

WATER QUALITY DATA MANAGEMENT DATABASE

by

Rushit Hila

A project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Civil and Environmental Engineering

Brigham Young University

2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Rushit Hila

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Gustavious P. Williams

Date

M. Brett Borup

Date

E. James Nelson

Accepted for the Department

Steven E. Benzley
Department Chair

ABSTRACT

WATER QUALITY DATA MANAGEMENT DATABASE

Rushit Hila

Department of Civil and Environmental Engineering

Master of Science

One of the major tasks of State, Federal, and local water management agencies across the country is to monitor water quality in reservoirs, streams, and other water bodies. This information is used to characterize the state of the water body, assess trends, predict emerging problems, and plan ways of mitigating any potential problems. The challenge for these agencies, sometimes, is presenting information in an easily comprehensible format to policy makers or the public when seeking funding or educating the general public or students on water quality management issues. Data presentation for analysis is also important as sensors evolve and generate increasing amounts of data.

As part of the research group studying the water quality in Deer Creek reservoir we are trying to gain a better understanding of the physical, chemical, and biological properties of the lake. To be able to achieve a deeper understanding of the situation we complete regular field trips to collect water quality data. We use several sampling tools such as the HYDROLAB DS5 and YSI which helped us to measure different parameters such as temperature, depth, TDS, salinity, turbidity, pH, dissolved oxygen, conductivity, etc. These parameters are measured by sondes that are lowered through the water column, resulting in large data sets that provide vertical profiles. Traditional data management and visualization systems for surface water use point sources for data are not designed for three-dimensional spatial data with time. For appropriate data visualization and analysis, the data need to be able to be organized and viewed by time, location, elevation, or depth.

As part of my project I built a database to store collected data and also historic data available from online servers such as STORET supported by EPA. I developed tools and methods that provided a quick method to enter the data to this database. I also developed tools to quickly access the data stored in our database. These tools were designed to quickly and efficiently analyze and visualize the data.

ACKNOWLEDGMENTS

I want to acknowledge and thank my advisor Dr. Williams and the committee members Dr. Nelson and Dr. Borup for their support. I want to thank fellow students working on the Deer Creek project for their help and feedback on the database design.

I also wish to thank my wife, Janice and my parents for being a great support and motivation.

TABLE OF CONTENTS

TABLE OF CONTENTS	vii
LIST OF FIGURES	ix
1 Introduction.....	11
2 Data Collection	13
3 Storage Options.....	15
4 Database schema	17
5 Automating database population.....	21
5.1 Template Sheet	23
5.2 Measurements sheet.....	24
5.3 Sample sheet	25
6 GIS Implementation and tools.....	27
7 Conclusions.....	29
References.....	31
Appendix A. Conversion functions for excel.....	33
Appendix B	35
Appendix C Location ID's	37

LIST OF FIGURES

Figure 2-1 Monitoring points and their locations.	14
Figure 4 1. Adding new parameter to the measurements table.....	17
Figure 4-1 Access tables and relationships.....	18
Figure 4-2 Access Database tables used to store and organize data.....	19
Figure 5 1 Excel template for preparing raw data.....	22
Figure 5 2 Raw measurements sheet.....	25

1 Introduction

Storing and analyzing water quality data from modern probes can be a complicated process due to the large amount of data gathered and the ability to gather vertical profiles, which have x, y, z, and time coordinates. In addition these vertical profile in some cases need to be treated as coordinated measurements, and in other cases single values at specified depths or elevations need to be accessed for analysis. Over a period of several months our research groups has completed several trips to Deer Creek reservoir to collect data using electronic probes that generate large amounts of data, these data collections efforts are continuing. The ability to store and query these data has been the focus for the database I designed. During these trips we measured water quality parameters such as temperature, TDS, salinity, turbidity, pH, dissolved oxygen, conductivity and also geographic coordinates associated with these parameters. These readings were taken starting at surface of water and were continually recorded to the bottom of the lake creating a vertical profile. Our data were collected at several different locations through the lake. After collecting the first data sets we realized that we needed to develop a way to effectively store and retrieve the information. Also, we wanted to develop queries that would help us analyze the data accurately and to determine water quality based on our data. The tools that we have developed are intended to be useful in

other studies that have similar datasets. We found that traditional tools were not well suited for vertical profile data collected over time.

