Lab Report – Final Project

Title: Modelling Large Lake Responses to Climate Change
Notice: Dr. Bryan Runck
Author: Mary Heise
Date: May 11, 2021

Project Repository: https://github.com/Tuletara/GIS5572.git

Abstract

Climate change has unseen impacts on lakes that must be better understood in order to manage aquatic resources as well as predict overall climate consequences. The logistics of collecting in-situ data for variables well below the surface limits the temporal and physical range of data. Surface data is fairly abundant and useful but cannot be interpolated to show truly dynamic large lake processes. The quantity of data available offers challenges for harvest and integration into one singularly conclusive model that is best handled through ETLs. Point data can then be used to create an interpolation surface of water temperature throughout the lake basin. While this study is concerned with temperature data, there are many more factors to consider to fully capture the complex relationship of lake dynamics. Kriging was used here to take advantage of the tool's 3 dimensional aspect, but the results may have low accuracy due to the skewed nature of this data. Further statistical manipulation and a broader range of data could improve the predicted surface. 2 dimensional static outputs of lake temperatures were also used to get a snapshot in time and location.

Problem Statement

Climate change is affecting one of the largest and most important resources on Earth – water. The nature and implications of these changes must be understood in order to mitigate disastrous consequences that could leave millions of people without safe drinking water and harm entire ecosystems. Water has also been a stabilizing force in the face of climate change as it moderates the impacts of extreme shifts in temperature and precipitation, allowing for a buffer of land temperature and water storage. This has protected land dwellers from the brunt of climate change, but what level of these moderation services remain in the planet's oceans, lakes, and streams? What are the largely unseen impacts to aquatic ecosystems as the water continually warms?

The long term goal of this study aims to assess the current conditions of the world's largest surface area freshwater lake as well as predict future impacts of global warming. Lake Superior is situated in North America and is one of five in the chain of Great Lakes, Figure 1. The challenges of assessing and predicting these changes stem from the difficulties of collecting reliable and consistent data over large scale variables and complex inter-relationships. Table 1 lists many of the necessary data variables needed to create a fully comprehensive model. Historical and contemporary data necessary for stochastic modelling is extremely inconsistent in accessibility, format, quality, spatial and temporal range. That being said, much data is available through government and research agencies but it's largely restricted to surface measurements.

Given the complexity and scope of this analysis, this study has been separated into phases. This report specifically addresses phase one: data collection and a preliminary three-dimensional prediction surface using a subset of the data. The challenges of phase one include: 1) understanding the contents and quality of available data, 2) accessing the numerous and voluminous datasets, 3) standardizing data, and 4) determining the best processing parameters for interpolation. There are a multitude of agencies offering historic and real time water quality measurements with different options for viewing or downloading. Metadata does not always accurately indicate what information is within the dataset, often requiring an initial download before discovering if it's useful or not. This demands dedicated time to accessing, sampling, and processing test data. Therefore, an Extract-Transform-Load (ETL) protocol is the most efficient way to collect lake variable data.

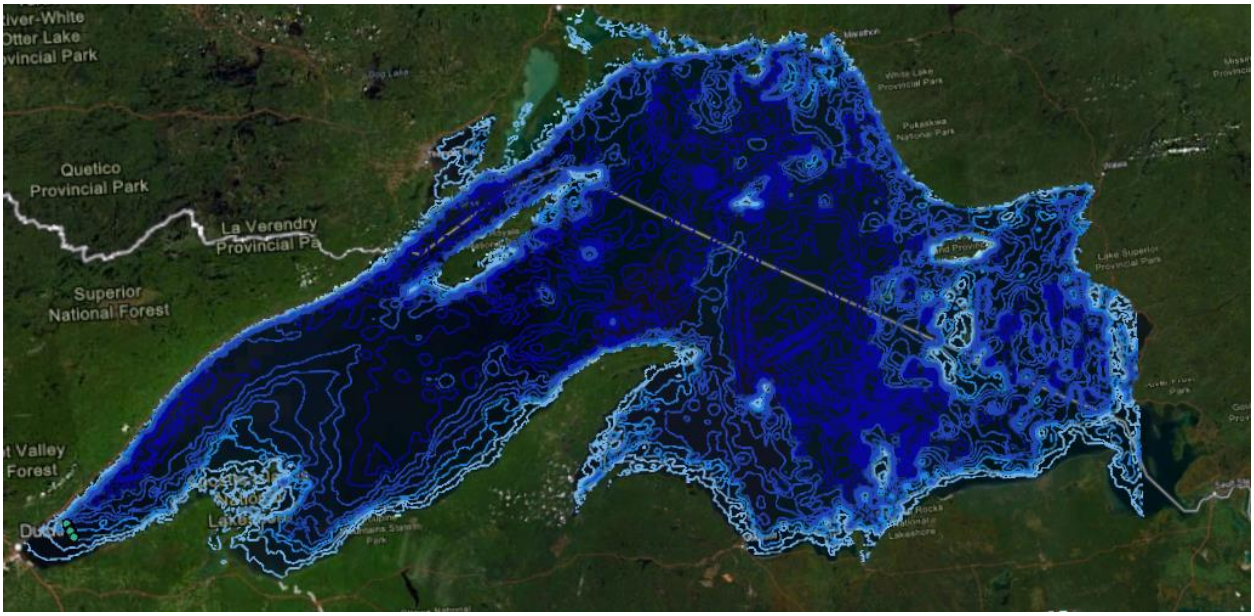

Figure 1. Bathymetric representation of Lake Superior, the western most lake of the North American Great Lakes.

Table 1. Data inputs required for the most comprehensive model. This report accesses the top 4 variables.

