Data Management and Sharing within a Large-Scale, Heterogeneous Sensor Network using the CUAHSI Hydrologic Information System

Jeffery S. Horsburgh

Amber Jones and the

iUTAH Cyberinfrastructure Team

What is CUAHSI?

- Consortium of Universities for the Advancement of Hydrologic Science, Inc.
- Founded in 2001, supported by the National Science Foundation
- Research organization representing more than 130 U.S. universities
- Develop infrastructure and service for advancement of water science in the U.S.

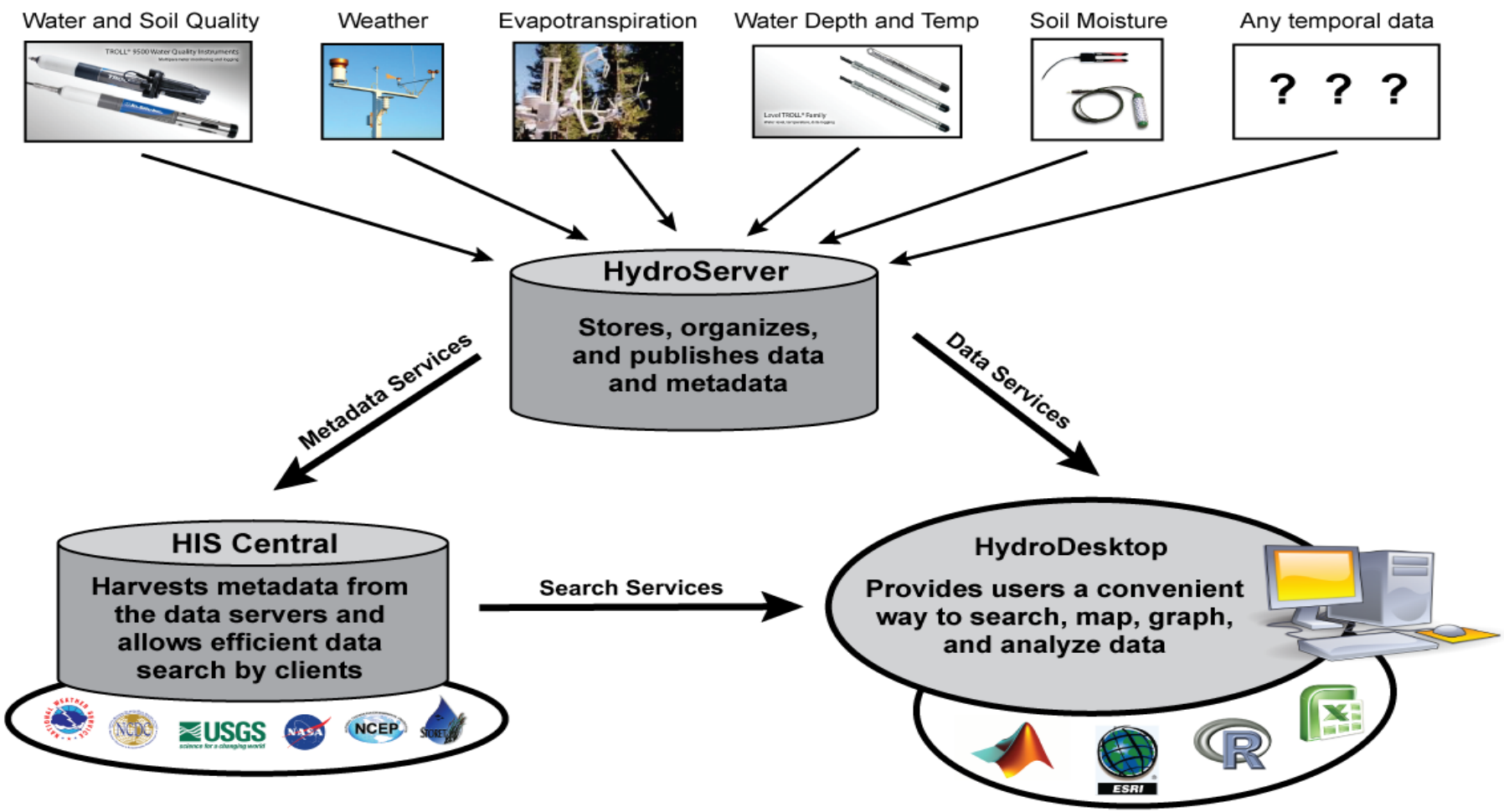
CUAHSI
universities allied for water research

What is the CUAHSI HIS?
http://his.cuahsi.org

- An Internet-based system for sharing hydrologic data
- Developed between 2007 and 2013 with support from the National Science Foundation
- Focused on time series from fixed monitoring sites
- Software stack and functionality now supported operationally by the CUAHSI Water Data Center - https://www.cuahsi.org/wdc

CUAHSI HIS
Sharing hydrologic data

CUAHSI HIS Service Oriented Architecture Enabling Water Science Data Discovery



iUTAH: innovative **U**rban **T**ransitions and **A**ridregion **H**ydro-sustainability

- Statewide effort
- $20 million competitive award from NSF EPSCoR
- Research capacity building
- Interdisciplinary and multi-institution
- Focused on sustainable management of Utah's water resources

Support:
EPS 1208732

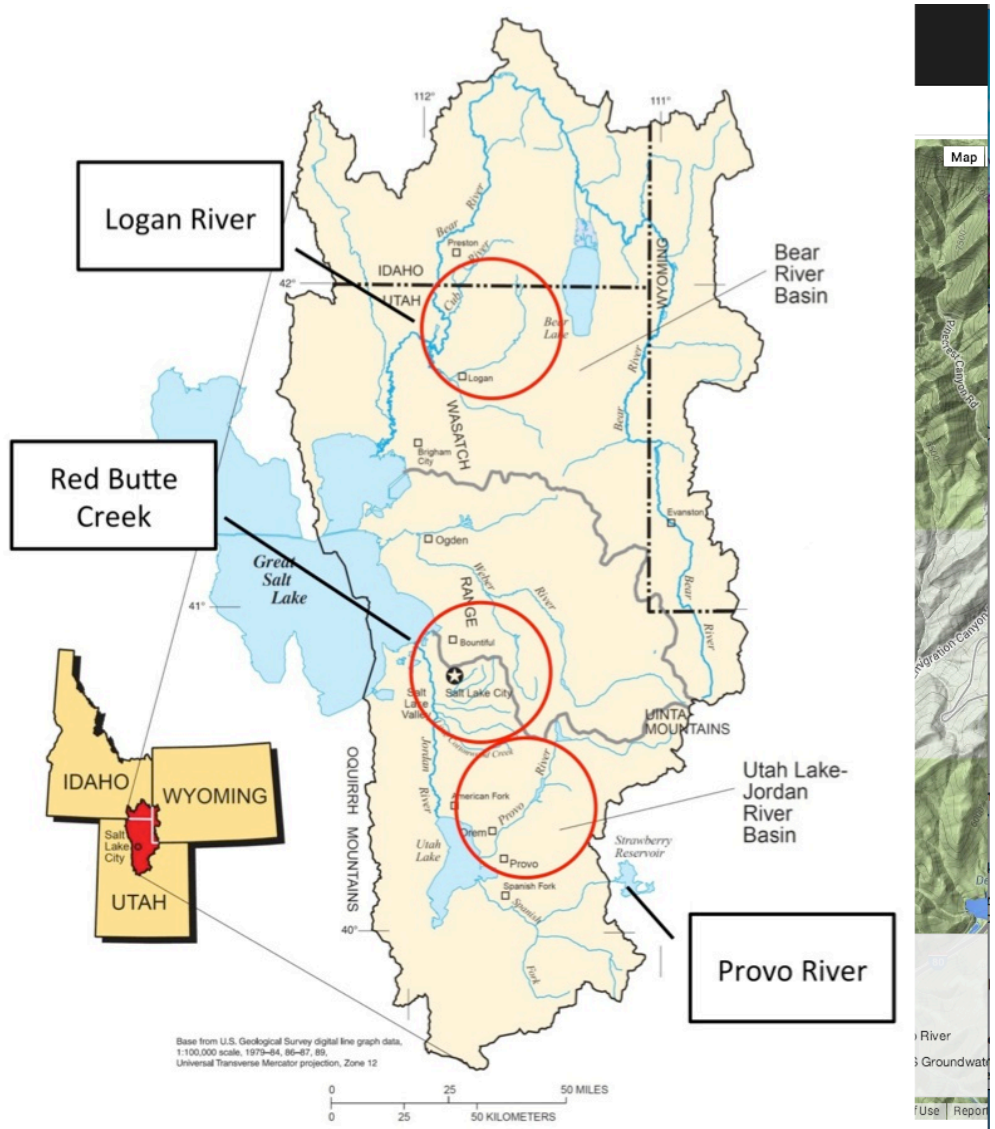
iUTAH Research Focus Areas

1. <u>Eco-hydrology</u>
   Expand Utah's capacity in the natural sciences through instrumentation of 3 watersheds.

2. <u>Social and Engineered Water Systems</u>
   Studying demographic characteristics, water use behaviors, water infrastructure, and other measures of urban form.

3. <u>Interdisciplinary Modeling and Visualization</u>
   Development of interdisciplinary models of socio-eco-hydrological systems to determine how changes in water availability and use alter water quantity and quality.

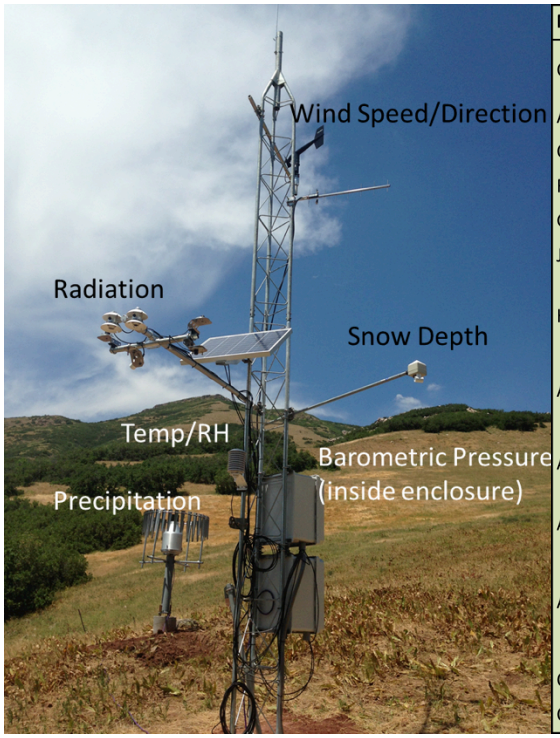
Gradients Along Mountain To Urban Transitions (GAMUT) Network



- Ecohydrologic observatory deployed in 3 watersheds in northern Utah, USA: Logan River, Red Butte Creek, Provo River
- Watersheds with similar water source (high elevation snow) but different land use transitions
- Measures aspects of water inputs and outputs and water quality over mountain-to-urban gradient
- Mix of aquatic and terrestrial *in situ* and re-locatable sensors

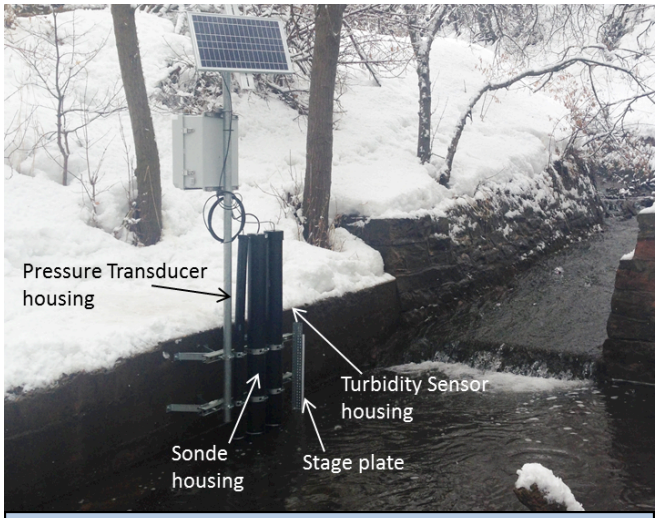
Gradients Along Mountain to Urban Transitions (GAMUT) Network

Climate/Terrestrial Sites



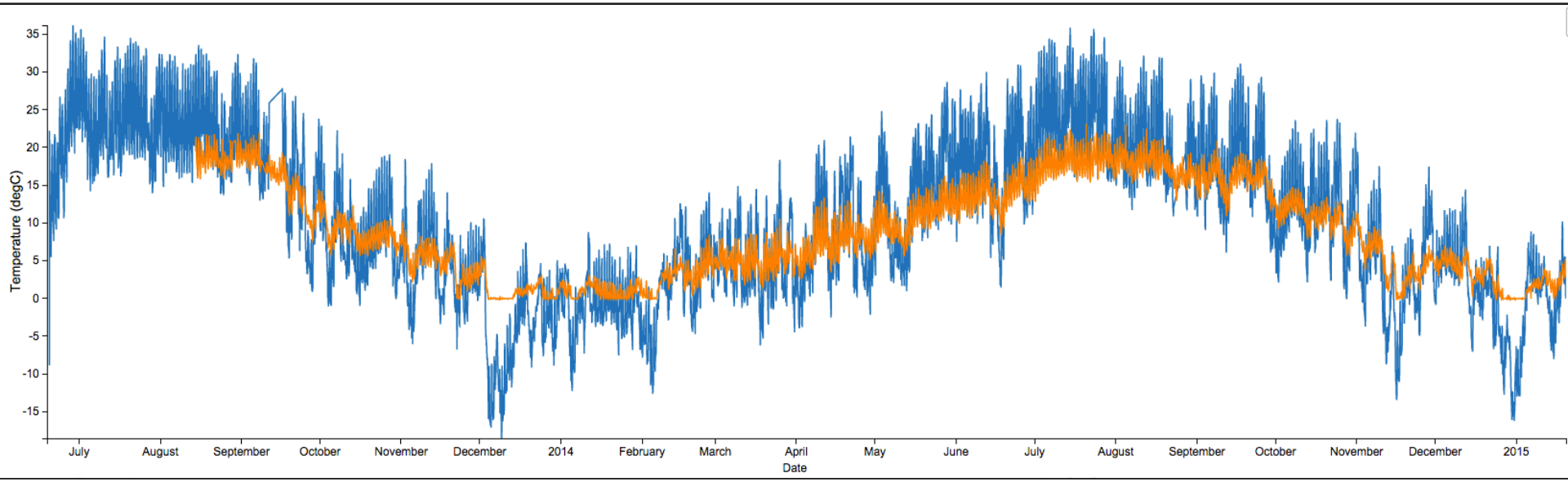
Manufacturer	Instrument	Variables Measured
Campbell	HC2S3	Air Temperature and Relative Humidity
Apogee	ST110	Air Temperature
Campbell	CS106	Barometric Pressure
RM Young	5303	Wind Speed/Direction
Geonor	TB-200	Precipitation
Judd	Depth Sensor	Snow Depth
Hukseflux	NR01	Incoming and Outgoing Shortwave and Longwave Radiation
Apogee	SP-230	Incoming Shortwave Radiation
Apogee	SQ-110	Incoming and Outgoing Photosynthetically Active Radiation
Apogee	SI-111	Surface Temperature
Acclima	ACC-SEN-SDI	Soil Moisture, Temperature, and Conductivity at 5 cm, 10 cm, 20 cm, 50 cm, 100 cm below ground
Campbell	CS210	Enclosure Humidity
Campbell	18166	Enclosure open door sensor

Aquatic Sites



Manufacturer	Instrument	Variables Measured
YSI	599100-01	Dissolved Oxygen
YSI	599870-01	Specific Conductivity and Water Temperature
YSI	599795-02	pH
YSI	599101-01	Fluorescent Dissolved Organic Matter (fDOM)
YSI	599102-01	Blue Green Algae and Chlorophyll *a*
Campbell	CS451	Water Depth and Water Temperature
FTS	DTS-12	Turbidity and Water Temperature

The Data Deluge



One year = 35,040 observations
So far (~1.5 years) = 55,000+ observations

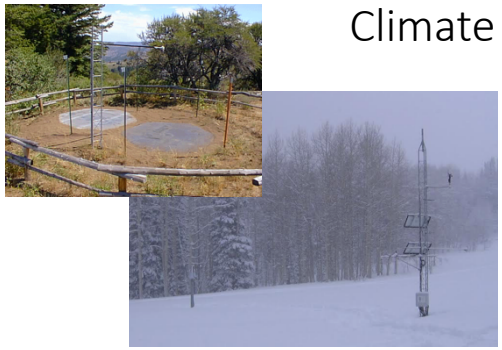
Times 14 Aquatic Sites with ~26 Variables
Times 14 Climate Sites with ~74 Variables
Plus different versions of the data (raw versus checked) = **43,400,000+** observations

You need some infrastructure to manage and share the data.

