



Theses and Dissertations

2021-06-07

Regional Water Quality Data Viewer Tool: An Open-Source to Support Research Data Access

Danisa Dolder
Brigham Young University

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Engineering Commons](#)

BYU ScholarsArchive Citation

Dolder, Danisa, "Regional Water Quality Data Viewer Tool: An Open-Source to Support Research Data Access" (2021). *Theses and Dissertations*. 9196.
<https://scholarsarchive.byu.edu/etd/9196>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

Regional Water Quality Data Viewer Tool: An Open-Source
to Support Research Data Access

Danisa Dolder

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Master of Science

A. Woodruff Miller, Chair
Gustavious P. Williams
E. James Nelson

Department of Civil and Environmental Engineering
Brigham Young University

Copyright © 2021 Danisa Dolder

All Rights Reserved

ABSTRACT

Regional Water Quality Data Viewer Tool: An Open-Source to Support Research Data Access

Danisa Dolder

Department of Civil and Environmental Engineering, BYU
Master of Science

The majority of this thesis consists of a paper submitted to the Hydrology journal.

Water quality data collection, storage, and access is a difficult task and significant work has gone into methods to store and disseminate these data. We present a tool to disseminate research in a simple method that does not replace but extends and leverages these tools. In the United States, the federal government maintains two systems to fill that role for hydrological data: the U.S. Geological Survey (USGS) National Water Information System (NWIS) and the U.S. Environmental Protection Agency (EPA) Storage and Retrieval System (STORET), since superseded by the Water Quality Portal (WQP). The Consortium of the Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) has developed the Hydrologic Information System (HIS) to standardize search and discovery of these data as well as other observational time series datasets.

Additionally, CUAHSI developed and maintains HydroShare.org as a web portal for researchers to store and share hydrology data in a variety of formats including spatial geographic information system data. We present the Tethys Platform based Water Quality Data Viewer (WQDV) web application that uses these systems to provide researchers and local monitoring organizations with a simple method to archive, view, analyze, and distribute water quality data. WQDV provides an archive for non-official or preliminary research data and access to those data that have been collected but need to be distributed prior to review or inclusion in the state database. WQDV can also accept subsets of data downloaded from other sources, such as the EPA WQP. WQDV helps users understand what local data are available and how they relate to the data in larger databases. WQDV presents data in spatial (maps) and temporal (time series graphs) forms to help the users analyze and potentially screen the data sources before export for additional analysis. WQDV provides a convenient method for interim data to be widely disseminated and easily accessible in the context of a subset of official data. We present WQDV using a case study of data from Utah Lake, Utah, United States of America.

Keywords: web application, water quality, Utah lakes, research data access, HydroShare

ACKNOWLEDGEMENTS

I thank Dr Miller, Dr Williams, and Dr Nelson who served as my committee for the invaluable assistance during the process of completing this thesis. I would also like to thank Giovanni Romero and Jorge Luis Sanchez, graduate students from the BYU Hydroinformatics laboratory, who unconditionally helped me when I needed it.

I want to express my gratitude to my parents Herman and Mariluz Dolder who have always encouraged me to study and improve myself, teaching me that an education is the best legacy they can give their children. I appreciate their support throughout all my years of study. Also, to my siblings for their help in all aspects. I especially thank my daughter Nahiara and son Dan for their unconditional support, patience and sacrifice during this time.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
1 Introduction	1
2 Research Work	2
2.1 Introduction	3
2.1.1 Background and Need.....	3
2.1.2 Application Objectives.....	6
2.2 Methods: Application Design and Capabilities.....	8
2.2.1 Application Design	8
2.2.2 WQDV Data Schema.....	9
2.2.3 Data Access.....	11
2.2.4 Visualization	13
2.2.4.1 Spatial Map	13
2.2.4.2 Time Series.....	13
2.2.5 Downloading Data	14
2.2.5.1 Download Data by Station	14
2.2.5.2 Download Plot Data	15
2.3 Case Study.....	15
2.3.1 Utah Water Quality	15
2.3.2 Utah Water Quality Data	16
2.3.2.1 Brigham Young University Data	16
2.3.2.2 State of Utah Data	17
2.4 Case Study Conclusions	18
2.4.1 Background.....	18
2.4.2 Case Study Data.....	18
2.4.3 Data Limits.....	22
2.4.3.1 Minimum Limit Value	22
2.4.3.2 Maximum Value.....	23
2.4.4 Data Visualization.....	23
2.4.4.1 Maps.....	25
2.4.4.2 Graphs	26

2.5	Conclusions	28
3	COMMENTS.....	30
	References.....	31

LIST OF TABLES

Table 2-1: Metadata for the WQDV schema	10
Table 2-2: Utah Lake BYU measurements uploaded to the WQDV database.	19
Table 2-3. Utah Lake AWQMS measurements included in WQDV for our case study.	21
Table 2-4. Available data for the selected parameters for at least one in-lake station	24

LIST OF FIGURES

Figure 2-1: WQDV Home Page showing in-lake, out-of-lake, and BYU measurements locations as light blue, dark blue, and “Y” icons, respectively.....	11
Figure 2-2: Utah Lake BYU Stations.....	19
Figure 2-3: Utah Lake AWQMS Stations, blue dots are in-lake stations, black dots are out-of-lake stations.....	20
Figure 2-5: Timeseries for Total Nitrogen at BYU-101 station. Data January 2019: (a) the minimum value equal to 0; (b) the minimum value equal to reporting limit.....	22
Figure 2-6: Timeseries for Total Nitrogen at BYU-104 station. (a) All data; (b) <1 Standard Deviation.....	23
Figure 2-7: Utah Lake (a) All lake stations; (b) Stations with Total Nitrogen data from AWQMS.	25
Figure 2-8: Utah Lake stations from AWQMS with Total Nitrogen data of all periods.	26
Figure 2-9: Utah Lake stations from AWQMS with Total Nitrogen data of part of the period. ..	27
Figure 2-10: Utah Lake stations from AWQMS with Total Nitrogen data of only 2 samples.....	27

1 INTRODUCTION

As part of my research work, I submitted a paper to the Hydrology journal published by MDPI. The title of this paper was: Introducing an Open-Source Regional Water Quality Data Viewer Tool to support Research Data Access. The authors of this paper were: Danisa Dolder, Gustavious P. Williams, A. Woodruff Miller, E. James Nelson, Norman L. Jones, and Daniel P. Ames.

This thesis is structured as follows: Chapter 1 provides some background on my work, Chapter 2 is a copy of the submitted journal article, and Chapter 3 provides comments on my work.

2 RESEARCH WORK

This Chapter is a copy of the paper I submitted to the Hydrology journal published by MDPI. The abstract of that paper is:

Water quality data collection, storage, and access is a difficult task and significant work has gone into methods to store and disseminate these data. The Water Quality Data Viewer (WQDV) tool is presented to disseminate research in a simple method that does not replace but extends and leverages these tools. In the United States, the federal government maintains two systems to fill that role for hydrological data: the U.S. Geological Survey (USGS) National Water Information System (NWIS) and the U.S. Environmental Protection Agency (EPA) Storage and Retrieval system (STORET), since superseded by the Water Quality Portal (WQP). The consortium of the Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) has developed the Hydrological Information System (HIS) to standardize search and discovery of these data as well as other observational time series datasets. Additionally, CUAHSI developed and maintains HydroShare.org as a web portal for researchers to store and share hydrology data in a variety of formats including spatial geographic information system data. The Tethys Platform based WQDV web application uses these systems to provide

researchers and local monitoring organizations with a simple method to provide researchers and local monitoring organizations with a simple method to archive, view, analyze, and distribute water quality data. WQDV provides an archive for non-official or preliminary research data and access to those data that have been collected but need to be distributed prior to review or inclusion in the state database. WQDV presents data in spatial (maps) and temporal (time series graphs) form to help the users analyze and potentially screen the data sources before export for additional analysis. WQDV provides a convenient method for interim data to be widely disseminated and easily accessible in the context of a subset of official data. This thesis presents WQDV using a case study of data from Utah Lake, Utah, United States of America.

