Final project report

Title: Assessment of the impacts of BP oil spill in the Gulf of Mexico on water dissolved oxygen, oxygen saturation, and conductivity
Notice: Dr. Bryan Runck
Author: Diego Osorio
Date: 12/21/2022

Project Repository: https://github.com/osori050/GIS5571_project
Google Drive Link:
Time Spent: 12 days

Abstract

In 2010, an explosion occurred in the Deepwater Horizon semi-submersible mobile offshore drilling unit leading to an oil spill of roughly 210 million gallons into the Gulf of Mexico. To assess the impact of said catastrophe, a NOAA dataset is acquired to analyze trends of dissolved oxygen, oxygen saturation, and conductivity over that area in 2010. Inverse distance weighting (IDW), global polynomial interpolation (GPI), local polynomial interpolation (LPI), and radial basis function (RBF) are used to model these water variables across the area of interest. LPI order 0 yields the lowest RMSE for the 3 variables, whereas IDW has the highest RMSE values. Likewise, dissolved oxygen and conductivity have low RMSE, while oxygen saturation has RMSE over 16. This means oxygen saturation is not a reliable parameter. Nevertheless, although the RMSE is low, dissolved oxygen and conductivity trends also introduce high uncertainty since they may have been affected by other factors such as pre-existing hypoxia and influx of freshwater from rivers respectively.

Problem Statement

In 2010, the Deepwater Horizon semi-submersible mobile offshore drilling unit suffered an explosion that killed 11 workers, injured 17 others, and spilled around 210 million US gal (780000 m^3) of oil into the ocean over 87 days. This catastrophe caused the biggest oil leak in history and severely impacted the quality of water in the Gulf of Mexico (Deepwater Horizon oil spill, 2022; Deepwater Horizon explosion, 2022). Thus, this project aims to assess the impact of the BP oil spill in the Gulf of Mexico on dissolved oxygen (mg/L), oxygen saturation (%), and conductivity (S/m) from May to October 2010.

Oxygen is a highly relevant variable since low levels (hypoxia) lead to dead zones where marine life cannot exist. The introduction of oil into the ocean may make oxygen levels drop as microorganisms use oxygen to consume the carbon compounds in oil (NOAA, 2010). Conductivity is also an indicator of ecological disturbance (EPA, 2022): seawater is a very good conductor thanks to the dissolve ions; therefore, low conductivity in seawater may be caused by the introduction of a non-conductive pollutant such as oil. That is, the lower the conductivity, the higher the concentration of oil. Figure 1 shows roughly the area affected by the spill (oil plume trajectory), and Figure 2 shows the locations where the physicochemical variables were measured.