In the recent years ArcGIS is increasingly becoming the industry standard for visualizing and processing geographic data and managing geodatabases because of its flexibility and the tools it affords. It can be customized using several programming languages to take full advantage of its capabilities and streamlining the preprocessing of data for analysis with other programs. For this reason our database was designed to work with custom made ArcGIS tools that we also have developed.

The creation of these tools arose from a research being undertaken by the research group at BYU investigating how the various chemical, biological, and reservoir management contribute to the formation of large algal blooms on Deer Creek Reservoir.

2 Data Collection

The data, which are the focus of this case study, were collected at Deer Creek reservoir which is located upstream of the Provo canyon in Utah. This field study collects both point samples and vertical profile data. These samples were taken at several locations on the reservoir, and the more samples that are taken, the more accurate the analysis would be. In our study several monitoring points are located at the upstream of Deer Creek reservoir, two monitoring points are located at middle of the reservoir, one monitoring point was located on one of the tributaries, and the last monitoring point is by the reservoir outlet.

As part of this study we monitored pH, ORP, DO, NO₃, turbidity, salinity, chlorophyll, temperature and other parameters (see appendix for a complete list). Data collection was accomplished with a HYDROLAB DS5 sonde and YSI probe for vertical profiles, and various methods, including water samples returned to the laboratory, for point data. These parameters were measured at predefined points to appropriately reflect the spatial distribution within the lake. The position of the inflow into the lake changed periodically so it was critical to locate this position and measure the parameters at the inflow.

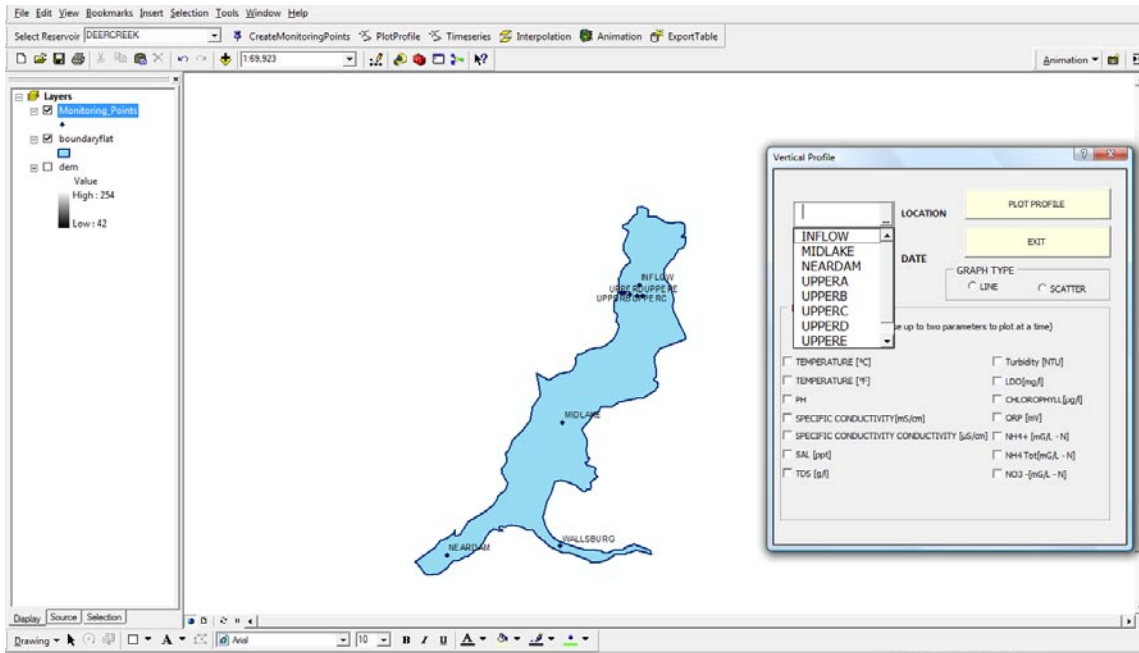


Figure 2-1 Monitoring points and their locations.