#	Requirement	Defined As	Spatial Data	Attribute Data	Dataset	Preparation
1	Sea Surface Temperature	Heat	Lake Superior	Temp, time, location	NOAA COBE SST	Clean, average values, transform into raster
2	Column Temp	Heat & distance	Lake Superior	Temp, depth, time, location	LLO GLOS Glider	Clean, transform & stack raster
3	Weather Variables	Air temps, precip, winds	Lake Superior	Precip, wind speed & direction, temp	NOAA	Clean, consolidate
4	Bathymetry	Basin topography	Lake Superior	DEM	NOAA	None
5	Lake Currents	Flow rates & direction/depth	Lake Superior	Distance, direction, flow rate	GLERL	-
6	Historical Land Weather	Weather	Lake Superior	Temp, precip, winds	NWS	-
7	Ice Cover	Dates of ice cover	Lake Superior	Dates, percent cover, thickness	NOAA NCEI	-
8	Stream Inputs	Stream water contributions	Lake Superior	Flow rates, temp	Stream Watch	-

Input Data

The National Oceanic and Atmospheric Administration (NOAA) agency provides different portals for various datasets and sources. The Glider data originated from the Large Lakes Observatory underwater remote collection drone which has different datasets for each collection mission that occurred during the summers between 2012 and 2018. The Gliders sampled several water qualities but the data pulled for the initial model was reduced to temperature, column depth, current speed and direction. The data was pulled from NOAA's gridded (grddap) ERDDAP API as multiple netCDF files. Likewise, ERDDAP's tabledap portal provides in situ surface temperature data collected by the Great Lakes Observing System (GLOS) Buoy program. This includes datasets for 13 buoys which lists hourly air temperature and wind readings. NOAA also facilitates an FTP for thousands of datasets collected by the COBE satellite and its derived sea surface temperature (SST) data. The specific subsets downloaded for this study include monthly mean temperatures for 2006 – 2009. Table 2 refers to the actual data used for this phase of the study.

Table 2. Data pulled with ETL for phase one.

#	Title	Purpose in Analysis	Temporal	Format	Portal	Link to Source
1	LLO Glider	Depth temps, current inputs	2012-2018, Trajectory	esricsv, .nc3	IOOS/ERDDAP	LLO Glider
2	COBE SST	Monthly mean satellite derived sst	2006 - 2009	.nc	NOAA FTP	COBE SST
3	GLOS Buoy	Surface air temp, winds	2015 + Seasonal - hourly	csv	GLOS	GLOS Buoy

Methods

An ETL was designed to leverage ERDDAP's API with Python requests and erddapy modules; as well as to perform some basic standardization. Null values were removed using Pandas. Since each source varied widely in format and quality, even within the same datasets, manual clean-up was performed directly on the CSV files to match columns headers.

Once consolidated, the glider data consisted of more than 450,000 sample points. This led to excessive draw requests, run-time errors, and crashes. To work around this issue, a small subset of data was used to create a preliminary prediction surface using 3D Empirical Bayesian Kriging (EBK). EBK was processed in both ArcGIS Pro and arcpy. The first step requires converting the csv table into points with the **XY Table To Point** tool and then projected into NAD 1983 UTM Zone 15 coordinates with the **Project** tool. In order to work with the points in 3D, they were brought into a **New Local Scene** within ArcGIS Pro - from the Insert tab select New Map in the Project group. To make the densely packed vertical layers more accessible, a vertical exaggeration of 10 was applied to the point layer by right clicking the points layer in the Contents pane and changing the exaggeration level under the Elevation group. This spreads out the vertical distribution making each point more visible and clickable to view attribute information with the **Identify** tool.

A look at the layer statistics is helpful to better understand the data. From a right click on the layer in the Contents pane, selecting Create Chart and then Histogram enables the **Chart Properties** tool. The Statistics breakdown for Temperature of this subset of 5,598 points shows a mean value of 6.37 degrees Celsius with min and max values of 3.82 and 9.63. As seen in Figure 2, the distribution is skewed regardless of applying either a Logarithmic or Square Root transformation, available in the chart properties dialogue box. This is problematic as all Kriging methods assume a normal distribution. However, EBK is the only option to interpolate three dimensions so this was still used as a test case (esri, "Interpolate Shape Documentation").

Interpolation surfaces can be created using the Pro GUI Geoprocessing tool or using the Geostatistical Wizard found under the Analysis tab in the Workflow group. ArcPy is another option and the specific code used here is available in the GitHub repo. Two EBK surfaces were generated using an Exponential Semivariogram Model Type, selected by expanding the Advanced Model Parameters section in Pro, one with no transformation and the other with a Log Empirical transformation. A third surface was created as a Power Semivariogram Model Type without any transformation. A shapefile of a static representation of a given depth temperature profile was captured as contour lines and filled polygons using the **GA Layer to Contour** tool; accessible in Pro by right clicking the EBK layer, selecting Export Layer, and then To Contours. The code is also available in GitHub within the Lake_Superior_GIS notebook.

Cross Validation was used to gauge the accuracy of each EBK iteration; accessed by right clicking on the model in the Contents pane and selecting Cross Validation from the menu. This method removes a sample point from the analysis and compares the actual values for that point to the previously predicted value. This is further discussed in the results validation section below.