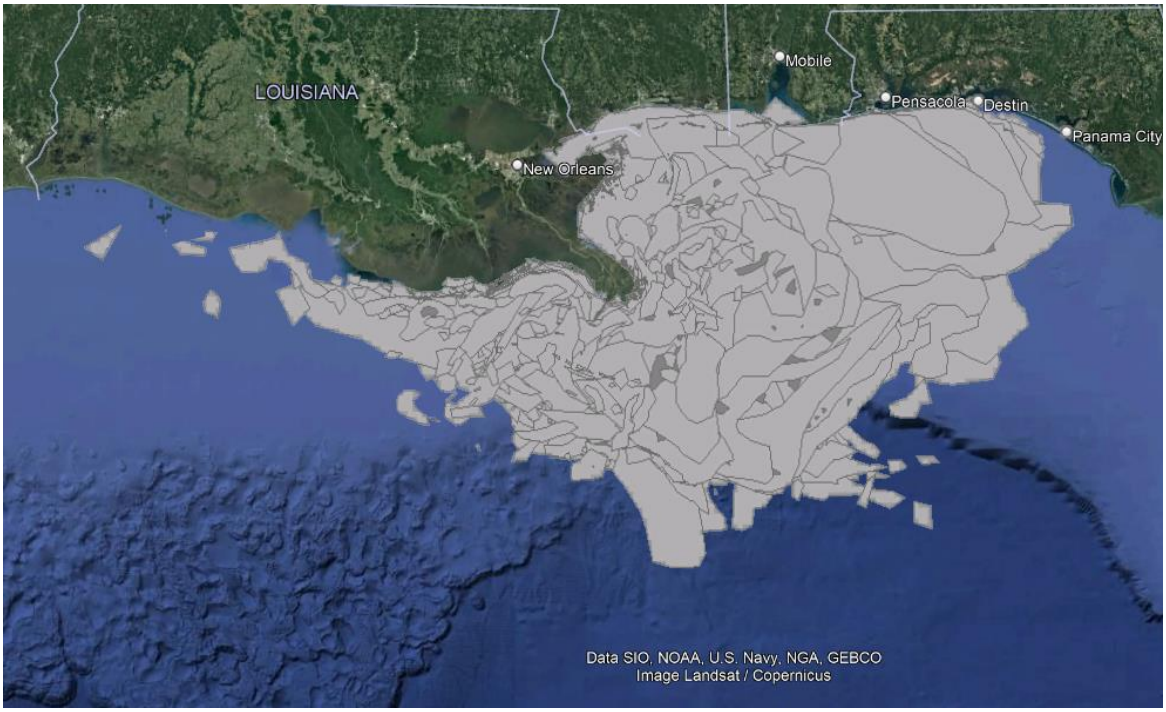


Figure 1. Oil spill plume (ESRI, n.d.)

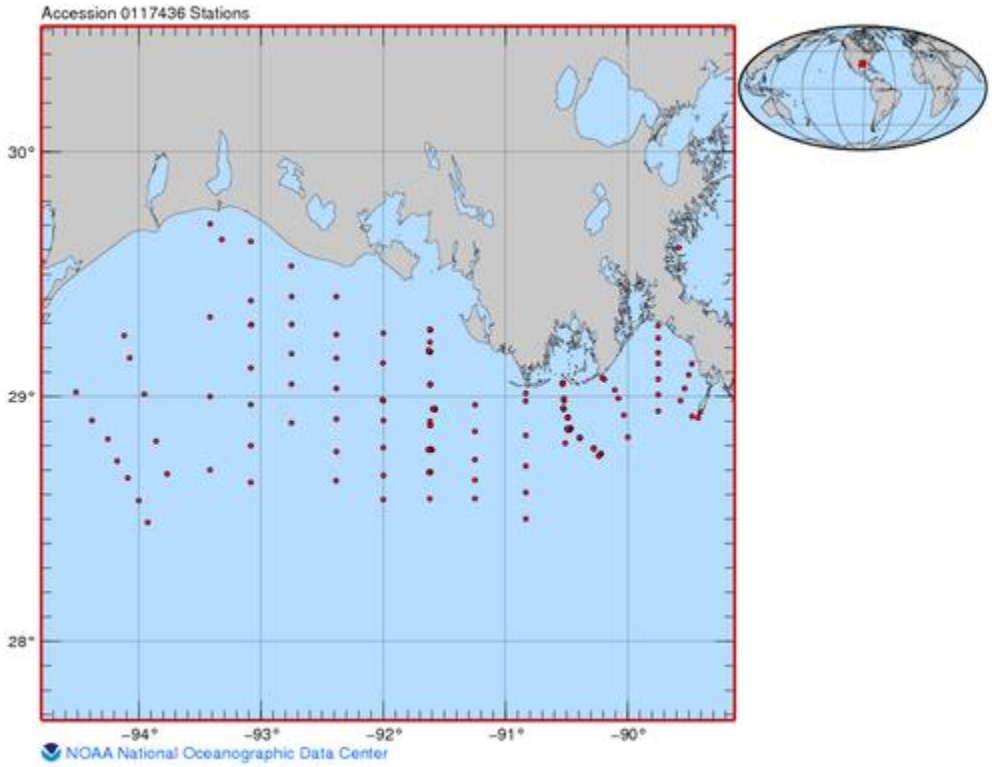


Figure 2. Water variables monitoring station locations (NOAA, 2021)

Table 1. Requirements for the oil impact on water quality

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset	Preparation
1	Area affected by the oil spill	Projection of the trajectory of the spill plume	Polygon geometry	Dates and area	NOAA Data	
2	Water physicochemical variables	Raw input dataset	Tables	Physicochemical variable readings and dates	NOAA Data	Transformation into GIS layers
3	US Borders	US mainland	Polygon geometry		Census data	Projection to WGS 1984

Input Data

First, ESRI has created a layer on ArcGIS Online that shows the daily projection of the oil spill plume in the Gulf of Mexico from May 2 to August 5 of 2010 based on NOAA data acquisition. The full extent is from -85.7202 to -93.4310 latitude, and from 27.2551 to 30.4219 longitude (datum WGS 1984). Additionally, there is a collection of physical, chemical, and biological data in the Gulf of Mexico from February 2 to October 28 of 2010 by NOAA recorded in a Microsoft Access database and text files. A polygon shapefile with the US mainland from the US Census Bureau is also included to act as a barrier so the analyses are performed only over water.