In figure 2-1 shows the reservoir and the monitoring points used in this case study. Figure 2-1 also shows one of the dialogs used inside ArcGIS to access the data base I developed as part of this case study. These tools can also be used for river or other data sets that have a spatial component as well.

3 Storage Options

The spatial, both horizontal and vertical, and temporal aspects of our data required developing an appropriate database structure. We needed to be able to store and query data by location, time, depth, or elevation. Especially difficult was the ability to store and retrieve vertical profiles as associated data, in addition to individual measurement points. Because we have a large dataset that is continually updated, we needed to store these data in one main location to manage the data efficiently. Our data also needed to be organized so we could quickly analyze it by different methods such as profiles, depth, elevation, or time. Our data could have been stored in several structures but we designed our storage structure in a way to be efficient for data entry, management, and export. We also made the database structure compatible with ArcGIS.

ArcGIS has the ability to connect to three different types of databases; the ArcSDE geodatabase, a file geodatabase, and the personal geodatabase. In single project or area cases similar to ours where data sets are collected data are uploaded a few times per week and there are not a large number of queries, a personal geodatabase is efficient and simple to implement. ArcGIS personal databases can be created and managed by Microsoft Access. Using Access for database management has some advantages. Access allows us to easily manage the database on a Windows system. It also allows the use of Excel for a data entry, management, and visualization interface. The disadvantage is that

once the file reaches a two GB size the performance of the database starts to decrease but in our case is not a concern, since we would not be running a large number of queries. Even though our data sets are large, we do not expect them to reach this size. In addition, most of our analysis time is spent in ArcGIS so slightly slow queries are not a major concern, even if a query took a few 10s of seconds, this would not significantly affect the use of these tools. In cases when much larger data needs to be managed a better option would be to use the ArcSDE geodatabase which can be managed from an Oracle, Microsoft SQL server, or IDM DB2.

4 Database schema

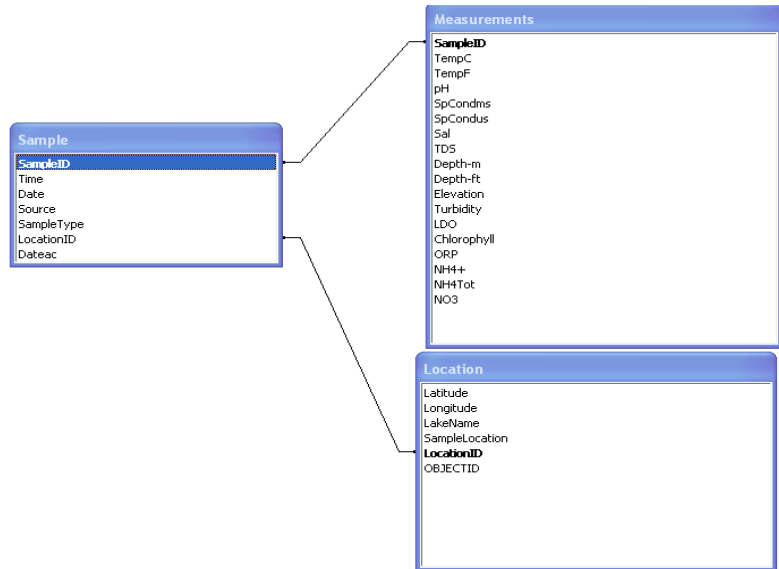
Our data was stored in an Access database and organized into a schema with three main tables. As shown in Figure 4.1 the tables contained information for the locations, measurements, and samples. The sample table includes information about the time, date, source of the sample. The location table includes information about the latitude, longitude, lake name, and sample location. This not only allows the user to better manage as many sample points