2.1 Introduction

2.1.1 Background and Need

Water quality data collection, storage, and access are difficult tasks because of the large number of data types, field sample types, laboratory analysis methods, and other critical information such as detection limits or other quality control information [1]. The way in which data are collected, organized, distributed, and managed is crucial for supporting water quality analysis [2,3]. Engineers and researchers generally face many challenges when trying to archive and use such data, including: a lack of data or data gaps in time; data are difficult to find or are not published or stored in an accessible manner; data access is difficult or data are required to be preprocessed or quality controlled before they can be archived or used; and data are inadequately

documented [2]. These are examples of challenges that affect the development of water resource information systems [3]. According to Horsburgh, *et al.* [4] a need exists for new methods to organize environmental data to allow researchers to publish their data so they can be easily distributed and accessed by others.

In the United States of America (U.S.), the federal government has sponsored and maintains large data repositories for both hydrologic and water quality data [1,5]. Beran and Piasecki [5] provide an overview of the two main government-operated systems for hydrological data, the United States Geological Survey (USGS) National Water Information System (NWIS) (<http://waterdata.usgs.gov/nwis>) and the U.S. EPA's Storage and Retrieval System (STORET) (www.epa.gov/storet/). They note that these systems provide data accessibility through a web interface and play a critical role in national access to hydrological data. STORET mostly focuses on water quality data with some flow data, while NWIS includes groundwater and streamflow data, with some overlap between the two services [5]. Beran and Piasecki [5] state that while these services exist, identifying the correct service, accessing the database, and acquiring the data is complicated and time consuming. In general, for the NWIS system the USGS does its own data collection, while the EPA STORET receives data from other organizations such as state environmental agencies and other environmental organizations [5]. However, not all state or local government organizations in the U.S. submit data to STORET as it is not mandatory [5].

In 2009, EPA and the USGS implemented the Water Quality Exchange (WQX) framework for submitting data to the STORET Data Warehouse [6,7]. Water quality data, which are collected by large numbers of research groups and organizations, often have non-standard or inconsistent data descriptions and different methods of data access and distribution [1].

Aggregating water quality data from these different sources can require significant effort and

requires a wide range of expertise [1,8]. To address these issues the U.S. EPA started the Open Water Data Initiative (OWDI) to provide a platform to support access and storage of water quality data from many different providers and users [9]. Since then, the U.S. EPA has partnered with the USGS and the National Water Quality Monitoring Council to develop and implement the Water Quality Portal (WQP) [1]. Researchers note that the WQP system is needed because, while systems such as NWIS, STORET, and CUAHSI-HIS provide high-quality hydrologic data such as streamflow, water quality data are often collected by small, diverse monitoring organizations and have a broad range of access ranging from web services to spreadsheet files only available on direct request, making aggregation and dissemination difficult [1].

The Consortium of the Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) Hydrologic Information System (HIS) project was developed to provide a framework for distributed data storage [10,11]. The CUAHSI-HIS system created an extensive catalog of data sources for search indexes that included the data available in both NWIS and STORET. CUASHI performed a survey in 2004 and found that the foremost need of hydrologists was “...better and easier access to hydrological data”[12]. CUAHSI-HIS was developed following a distributed web services approach with well described metadata to support modeling and data access, description, and storage [4,13,14].

Our work extends this concept of distributed access to water quality data to provide a small web-based application that provides tools for archiving, accessing, and disseminating water quality data created by local research or monitoring groups. It uses the HydroShare resource developed by CUASHI and is designed to ingest, archive, present, and disseminate these water quality data in an easy-to-use accessible manner that is under the control of the local group. It is not meant to replace comprehensive water quality databases, but to act as a tool to

help users better understand their own data, disseminate these data to other interested users, and show how these data fit with previously collected data available from other sources. We expect that many data that are entered into our system will at some point be submitted to larger, more comprehensive systems.

2.1.2 Application Objectives

To provide researchers, managers, and the concerned public with better access to the water quality data collected by our research group, we developed an open-source application called Water Quality Data Viewer (WQDV). Our primary objective was to provide an archive for our non-official research data and to provide others with access to these data that we collected. These data need to be distributed to others working in the same area and other interested groups prior to review and inclusion in a comprehensive, official database.

We use the WQDV as a tool to organize and provide access to our research data across multiple research groups and to distribute these data to our collaborators and to other interested groups. Previous methods we used to archive and distribute research data were ad hoc, and mostly consisted of large numbers of un-organized spreadsheets or reports distributed on request or informally.

We developed WQDV for supporting any research or monitoring projects that collect water quality data. As we discussed, it is not meant to replace more comprehensive databases, but rather compliment them. The data stored in WQDV could later be submitted to more comprehensive databases such as EPA STORET or the WQP. WQDV provides the capability to quickly consolidate, review, analyze, and disseminate research or monitoring data directly under the control of the local group. In addition to methods to store, preview, analyze, and disseminate, locally-collected data, it provides users with methods to add data from other water quality

databases, such as STORET or WQP, so that the local data can be presented in context. It provides several graphical tools to present, access, and visualize data to help users better understand what data are available and deliver preliminary analysis.

The secondary objective of creating WQDV was to help users understand how our data fit or correlate with data available in the official State of Utah Ambient Water Quality Data Management System (AWQMS) water quality database. WQDV can download and include a limited focused subset of data available from official databases. It provides tools that present these data alongside local data to provide a complete picture of the local data and how it integrates with or extends other available data. This allows other researchers or monitoring groups to not only access and disseminate their data but helps them understand how their data fits within the context of other data. Most comprehensive water quality databases are quite complex with a large number of different parameters, measures, and sites; WQDV allows users to only include a focused subset of these data to help present their data in a more complete manner.

In this paper we present WQDV as we configured it for our work on Utah Lake, Utah, U.S.A., as a case study. Based on our experience, our needs and requirements are common among researchers and WQDV can be easily modified to fit other locations. We present this work as a focused case study using our data, as it highlights the issues that WQDV helps researchers address and demonstrates the WQDV capabilities. While in this case study we will reference the State of Utah AQWMS database, other users could replace AWQMS with STORET or other government or local databases.

2.2 Methods: Application Design and Capabilities

2.2.1 Application Design

We used Tethys Platform to develop WQDV. Tethys is an open-source platform developed to facilitate the creation of water resources web applications (apps) [15]. Tethys Platform provides a suite of web development components for spatial data management, mapping/visualization, and user authentication and permissions management. It is based primarily on the Python language which makes it easy to connect to custom scripts for specific applications. The objective of the Tethys Platform project is to lower the technical barrier for the development of web-based water resource management tools.