Table 2. Input data for oil spill analysis

#	Title	Purpose in Analysis	Link to Source
1	NOAA_Gulf_Oil_Spill	Analysis of the extension of the oil spill plum in the Gulf of Mexico over time	http://maps1.arcgisonline.com/arcgis/rest/services/NOAA_Gulf_Oil_Spill/MapServer
2	Physical, chemical, and biological data collected in the Gulf of Mexico from 02 Feb 2010 to 28 Oct 2010 (NCEI Accession 0117436)	Assessment of physicochemical parameters	https://www.ncei.noaa.gov/archive/archive-management-system/OAS/bin/prd/jquery/accession/download/117436
3	CB 2018 US Nation 5m	Barrier for interpolation	https://www2.census.gov/geo/tiger/GENZ2018/shp/cb_2018_us_nation_5m.zip

Methods

The water variables are downloaded through a GET request from NOAA API as a tar file which is later decompressed. Since the data is stored in a Microsoft Access database, the pyodbc package in Python can be used to access the data. Nonetheless, as shown in Figure 3, the code generates an error message as the Microsoft Access Driver is not present.

```
try:
    con_string = r'Driver={Microsoft Access Driver (*.mdb, *.accdb)};DBQ=E:\ArcGIS_1\Project\BP\0117436\2.2\data\0-data\NGOMEX_2010_Data_NODC_v2.accdb;'
    conn = pyodbc.connect(con_string)
except pyodbc.Error as e:
    print('Error in Connection', e)
```

```
Error in Connection ('IM002', '[IM002] [Microsoft][ODBC Driver Manager] Data source name not found and no default driver specified (0) (SQLDriverConnect)')
```

Figure 3. Block of code to access MS Access databases

Hence, the data needs to be pulled from the text files. Since the data in the files are separated by symbols (- and |), which altogether visually recreate cells as shown in Figure 4, a cleaning process is performed by opening the files and writing to a new file the lines without the dashes. Next, the latter file is opened as a pandas DataFrame where unwanted columns and rows are dropped, and whitespaces are removed. This is done to obtain the DataFrames with the station locations and the water variable readings. A new column with the geometry is added to the station DataFrame to create then a station GeoDataFrame.

```
|        ID        |     Station ID     |      Station       |        Date        |       depth        |   Inorganic SPM    |    Organic SPM     |      Total SPM     |
|                1 | 42                 | C5                 |          05/17/10 |               0.00 |               2.20 |               4.40 |               6.60 |
|                2 | 40                 | C6C                |          05/17/10 |               0.00 |               5.00 |               3.80 |               8.80 |
|                3 | 40                 | C6C                |          05/17/10 |               3.20 |               4.00 |               2.00 |               6.00 |
```

Figure 4. Raw data in text files

As the stations took several measures at different depths, only the readings at the minimum depth are considered. This is because oil tends to float over water, so the main impacts are expected to occur at the surface level. Additionally, on one given date, several stations took measures, and sometimes, one station took measures on more than one date. Therefore, a unique ID is needed to filter the DataFrame to get only the reading at the minimum depth at each station on each date. This is done by adding a new column and populating the rows with their corresponding date and station ID information. An auxiliary DataFrame is then created by grouping the rows by 'Date & Station ID' and aggregating the depth by the minimum function as shown in Table 3.

Table 3. First stage of auxiliary DataFrame

	Date & StationID	Depth (m)
0	10/27/2010 203	0.03
1	10/27/2010 204	0.05
2	10/27/2010 205	0.07
3	10/27/2010 206	0.04
4	10/27/2010 207	0.03

This depth value is converted to a string and concatenated with the 'Date & Station ID' field to create the unique ID. The same process is done on the water variables DataFrame. Later, a join keeping only features in common is run so all the readings at the other depths are left out. Table 4 shows what the two DataFrames look like after creating the unique ID (Join key).

Table 4. DataFrames after creating the unique ID to filter the readings at the minimum depth at each station on each date. Top: auxiliary DataFrame; bottom: water variables DataFrame

	Date & StationID	Depth (m)	Join key
0	10/27/2010 203	0.03	10/27/2010 203 0.03
1	10/27/2010 204	0.05	10/27/2010 204 0.05
2	10/27/2010 205	0.07	10/27/2010 205 0.07
3	10/27/2010 206	0.04	10/27/2010 206 0.04
4	10/27/2010 207	0.03	10/27/2010 207 0.03