Challenges to Managing Sensor Data

- Volume of data
- Data heterogeneity
- Multiple watersheds
- Multiple institutions
- Multiple personnel
- Scale
- Data quality assurance and quality control
- Standardize data editing
- Synchronize timing, data access, equipment tracking

Rain, Snow, and Climate



Sapflux

Soil water and chemistry

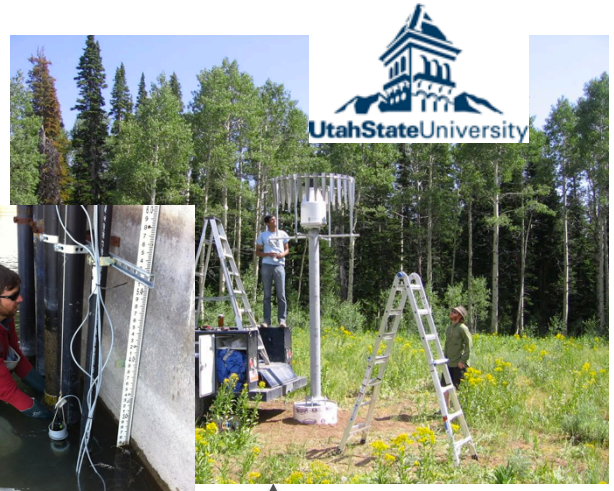
Water quality

Water quantity

Groundwater

Mobile Platforms

Challenges to Managing Sensor Data



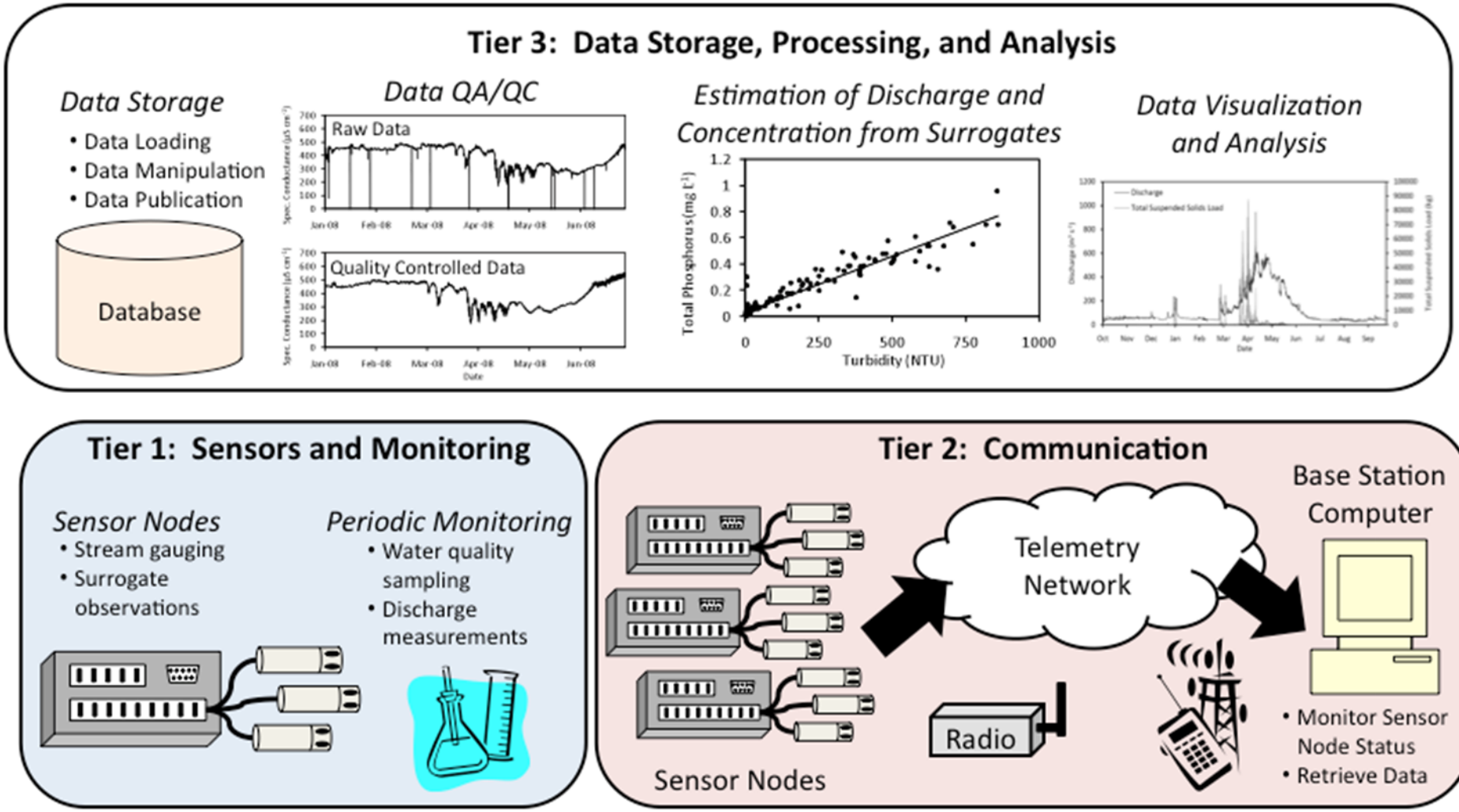
EPSCoR iUTAH

Modeling and Data Federation

innovative Urban Transitions and Aridregion Hydro-sustainability

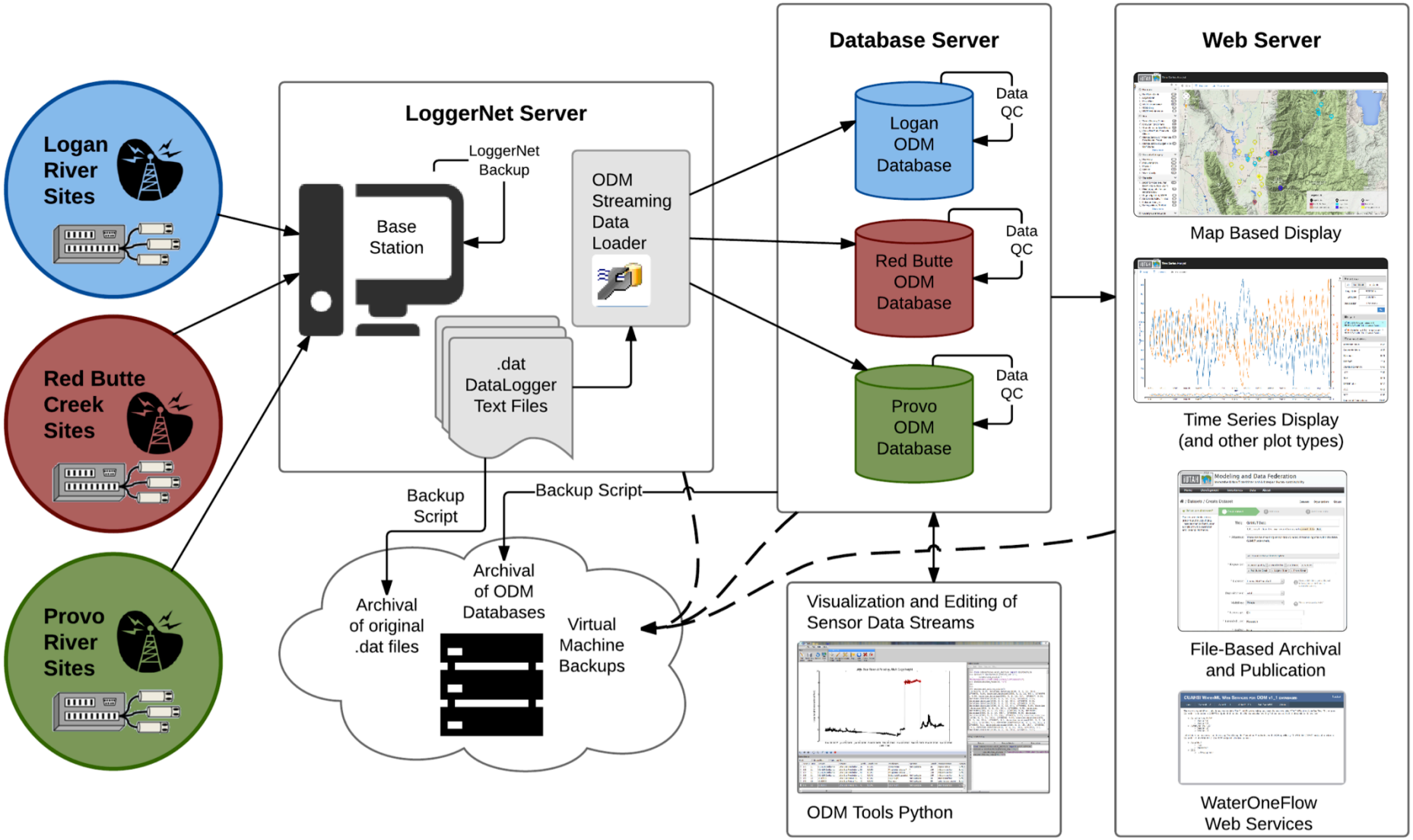
Utah Water Research Laboratory UWRL

Cyberinfrastructure

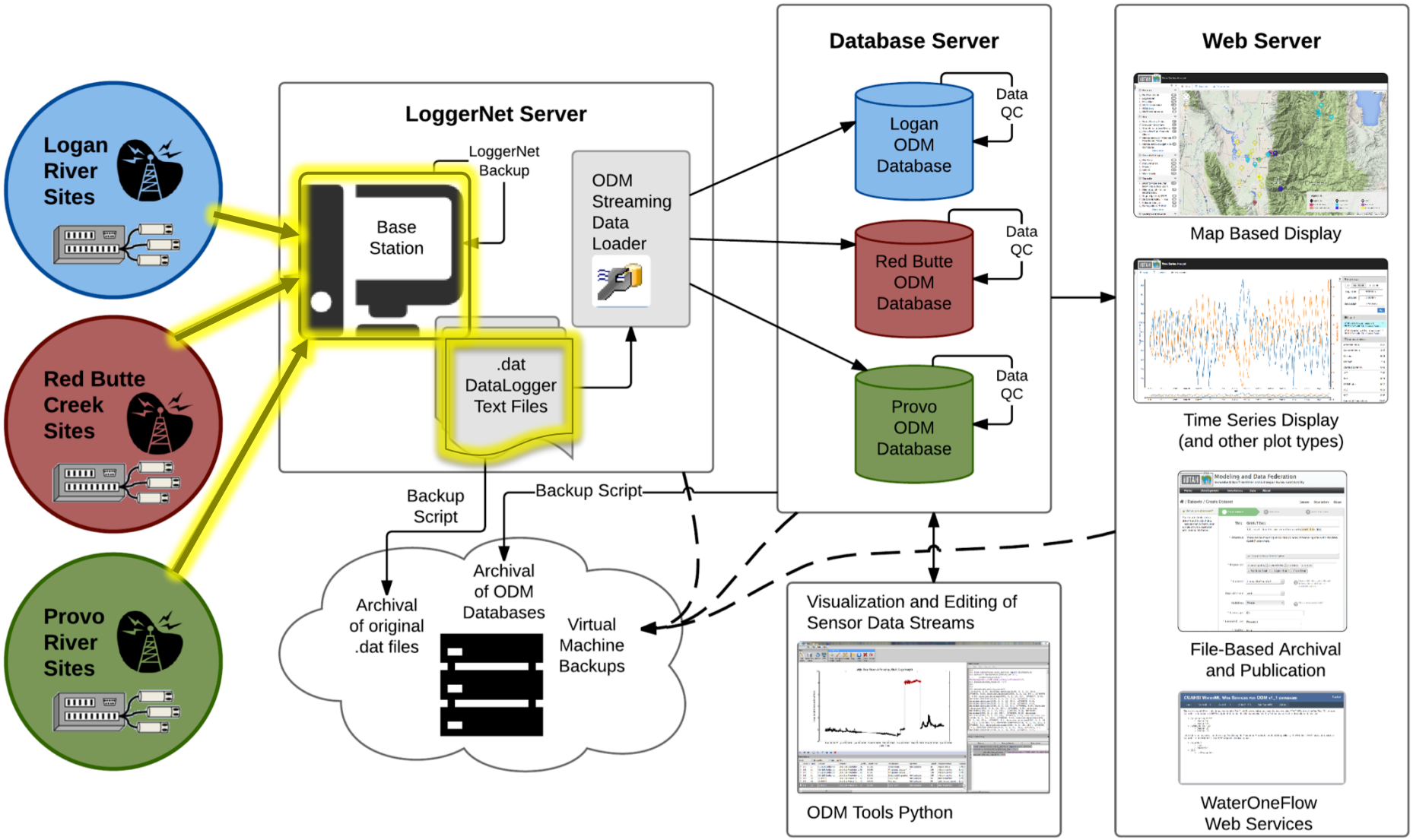


Horsburgh, J. S., A. Spackman Jones, D. G. Tarboton, D. K. Stevens, and N. O. Mesner (2010), A sensor network for high frequency estimation of water quality constituent fluxes using surrogates, *Environmental Modelling & Software*, 25, 1031-1044, doi:10.1016/j.envsoft.2009.10.012.