Field Name	Data Type	
SampleID	AutoNumber	
TempC	Number	Temperature of sample in Fahrenheit
TempF	Number	Temperature of sample in Celsius
pH	Number	pH of water
SpCondms	Number	Units in mS/cm
SpConduS	Number	Units in µS/cm
Sal	Number	Units in ppt
TDS	Number	g/l
Depth-m	Number	Units in meters
Depth-ft	Number	Units in feet
Elevation	Number	Units feet
Turbidity	Number	Unit in NTU
LDO	Number	Units in mg/l
Chlorophyll	Number	Units in µg/l
ORP	Number	mV
NH4+	Number	mg/L -N
NH4Tot	Number	mg/L -N
NO3	Number	mg/L -N

Figure 4 1. Adding new parameter to the measurements table.

as needed, but the database can also store information for other lakes or rivers as well. The measurements table stores the specific sample parameter readings such as temperature, pH, conductivity, salinity, TDS, depth, turbidity, chlorophyll, ORP , and NH4+. Other measurements or parameters can be easily added to this table. This is done by adding a new field to the Measurements table in the Access database and adding a corresponding entry in the Excel spreadsheet that is used to enter data into the Access database. The Excel spreadsheet has a Measurements sheet; to add the data to the Access database you must use the same name you used in the Excel spreadsheet. As shown in Figure 4.1, the new parameter names are entered into

the first row column R, this is only done when a new parameter is added to the database and spreadsheet. The parameter list in the Excel data importation spreadsheet must match the table structure of the Access data base.



However, for most field sampling

studies, these lists will be **Figure 4-1 Access tables and relationships**

relatively constant, typically only changing near the start of a project.

The schema and relationships for the Access database is a relatively simple design. This simple design supports the field sampling requirements, makes additions and changes relatively quick, and allows ArcGIS tools to be easily connected to the database. The three tables are related to each other using the SampleID so that data are not repetitive. The database schema is shown in the figure 4.2.

Our database has three tables which contain information for Measurements, Location, and Sample information. These tables contain the majority of the data collected in the field. The database includes several other tables as shown in figure 4.3. These tables are designed to make the database functional with the ArcGIS tools that we have developed. These tables are part of the database schema but do not require the user to update or edit them when more sampling points or parameters are added. The database is designed to store the data collected in the field; provide a way to query the data by location, depth, elevation, parameter, or time; and integrate easily with the ArcGIS tools that were developed. Many of the structures in the database are

specifically designed to make interaction with ArcGIS possible. These portions of the database should not be changed unless the user has a specific reason for doing so.

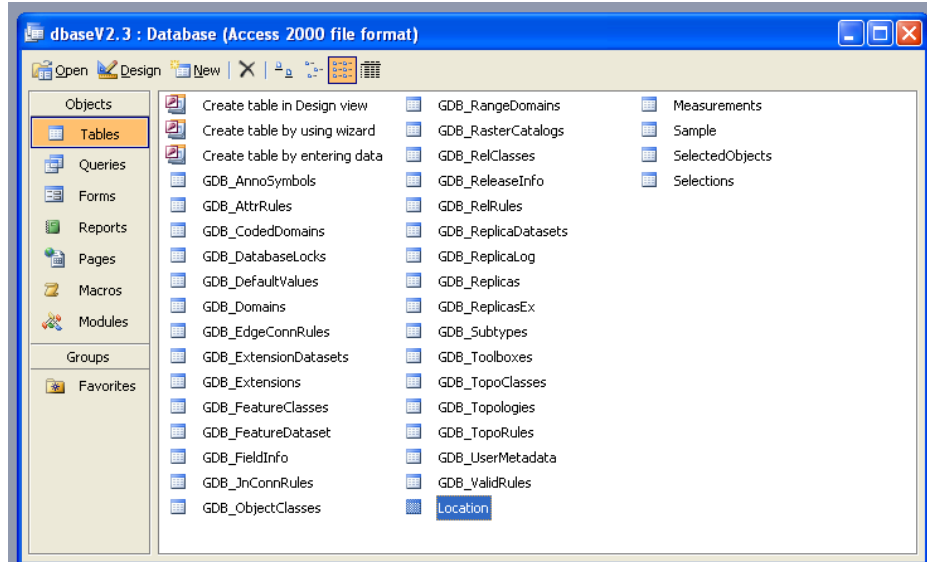


Figure 4-2 Access Database tables used to store and organize data.