	StationID	Date	Station	Scan	Depth (m)	Alt (m)	Irradiance (%)	Transmiss (%)	Fluorescence (RFU)	Temperature (C)	Conductivity (S/m)	Salinity (PSU)	Density (kg/m^3)	Oxygen (mg/L)	Oxygen (%)	Date & StationID	Join key
0	17	3/23/2010	C1	832	0.26	4.13		22.67	0.90	17.15	3.46	25.78	18.42	8.36	100.93	3/23/2010 17	3/23/2010 17 0.26
1	17	3/23/2010	C1	902	1.05	3.81	68.13	23.29	0.84	17.10	3.49	26.08	18.66	8.38	101.27	3/23/2010 17	3/23/2010 17 1.05
2	17	3/23/2010	C1	912	1.12	3.69	57.72	23.29	0.85	17.10	3.48	26.05	18.64	8.40	101.43	3/23/2010 17	3/23/2010 17 1.12
3	17	3/23/2010	C1	952	1.24	3.32	44.52	23.09	0.88	17.09	3.47	25.98	18.59	8.35	100.70	3/23/2010 17	3/23/2010 17 1.24
4	17	3/23/2010	C1	960	1.30	3.35	43.77	22.78	0.89	17.08	3.48	25.99	18.60	8.30	100.18	3/23/2010 17	3/23/2010 17 1.3

The output DataFrame from the previous join is joined once again with the stations GeoDataFrame which is finally converted into a shapefile setting WGS 1984 as its spatial reference. Figure 5 illustrates the conceptual model of the water variable data transformation.

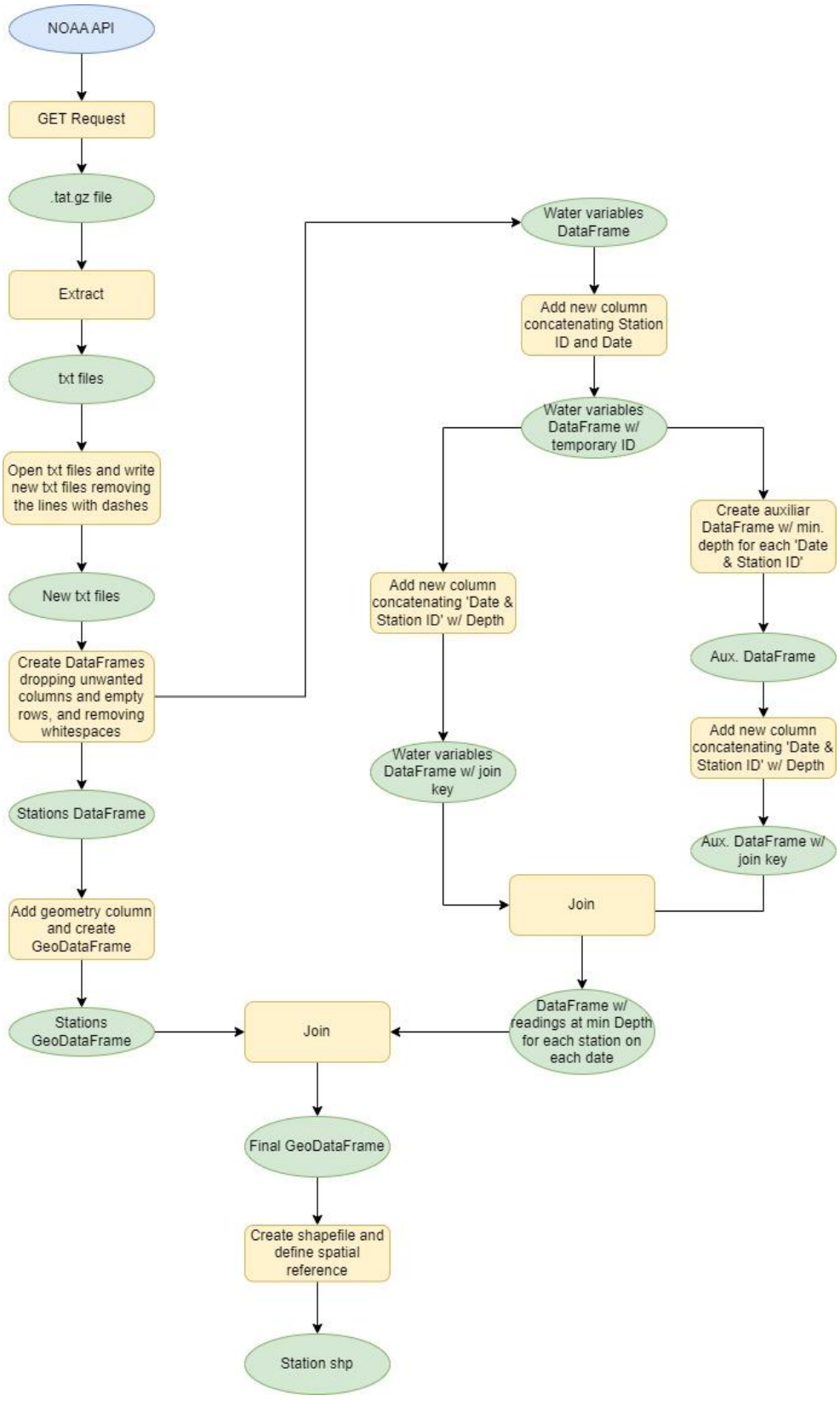


Figure 5. Diagram of the process to transform raw water datasets into a spatial dataset