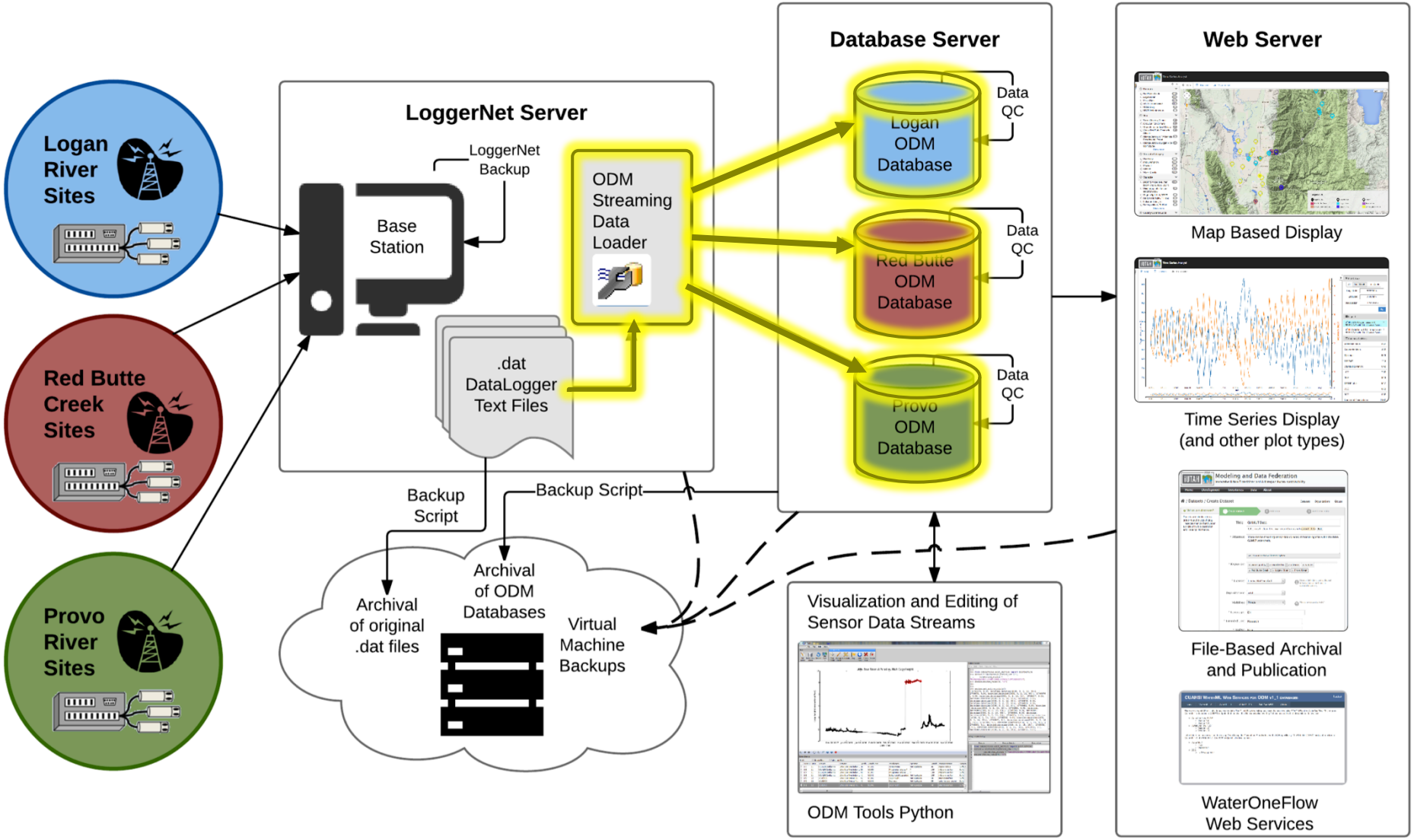
GAMUT Data Workflow



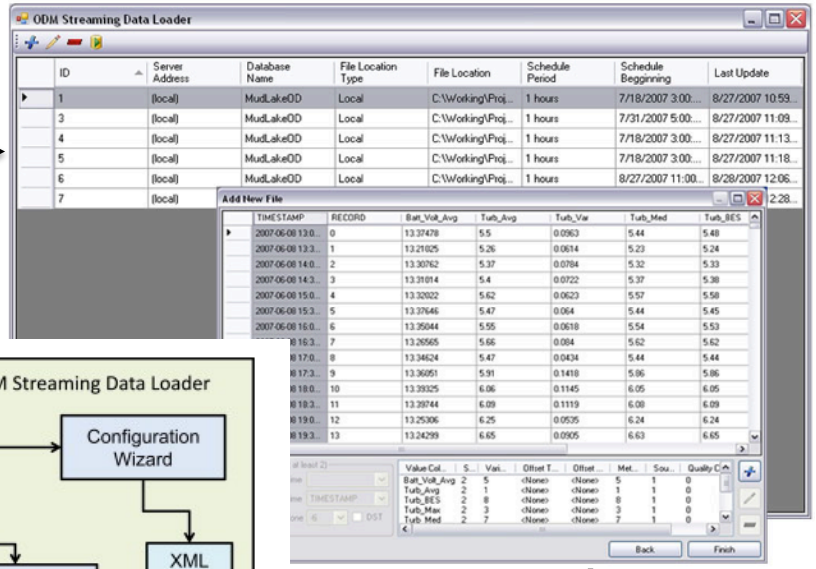
Sensor Data Acquisition

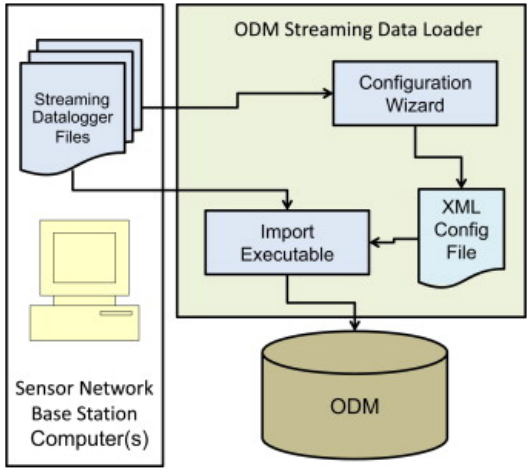


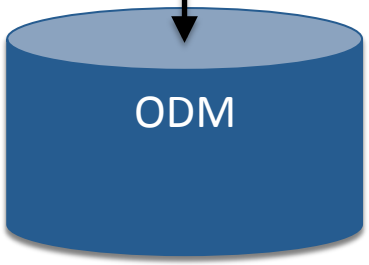
Data Loading and Storage



Data Loading and Storage
ODM Streaming Data Loader







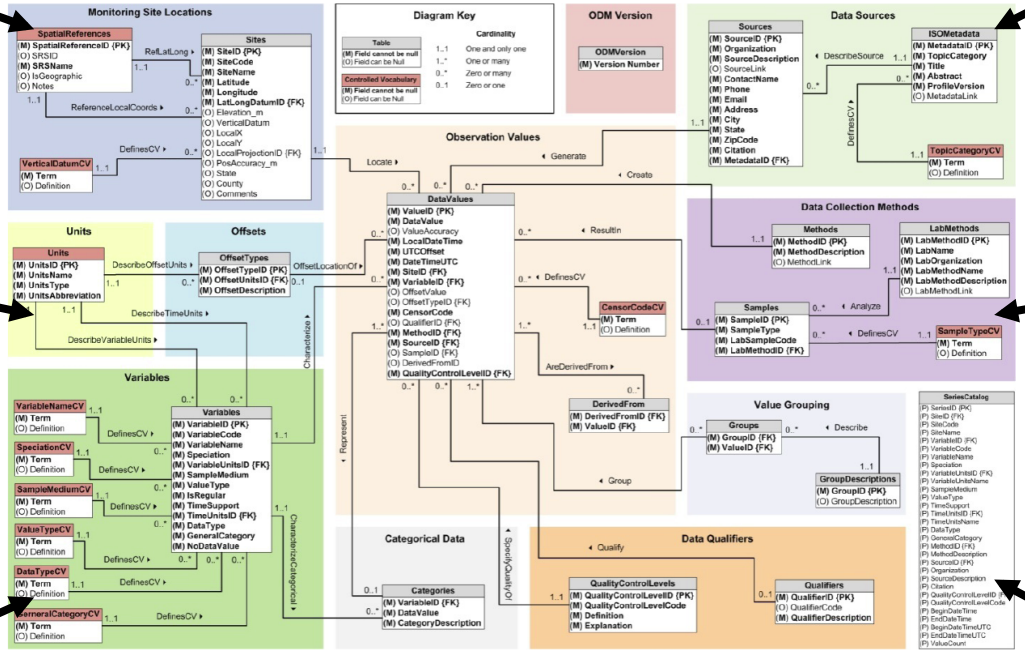
Observations Data Model (ODM)

Streamflow

Groundwater levels

Precipitation & Climate

Soil moisture data



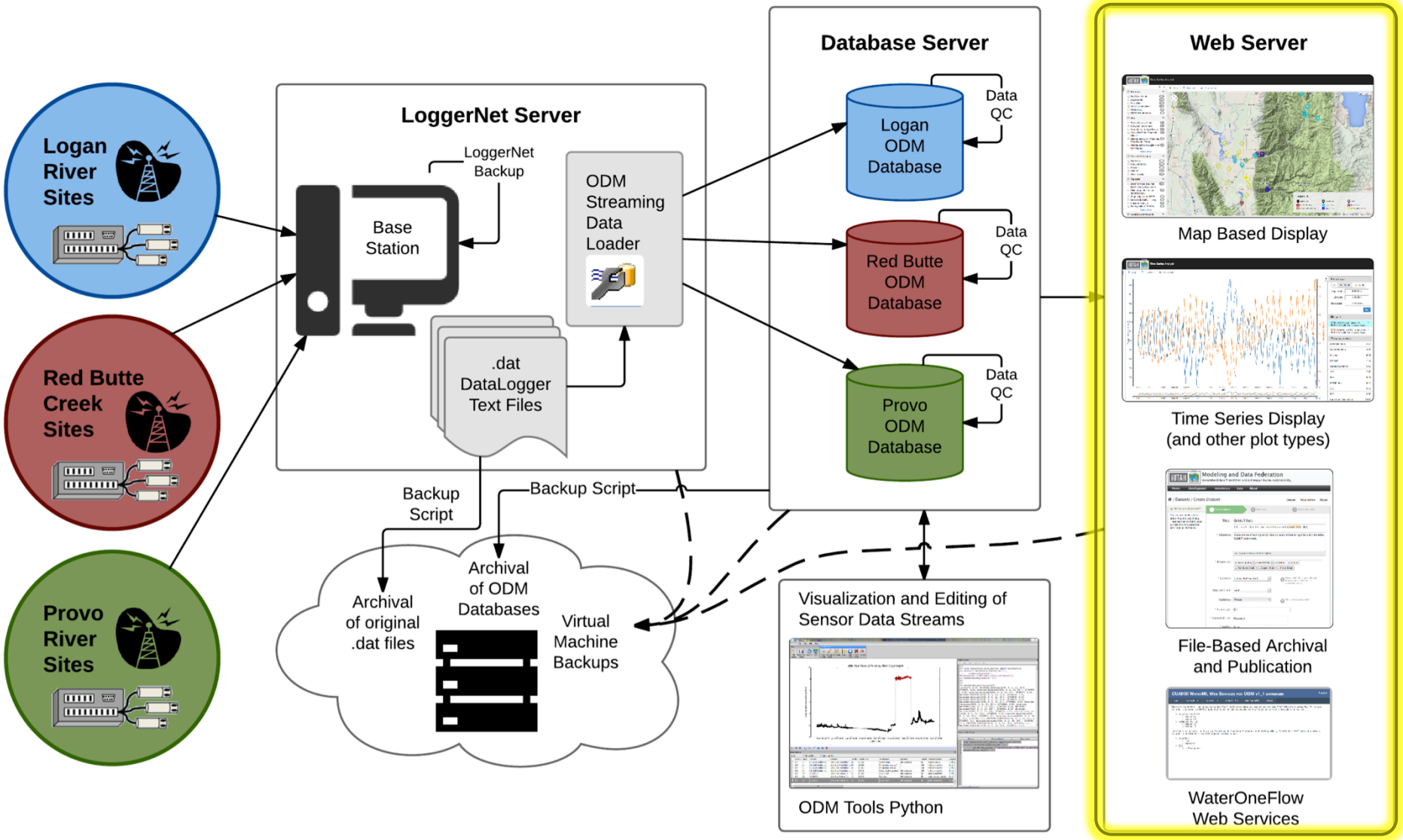
Water Quality

Flux tower data

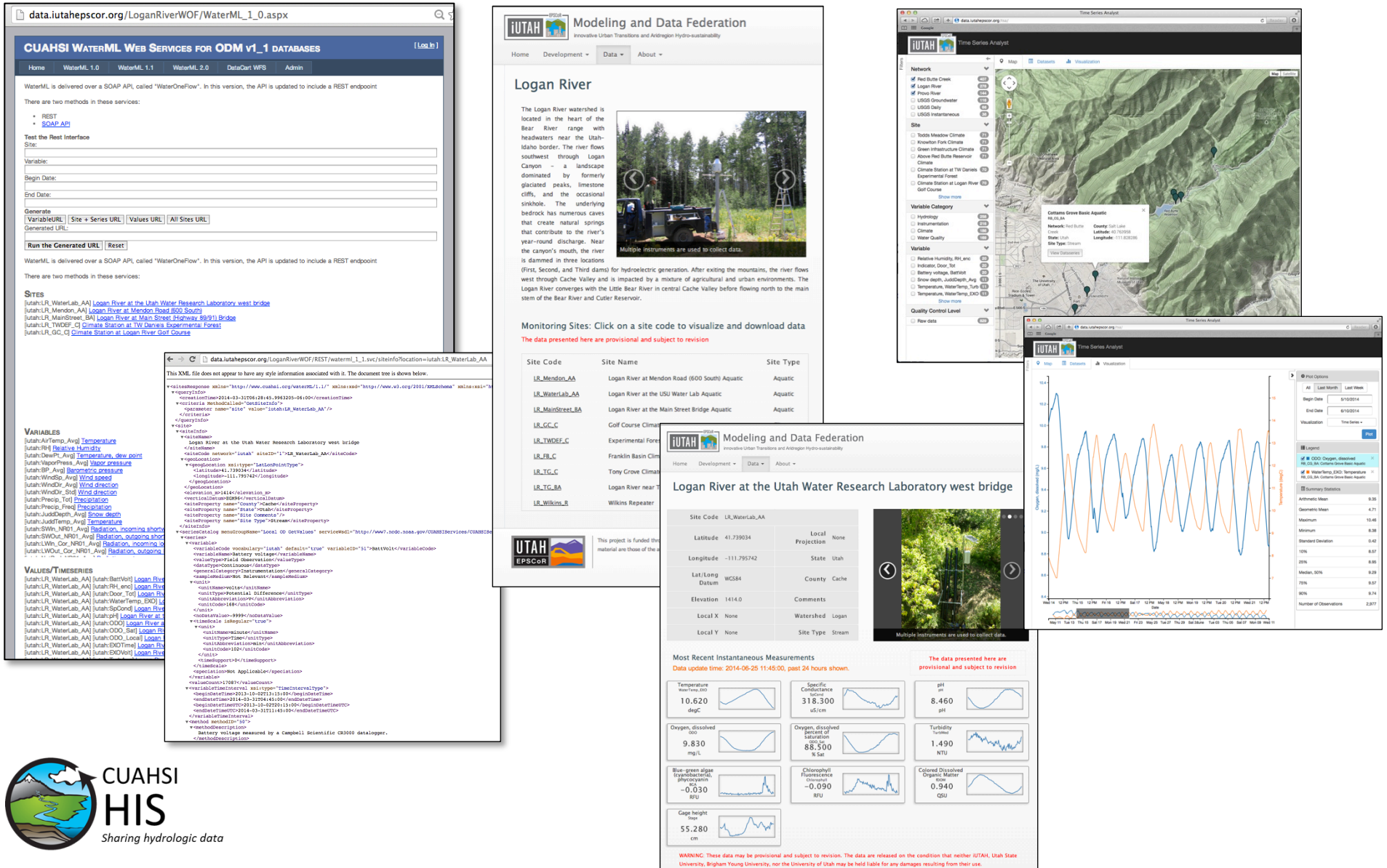
- A relational database at the single observation level
- Metadata for unambiguous interpretation
- Traceable heritage from raw measurements to usable information
- Promote syntactic and semantic consistency
- Cross dimension retrieval and analysis

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resources Research*, 44, W05406, doi:10.1029/2007WR006392.