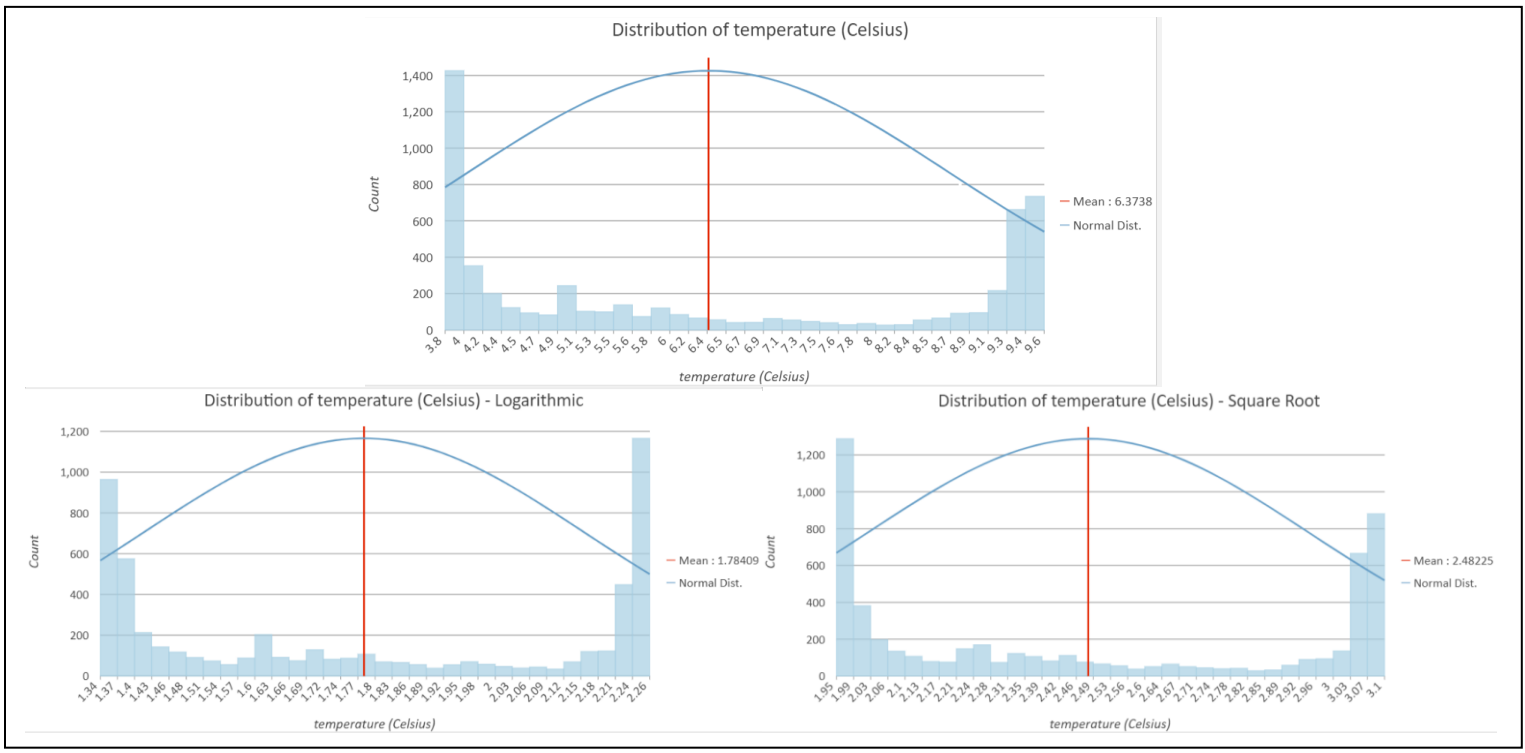

Figure 2. Comparison of distribution curves with no transformation (top), Logarithmic (bottom left) and Square Root (bottom right) transformations. Neither manipulation achieved a normal distribution necessary for a Kriging Interpolation method.

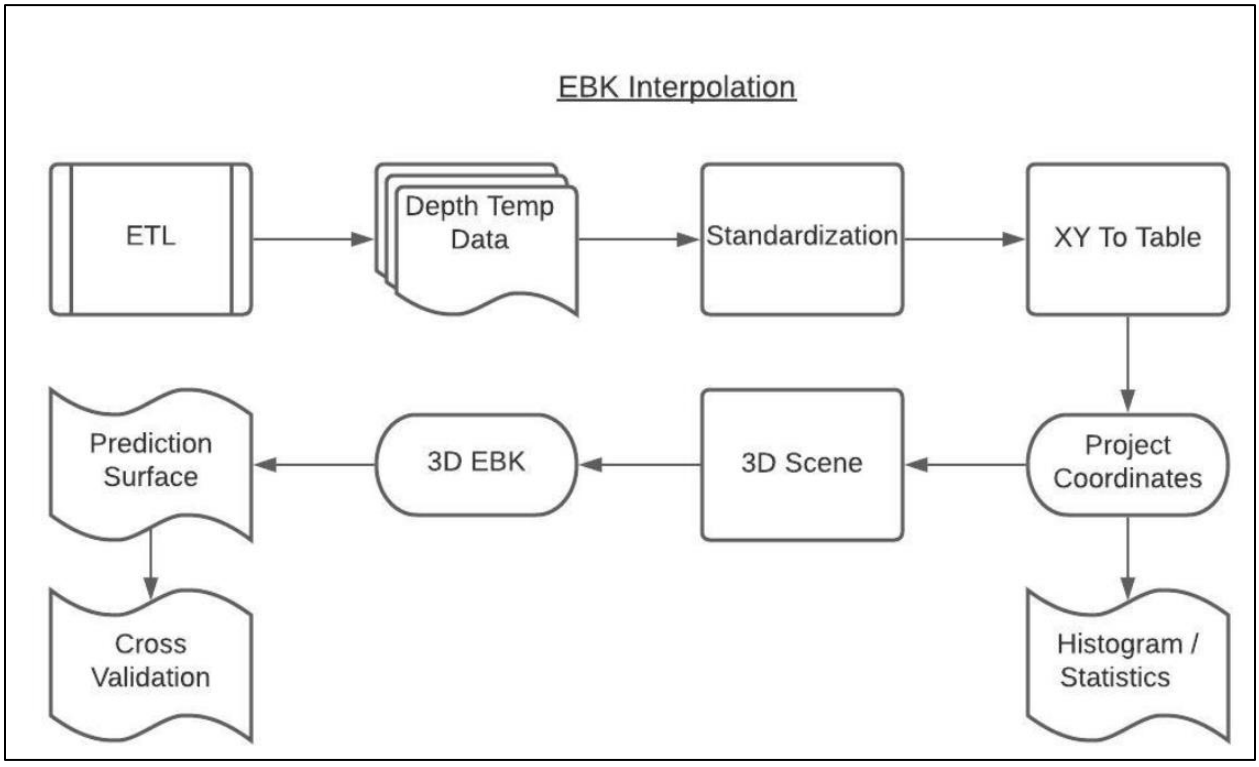

Figure 3. Workflow used to generate interpolation surfaces for column temperatures of lake water.

