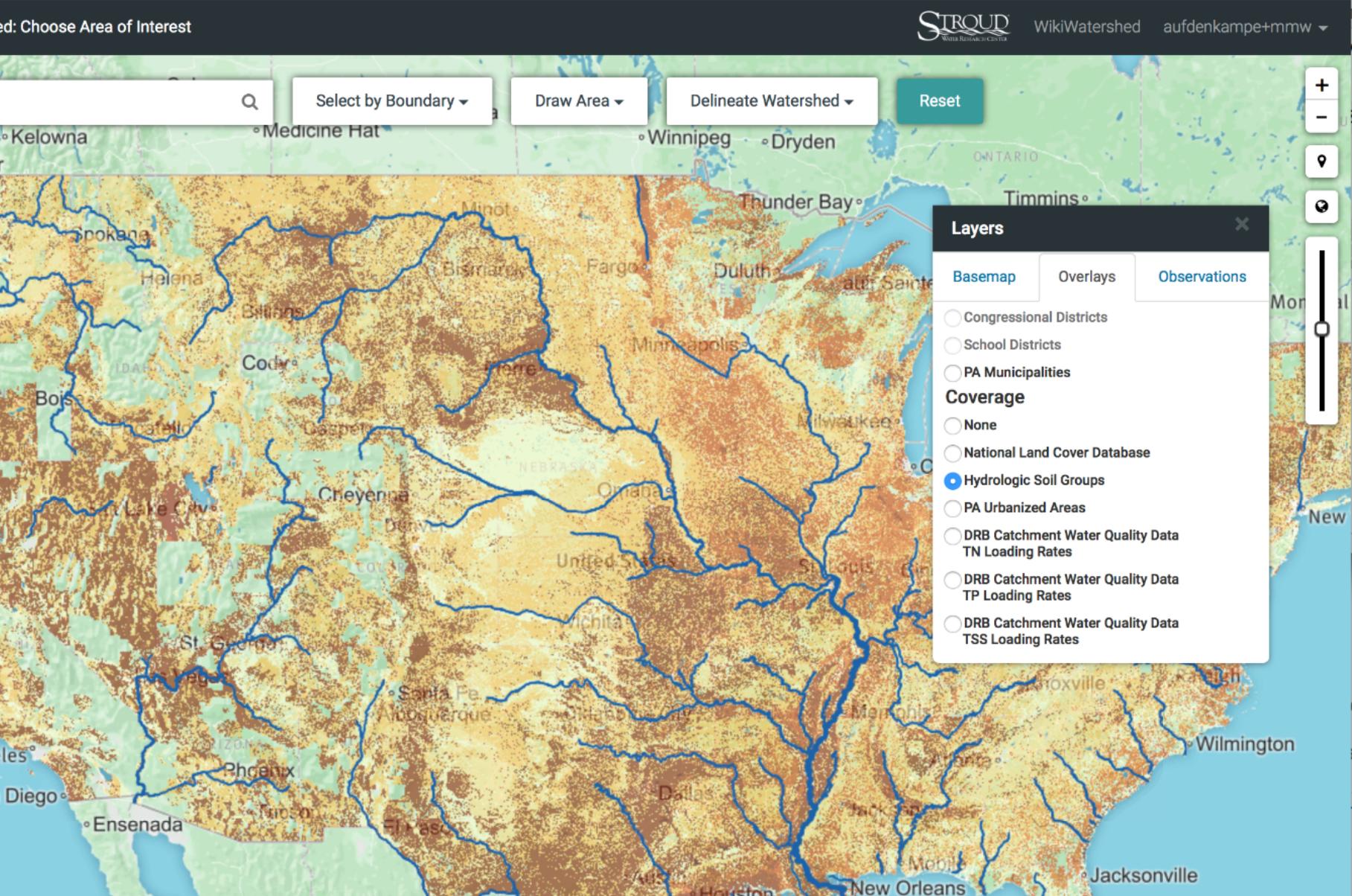
We have prototyped the ingestion of point observation data from many sources (USGS, CUAHSI WDC, System for Earth Sample Registration/IGSN and MG-RAST at <http://portal.bigcz.org>.

We have also developed the core geoprocessing within our, Model My Watershed project, based on: Amazon EC2 & Apache Spark; Amazon S3 Elastic File System; GeoTrellis Geoprocessing; Python, PostGIS, PostgreSQL; Django Python Web Framework.

Anyone can now perform 30 m on-the-fly watershed delineation anywhere in the continental USA at <https://app.wikiwatershed.org>. Code at

<https://github.com/WikiWatershed>

We are in the process of combining all of these capabilities.



Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., ... & Zaslavsky, I. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations. Environmental Modelling & Software, 79, 55–74. DOI: 10.1016/j.envsoft.2016.01.010



NSF Awards:  
GEO-EAR 1153164,  
1224638,  
1332257  
CISE-ACI 1339834

**LimnoTech**  
Water Environment Scientists Engineers

**Utah State University**  
WATER  
UNIVERSITY OF WASHINGTON



**SDSC**  
SAN DIEGO SUPERCOMPUTER CENTER

**STROUD**  
WATER RESEARCH CENTER

**azavea**