Web-Based Data Access

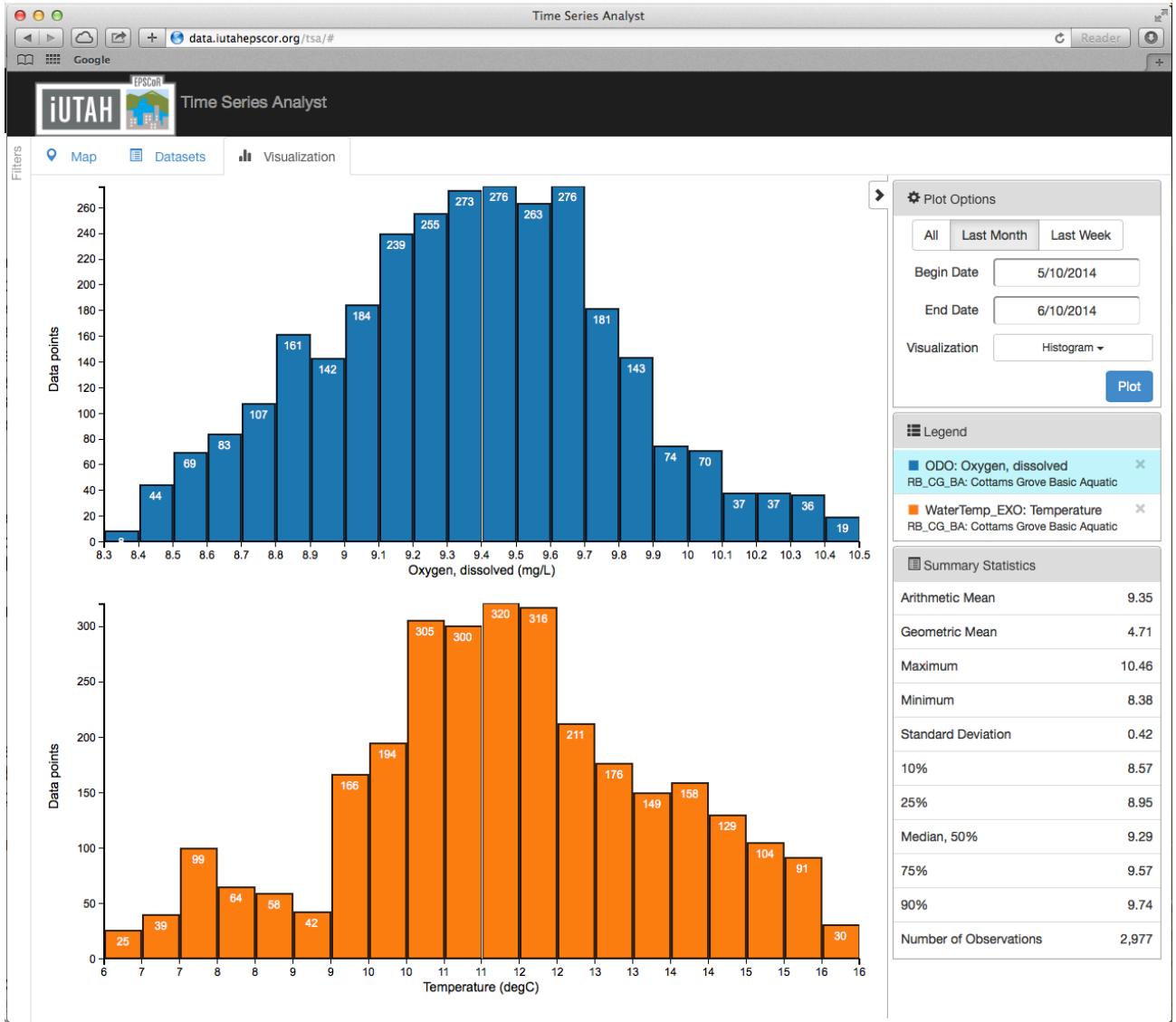


Web-based Data Access

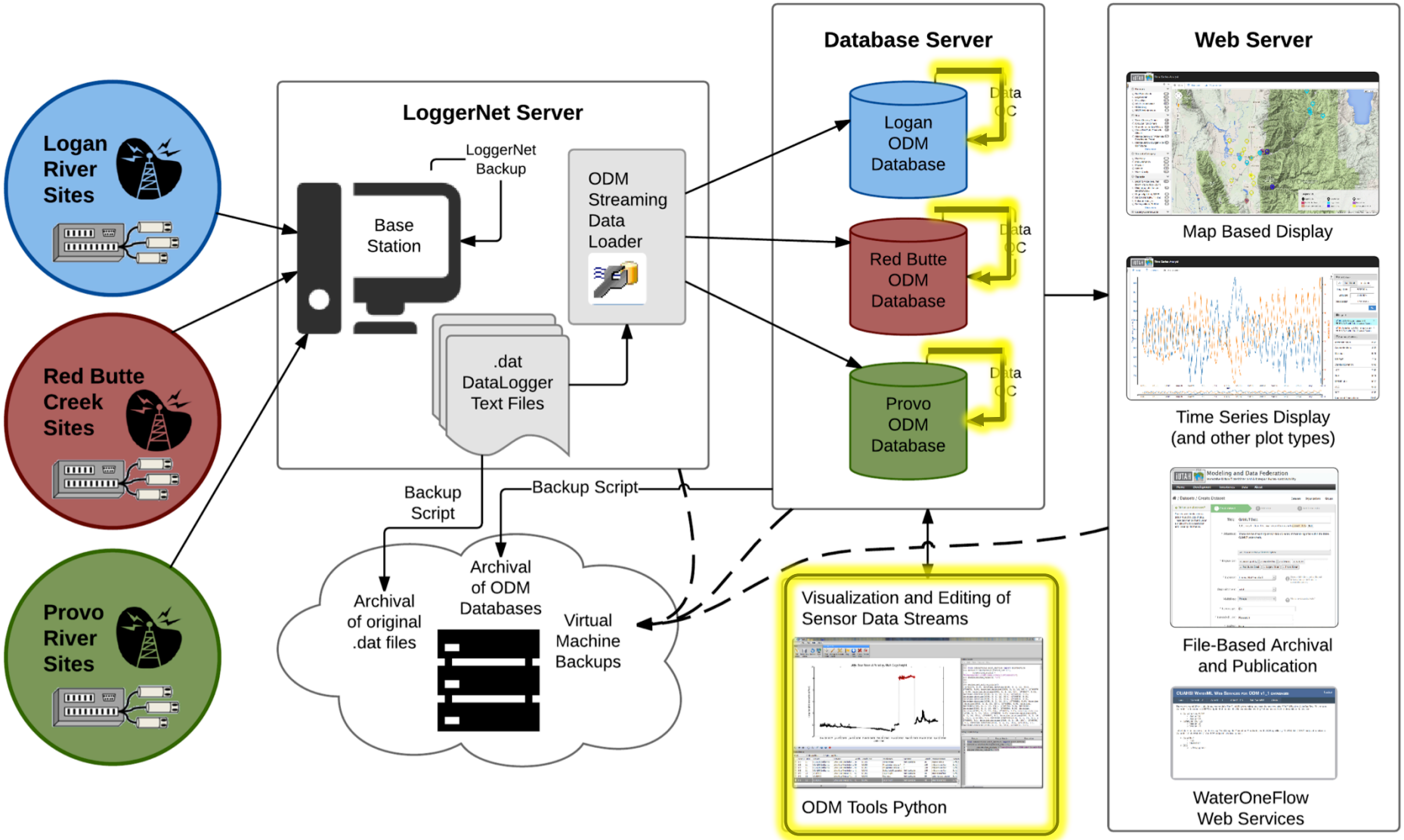


CUAHSI HIS
Sharing hydrologic data

Enhanced Web-Based Time Series Data Access and Visualization



Sensor Data QAQC



Development of a QAQC Plan

Quality Assurance:

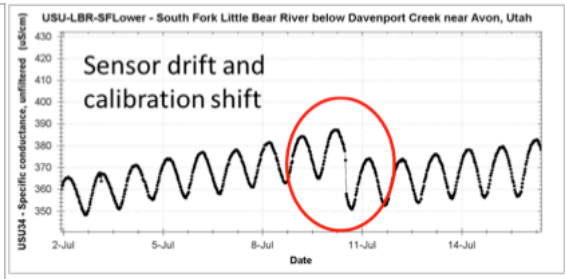
"…protocols developed and adhered to in a way that minimizes inaccuracies in the data produced … produces high-quality data while minimizing the need for corrective measures to improve data quality."

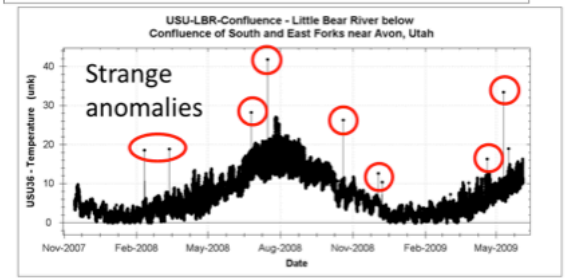
- Site Standardization
- Data Curation (Datalogger Programs and Files, Data Averaging, Database Structure, Equipment Management)
- Replicate Sensors
- Factory Maintenance
- Field Maintenance Schedule and Procedures
- Field Calibration Schedule and Procedures
- Manual Data Monitoring
- Automated Data Monitoring and Alerts
- Recording Events

Quality Control:

"occurs after the data are generated and tests whether they meet the necessary requirements for quality outlined by the end users."

- Data Qualifiers and Flagging
- ODM Tools Python (data management software)
- Quality Control Levels
- Data Processing Steps





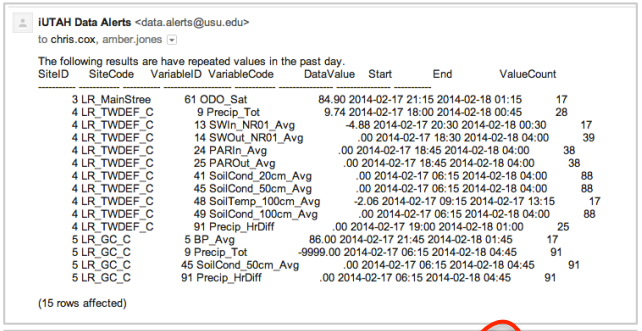
Campbell, J. L., Rustad, L. E., Porter, J. H., Taylor, J. R., Ethan, W., Shanley, J. B., … (2013). Quantity is Nothing without Quality. *BioScience*, *63*(7), 574–585. doi:10.1525/bio.2013.63.7.10

QAQC: Automated Alerts
(Enhancements to ODM)

Technicians receive email alerts daily

1. Power: battery voltage < 12 volts
2. Persistence: value of a variable is unchanging
3. Updates: data are not being reported
4. NaNs: sensor is reporting “NaN” values

Additional alerts will be implemented as needed (e.g., variable-specific range checks, internal consistency, spatial consistency).

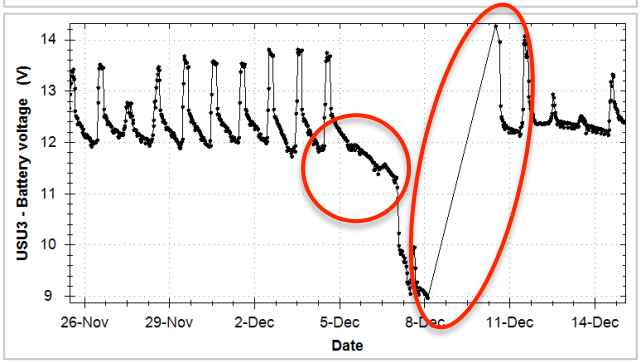


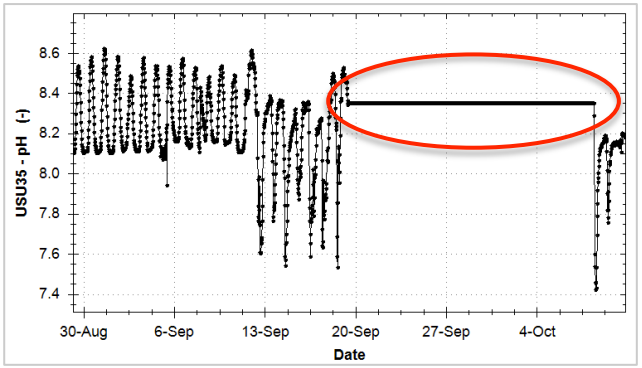
iUTAH Data Alerts <data.alerts@usu.edu>
to chris.cox, amber.jones

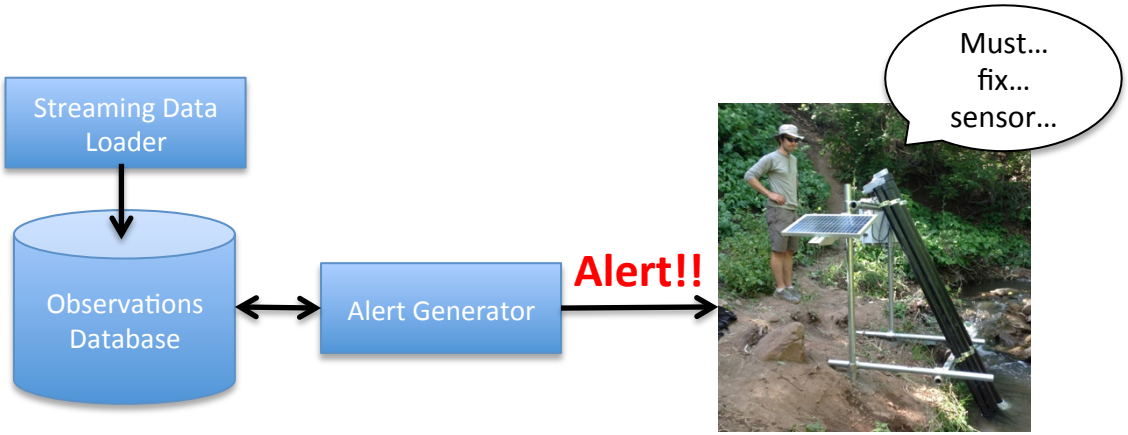
The following results are have repeated values in the past day.

SiteID	SiteCode	VariableID	VariableCode	DataValue	Start	End	ValueCount
3	LR_MainStree	61	ODO_Sat	84.90	2014-02-17 21:15	2014-02-18 01:15	17
4	LR_TWDEF_C	9	Precip_Tot	9.74	2014-02-17 18:00	2014-02-18 00:45	28
4	LR_TWDEF_C	13	SWIn_NR01_Avg	-4.88	2014-02-17 20:30	2014-02-18 00:30	17
4	LR_TWDEF_C	14	SWOut_NR01_Avg	.00	2014-02-17 18:30	2014-02-18 04:00	39
4	LR_TWDEF_C	24	PARIn_Avg	.00	2014-02-17 18:45	2014-02-18 04:00	38
4	LR_TWDEF_C	25	PAROut_Avg	.00	2014-02-17 18:45	2014-02-18 04:00	38
4	LR_TWDEF_C	41	SoilCond_20cm_Avg	.00	2014-02-17 06:15	2014-02-18 04:00	88
4	LR_TWDEF_C	45	SoilCond_50cm_Avg	.00	2014-02-17 06:15	2014-02-18 04:00	88
4	LR_TWDEF_C	48	SoilTemp_100cm_Avg	-2.06	2014-02-17 09:15	2014-02-17 13:15	17
4	LR_TWDEF_C	49	SoilCond_100cm_Avg	.00	2014-02-17 06:15	2014-02-18 04:00	88
4	LR_TWDEF_C	91	Precip_HrDiff	.00	2014-02-17 19:00	2014-02-18 01:00	25
5	LR_GC_C	5	BP_Avg	86.00	2014-02-17 21:45	2014-02-18 01:45	17
5	LR_GC_C	9	Precip_Tot	-9999.00	2014-02-17 06:15	2014-02-18 04:45	91
5	LR_GC_C	45	SoilCond_50cm_Avg	.00	2014-02-17 06:15	2014-02-18 04:45	91
5	LR_GC_C	91	Precip_HrDiff	.00	2014-02-17 06:15	2014-02-18 04:45	91

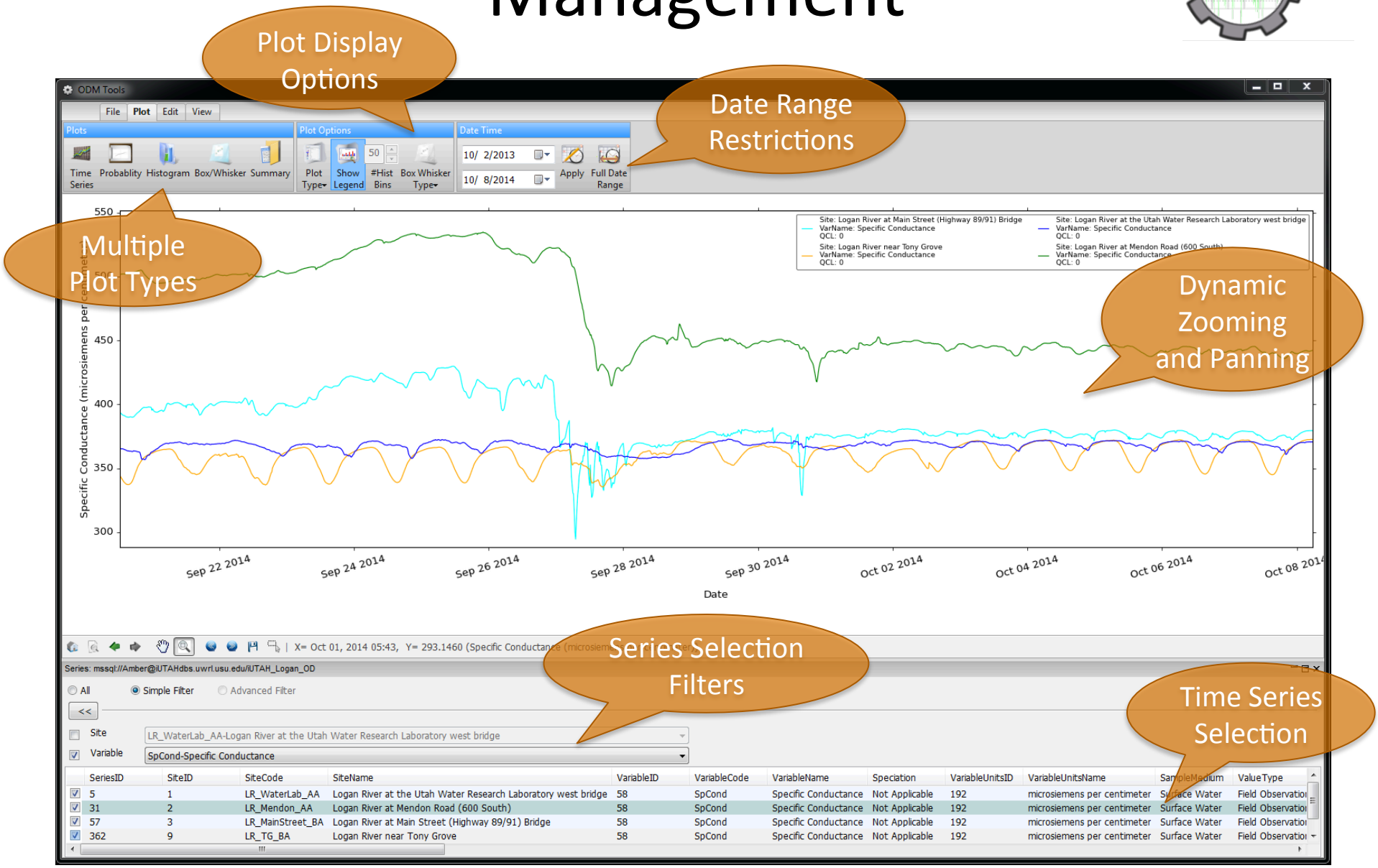
(15 rows affected)