Results

Figure 4 shows the 3D points generated from the csv subset data with appropriate symbology applied. An animated timeseries of the interpolated surface was successfully created from these points that visualizes each depth layer in 3D shown in Figure 5. However, given the numerous vertical points even in the small subset test data, the timeseries has several steps to iterate through to show all elevation transects. This takes considerable time and computation to iterate through each surface layer. While the settings can be changed under the Range tab in Pro to speed up the animation or group multiple intervals together, this diminishes the visibility of the results. It would be worth finding an acceptable averaging range to summarize both temperature and depth in order to compress and simplify the stack of interpolation layers. This warrants further analysis, especially when applying manipulations over multiple years and datasets.

A few static layers showcasing depth-specific temperature conditions derived from the EBK surface were created and exported as shapefiles. Figure 6 shows contour lines and filled polygons for depths of 2.17 and 70.27 meters respectively. There are other export options available, such as **To Multidimensional Raster**, but these would be best pursued only after an accurate model surface is achieved.

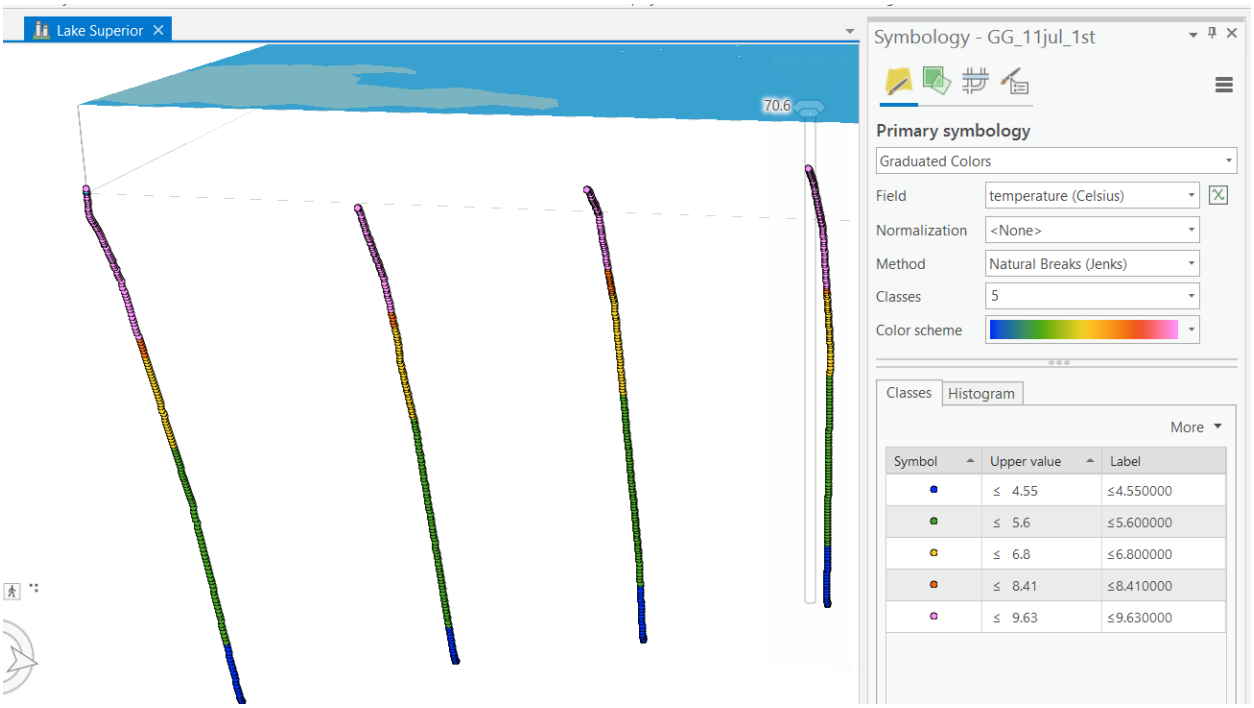

Figure 4. A visual of some of the subset temperature and depth data points brought into a 3D scene symbolized to show temperature changes with depth.