We used HydroShare to store the data for WQDV. Tethys has tools to easily access data stored in HydroShare. HydroShare was developed for sharing hydrologic data among the scientific community [11,16-19]. HydroShare evolved from the CUASHI-HIS project. CUAHSI-HIS is comprised of a relational database schema called the Observations Data Model (ODM) [4,20], data servers called HydroServers [21], client tools for accessing data from these servers, including HydroDesktop [22], and a central catalog called HIS Central, which stores searchable metadata and supports data discovery services. The CUAHSI HIS was built on a community-controlled shared vocabulary for hydrologic terms and formal protocols for communication between system components [23]. We decided to use HydroShare as the backend for WQDV because it uses a standard protocol and is widely used by both the CUASI-HIS organization and others including the World Meteorological Organization (WMO) Hydrological Observing System (HOS) data broker [24].

Another benefit of using HydroShare is that users can manage data access. For example, HydroShare provides several different access levels: some users may be allowed to upload,

delete, modify, or access data, while others may only upload or access data, while most may only access data. This provides a minimal level of quality control for the data in WQDV.

2.2.2 WQDV Data Schema

For the WQDV database schema, we followed the metadata field names from the AWQMS database to allow us to easily integrate our local data with the state data. While this structure is based on AWQMS, it is a general structure for water quality data and easily used by others. While we wanted to follow the AWQMS format, a download from AWQMS includes 57 different columns of metadata. While information contained in these columns is important for the official archive, much of this information is related to quality assurance/quality control (QA/QC) procedures such as who collected the data, which laboratory did the analysis, what methods were used, etc. We do not have much of this information for our research data and choose not to include these fields in WQDV. This makes it possible to focus on the information required to analyze lake or reservoir water quality for a particular research project, without the overhead required by official databases to ensure that data are certified and have undergone extensive review.

The WQDV schema adopts the basic structure of a more comprehensive water quality database, but with limited metadata. Specifically, we used the following metadata for the WQDV schema (the column names match those in AWQMS):

Table 2-1: Metadata for the WQDV schema

<i>Name</i>	<i>Type</i>	<i>Description</i>
Activity Start Date	Date	Date when the example was taken
Monitoring Location ID	Variant	Number assigned to the station
Monitoring Location Name	String	Name assigned to the station
Monitoring Location Latitude		
Monitoring Location Longitude		
Monitoring Location type		
Characteristic Name	String	Parameter Name
Sample Fraction	String	Total or dissolved parameter measurement differentiated
Result Value	Single	Left blank if below the detection limit or lower than the reporting value
Result Unit	String	
Detection Condition	String	Some methods provide concentrations below the detection limit. Many water quality databases, including AWQMS, allow a measurement to have a lower reporting value that is above the detection limit. This field indicates if the reported value is below the detection limit or below the reporting limit
Detection Limit Value	Single	For data below the detection limit or reporting limit, most water quality databases, including AQWMS, leave the results "Results Value" field empty, users can either ignore the measurement, use the detection, reporting, or some other value as appropriate when this occurs
Detection Limit Unit	String	Units of the Detection Limit Value

This schema allows users with the ability to enter research data in a relatively simple format. WQDV uses the "Monitoring Location Type" field, such as BYU_Dust data, or In_Lake data, to select different icons to represent the different location types. This set of icons is easily configured in WQDV and other researchers may choose to provide additional icons or data categories to display on the map.

For uploading data, WQDV uses a comma-separated values (CSV) file with columns reflecting the metadata listed above. Columns such as “Detection Condition”, “Detection Limit Value” and “Detection Limit Unit” can be blank. Once the file is created, a user with administrator privileges can then upload these data to WQDV and they can be added to the display.

2.2.3 Data Access

Data access using the WQDV application follows a simple graphical process with pull down menus (Figure 2-1). Data are organized by “lake” or reservoir, so the first step in accessing data is lake selection, followed by data source selection, then parameter selection. This allows a researcher to quickly see which measurement locations have data in these collections.

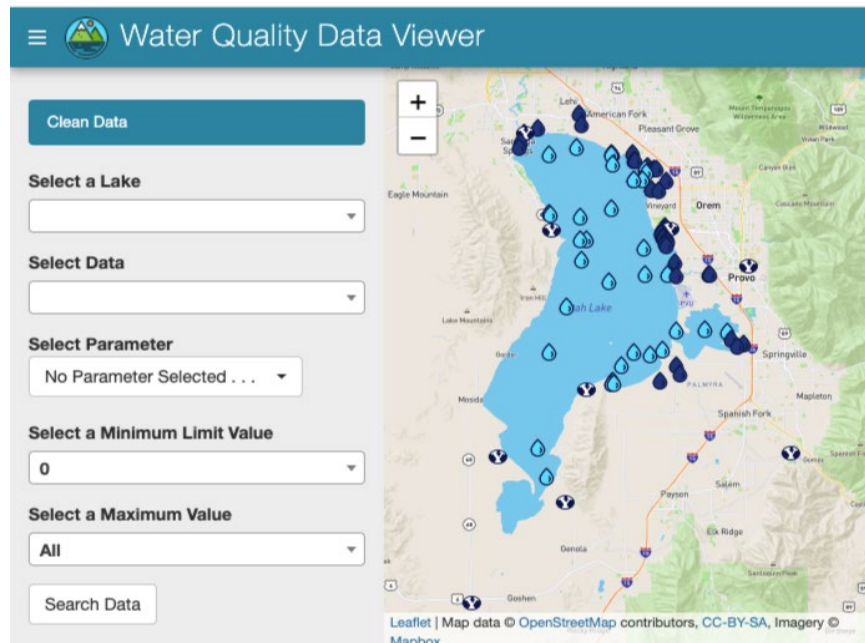


Figure 2-1: WQDV Home Page showing in-lake, out-of-lake, and BYU measurements locations as light blue, dark blue, and “Y” icons, respectively.

Once a parameter is selected, the “Sample Fraction” (e.g., total or dissolved concentrations) can be selected. Next the minimum and maximum data values for display can be selected. This allows users to exclude outliers from the plotted data shown in the interface but does not remove data from the database or downloaded data for distribution. Once these parameters have been selected, WQDV searches the database for data that meet the criteria and provides those data for either display or download.

2.2.3.1 Minimum Limit Values

For many parameters, when a sample is analyzed, test results for the parameter might be below the detection limit and the “Result Value” column is empty but is still added to the database. This is common practice so the data show that a sample was taken on that date but found to have concentrations below the detection limits. These “below-detection-limit” results are typically filled with some number. WQDV gives the option to fill such data with a 0, with $\frac{1}{2}$ the detection limit value, or with the detection limit value. When the user selects one of these options, WQDV uses the selected value to display the data. Data are only replaced if the ‘Result Value’ column is empty (i.e., was below the detection limit).

2.2.3.2 Maximum Values

Many datasets include outliers. WQDV has capabilities to exclude these outliers from the plots if desired. This is important for plots, as the plots scale to the data extent, and if there is an outlier, the plot scales can make data evaluation difficult by hiding variations in lower values. WQDV has an option for removing values that are above 1, 2, 3, or 4 standard deviations.

The standard deviation formula used is:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \quad (2-1)$$

Where: σ is the standard deviation; x_i = the value of the dataset; n = the total number of data points; and μ = the arithmetic mean of the data. WQDV uses the standard deviation values computed on the data in the database to filter outlier values.

2.2.4 Visualization

2.2.4.1 Spatial Map

We designed the WQDV interface to allow a user to quickly determine which stations have data for the parameter of interest from either all data sources or a selected data source. In addition, the icons present the data source visually.

The initial map presented when first accessing WQDV shows all the measurement locations. Then, after selecting a data source and a parameter of interest, and searching the database, the map only shows sample locations that have results for the selected criteria. WQDV can also show either all the stations with data from a selected source, or all the stations with data related to a selected parameter by selecting all data sources.