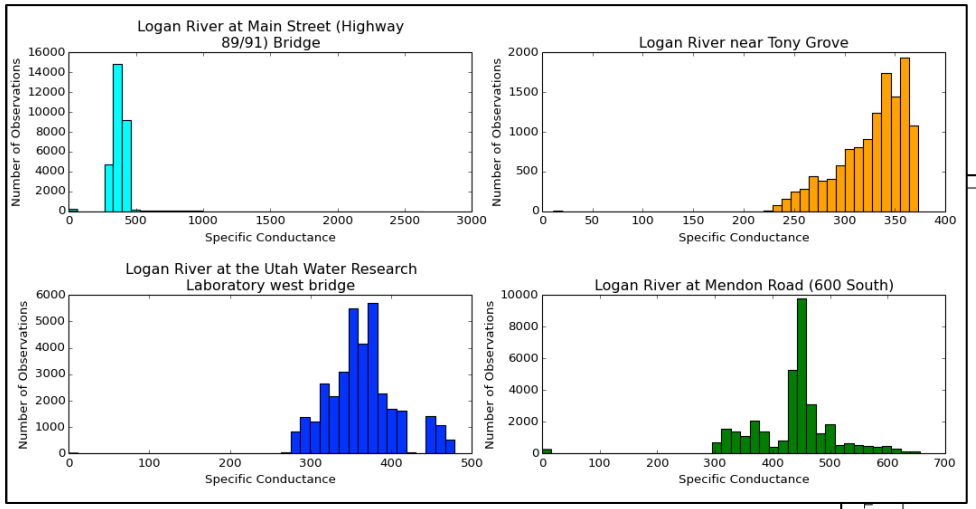


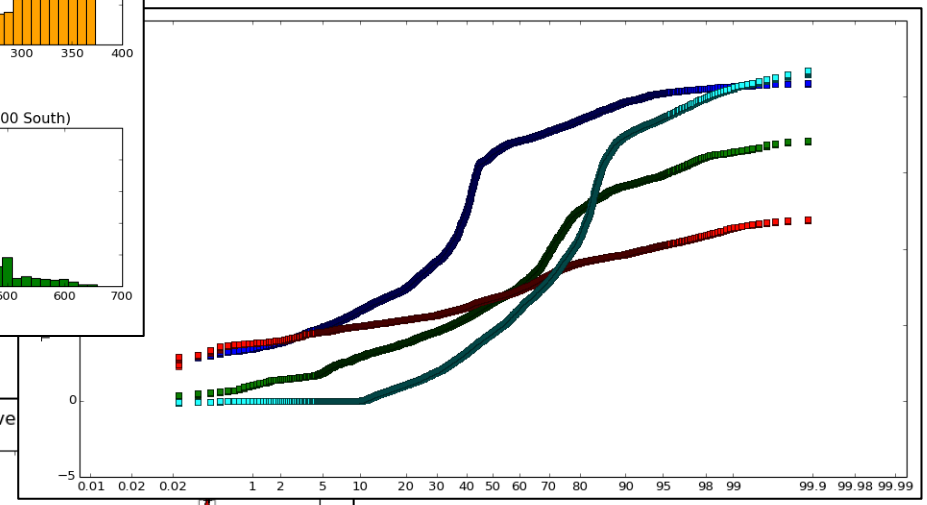


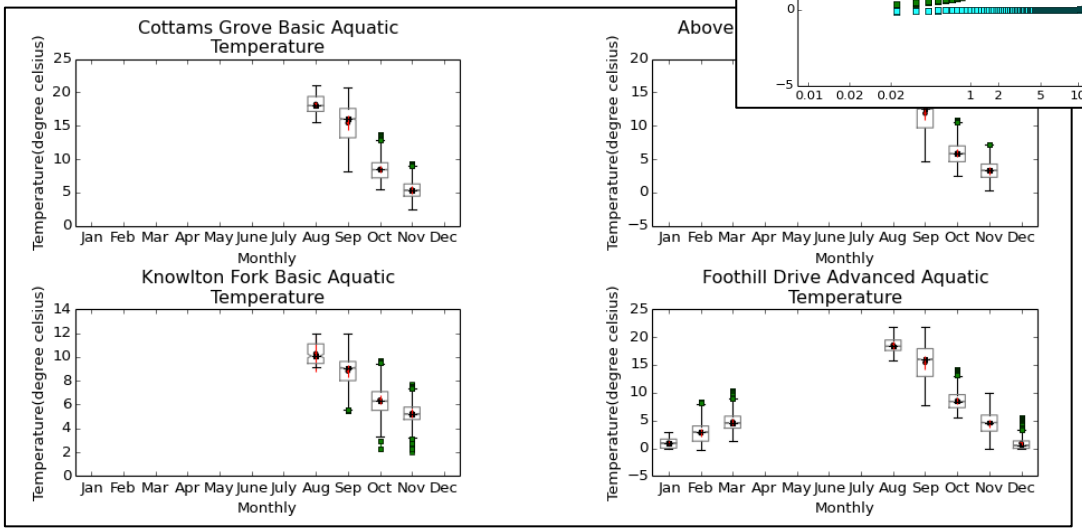
QAQC: Data Visualization and Management



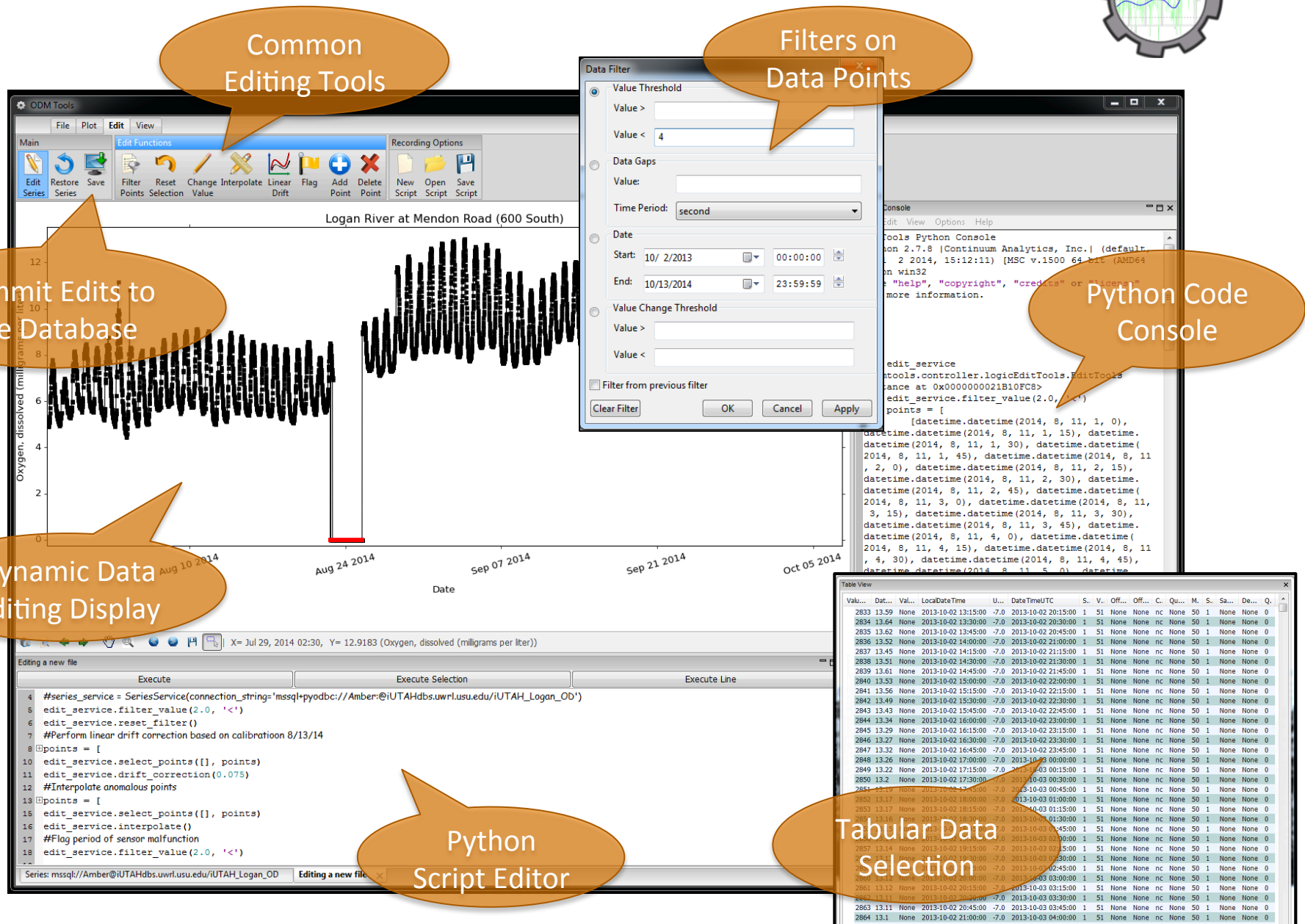
QAQC: Data Visualization and Management



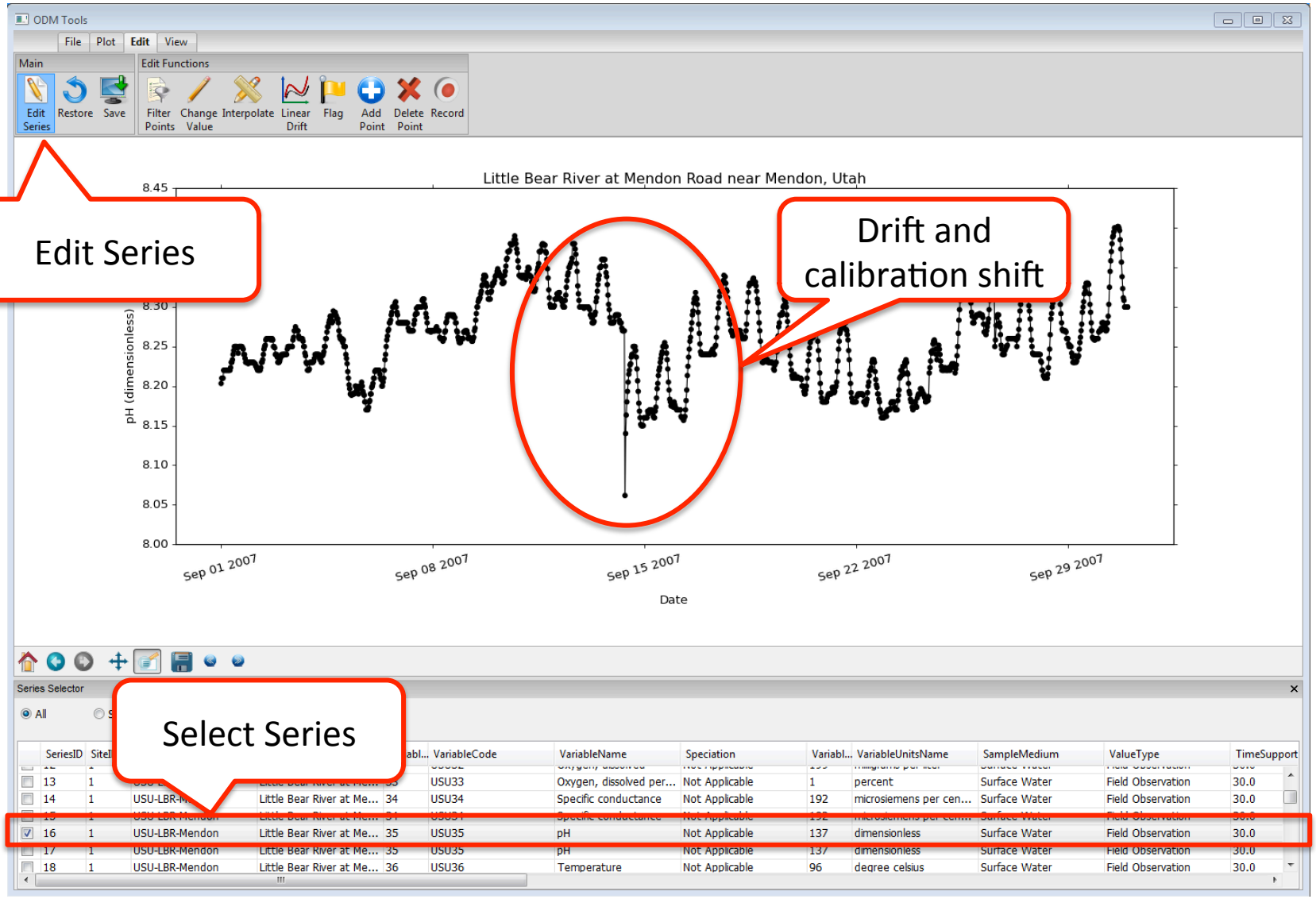




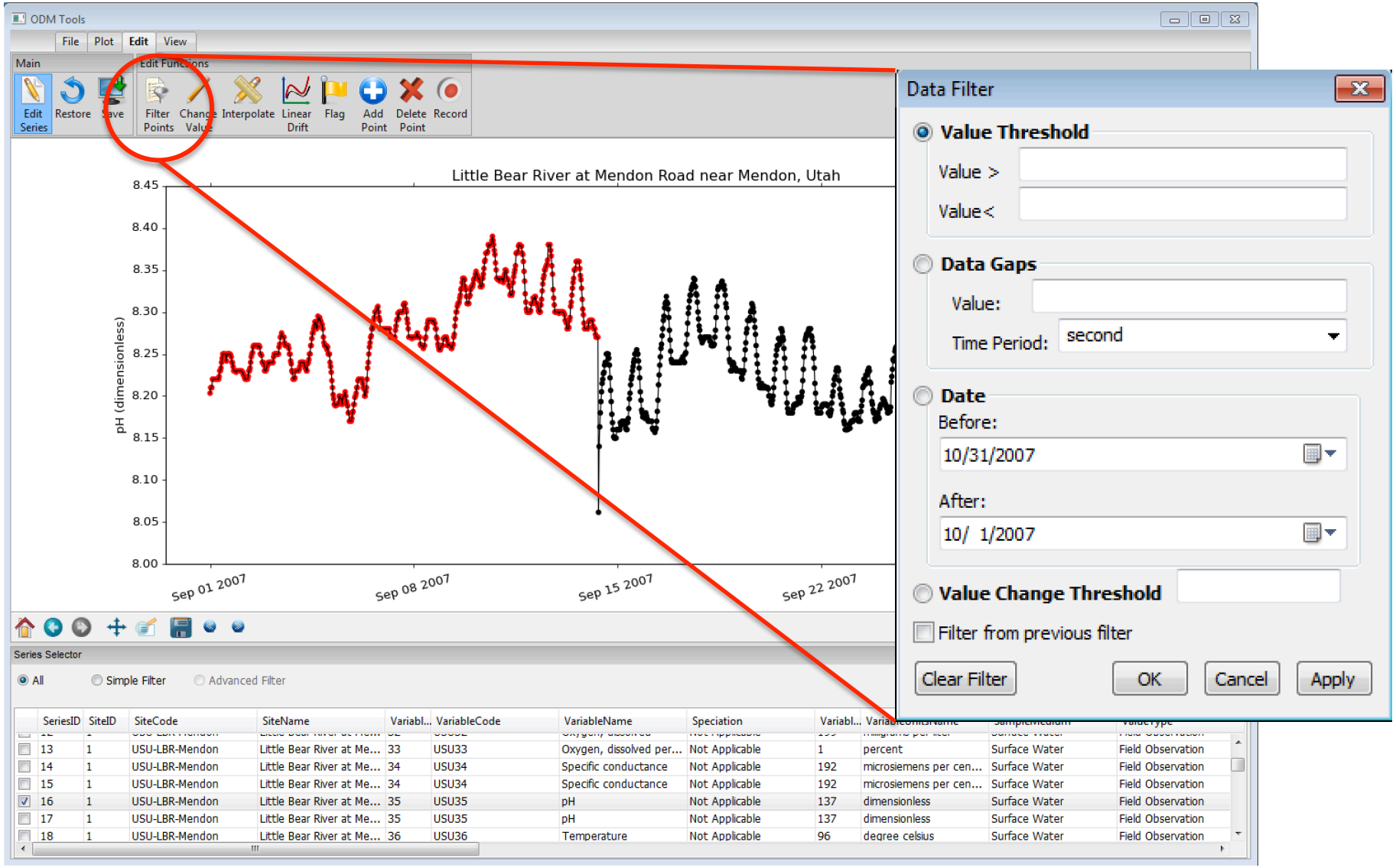
QAQC: Post Processing



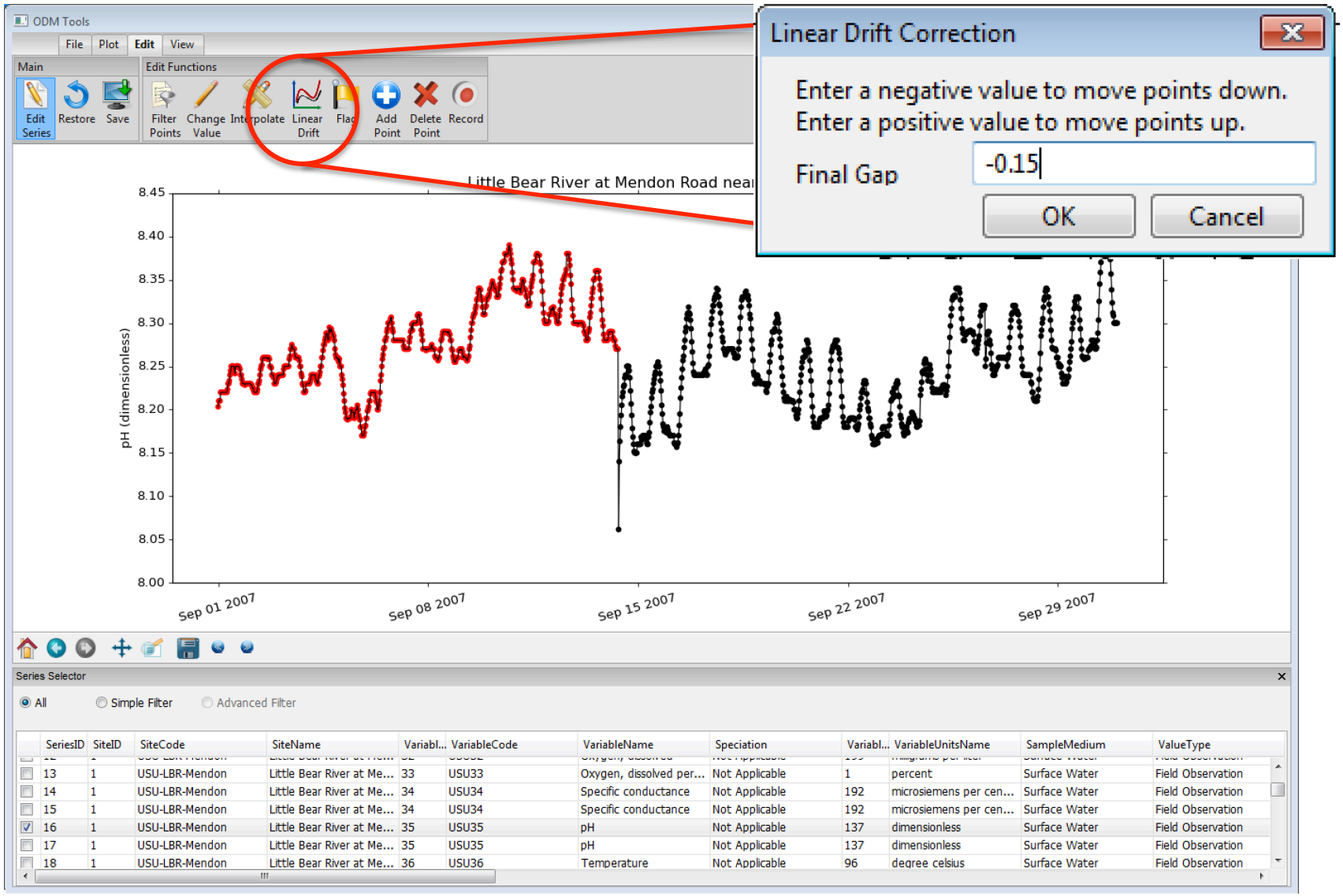
Step 1: Select a Time Series for Editing