To model the water variables over the Louisiana continental shelf, 4 interpolators are used to compare head to head and decide which is the best fit. These interpolators are inverse distance weighting (IDW), radial basis function (RBF), global polynomial interpolator (GPI), and local polynomial interpolator (LPI). They are selected based on decision trees to have a representation of each kind, when possible, in the different categories as shown in Table 5. From Lab 3, it was found that alpha 2 and order 2 are the best parameters for IDW and GPI respectively and thus, those are the parameters used here. The parameters utilized for RBF and LPI are the default ones.

Table 5. Classification of the selected interpolators (ESRI, 2021)

Category	**Kind**	**IDW**	**RBF**	**GPI**	**LPI**
Type of information	One prediction per location	X	X	X	X
	Quantile value				
	Many predictions per location*				
	Predicted values and errors				
Measurement of spatial autocorrelation	Yes				
	No		X	X	X
	Implicit	X			
Output type	Prediction	X	X	X	
	Prediction error				X
	Probability				
	Full distribution of possible values*				
Level of assumptions	Few	X	X	X	
	Intermediate				X
	Many				
Type of interpolation	Exact	X	X		
	Inexact			X	X
Smoothness of the output	Smooth		X	X	
	Intermediate				X
	Not smooth	X			
Uncertainty of predicted values	Yes				X
	No	X	X	X	

Processing speed	Slow				
	Intermediate				
	Fast	X	X	X	X

The water variables are interpolated by utilizing the Geostatistical Analysis Tools of arcpy which yield two outputs for each interpolator. The first output is a geostatistical layer used to carry out one-value cross-validation to obtain the root mean square error (RMSE). The other output is the interpolation raster. Due to the proximity of the stations to the coast and the irregularity of the coastline, the interpolations predict values over land areas as well. Hence, a mask is needed to erase those predictions. This is done by making a GET request through the US Census Bureau API which returns a zip file. The zip is decompressed, and the shapefile is projected to WGS 1984. Then, by using the outside extraction area within the Extract By Mask tool, the final interpolation rasters are generated with the predicted values only over the water. Figure 6 shows the diagram of the interpolation process.