2.2.4.2 Time Series

WQDV provides tools to easily view the time history of a parameter at any location presented on the map. Clicking on any of the stations displayed on the map after parameter selection will generate a time series graph of the selected water quality parameter. As additional stations are selected, additional plots (or lines) are added to the graph. This capability supports

comparing data at different stations, both in value and the amount of data available. Data from a selected station can be removed from the plot by selecting the legend entry. A single station can be selected by double clicking the legend, which removes the other stations from the graph.

WQDV has the capability to filter the data on the graphs based on the time period.

WQDV has pre-determined time frames which include: the last 6 months of data and the last 12 months of data. WQDV provides the ability to select a custom range by using a slider at the bottom of the graph. The slider shows over what time period the datasets included in the graph are available.

Once data have been selected and filtered on a graph, they can be downloaded. This provides a visual way to search, subset, and download data for additional analysis. By only exporting or downloading the data used to create the graphs, rather than the full data set, a researcher can quickly evaluate the data available at each station, then only download data of interest.

2.2.5 Downloading Data

2.2.5.1 Download Data by Station

After selecting a subset of the available data using the map, the WQDV application provides an efficient method to download the selected data by selecting the “Download Parameter Data” button. This generates a .csv file with the time series data for all the stations shown on the map. This visual search provides a more efficient interface to the database and allows a researcher or user to quickly identify and acquire data of interest.

2.2.5.2 Download Plot Data

As discussed, WQDV can create time series plots by clicking on a station icon, with additional stations added when clicked. Additional stations can be selected, and WQDV will add those data to the plot. Once created, WQDV provides a way to copy the graph or download the data used to create the graph. This capability provides a way to further subset the data by station, value, and time. WQDV provides a clean graph that can be used in a report and a .csv file of the data used to create the graph that can be further analyzed or plotted in other software.

2.3 Case Study

2.3.1 Utah Water Quality

Fresh water bodies are a critical part of our infrastructure. One of the great management challenges in maintaining water quality is understanding and controlling the amount of nutrients in the water. For many surface water impoundments, nutrients are a greater concern than toxic chemicals and affect many more water bodies. Nutrients reach water bodies in several ways that include atmospheric transport, groundwater, stream discharge, surface runoff, sewage, or other process discharges. Although nutrients are essential for aquatic ecosystems, an excess of nutrients acts as pollution, creating algal blooms which are a major water quality issue [25]. An algal bloom is an excessive increase in the concentration of phytoplankton in a water body and is very common in lakes and reservoirs. These algal blooms act as pollution from both toxins that can be released by the blooms and by decay of the algal mass, resulting in decreased oxygen levels and other issues. This increase in pollution due to algal blooms in freshwater bodies is a worldwide problem [25].

The State of Utah relies on several freshwater bodies for irrigation, recreation, and potable water supply. The first settlements, which began in the mid-19th century, were located near the water resources necessary to live in this arid area. However, these settlements produced a change in land use, impacting ecosystems, including the aquatic habitat. The first recorded signs of water contamination in Utah Lake date from the year 1948, when contamination was found in different areas due to the lack of sewage treatment and was identified by algal blooms [26]. The first efforts to control algal blooms were focused on reducing phosphorus input since it is considered the limiting growth factor in most freshwater systems. At some locations, nitrogen inputs are more limited than phosphorus, principally in high and dry areas such as the intermountain west where Utah Lake is located [27]. To better understand these issues, identify the important catalysts, and create management plans requires researchers to have access to historical water quality data and to easily store, retrieve, and distribute research data.

One area we have been researching is the atmospheric deposition of nutrients in Utah Lake (i.e., dust). We have established a set of sampling stations around the lake and collect information on nutrient deposition at these stations from the atmosphere. In this case study we use WQDV to store, visualize, and distribute these data. We also added the 20 typical water quality data results from the AQWMS database to WQDV to present our research data in context.

2.3.2 Utah Water Quality Data

2.3.2.1 Brigham Young University Data

Brigham Young University (BYU) has been collecting data on different parameters from Utah lakes and reservoirs for various projects for over 40 years. Many of these data, while

valuable, will not be submitted to the state database, as they were collected using various research methods, or did not use full QA/QC processes. Based on our experience, this is not uncommon for many researchers. For us, the most important feature of a database system is data preservation for our own future research. A second goal is data distribution so that other researchers and interested parties can have access to our data and leverage our work. Many of our previous research efforts had minimal data archiving. This has resulted in data from the previous 40 years of research being scattered among reports, computer files, and databases. Based on our experience, this is not unusual. For most on-going research, we use data from the state AWQMS database along with data we have or are collecting. Access to both types of data supports a strong research program.

2.3.2.2 State of Utah Data

The Utah Department of Environmental Quality (DEQ) supports collection and distribution of water quality data for all the water bodies in the state. These state databases support research work such as trying to better understand the processes that lead to algal blooms and other water quality problems and explore mitigation measures. The state follows goals set by the United States Environmental Protection Agency (EPA) in the way they archive and manage these data. The state of Utah publishes these data in the AWQMS which can be accessed through a web page. Because AWQMS is the official database, it is large and complex. Data submitted and archived in AWQMS are also required to undergo significant quality assurance and quality control (QA/QC). Researchers often collect data which do not meet these QA/QC standards or do not bother to go through the processes required to submit their data to the official database.

2.4 Case Study Conclusions

2.4.1 Background

We used the WQDV application to integrate BYU research data with data from the state of Utah AWQMS database and show how WQDV could be used to explore, access, and distribute these data. These data were selected with a focus on water quality. Here we present this combined data set and demonstrate how the WQDV application can be used to analyze the data. Utah Lake is the focus of our research group, we have a significant amount of data that cannot be found in the AWQMS database. These results show how the WQDV application can be used to archive, access, and analyze our local research data but also analyze data from the State AWQMS database on the same platform.

In this section, we present several of the capabilities of the WQDV application including spatial visualization of the measurement locations; downloading measurement data from WQDV; visualizing and analyzing the time history of a parameter at a measurement location; demonstrating using data limits for analysis; and downloading graphs and other visualizations.

2.4.2 Case Study Data

We added BYU data to the WQDV database from nine stations. These are data collected by BYU researchers over the last four years that have not been submitted to the AWQMS database. Figure 2-2 shows the location of the stations for the BYU data.

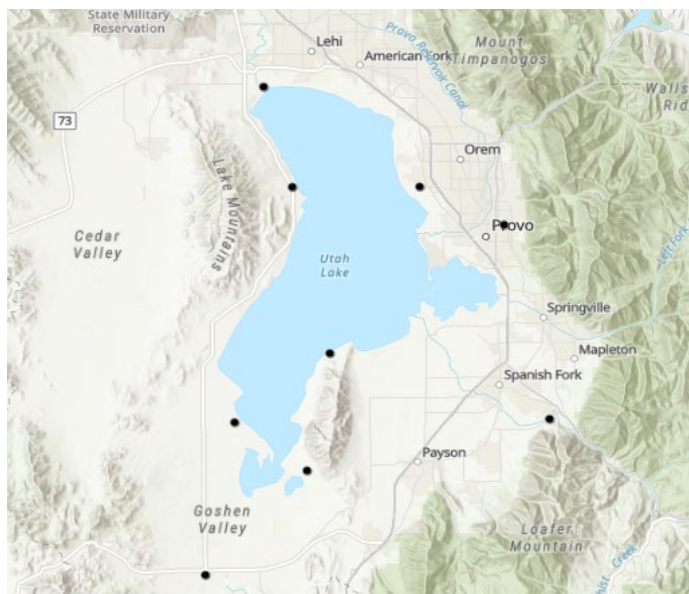


Figure 2-2: Utah Lake BYU Stations

Table 2-2: Utah Lake BYU measurements uploaded to the WQDV database. lists the 3 parameters and the number of measurements for each parameter in the BYU data set. These data are from October 31st, 2016 to June 8th, 2020. We uploaded a total of 917 measurements of Utah Lake research data collected by BYU researchers.