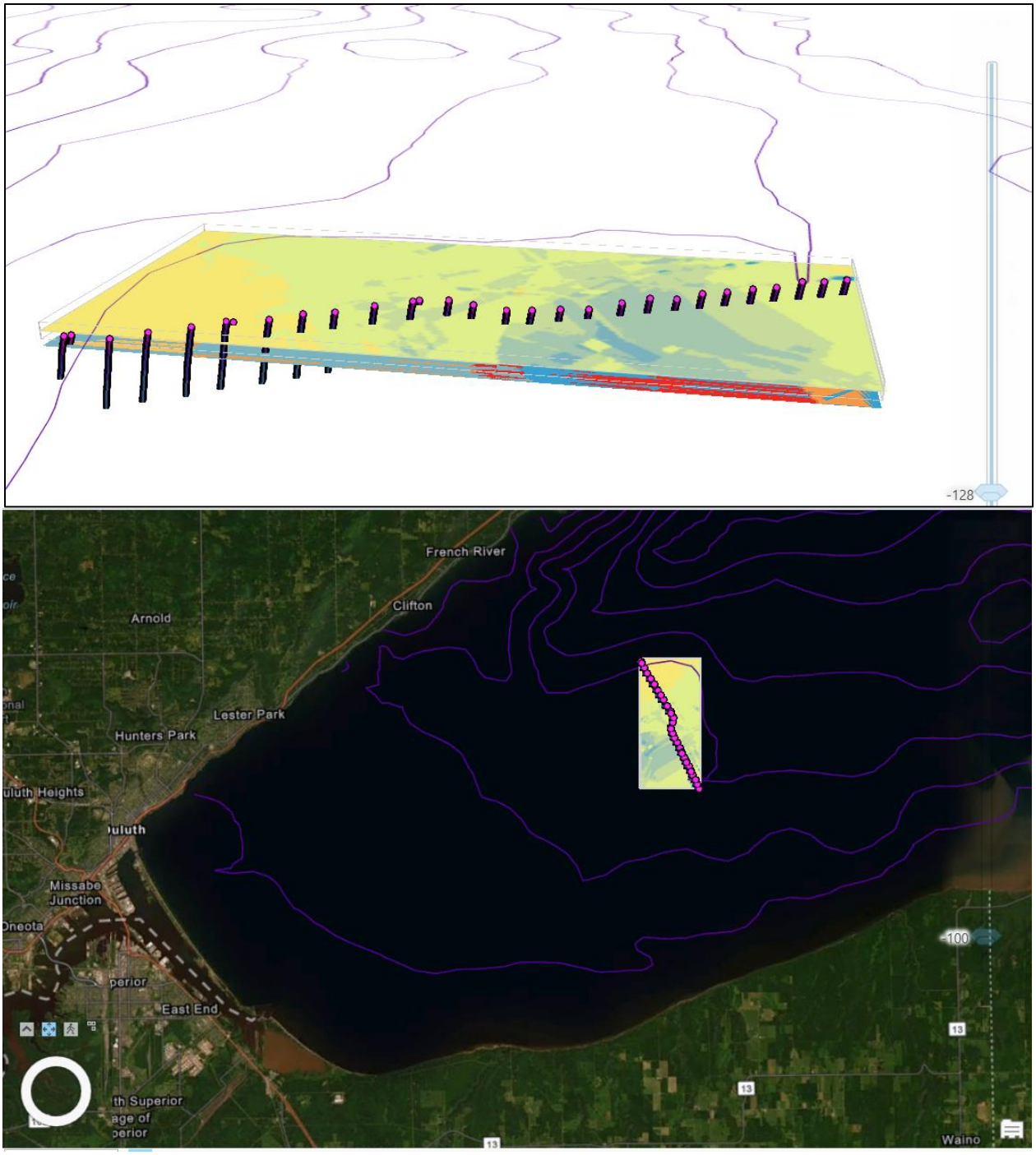

Figure 5. An angled view of the depth intervals to showcase the dimensionality (top) and a more basic overview of the prediction surface (bottom).

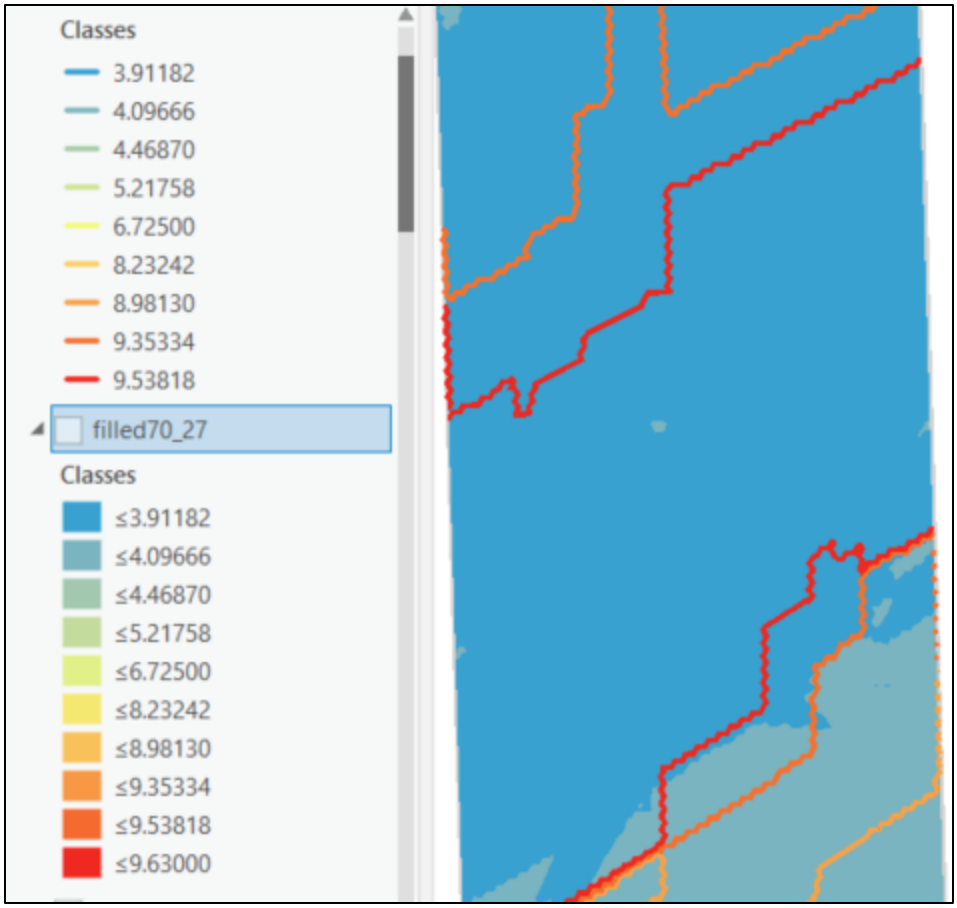

Figure 6. GA Layer To Contour of lines at a depth of 2.17 meters and filled polygons (the base surface) at 70.27 meters.