5 Automating database population

Often the data that are recorded and retrieved from the monitoring tools are obtained formatted as a .txt or as an Excel file. These data are human-readable, but in this form are not well suited for analysis and visualization. I developed an Excel spreadsheet which is used to populate data for the Access database. This spreadsheet was designed to be easy to use, after the data are collected; the resulting data is opened using a generic Excel spreadsheet. The data are then selected and copied to the Excel Database Input spreadsheet by column. This spreadsheet is organized so that the data are automatically imported into the Access database correctly. This file is named ImportTemplateVersionxx.exl. Figure 5.1 shows how the data are organized in this spreadsheet. The first row contains the headers describing each reading that has been taken and the rows below contain the measurements.

The ImportTemplateVersionxx.exl provides a tool to implement an automated process to both check the data for consistency and make the importing of field data efficient. As described above, the data are copied into the preformatted Excel file where the data are rearranged and formatted for the Access database. Once the data are formatted, they can be easily checked for quality control. This is not an in-depth review, but just a quick scan for outliers or for data that are not consistent with expectations, such as data that are out of range. For example, during one data collection event, the DO values were all recorded as zero due to a mechanical error, this error was identified during data importation. Once the user is satisfied, the user can export the

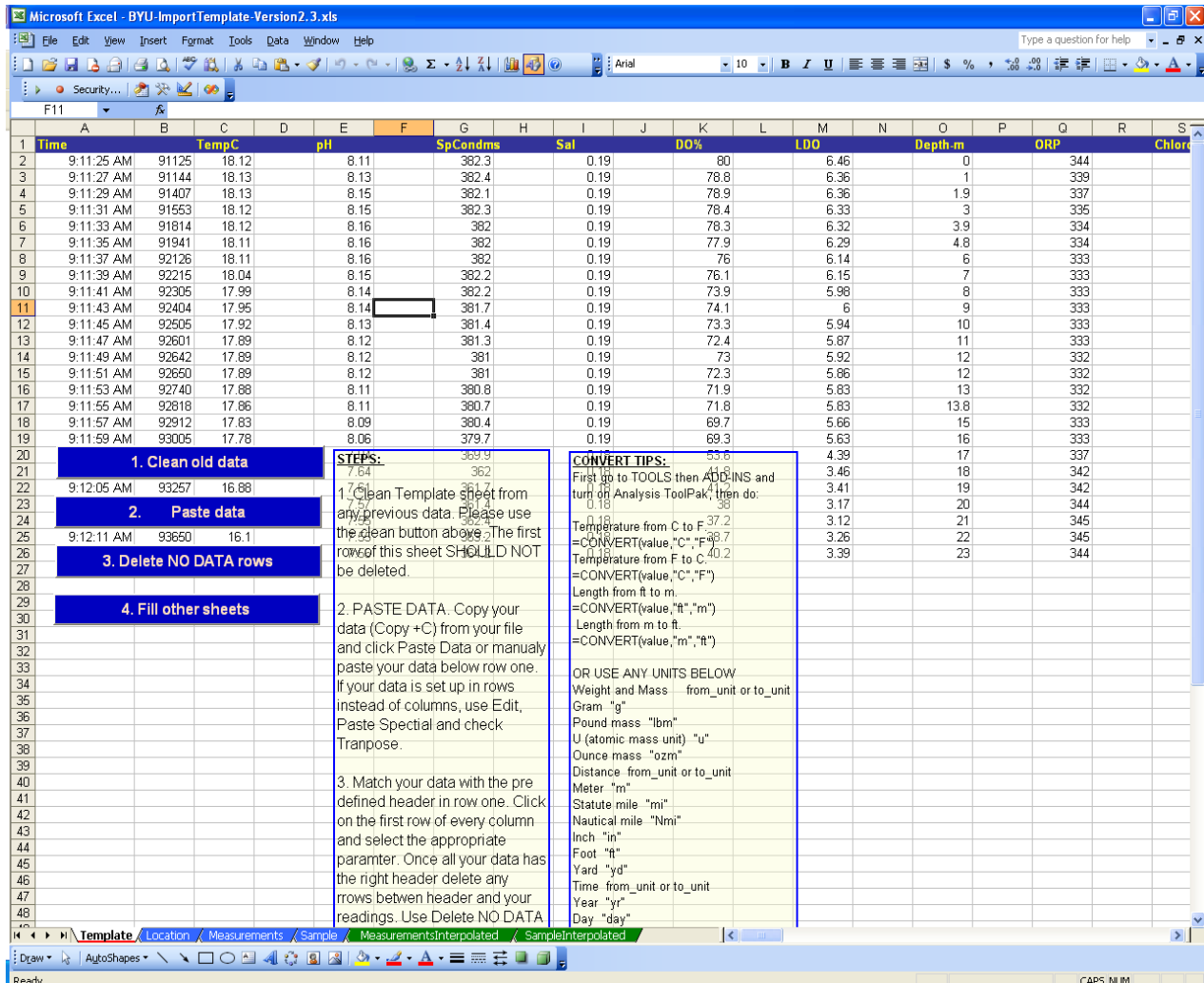


Figure 5 1 Excel template for preparing raw data.

data to the Access database. The spreadsheet and database are designed so that the user does not need to have any advanced knowledge of Access. VBA code within the Excel spreadsheet is used to link the various tables. Exporting data to the Access database is done by running the code which organizes and formats the data within spreadsheet, then exports these data to the appropriate Access tables.

Figure 5.2 shows the Excel template used to export the field data into the Access database. Collected data are passed into the other sheets within the Excel spreadsheet to organize and format the data for the Access database.

The ImportTemplate.xls contains six separate sheets. These sheets are explained in the sections below. Except for when the user adds a new parameter to the database and import spreadsheet, only the first sheet is edited. This is the sheet that receives the raw data.

5.1 Template Sheet