Table 2-2: Utah Lake BYU measurements uploaded to the WQDV database.

Measurement Type	Number of Measurements
Nitrogen	397
Orthophosphate	102
Phosphate-phosphorus	418
TOTAL Measurements	917

The state of Utah maintains the AWQMS, which, for samples within Utah Lake, has records for over 83 different analytes, with 77 above detection limits. AWQMS records data from a very large number of sample locations, many with results from only a single sample.

These data are extremely valuable, but due to the amount of data, researchers can spend significant effort on identifying and cleaning a data set for study, making the database difficult to access and analyze.

For our analysis, we selected and downloaded data from the AWQMS database. We then deleted 37 of the 57 data columns in the AQWMS database, we kept the 20 columns included in WQDV. This resulted in 36 in-lake stations and 42 out-of-lake stations from the AWQMS database. Figure 2-3 shows both the in-lake and out-of-lake stations as blue and black dots, respectively. The out-of-lake stations were the stations nearest to the lake in the AWQMS database.

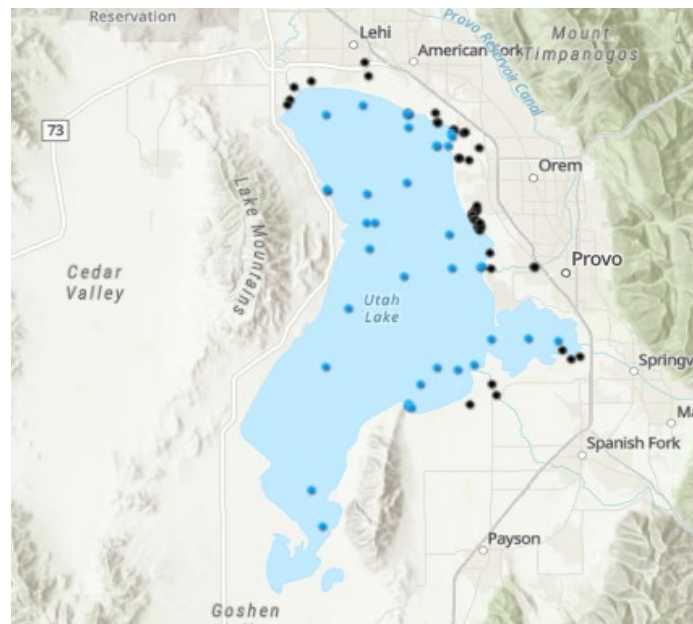


Figure 2-3: Utah Lake AWQMS Stations, blue dots are in-lake stations, black dots are out-of-lake stations.

Table 2-3. Utah Lake AWQMS measurements included in WQDV for our case study. lists the 20 parameters we selected and the number of measurements for each parameter. These data are from January 1st, 1989 to December 31st, 2019.

Table 2-3 separates the in-lake and out-of-lake stations. We downloaded a total of 35,897 samples of Utah Lake data from the AWQMS database for inclusion in the WQDV application.

For our case study, the stations for the three different location types: “In-lake”, “Out-of-lake”, and “BYU” are represented with a different icon (see Figure 2-1).

Table 2-3. Utah Lake AWQMS measurements included in WQDV for our case study.

Measurement Type	Number of Measurements	
	In-lake Stations	Out-of-lake Stations
Ammonia-nitrogen	481	1439
Chlorophyll a	47	32
Chlorophyll a, corrected for pheophytin	426	43
Chlorophyll a, free of pheophytin	218	66
Chlorophyll a, uncorrected for pheophytin	790	89
Depth, Secchi disk depth	3082	1
Dissolved oxygen (DO)	5618	2033
Inorganic nitrogen (nitrate and nitrite)	896	2061
Magnesium	757	1209
Nitrate	71	60
Nitrite	1	2
Nitrogen	1011	1221
Orthophosphate	7	41
pH	5991	2682
Phosphate-phosphorus	2184	2518
Specific conductance	6019	2835
Temperature, water	5708	2048
Total dissolved solids	717	1551
Total suspended solids	939	1509
Turbidity	417	805
Total Measurements	13652	22245

2.4.3 Data Limits

During the search process, WQDV can change how minimum values are reported and filter the maximum data to exclude outliers.

2.4.3.1 Minimum Limit Value

As noted above, in the AWQMS database, the results field for a measurement can be blank due to the value being below the detection limit or reporting value. Figure 2-4 shows how the minimum value changes according to the selection options. In this graph, a single point in time has been selected on the graph. When this is done, the graph reports the values at that point. The value for the station at BYU-101, which is the lowest value at this point and colored blue on the graph, changed from 0 (left panel), to 0.1 (right panel) which is 0 and the detection limit, respectively, for the total nitrogen parameter plotted on these graphs. If these data points were not filled, the graph would not have a point at this location as the cell is empty in the database, that is the measurement at this point in time would not be included.

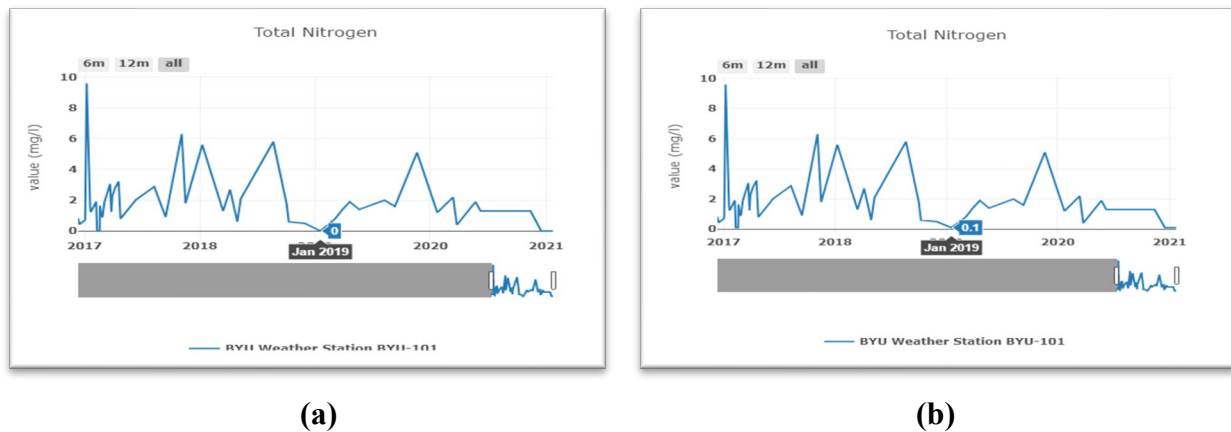


Figure 2-4: Timeseries for Total Nitrogen at BYU-101 station. Data January 2019: (a) the minimum value equal to 0; (b) the minimum value equal to reporting limit.

2.4.3.2 Maximum Value

Figure 2-5 shows how the maximum value changes according to the selection options and how removing the outlier makes data evaluation easier. In panel (a) of Figure 2-5 BYU-104 has a data value of 79.2, well outside the range of the other data. In panel (b) of Figure 2-5 this data point has been filtered, allowing the user to better analyze the other data. When the outlier was included in the plot, important variations in the data were difficult to see in the plots.