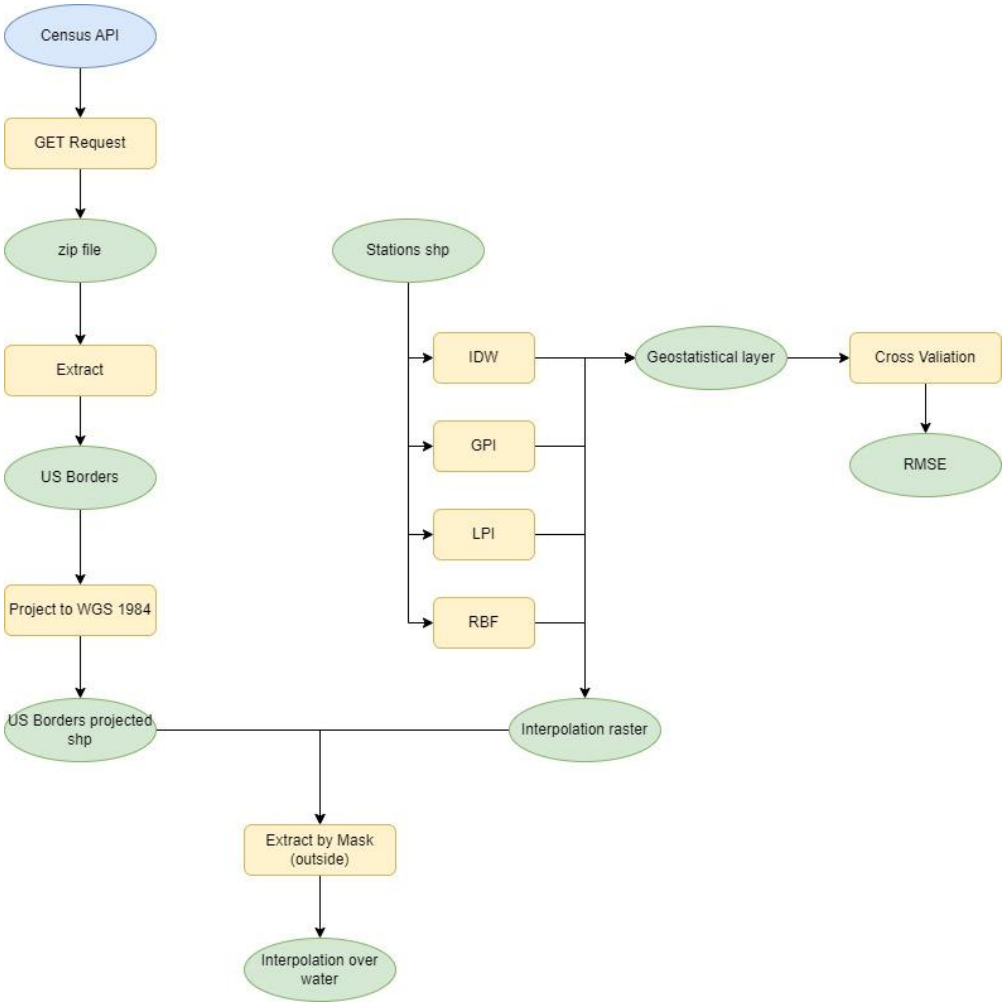


Figure 6. Diagram of interpolation over water by using IDW, GPI, LPI, and RBF methods

As the NOAA's Gulf Oil Spill layer lies on ArcGIS REST API, a request is needed by querying the data. Figure 7 shows the graphic version of the parameters used in the query.