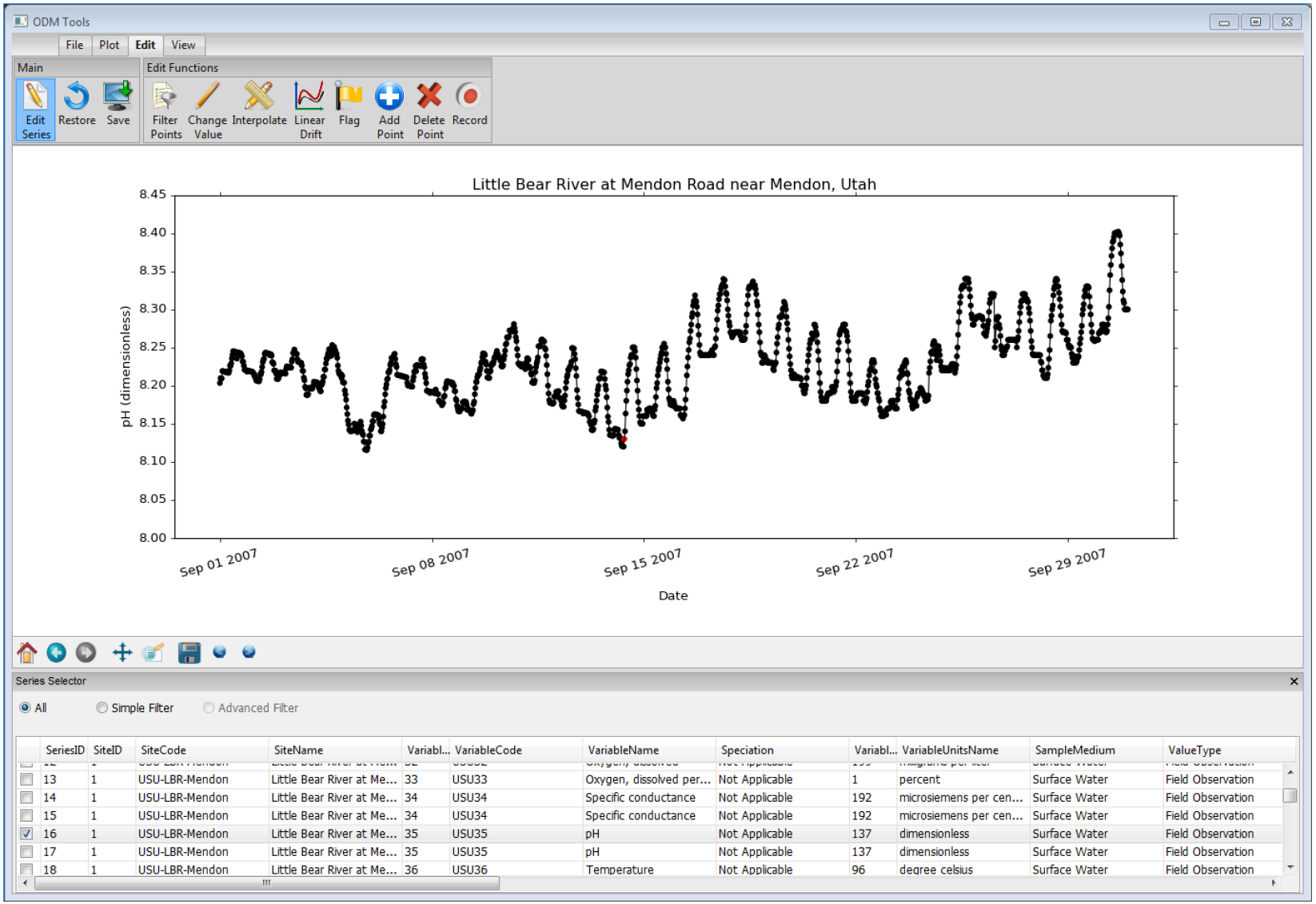
Step 2: Select Data to Edit



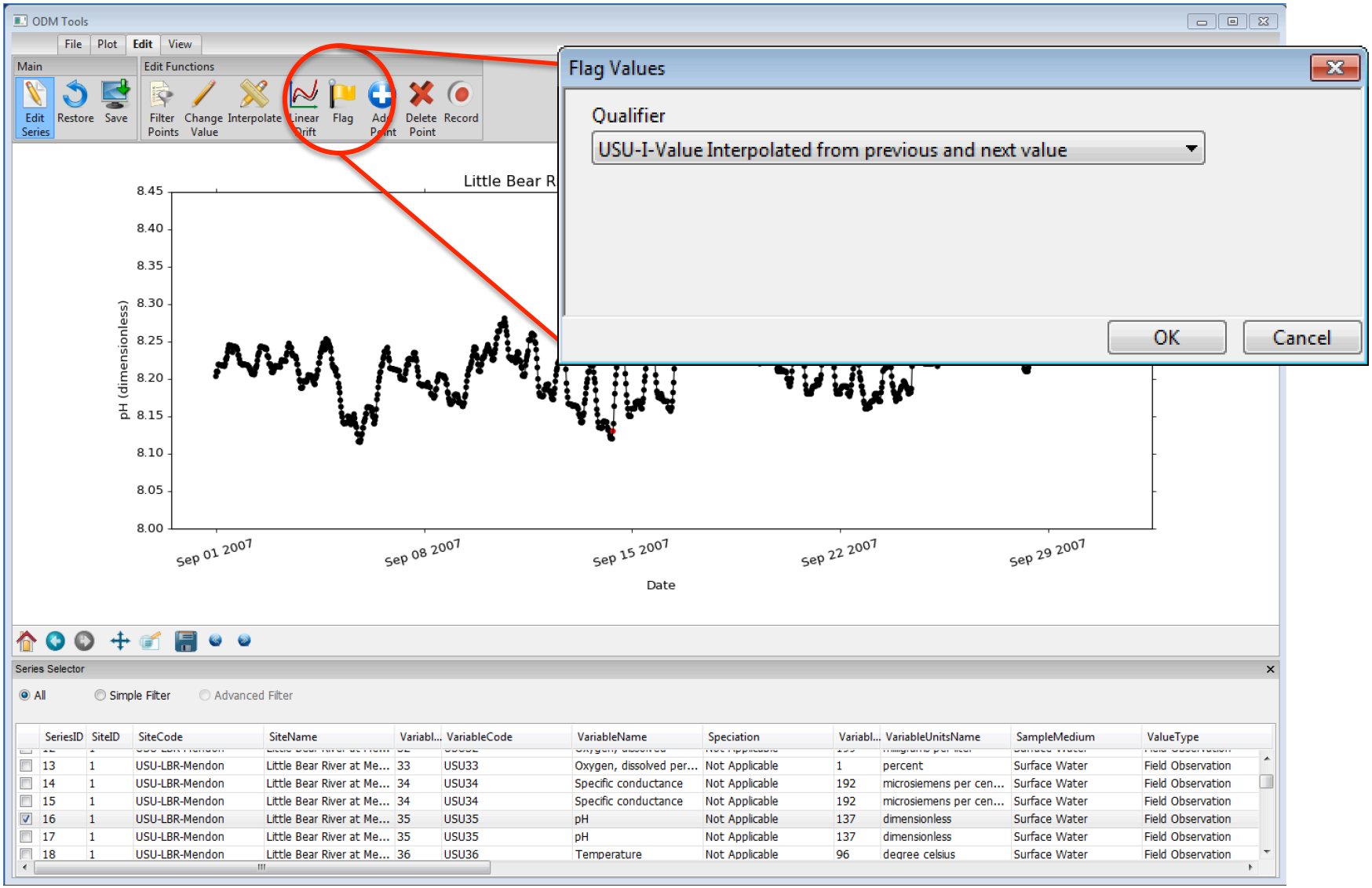
Step 3: Linear Drift Correction



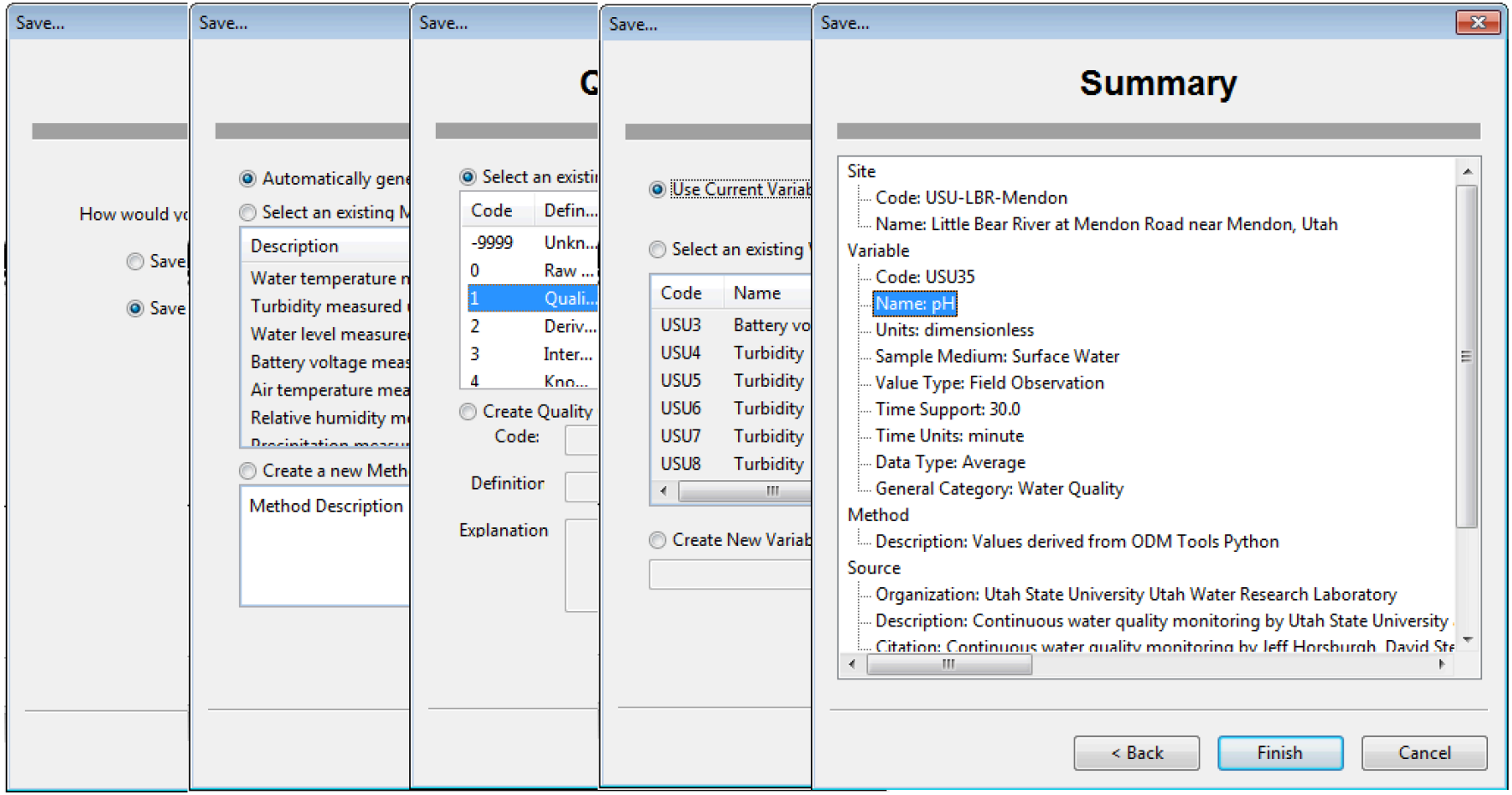
Step 4: Interpolate



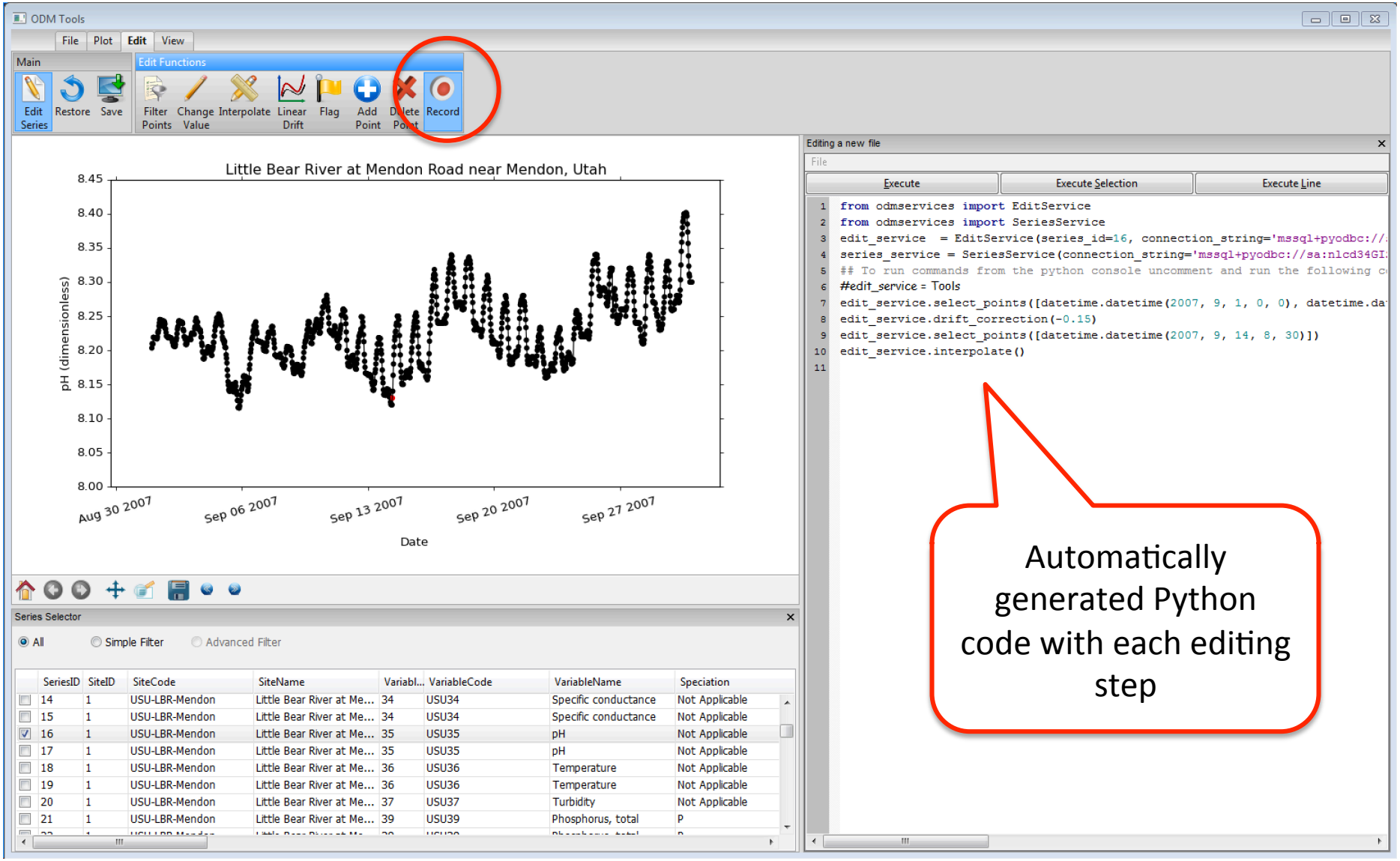
Step 5: Flag


ODM Tools
File
Plot
Edit
View
Main
Edit Series
Restore
Save
Edit Functions
Filter Points
Change Value
Interpolate
Linear Drift
Flag
Add Point
Delete Point
Record
Flag Values
Qualifier
USU-I-Value Interpolated from previous and next value
OK
Cancel
Little Bear R
pH (dimensionless)
8.45
8.40
8.35
8.30
8.25
8.20
8.15
8.10
8.05
8.00
Sep 01 2007
Sep 08 2007
Sep 15 2007
Sep 22 2007
Sep 29 2007
Date
Series Selector
All
Simple Filter
Advanced Filter
SeriesID SiteID SiteCode SiteName VariableCode VariableName Speciation VariableUnitsName SampleMedium ValueType
13 1 USU-LBR-Mendon Little Bear River at Me... 33 USU33 Oxygen, dissolved per... Not Applicable 1 percent Surface Water Field Observation
14 1 USU-LBR-Mendon Little Bear River at Me... 34 USU34 Specific conductance Not Applicable 192 microsiemens per cen... Surface Water Field Observation
15 1 USU-LBR-Mendon Little Bear River at Me... 34 USU34 Specific conductance Not Applicable 192 microsiemens per cen... Surface Water Field Observation
16 1 USU-LBR-Mendon Little Bear River at Me... 35 USU35 pH Not Applicable 137 dimensionless Surface Water Field Observation
17 1 USU-LBR-Mendon Little Bear River at Me... 35 USU35 pH Not Applicable 137 dimensionless Surface Water Field Observation