Results Verification

Cross Validation for the three iterations is shown below. The Mean and Mean Standard Errors indicates whether the model tends to predict values that are beyond the normal range of the data. Values close to zero signify less biased results. The Root-Mean-Square (RMS) measures the average level of departure from actual values. The target value for the Average Standard Error is one that closely matches the RMS value while an RMS Standardized value as close to one as possible indicates high accuracy. The 90 and 95 Percent Intervals criteria pertain to confidence intervals, so the nearer the values are to 90 and 95 indicates a more consistent prediction. The average Continuous Ranked Probability Score (CRPS) considers the deviation between each predicted cumulative distribution function to the actual data. A smaller value indicates a more accurate result (esri, “Cross Validation Documentation”).

As seen in Table 3, none of the three predictions achieved very good results according to the cross validation process. This can largely be attributed to the skewed distribution of the data points. As previously mentioned, this is not the most suitable modelling method for this scenario.

A scatter plot of all subset points shown in Figure 7, indicates a general relationship of cooler temperatures at increased depths. This relationship allows for a rough comparison of the interpolated surfaces and overall temperature. The deeper layers should predict cooler temps than those found in shallow waters and this is generally true of all three EBK surfaces. It is important to consider how lake currents, winds, basin depth, and stream or industrial inputs will alter micro zones throughout the lake. These factors must be incorporated in future phase studies to get a more accurate representation.

Table 3. Comparison of three iterations of EBK methods and a ranking of overall accuracy. Target values are specified in the grey header boxes with 0 indicating as-low-as-possible rather than exactly 0. The best results for each iteration are in bold.

Parameters	**Mean/Mean Standard Errors**	0	**RMS / RMS Standardized / Ave Standard Error**	0 / 1 / =RMS	**90/95 Percent Interval**	90/95	**CRPS**	0	**Rank**
Power	**-0/0.00014**		0.022 / **0.64** / 0.038		**95.89/97.62**		0.0088		1
Log Empirical Transformation	-0/-0.0033		0.022 / 0.567 / 0.05		97.12/98.52		**0.0093**		3
Empirical - No Transformation	-0/0.00132		**0.022** / 0.626 / **0.0377**		96.03/97.67		0.0088		2

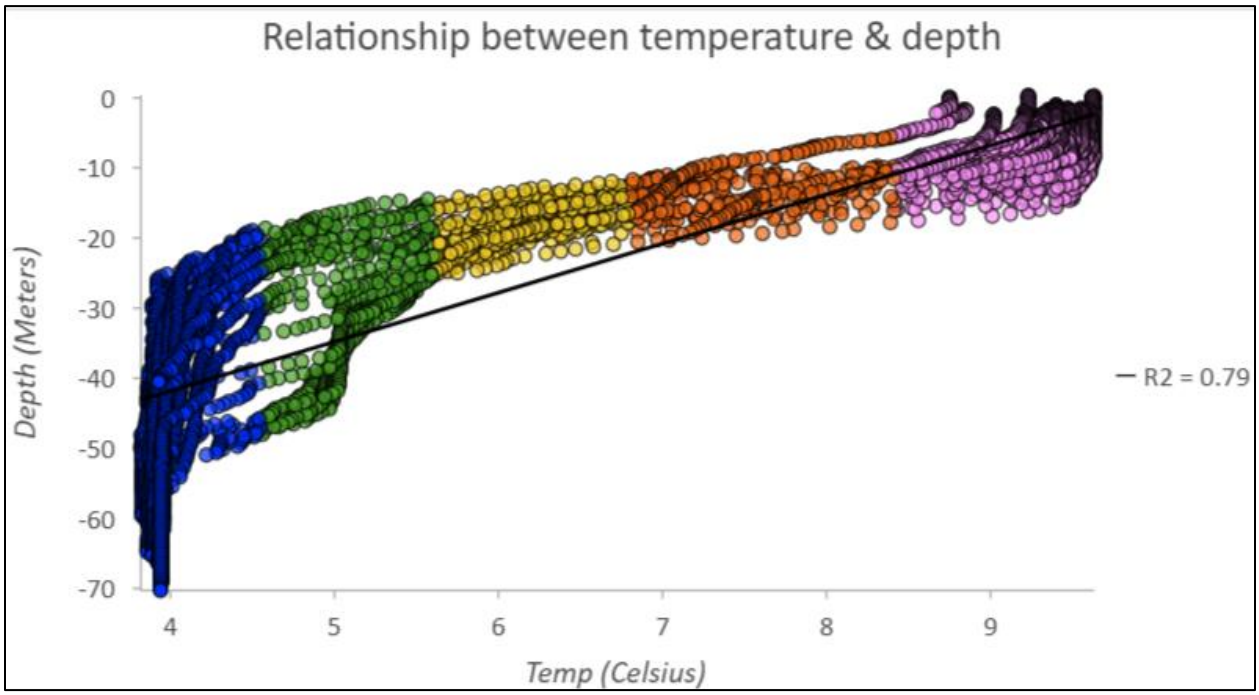

Figure 7. Scatter plot of subset points shows that temperature generally decreases with depth.

Discussion and Conclusion

There is an abundance of data for this study. However, the majority of that data is not specific to the variables of most interest here, that being water temperatures at depth. The wealth of data also creates a wealth of logistical work to discover, download, clean, and incorporate disparate sources and variables into a functioning informative dataset. An ETL was essential to maximize this process for large data repositories which often have streamlined, though complex, options for downloading. The depth data that does exist is so temporally and geographically restricted that it's impossible to make any useful model for the lake at large. It could potentially make an accurate snapshot of lake temperatures for the western most tip of the lake where most of the in-situ glider data was collected.