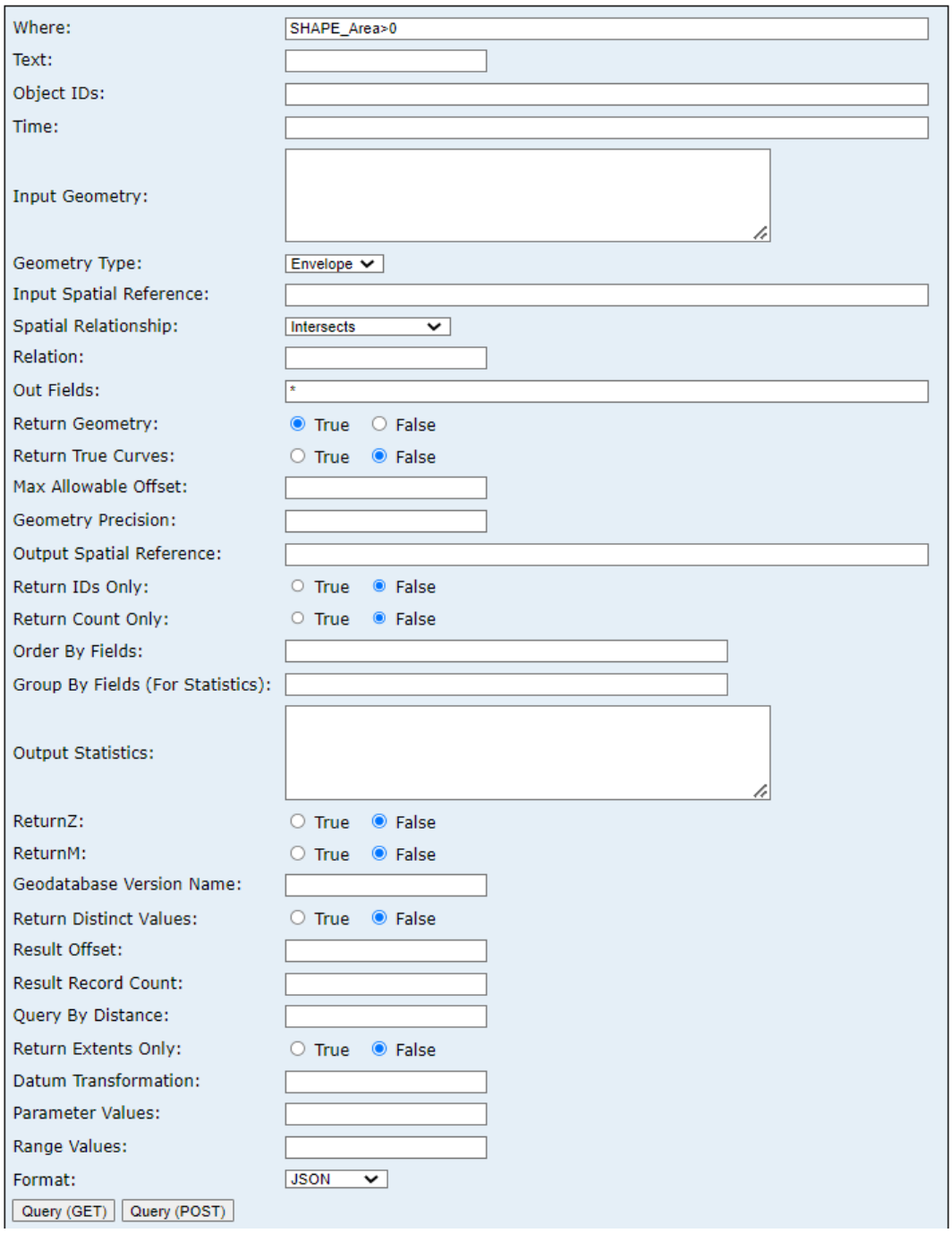


Figure 7. ArcGIS REST API query

The data is retrieved with a GET request and the content is written in a JSON file which is then converted to a shapefile by using the tool JSON To Features. One of the objectives is to see the stations that overlap both in space and time with the oil plume trajectory to analyze the direct impacts of the oil spill on the water variables over time. For that reason, an intersection is carried

out between the two datasets. The output shapefile is then read as a GeoDataFrame and the rows that do not match both date fields (stations and plume) are dropped. Figure 8 illustrates the conceptual diagram of the spatiotemporal intersection.