The template sheet is where the user copies the data from the field collections or probes. The ImportTemplate first converts the depth readings to elevations and also creates a data set with values interpolated to a regular spacing. Both the interpolated and measured data are imported into the Access database. These processed data are used to generate the other sheets.

As noted above, except when adding a new parameter or location, this is the only sheet the user edits. To use the ImportTemplate and place data into this sheet is relatively simple as explained in the following instructions:

1. **CLEAN DATA.** The first step is to Clean the Template sheet to remove previous data. DO NOT delete the previous data, instead use the “clean” button at the top of the sheet. Using this function retains the sheet programming used to convert and format the data. The first row of this sheet SHOULD NOT be deleted.
2. **PASTE DATA.** Copy your data (Copy +C) from your file and click Paste Data or manually paste your data below row one. If your data is in rows instead of columns, use Edit, Paste Special and check Transpose.
3. **DEFINE DATA.** Define your data using the selection boxes in row 1. To do this, click on the first row of each column and select the appropriate parameter from the predefined list. This makes sure that the spreadsheet data matches the structure and definitions in the database. Once all your data has the correct header title, delete any

rows between header and your readings. To do this, use the “Delete NO DATA rows” button. This allows the data columns that are imported to be in any order.

4. CHECK DATA. Any columns that do not have a header selected will not be imported. At this point the user should make sure the data are in the correct units. There is a box titled “conversion tips” to that provide guidance on using built-in Excel functions to calculate most standard conversions.
5. PROCESS DATA. Once the data are clean, open the “Measurements Interpolated” sheet. The user should follow instructions on the sheet to interpolate data to regular elevations for the data base. The data base will include the uninterpolated data also. Interpolating to regular elevations, for example every meter, allows data to be plotted and compared at a single elevation over time.
6. IMPORT DATA. Use this buttons to import each sheets into the Access database.

5.2 Measurements sheet

The "measurement sheet" is used to import the water quality parameters from the raw data pasted into the first sheet. The first row of this sheet should not be changed unless the Access database schema is also modified to reflect the changes. This would be done if you added a new parameter to the database. After entering the data, the user should check the content of this sheet and make sure the data are identified with the correct header. In case no data are displayed, go back to the "Template" sheet and verify step four was completed correctly.