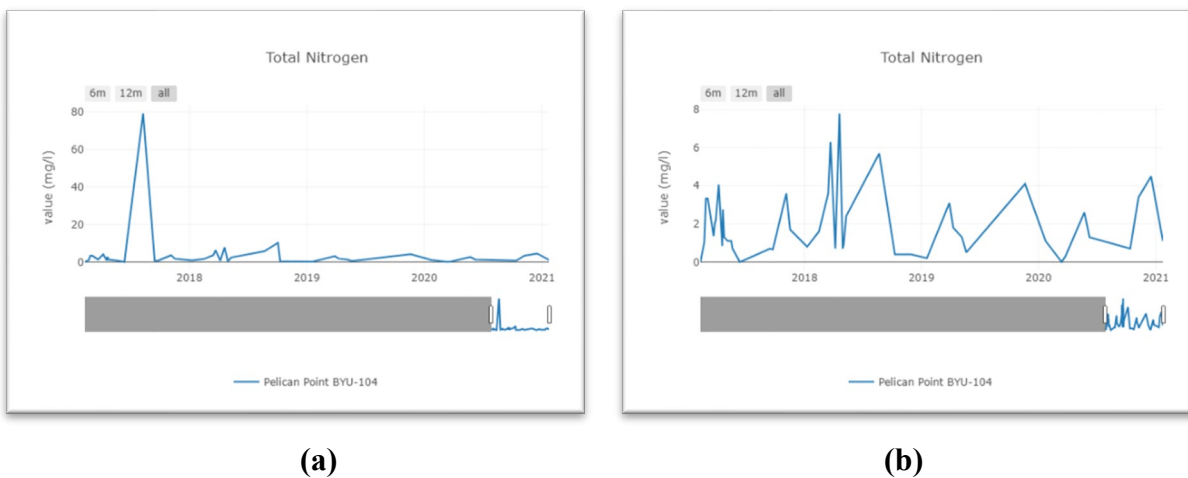


Figure 2-5: Timeseries for Total Nitrogen at BYU-104 station. (a) All data; (b) <1 Standard Deviation.

2.4.4 Data Visualization

Even though there are many data, the various stations or measurement locations are not consistent with how the data were collected over time. Table 2-4 shows that for any given analyte, there are large gaps in time. This issue is exacerbated when looking at individual stations, many of which only have minimal, i.e., 1 or 2, measurements. Table 2-4 shows the Utah Lake data accessed from the AWQMS database for the years 2010-2019 in which at least one

station had a measurement for the analyte. As noted, the data are much sparser at any individual station.

Table 2-4. Available data for the selected parameters for at least one in-lake station in Utah Lake for the years 2010-2019

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Ammonia-nitrogen	█	█	█	█	█	█	█	█	█	█
Chlorophyll a		█	█							
Chlorophyll a, corrected for pheophytin	█	█	█	█	█	█	█	█	█	█
Chlorophyll a, free of pheophytin								█	█	█
Chlorophyll a, uncorrected for pheophytin		█	█	█	█	█	█	█	█	█
Depth, Secchi disk depth	█	█	█	█	█	█	█	█	█	█
Dissolved oxygen (DO)	█	█	█	█	█	█	█	█	█	█
Inorganic nitrogen	█	█	█	█	█	█	█	█	█	█
Magnesium	█	█	█	█	█	█	█	█	█	█
Nitrate										
Nitrite										
Nitrogen	█	█	█	█	█	█	█	█	█	█
Orthophosphate										
pH	█	█	█	█	█	█	█	█	█	█
Phosphate-phosphorus					█	█	█	█	█	█
Specific conductance	█	█	█	█	█	█	█	█	█	█
Temperature, water	█	█	█	█	█	█	█	█	█	█
Total dissolved solids	█	█	█	█	█	█	█	█	█	█
Total suspended solids	█	█	█	█	█	█	█	█	█	█
Turbidity				█	█	█	█	█	█	█

Table 2-4 demonstrates one of the issues with access to data in the AWQMS database. In our example, Utah Lake has many stations with analyte measurements, many of these sites only have a few measurements. Even a given analyte measured at any location in the lake might have large gaps. While it is not difficult to download all these data and then screen the various locations or analytes to determine if there are data for a given time, the WQDV application provides a visual interface that allows researchers to more quickly understand which data are

available, both in space - using the map interface, and in time - by generating time-plots at individual stations.

2.4.4.1 Maps

To analyze the data of Table 2-4, we selected AWQMS data as the data source and total nitrogen as the parameter to explore. Figure 2-6a shows Utah Lake with all the stations. These include in-lake stations (light blue), out-of-lake stations (dark blue), and stations with BYU data (dark blue with a Y). In the next panel, Figure 2-6b, we have selected the data source (in this case AWQMS), and total nitrogen as the parameter. Once we click the “search data”, seen in the bottom portion of the menu, the map shows only the Utah Lake stations that have at least one sample of Total Nitrogen from the AWQMS data (Figure 2-6b). This example shows a subset: data from only one source and one parameter.

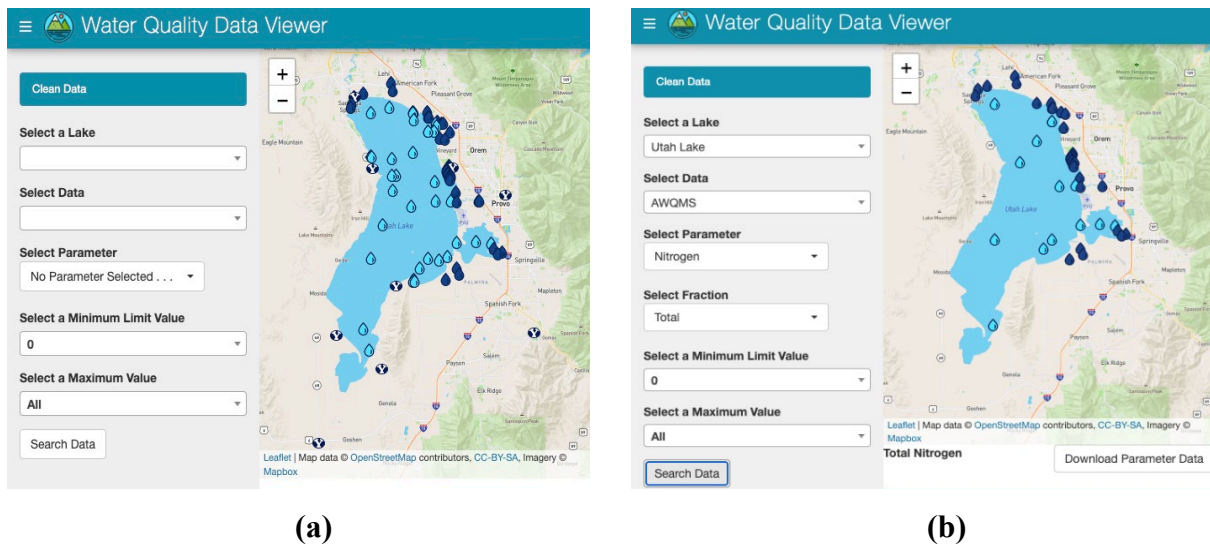


Figure 2-6: Utah Lake (a) All lake stations; (b) Stations with Total Nitrogen data from AWQMS.