18 1 USU-LBR-Mendon Little Bear River at Me... 36 USU36 Temperature Not Applicable 96 degree celsius Surface Water Field Observation

Step 6: Save Modified Data Series



Recording Edits

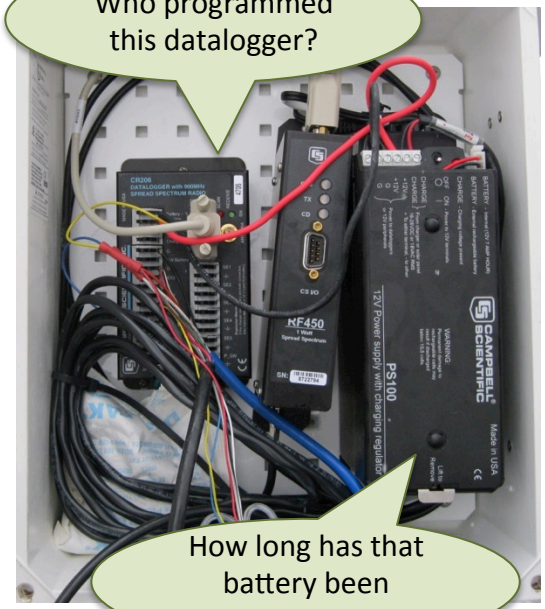


Monitoring Equipment Management

When was the last time we cleaned this sonde?

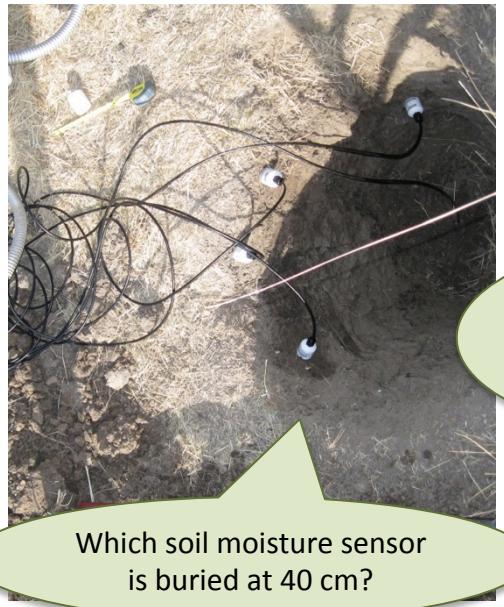
What were the field conditions of our discharge measurement?

Which sensors were deployed at this site? Who installed them?



Who programmed this datalogger?

How long has that battery been deployed?

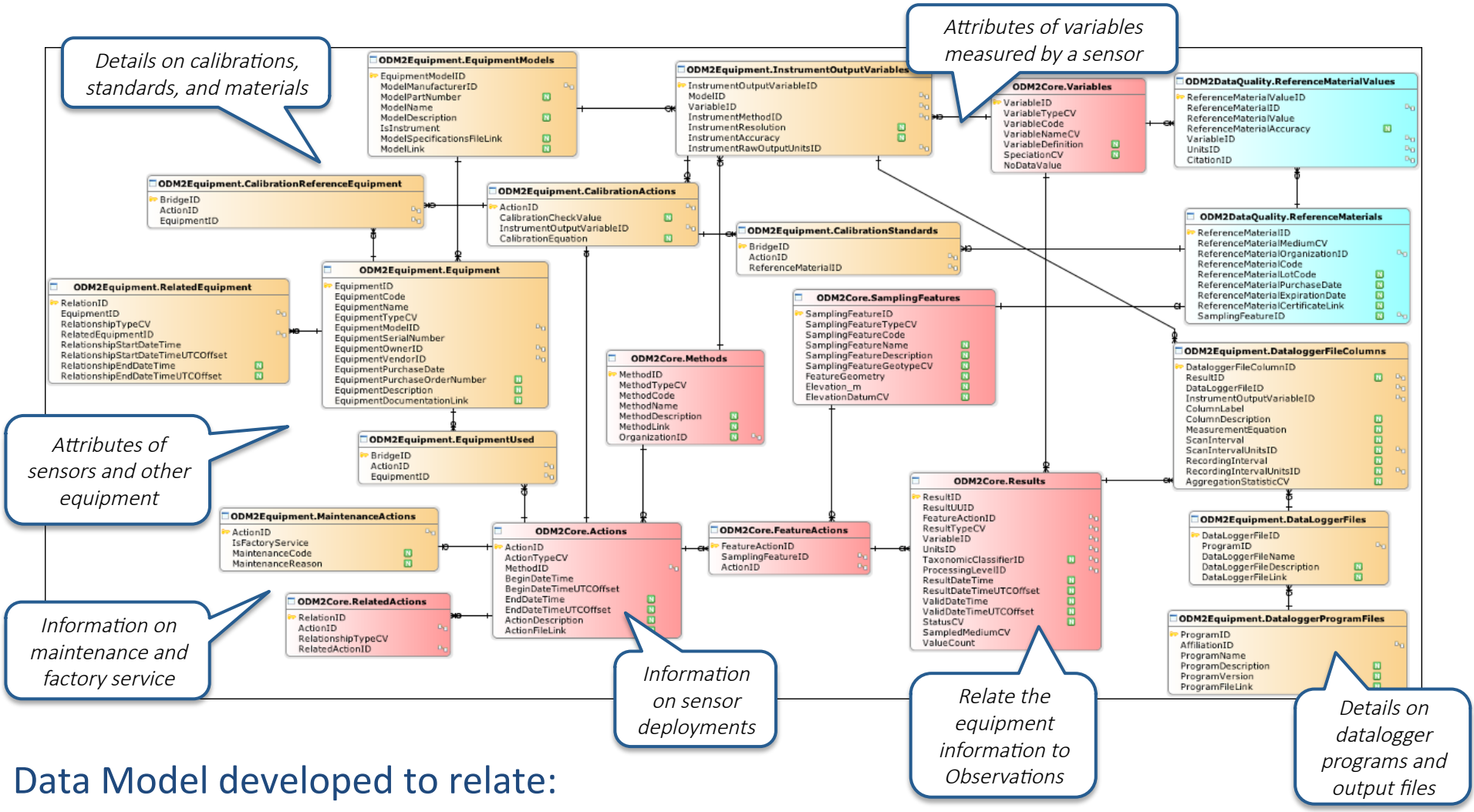


Which soil moisture sensor is buried at 40 cm?

Has this turbidity sensor been serviced at the factory?

What is this dissolved oxygen sensor's calibration history?

Monitoring Equipment Management

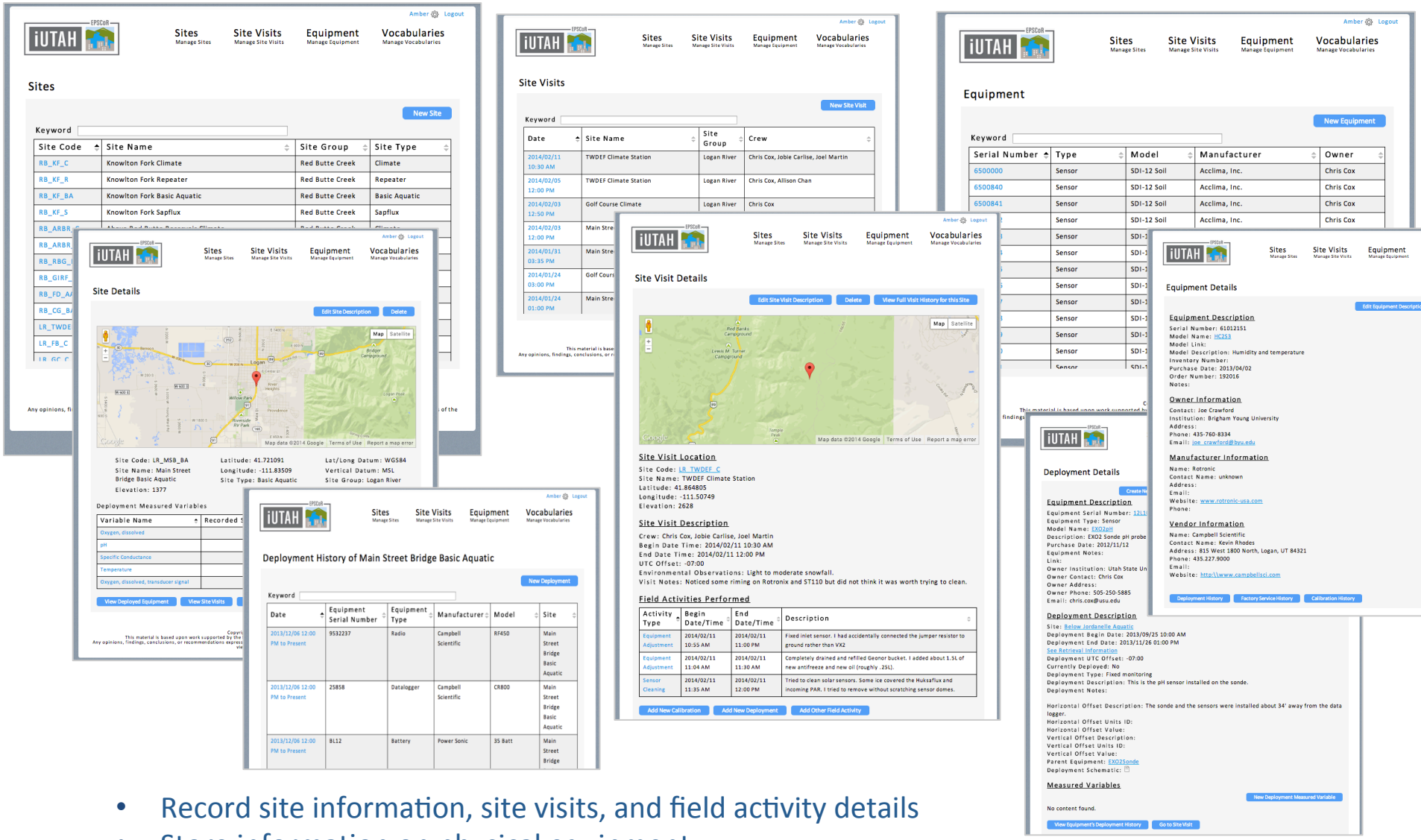


Data Model developed to relate:

- Equipment
- Field Activities
- Deployments
- Calibrations
- Measured Variables
- Datalogger Programs

Database serves as underlying structure to web interface.

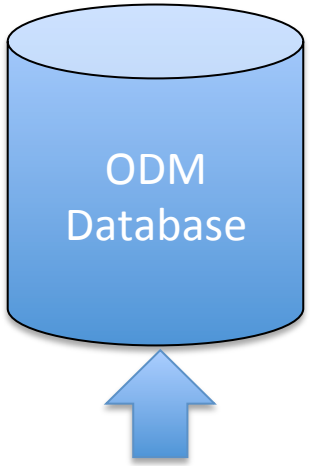
Monitoring Equipment Management



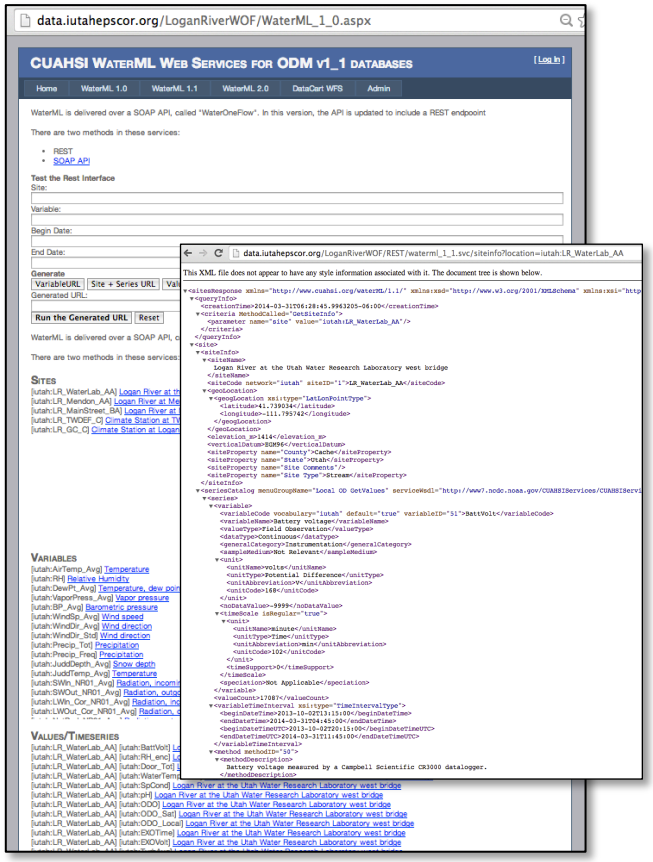
- Record site information, site visits, and field activity details
- Store information on physical equipment
- Track equipment deployments, calibrations, service events

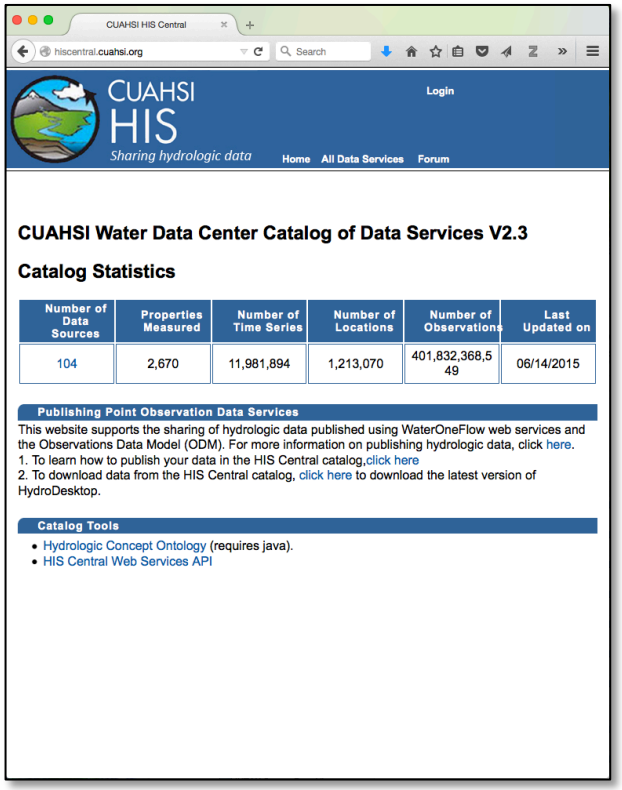
Publishing Data with the CUAHSI Water Data Center