1	TempC	TempF	pH	SpCondms	SpConduS	Sal	TDS	Depth-m	Depth-ft	Elevation	Turbidity	LDO	Chlorophyll	ORP
2	16		7.56	364.2		0.18	0.2331	23		5289.652	13.7	3.39	0	0
3	16.1		7.55	363.2		0.18	0.2325	22		5292.933	0	3.26	0	0
4	16.23		7.55	362.4		0.18	0.232	21		5296.214	0	3.12	0	0
5	16.6		7.57	361.4		0.18	0.2313	20		5299.495	0	3.17	0	0
6	16.88		7.61	361.7		0.18	0.2319	19		5302.776	0	3.41	0	0
7	16.94		7.64	362		0.18	0.2317	18					0	0
8	17.41		7.84	369.9		0.18	0.2371	17					0	0
9	17.78		8.06	379.7		0.19	0.243	16					0	0
10	17.83		8.09	380.4		0.19	0.2435	15		5315.899	1.3	5.66	0	0
11	17.86		8.11	380.7		0.19	0.2436	13.8					0	0
12	17.88		8.11	380.8		0.19	0.2437	13					0	0
13	17.89		8.12	381		0.19	0.2438	12					0	0
14	17.89		8.12	381		0.19	0.2442	12					0	0
15	17.89		8.12	381.3		0.19	0.244	11					0	0
16	17.92		8.13	381.4		0.19	0.2441	10					0	0
17	17.95		8.14	381.7		0.19	0.2443	9					0	0
18	17.99		8.14	382.2		0.19	0.2446	8					0	0
19	18.04		8.15	382.2		0.19	0.2449	7					0	0
20	18.11		8.16	382		0.19	0.2444	6					0	0
21	18.11		8.16	382		0.19	0.2445	4.8					0	0
22	18.12		8.16	382		0.19	0.2445	3.9		5352.316	0	6.32	0	0
23	18.12		8.15	382.3		0.19	0.2447	3		5355.269	0	6.33	0	0
24	18.13		8.15	382.1		0.19	0.2446	1.9		5358.878	0	6.36	0	0
25	18.13		8.13	382.4		0.19	0.2447	1		5361.831	0	6.36	0	0
26	18.12		8.11	382.3		0.19	0.2447	0		5365.112	1.99	6.46	0	0
27														

Import Measurements to testbase

IMPORTANT:
PLEASE DON'T MESS WITH THE CODE .
UNLESS YOU KNOW WHAT YOU ARE
DOING.
IF YOU HAVE ANY QUESTIONS CONTACT:
Rushit or Caleb
tepelena@gmail.com
calebgh@gmail.com

Figure 5 2 Raw measurements sheet

Some columns may not contain any data. If a column does not contain data it is because that parameter was not selected on the raw data column headings – it was not collected. Some of these columns can be manually filled, but the columns do not need to contain data. For example if depth is only recorded in meters you can convert it to feet. The Appendix contains descriptions of quick conversion functions that can be used to assist in this process if required.

5.3 Sample sheet

Sample sheet contained in the spreadsheet is necessary for the database to function properly. It includes information on the time, data, and source of data, sample type, and location ID. The “Location ID” is an important feature of this table. Every sample location is identified by a unique number that identifies that location. This information is stored in the Location table in the Access database and included in the Excel Sample sheet for reference. Additional sample locations must be added to the Locations table in the Access database. Appendix C contains the

location IDs that we have been using in our field case study. When a new location point is added to the database the next available number is automatically assigned as the new Location ID for that location. To verify the available sample locations and their ID numbers check the Location table in the Access database.

6 GIS Implementation and tools

We implemented connections between the database and the ArcInfo GIS program. Within the GIS system we developed special tools to select, query, plot, and analyze these data. GIS systems provide tools that make management and visualizing geographic data less time consuming. Extending these capabilities to work on specific data types further simplifies using these tools. Caleb helped in the design of custom tools for water quality monitoring for ArcGIS which were primarily developed with Visual Basic and compiled into .dll files which could then be easily distributed. While the ArcInfo tools are an integral part of this effort, they were not part of my project and are not reported here.