2.4.4.2 Graphs

The WQDV application includes capabilities to visually select locations and parameters and graph the resulting data. For this case study, the AWQMS data we use is from January 1st, 1989 to December 31st, 2019. Figure 2-7, Figure 2-8, and Figure 2-9 show timeseries of some Utah Lake stations from AWQMS with total nitrogen data. From the time series graphs of each station, we can understand that not all the stations have the same number of measurements for total nitrogen. In this analysis, we found that some stations have total nitrogen data consistently from June 2009 to Sep 2009 (Figure 2-7), while others have started to collect total nitrogen data in May 2017 continuing through Sep 2019 (Figure 2-8). We also found that the station 4917470 has only two measurements for total nitrogen (Figure 2-9).

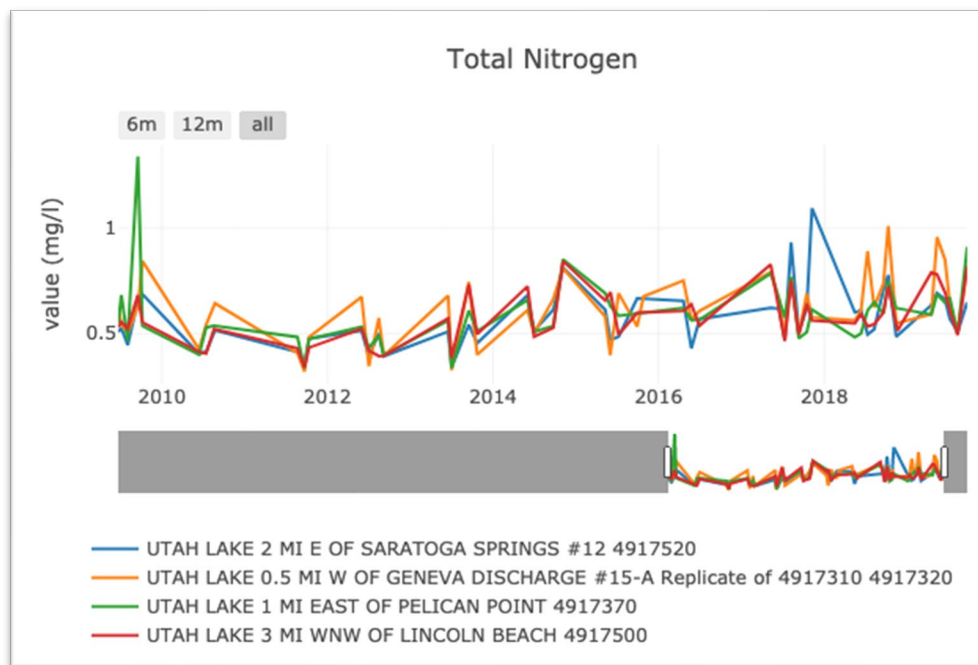


Figure 2-7: Utah Lake stations from AWQMS with Total Nitrogen data of all periods.

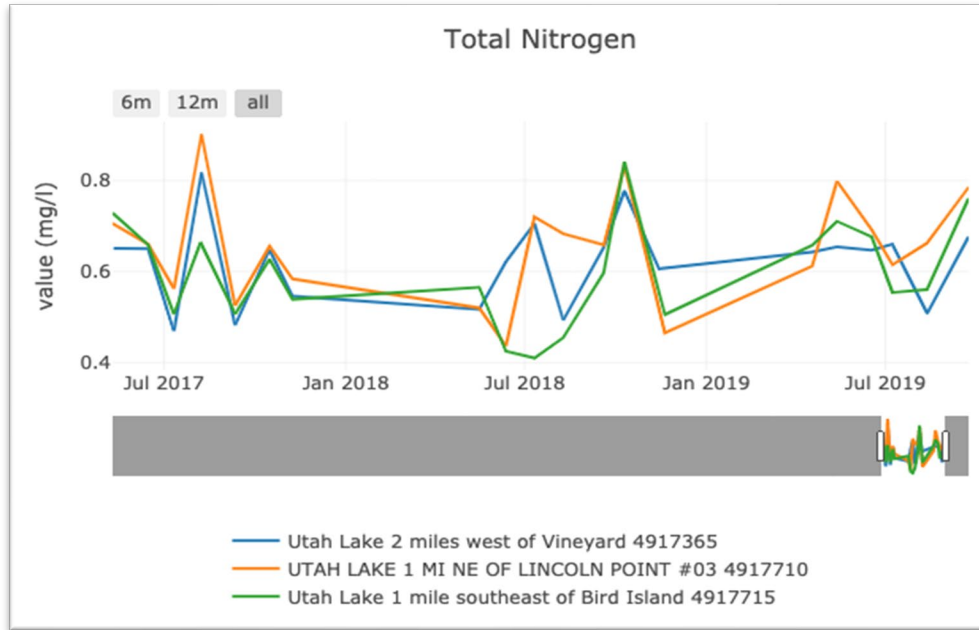


Figure 2-8: Utah Lake stations from AWQMS with Total Nitrogen data of part of the period.

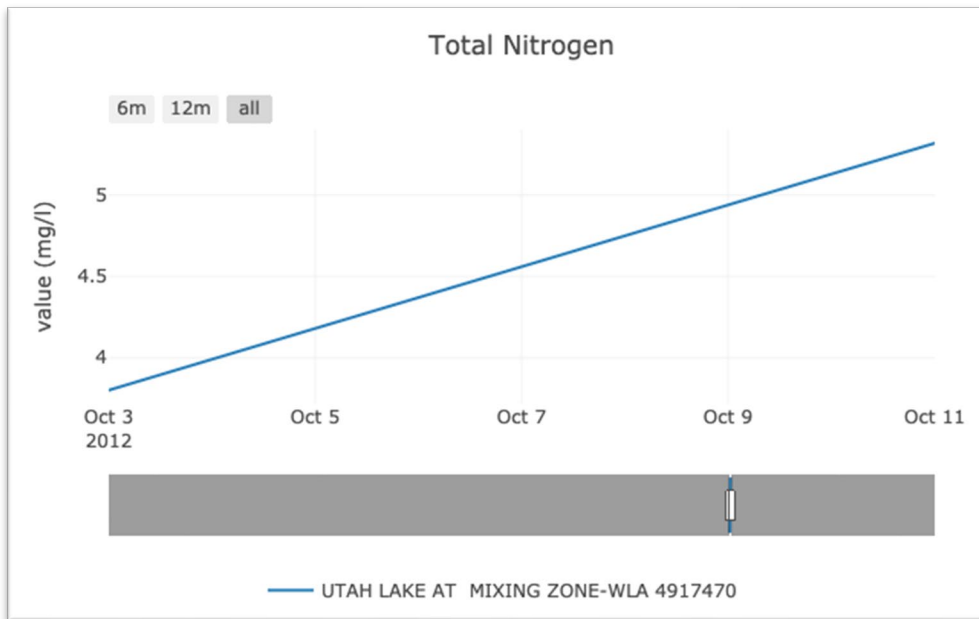


Figure 2-9: Utah Lake stations from AWQMS with Total Nitrogen data of only 2 samples.

2.5 Conclusions

We demonstrate the WQDV application using data from Utah Lake, from our research and the state data. Our research data would normally be difficult to find, access, and distribute, and the data from the AWQMS database, which are accessible, have a somewhat more complex access method. While this presentation focuses on the data from Utah Lake, WQDV can work with data from any location. To use the WQDV application for another lake or reservoir, the user would need to add the location to the menu, which requires adding a few lines to the code, and upload the associated data using the metadata.