1. Load your data into an ODM database
2. Deploy a WaterOneFlow web service for each ODM database on a web server
3. Register the web service with the CUAHSI Water Data Center



The iUTAH workflow focuses on this step

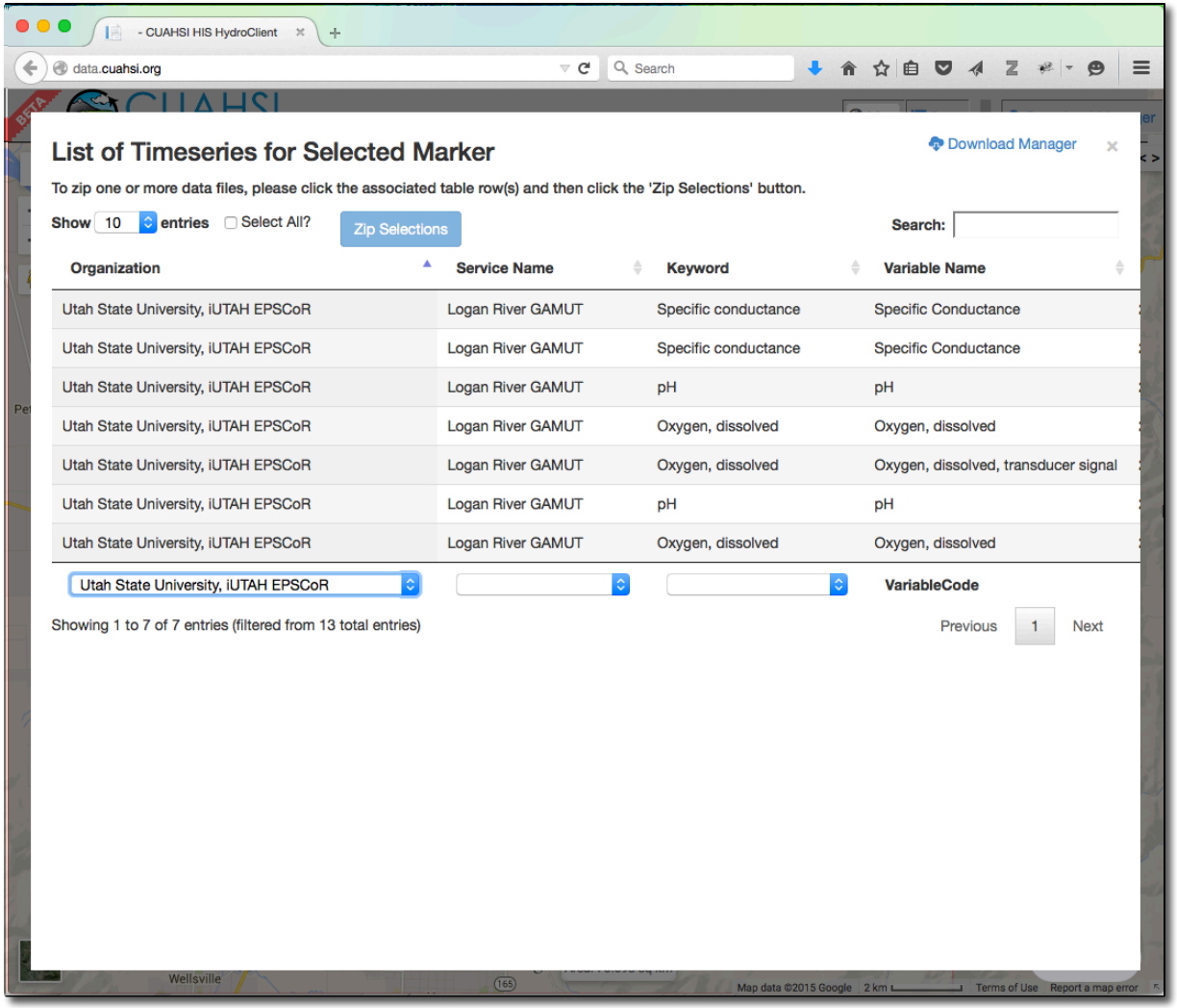




New CUAHSI HIS Data Search Web Client

http://data.cuahsi.org/ (Official release on June 22, 2015)

- Select a location
- Select a time period
- Select keywords
- Select data services
- Search map
- Select a site and download time series as CSV



CUAHSI HIS Resources

- CUAHSI Water Data Center: https://www.cuahsi.org/wdc
- CUAHSI HIS HydroServer Codeplex Website: http://hydroserver.codeplex.com/
- Contact: Jon Pollak
  - User Support Specialist
  - jpollak@cuahsi.org

CUAHSI HIS
Sharing hydrologic data

New Development in Open Source Code Repositories

- **WEBTSA** – Time series data visualization
  - https://github.com/UCHIC/WEBTSA
- **ODM Streaming Data Loader**
  - https://github.com/ODM2/ODM2StreamingDataLoader
- **ODM Tools Python** – Sensor Data Management
  - https://github.com/UCHIC/ODMToolsPython
- **ODM2 Sensor** – Sensor equipment management
  - https://github.com/UCHIC/ODM2Sensor
- **iUTAH Utilities** – Automated alerts, etc.
  - https://github.com/UCHIC/iUtahUtilities
- **iUTAH Data** – Data Website
  - https://github.com/UCHIC/iUTAHData

GitHub

More Resources

WATER RESOURCES RESEARCH, VOL. 44, W05406, doi:10.1029/2007WR006392, 2008

Click Here for Full Article

A relational model for environmental and water resources data

Jeffery S. Horsburgh,[1] David G. Tarboton,[1] David R. Maidment,[2] and Ilya Zaslavsky[3]

Received 30 July 2007; revised 24 January 2008; accepted 13 February 2008; published 8 May 2008.

[1] Environmental observations are fundamental to hydrology and water resources, and the way these data are organized and manipulated either enables or inhibits the analyses that can be performed. The Observations Data Model presented here provides a new and consistent format for the storage and retrieval of point environmental observations in a relational database designed to facilitate integrated analysis of large data sets collected by multiple investigators. Within this data model, observations are stored with sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and to provide traceable heritage from raw measurements to useable information. The design is based upon a relational database model that exposes each single observation as a record, taking advantage of the capability in relational database systems for querying based upon data values and enabling cross-dimension data retrieval and analysis. This paper presents the design principles and features of the Observations Data Model and illustrates how it can be used to enhance the organization, publication, and analysis of point observations data while retaining a simple relational format. The contribution of the data model to water resources is that it represents a new, systematic way to organize and share data that overcomes many of the syntactic and semantic differences between heterogeneous data sets, thereby facilitating an integrated understanding of water resources based on more extensive and fully specified information.

Citation: Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resour. Res.*, *44*, W05406, doi:10.1029/2007WR006392.

1. Introduction

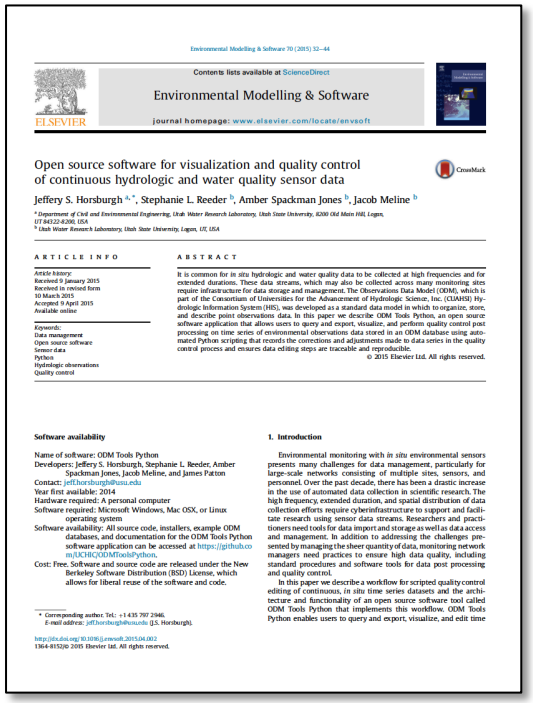
[2] Environmental observations are fundamental to hydrology and water resources, and the manner in which the data are collected, organized, and manipulated either enables or inhibits their scientific analysis [*Pokorný*, 2006; *Tomasic and Simon*, 1997]. When scientists and engineers want to search for and use environmental observations data, they are generally faced with the following problems [*Tomasic and Simon*, 1997]: (1) data are not sufficient or do not exist; (2) data are not published and are hard to locate; (3) data are not easy to access, they are either private or expensive, or require costly preprocessing before they can be used; (4) data are not easy to use because they are inconsistent or noncompatible; and (5) data are not adequately documented. Addressing these issues is one of the main challenges influencing recent developments in environmental information systems, which include water resources and hydrologic information systems [*Pokorný*, 2006; *Bouguettaya et al.*, 2001].

[3] Even for data sets that have been published for widespread use, points three through five above still apply. Generally, data sets published on public Web sites are in file-based systems that are different syntactically (e.g., file types, file formats, and data structure) and semantically (e.g., variable names, units, and descriptive metadata) from one data source to the next. In accessing these data archives, users are faced with the daunting task of navigating through directories and supporting files to find all of the metadata necessary for interpreting and using the data. There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. This need is being driven by the ever increasing number of environmental observations being produced as sensor technology improves, as the number, size, and complexity of environmental monitoring programs grow (including efforts to establish a national network of large-scale environmental observatories), and as engineers and scientists realize that it is as important to characterize the environment with observations as it is to describe it with models and simulations. It is critical that the data, when published, be carefully annotated with metadata so that they can be unambiguously interpreted and used.

[4] In this paper we present a logical database design for the Observations Data Model (ODM) that advances the information science knowledge base of water resources research. We describe a relational model that eases access to and manipulation of time series of observations from experimental sites and watersheds and facilitates data pub-

[1]Utah Water Research Laboratory, Utah State University, Logan, Utah, USA.
[2]Center for Research in Water Resources, University of Texas at Austin, Austin, Texas, USA.
[3]San Diego Supercomputer Center, University of California, San Diego, La Jolla, California, USA.

Copyright 2008 by the American Geophysical Union.
0043-1397/08/2007WR006392$09.00



Environmental Modelling & Software 70 (2015) 32–44

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Open source software for visualization and quality control of continuous hydrologic and water quality sensor data

Jeffery S. Horsburgh [a, *], Stephanie L. Reeder [b], Amber Spackman Jones [b], Jacob Meline [b]

[a] Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
[b] Utah Water Research Laboratory, Utah State University, Logan, UT, USA

ARTICLE INFO

Article history:
Received 9 January 2015
Received in revised form
10 March 2015
Accepted 9 April 2015
Available online

Keywords:
Data management
Open source software
Sensor data
Python
Hydrologic observations
Quality control

ABSTRACT

It is common for *in situ* hydrologic and water quality data to be collected at high frequencies and for extended durations. These data streams, which may also be collected across many monitoring sites require infrastructure for data storage and management. The Observations Data Model (ODM), which is part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS), was developed as a standard data model in which to organize, store, and describe point observations data. In this paper we describe ODM Tools Python, an open source software application that allows users to query and export, visualize, and perform quality control post processing on time series of environmental observations data stored in an ODM database using automated Python scripting that records the corrections and adjustments made to data series in the quality control process and ensures data editing steps are traceable and reproducible.

© 2015 Elsevier Ltd. All rights reserved.

Software availability

Name of software: ODM Tools Python
Developers: Jeffery S. Horsburgh, Stephanie L. Reeder, Amber Spackman Jones, Jacob Meline, and James Patton
Contact: jeff.horsburgh@usu.edu
Year first available: 2014
Hardware required: A personal computer
Software required: Microsoft Windows, Mac OSX, or Linux operating system
Software availability: All source code, installers, example ODM databases, and documentation for the ODM Tools Python software application can be accessed at https://github.com/UCHIC/ODMToolsPython.
Cost: Free. Software and source code are released under the New Berkeley Software Distribution (BSD) License, which allows for liberal reuse of the software and code.

1. Introduction

Environmental monitoring with *in situ* environmental sensors presents many challenges for data management, particularly for large-scale networks consisting of multiple sites, sensors, and personnel. Over the past decade, there has been a drastic increase in the use of automated data collection in scientific research. The high frequency, extended duration, and spatial distribution of data collection efforts require cyberinfrastructure to support and facilitate research using sensor data streams. Researchers and practitioners need tools for data import and storage as well as data access and management. In addition to addressing the challenges presented by managing the sheer quantity of data, monitoring network managers need practices to ensure high data quality, including standard procedures and software tools for data post processing and quality control.

In this paper we describe a workflow for scripted quality control editing of continuous, *in situ* time series datasets and the architecture and functionality of an open source software tool called ODM Tools Python that implements this workflow. ODM Tools Python enables users to query and export, visualize, and edit time

* Corresponding author. Tel.: +1 435 797 2946.
E-mail address: jeff.horsburgh@usu.edu (J.S. Horsburgh).

http://dx.doi.org/10.1016/j.envsoft.2015.04.002
1364-8152/0 2015 Elsevier Ltd. All rights reserved.

Environ Monit Assess (2015) 187:348
DOI 10.1007/s10661-015-4594-3

A data management and publication workflow for a large-scale, heterogeneous sensor network

Amber Spackman Jones · Jeffery S. Horsburgh · Stephanie L. Reeder · Maurier Ramirez · Juan Caraballo

Received: 21 January 2015 / Accepted: 5 May 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract It is common for hydrology researchers to collect data using in situ sensors at high frequencies, for extended durations, and with spatial distributions that produce data volumes requiring infrastructure for data storage, management, and sharing. The availability and utility of these data in addressing scientific questions related to water availability, water quality, and natural disasters relies on effective cyberinfrastructure that facilitates transformation of raw sensor data into usable data products. It also depends on the ability of researchers to share and access the data in useable formats. In this paper, we describe a data management and publication workflow and software tools for research groups and sites conducting long-term monitoring using in situ sensors. Functionality includes the ability to track monitoring equipment inventory and events related to field maintenance. Linking this information to the observational data is imperative in ensuring the quality of sensor-based data products. We present these tools in the context of a case study for the innovative Urban Transitions and Aridregion Hydrosustainability (iUTAH) sensor network. The iUTAH monitoring network includes sensors at aquatic and terrestrial sites for continuous monitoring of common meteorological variables, snow accumulation and melt, soil moisture, surface water flow, and surface water quality. We present the overall workflow we have developed for effectively transferring data from field monitoring sites to ultimate end-users and describe the software tools we have deployed for storing, managing, and sharing the sensor data. These tools are all open source and available for others to use.

Keywords Cyberinfrastructure · Sensor · Quality control · Data management · Hydrology · Data models · Observatory

Introduction

Advances in the development of in situ environmental sensors have led to the ubiquitous use of sensors and sensor networks in environmental monitoring (Martinez et al. 2004; Hart and Martinez 2006; Rundel et al. 2009). Researchers and practitioners are collecting data with in situ sensors at high frequencies, for extended durations, and with spatial distributions that generate volumes of data for which the deployment of cyberinfrastructure (CI) for data management is necessary. Additional challenges are presented by networks that consist of multiple data collection sites, sensors, and personnel (Rüegg et al. 2014). Consistency in data management across these factors can facilitate data integration.

CI integrates computing hardware, digitally enabled sensors, data observatories and experimental facilities,

A. S. Jones (✉) · S. L. Reeder · M. Ramirez · J. Caraballo
Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
e-mail: amber.jones@usu.edu

J. S. Horsburgh
Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA

Published online: 13 May 2015

Springer

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resources Research*, 44, W05406, doi:10.1029/2007WR006392.

Horsburgh, J. S., S. L. Reeder, A. Spackman Jones, J. Meline (2015). Open source software for visualization and quality control of continuous hydrologic and water quality sensor data, *Environmental Modelling & Software*, 70, 32-44, doi:10.1016/j.envsoft.2015.04.002.

Jones, A.S., J.S. Horsburgh, S.L. Reeder, M. Ramirez, J. Caraballo (2015). A data management and publication workflow for a large-scale, heterogeneous sensor network, *Environmental Monitoring and Assessment*, 187:348, doi:10.1007/s10661-015-4594-3.

Questions?

Utah Water Research Laboratory
UWRL

UtahState University

Jeff Horsburgh
jeff.horsburgh@usu.edu

http://data.iutahepscor.org

NSF

This project is funded by National Science Foundation grant EPS-1208732.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.