7 Conclusions

Excel and Access software were used to develop both the storage and analysis database and the data import tools for this project. These programs are widely available and familiar. I wrote VBA code to identify data outliers and format data automatically, automating many of the steps required to clean and verify data before importing it into the database. The database was designed to manage large amount of spatial data easily. These data differ from standard data because they have horizontal and vertical spatial coordinates and a temporal coordinate also. The Excel spreadsheet was designed and programmed to automate the cleaning, verifying and importing of collected data less time consuming and more accurate. The Access database and Excel template can be easily edited by someone who is familiar with Access and Excel and can write VBA code. The database was designed specifically for Deer Creek reservoir but the schema can be copied easily and be used in other studies as well. The custom made ArcGIS toolbar will be able to analyze new reservoirs without any modification other than the data.

The database is a powerful tool which when combined with the custom ArcGIS toolbar can be very useful to anyone who would like to analyze these data. Even though information can manually be extracted from the database by running simple queries, the custom ArcGIS tools allows us to create and graph vertical profile, plot time series, analyze, compare and interpolated collected data easily.

References

Bates, P. D., Stewart, M. D., Desitter, A., Anderson, M. G., Renaud, J. P., and Smith, J. A. “*Numerical simulation of floodplain hydrology.*” *Water Resources Research*, 2000

Bedient, P. B. and Huber, W. C. “*Hydrology and Floodplain Analysis*” 1988

David R. Maidment “*Arc Hydro: GIS for Water Resource*”s 2002

Appendix A. Conversion functions for excel

First go to TOOLS then ADD-INS and turn on Analysis ToolPak, then do:

Temperature from C to F. =CONVERT(value,"C","F")	Time from_unit or to_unit
Temperature from F to C. =CONVERT(value,"C","F")	Year "yr"
Length from ft to m. =CONVERT(value,"ft","m")	Day "day"
Length from m to ft. =CONVERT(value,"m","ft")	Hour "hr"
OR USE ANY UNITS BELOW	Minute "mn"
Weight and Mass from_unit or to_unit	Second "sec"
Gram "g"	Temperature from_unit or to_unit
Pound mass "lbm"	Degree Celsius "C"
U (atomic mass unit) "u"	Degree Fahrenheit "F"
Ounce mass "ozm"	Liquid Measure from_unit or to_unit
Distance from_unit or to_unit	Teaspoon "tsp"
Meter "m"	Tablespoon "tbs"
Inch "in"	Fluid ounce "oz"
Foot "ft"	Cup "cup"
Yard "yd"	U.S. pint "pt"
	Quart "qt"
	Gallon "gal"
	Liter "l"

Appendix B

Parameters stored in our database. More can be added.

TempC
TempF
pH
SpCondms
TDS
Depth-m
Depth-ft
Elevation
Turbidity
LDO
Chlorophyll
ORP
NH4+
NH4Tot
NO3
Time
Date
Source
SampleType
LocationID
Dateac
Latitude
Longitude
LakeName
SampleLocation
LocationID
Sal
SpCondu

Appendix C Location ID's

Latitude Longitude LakeName SampleLocation LocationID OBJECTID

40.46710312 -111.48183474 DEERCREEK INFLOW 1

40.4653505555556 -111.485621388889 DEERCREEK UPPERA 2

40.4654541666667 -111.484892777778 DEERCREEK UPPERB 3

40.4650238888889 -111.483845 DEERCREEK UPPERC 4

40.4647252777778 -111.482335277778 DEERCREEK UPPERD 5

40.4647127777778 -111.481190833333 DEERCREEK UPPERE 6

40.4380330555556 -111.498018888889 DEERCREEK MIDLAKE 7

40.4118133333333 -111.498406666667 DEERCREEK WALLSBURG 8

40.4098975 -111.5221375 DEERCREEK NEARDAM 9