Another snapshot of conditions is available with the Contour layers generated from the interpolated surface layer. While these shapefiles illustrate just a thin slice of lake temperature phenomena at a specific depth, they do provide a concise profile for a narrow parameter of interest. This could be useful for determining changes over time at the location of a waste water pipe for instance, or whether fluctuating stream inlets affect lake temperatures and at what range.

As it's been determined that Kriging is not the best method for interpolating this type of data, another methodology is needed. There are more complex models that should be considered from this point forward but these aren't necessarily available with ArcGIS or ArcPy. Certain statistical software such as Stochastic Analysis, Modeling and Simulation (SAMS) may be a beneficial integration. It's possible that a more complex transformation available with some statistics programs could even form the data into a normal distribution curve, thereby accomplishing a more accurate Kriging analysis (Fagherazzi et al.).

Another option is to work with alternative datasets. Using satellite imagery would create a far more accurate and continuous surface temperature map (Tavares et al.). Landsat and MODIS data has a high temporal resolution and therefore is ideal for assessing historical trends and current conditions. This would create a solid base layer analysis to build out an integrated stochastic model for the entirety of Lake Superior. Incorporating flow variables such as lake currents and winds could at least produce a dynamic model that temperature data could then be plugged into as it became available.

Visualizing such large data and surfaces proved rather challenging with 3D EBK. While aggregating values would improve the functionality, ArcGIS voxel layers is theoretically a better way to handle this type of analysis. An example of a voxel layer showing Relative Humidity (RH) data in Figure 8 demonstrates the visual ease of this method. While the output is easy to view and interpret, the process is not. All attempts to create this were unsuccessful, which may have been due to licensing restrictions. Even though 3 dimensional representation is a rather stunning analytical tool, resorting back to 2 dimensional contour lines can be an equally powerful and simple way to understand changing temperatures, see Figure 9.

This is a good start to a more in-depth analysis of changes occurring in the great Lake Superior. Despite the issues with statistical data distribution, 3D EBK does provide some insights as to what's happening below the surface. The ETL and collected data are ready to use for the next analysis phase and troubleshooting these initial steps should allow the next ones to run a bit smoother.

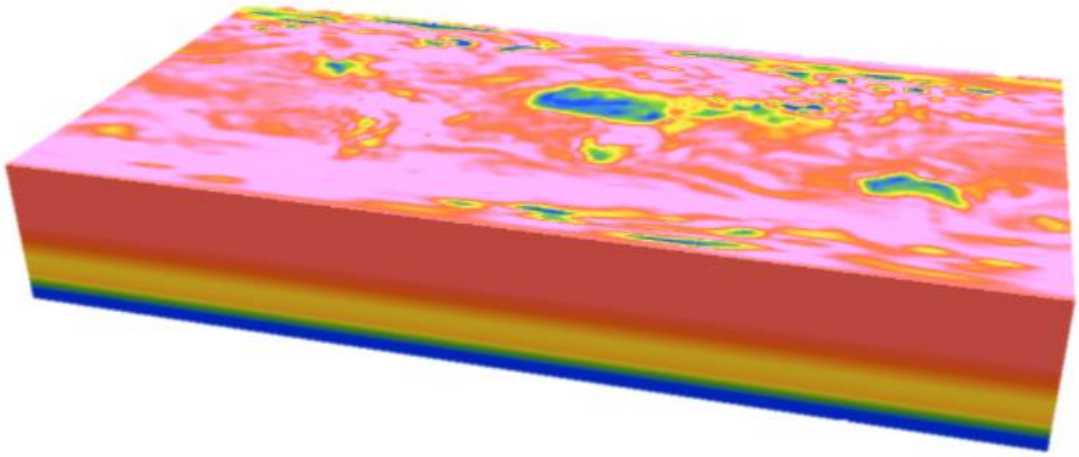

Figure 8. A voxel layer showing continuous 3 dimensional RH values. Source: esri ArcGIS tutorial (Nayak).

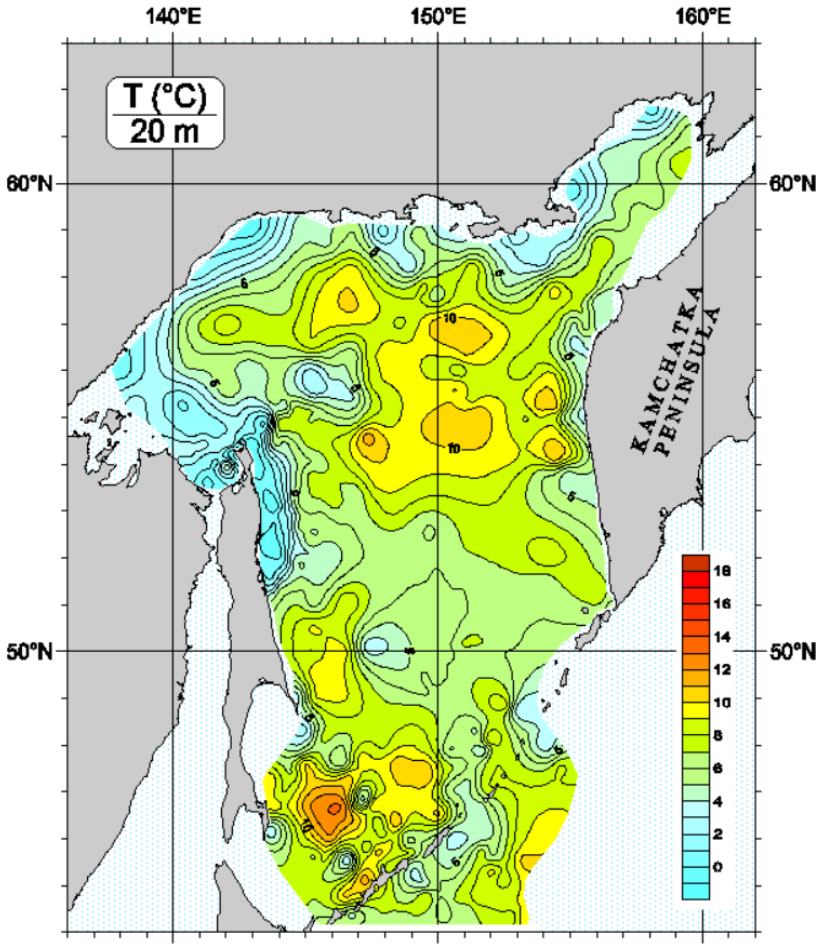

Figure 9. An example of depth specific lake temperatures represented as contour lines creates an easily interpretable layer. Source: (US Department of Commerce).