We developed the WQDV application to demonstrate how modern web-programing frameworks and approaches can help researchers archive, access, and distribute their data. We show how data from other databases can be loaded into WQDV to create a more comprehensive dataset for research use. WQDV helps efficiently distribute spatial environmental research data. These research data could be on an interim basis, as data are waiting to be accepted by official databases, or as a more permanent archive for research or other data that might not be submitted to official databases. The data are easily updated when the main database is changed.

In summary, the application solves two general problems:

- The difficulty in archiving and providing access to research data, or data gathered by other groups, to other users or interested parties.
- The difficulty of determining what data are available in a large, comprehensive database, such as the AWQMS database.

The WQDV application presents data using spatial and temporal visualizations. The spatial (i.e., maps) and temporal (i.e., time series plots) forms allow users to quickly understand what data are available, if those data are of interest, and to access and download those data.

For this demonstration we used BYU research data on and near Utah Lake which has not been submitted to the AWQMS database and included Utah Lake data from the AWQMS database maintained by the state of Utah. AWQMS is a comprehensive database of water quality data, but it is difficult to determine what data are available and to efficiently access and download those data. We showed how the WQDV application could be used to archive and widely distribute these data.

3 COMMENTS

In addition to the submitted journal article, I created a web application titled: “Water Quality Data Viewer”. It is available at <https://tethys-staging.byu.edu/apps/lake/data/>.

Source code for the application is available via a GitHub repository at https://github.com/BYU-Hydroinformatics/lake_parameters.

The data used for the application is available via a HydroShare Resource at <https://www.hydroshare.org/resource/cf0133c4d4a14a7f938918707abb4e05/>.

REFERENCES

1. Read, E.K.; Carr, L.; De Cicco, L.; Dugan, H.A.; Hanson, P.C.; Hart, J.A.; Kreft, J.; Read, J.S.; Winslow, L.A. Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research* 2017, *53*, 1735-1745.
2. Tomasic, A.; Simon, E. Improving access to environmental data using context information. *SIGMOD Rec.* 1997, *26*, 11–15, doi:10.1145/248603.248606.
3. Pokorný, J. Database architectures: Current trends and their relationships to environmental data management. *Environmental Modelling & Software* 2006, *21*, 1579-1586.
4. Horsburgh, J.S.; Tarboton, D.G.; Maidment, D.R.; Zaslavsky, I. A relational model for environmental and water resources data. *Water Resources Research* 2008, *44*.
5. Beran, B.; Piasecki, M. Availability and coverage of hydrologic data in the US geological survey National Water Information System (NWIS) and US Environmental Protection Agency Storage and Retrieval System (STORET). *Earth Science Informatics* 2008, *1*, 119-129, doi:10.1007/s12145-008-0015-2.
6. Slawewski, T.; Young, D.; Dean, B.; Bergenroth, B.; Sparks, K. Pilot implementation of the US EPA interoperable watershed network. *Open Geospatial Data, Software and Standards* 2017, *2*, 1-11.
7. Park, Y.S.; Engel, B.A.; Kim, J.; Theller, L.; Chaubey, I.; Merwade, V.; Lim, K.J. A web tool for STORET/WQX water quality data retrieval and Best Management Practice scenario suggestion. *Journal of environmental management* 2015, *150*, 21-27.
8. Soranno, P.A.; Bissell, E.G.; Cheruvilil, K.S.; Christel, S.T.; Collins, S.M.; Fergus, C.E.; Filstrup, C.T.; Lapierre, J.-F.; Lottig, N.R.; Oliver, S.K. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *GigaScience* 2015, *4*, s13742-13015-10067-13744.

9. Larsen, S.; Hamilton, S.; Lucido, J.; Garner, B.; Young, D. Supporting diverse data providers in the open water data initiative: Communicating water data quality and fitness of use. *JAWRA Journal of the American Water Resources Association* 2016, 52, 859-872.
10. Tarboton, D.G.; Horsburgh, J.S.; Maidment, D.R. CUAHSI community Observations Data Model (ODM) version 1.1 design specifications. *Des Doc* 2008.
11. Tarboton, D.G.; Idaszak, R.; Ames, D.; Goodall, J.; Horsburgh, J.S.; Band, L.; Merwade, V.; Song, C.; Couch, A.; Valentine, D. HydroShare: an online, collaborative environment for the sharing of hydrologic data and models. 2012.
12. Bandaragoda, C.; Tarboton, D.; Maidment, D. Hydrology's efforts toward the cyberfrontier. 2006.
13. Beran, B.; Piasecki, M. Engineering new paths to water data. *Computers & Geosciences* 2009, 35, 753-760.
14. Valentine, D.; Zaslavsky, I.; Whitenack, T.; Maidment, D. Design and implementation of CUAHSI WATERML and WaterOneFlow Web services. In Proceedings of the AGU Fall Meeting Abstracts, 2007; pp. IN53C-08.
15. Swain, N.R. Tethys Platform: A Development and Hosting Platform for Water Resources Web Apps. 2015.
16. Crawley, S.; Ames, D.P.; Li, Z.; Tarboton, D.G. HydroShare GIS: Visualizing Spatial Data in the Cloud. *Open Water Journal* 2017, 4, 3.
17. Heard, J.; Tarboton, D.G.; Idaszak, R.; Horsburgh, J.S.; Ames, D.; Bedig, A.; Castronova, A.M.; Couch, A. An architectural overview of HydroShare, a next-generation hydrologic information system. 2014.
18. Morsy, M.M.; Goodall, J.L.; Castronova, A.M.; Dash, P.; Merwade, V.; Sadler, J.M.; Rajib, M.A.; Horsburgh, J.S.; Tarboton, D.G. Design of a metadata framework for environmental models with an example hydrologic application in HydroShare. *Environmental Modelling & Software* 2017, 93, 13-28.

19. Sadler, J.M.; Ames, D.P.; Livingston, S.J. Extending HydroShare to enable hydrologic time series data as social media. *Journal of Hydroinformatics* 2016, *18*, 198-209.
20. Horsburgh, J.S.; Aufdenkampe, A.K.; Mayorga, E.; Lehnert, K.A.; Hsu, L.; Song, L.; Jones, A.S.; Damiano, S.G.; Tarboton, D.G.; Valentine, D. Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software* 2016, *79*, 55-74.
21. Conner, L.G.; Ames, D.P.; Gill, R.A. HydroServer Lite as an open source solution for archiving and sharing environmental data for independent university labs. *Ecological Informatics* 2013, *18*, 171-177.
22. Ames, D.P.; Horsburgh, J.S.; Cao, Y.; Kadlec, J.; Whiteaker, T.; Valentine, D. HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling & Software* 2012, *37*, 146-156.
23. Horsburgh, J.S.; Tarboton, D.G.; Hooper, R.P.; Zaslavsky, I. Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling & Software* 2014, *52*, 62-73.
24. Boldrini, E.; Mazzetti, P.; Nativi, S.; Santoro, M.; Papeschi, F.; Roncella, R.; Olivieri, M.; Bordini, F.; Pecora, S. WMO Hydrological Observing System (WHOS) broker: implementation progress and outcomes. In Proceedings of the EGU General Assembly Conference Abstracts, 2020; p. 14755.
25. Hallegraeff, G.M. A review of harmful algal blooms and their apparent global increase. *Phycologia* 1993, *32*, 79-99.
26. Probe Slated in Pollution of Utah Lake. *Deseret News* July 21 1948, p. 8.
27. Merritt, L. Utah Lake, a Few Considerations. *Unpublished letter*. 3p 2004.