References

Austin, Jay A., and Steven M. Colman. "Lake Superior Summer Water Temperatures Are Increasing More Rapidly than Regional Air Temperatures: A Positive Ice-Albedo Feedback." *Geophysical Research Letters*, vol. 34, no. 6, 23 Mar. 2007, 10.1029/2006gl029021. Accessed 18 May 2020.

esri. "Cross Validation (Geostatistical Analyst)—ArcGIS pro | Documentation." *Pro.arcgis.com*, 2021, pro.arcgis.com/en/pro-app/latest/tool-reference/geostatistical-analyst/cross-validation.htm.

---. "Interpolate 3D Oxygen Measurements in Monterey Bay | Learn ArcGIS." *Learn.arcgis.com*, 2020, learn.arcgis.com/en/projects/interpolate-3d-oxygen-measurements-in-monterey-bay/. Accessed 9 May 2021.

---. "Interpolate Shape (3D Analyst)—ArcGIS pro | Documentation." *Pro.arcgis.com*, 2021, pro.arcgis.com/en/pro-app/latest/tool-reference/3d-analyst/interpolate-shape.htm. Accessed 10 May 2021.

esri docs. "ArcGIS Desktop." *Help.arcgis.com*, help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//00460000000s000000.htm. Reading netCDF data using geoprocessing tools.

---. "Set the Time Properties on Data—ArcGIS pro | Documentation." *Pro.arcgis.com*, pro.arcgis.com/en/pro-app/latest/help/mapping/time/set-the-time-properties-on-data.htm.

EUMETSAT. “Visualising Data in NetCDF Format.” *Www.youtube.com*, 20 Sept. 2018, www.youtube.com/watch?v=XqoetylQAIY&t=1309s. Accessed 26 Apr. 2021. Overview on netCDF files, panoply app & xarray python library for explorations.

Fagherazzi, Laura, et al. *STOCHASTIC MODELING and SIMULATION of the GREAT LAKES -ST LAWRENCE RIVER SYSTEM PREPARED BY.* , 2005.

GeoDelta Labs. “Extracting Time Series Data from a NetCDF File into a CSV (Part 3).” *Www.youtube.com*, www.youtube.com/watch?v=hrm5RmsVXo0. Accessed 17 Apr. 2021.

---. “How to Plot NetCDF Data onto a Map Using Python (with Matplotlib Basemap Toolkit) (Part 4).” *Www.youtube.com*, 7 Mar. 2020, www.youtube.com/watch?v=r5m_aU5V6oY. Accessed 17 Apr. 2021.

Integrated Ocean Observing System. “IOOS Archive Data Portal.” *Www.ncei.noaa.gov*, 2021, www.ncei.noaa.gov/access/integrated-ocean-observing-system/. Accessed 27 Apr. 2021. Glider data portal.

Nayak, Neeti. “Three Ways to Prepare Your Data for Voxel Layer.” *ArcGIS Blog*, 9 Sept. 2020, www.esri.com/arcgis-blog/products/arcgis-pro/3d-gis/three-ways-to-prepare-your-data-for-voxel-layer/. Accessed 11 May 2021.

noaa. “COBE SST: NOAA Physical Sciences Laboratory.” *Psl.noaa.gov*, 2021, psl.noaa.gov/data/gridded/data.cobe.html. Accessed 26 Apr. 2021. Satellite sea surface mean temp data in .cn.

O’Reilly, Catherine M., et al. “Rapid and Highly Variable Warming of Lake Surface Waters around the Globe.” *Geophysical Research Letters*, vol. 42, no. 24, 16 Dec. 2015, 10.1002/2015gl066235. Accessed 11 May 2021.

Open Source Options. “Intro to NetCDF with Python (NetCDF4).” *Www.youtube.com*, www.youtube.com/watch?v=VH-PCQ991fw. Accessed 26 Apr. 2021. Short guide on using netCDF4 python library.

---. “NetCDF with Python (NetCDF4) - Metadata, Dimensions, and Variables.” *Www.youtube.com*, 26 May 2020, www.youtube.com/watch?v=-kHxOOGtPhI. Accessed 26 Apr. 2021. Basic guide for exploring netCDF4 file with netCDF4 python package.

Tavares, Matheus, et al. “Comparison of Methods to Estimate Lake-Surface-Water Temperature Using Landsat 7 ETM+ and MODIS Imagery: Case Study of a Large Shallow Subtropical Lake in Southern Brazil.” *Water*, vol. 11, no. 1, 18 Jan. 2019, p. 168, 10.3390/w11010168. Accessed 11 May 2021.

US Department of Commerce, NOAA National Centers for Environmental Information. “Contour Maps of Temperature.” *Www.nodc.noaa.gov*, 23 Mar. 2015, www.nodc.noaa.gov/OC5/okhotsk/temp_dep.html. Accessed 11 May 2021.

Self-score

Fill out this rubric for yourself and include it in your lab report. The same rubric will be used to generate a grade in proportion to the points assigned in the syllabus to the assignment.

Category	Description	Points Possible	Score
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	**28**
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	**20**
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	**26**
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	**20**
		100	**94**