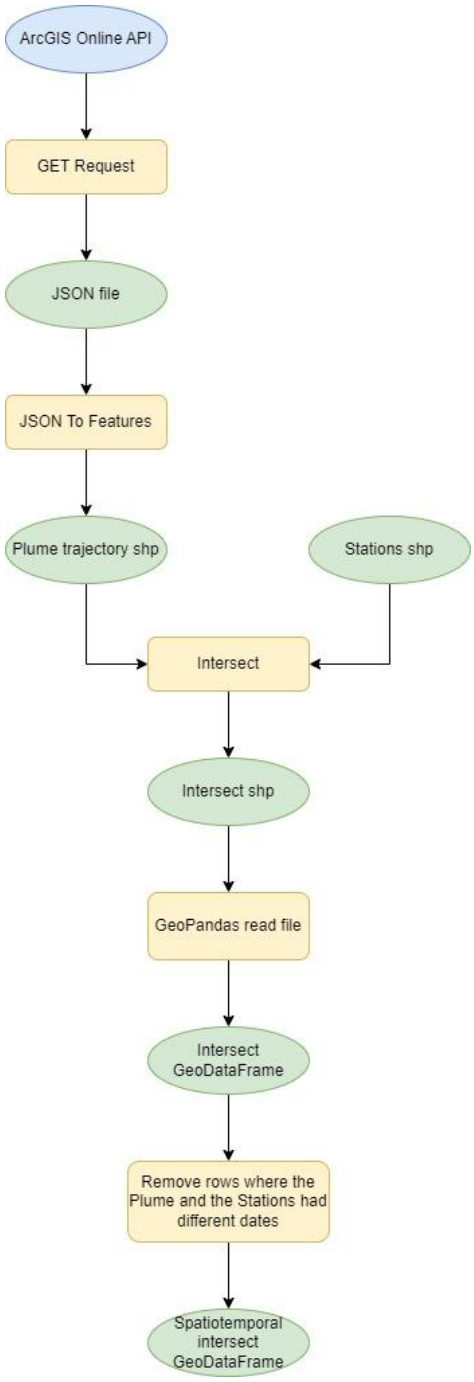


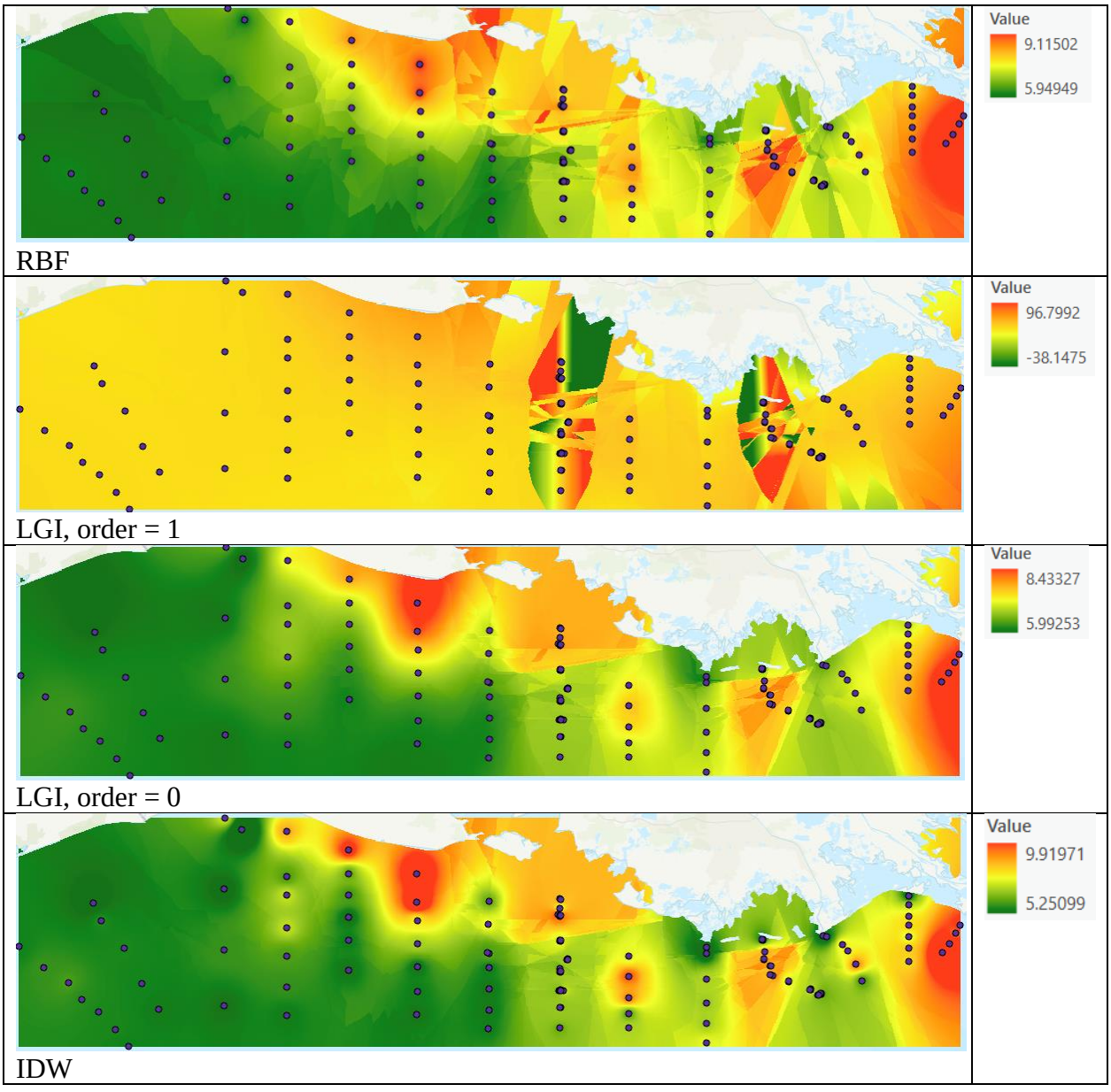
Figure 8. Diagram of spatiotemporal intersection

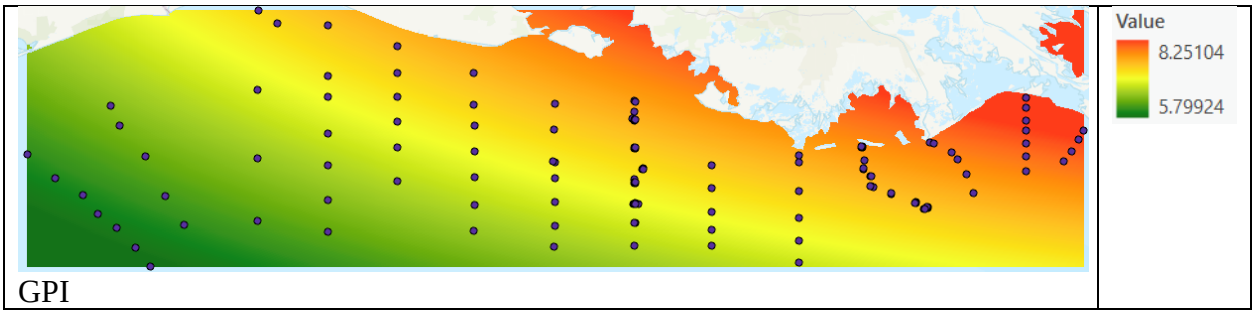
Unfortunately, the spatiotemporal intersection only occurs on two consecutive dates: July 30th and 31st. Since these results are not significant, this objective is dropped.

Results

Table 6 through Table 8 show the results of the interpolations for each water variable. In Table 6, two interpolations with LPI are displayed to compare the outputs when the order is changed. Initially, order 1 is used, as it is the default value in the tool, but LPI yields an output with negative values which is physically impossible since negative concentrations do not exist. Also, the upper range value is around 10 times more than expected, in comparison to the other interpolators. Therefore, the order is changed to 0 as it produces more reasonable results. Table 7 and Table 8 only include LPI order 0.

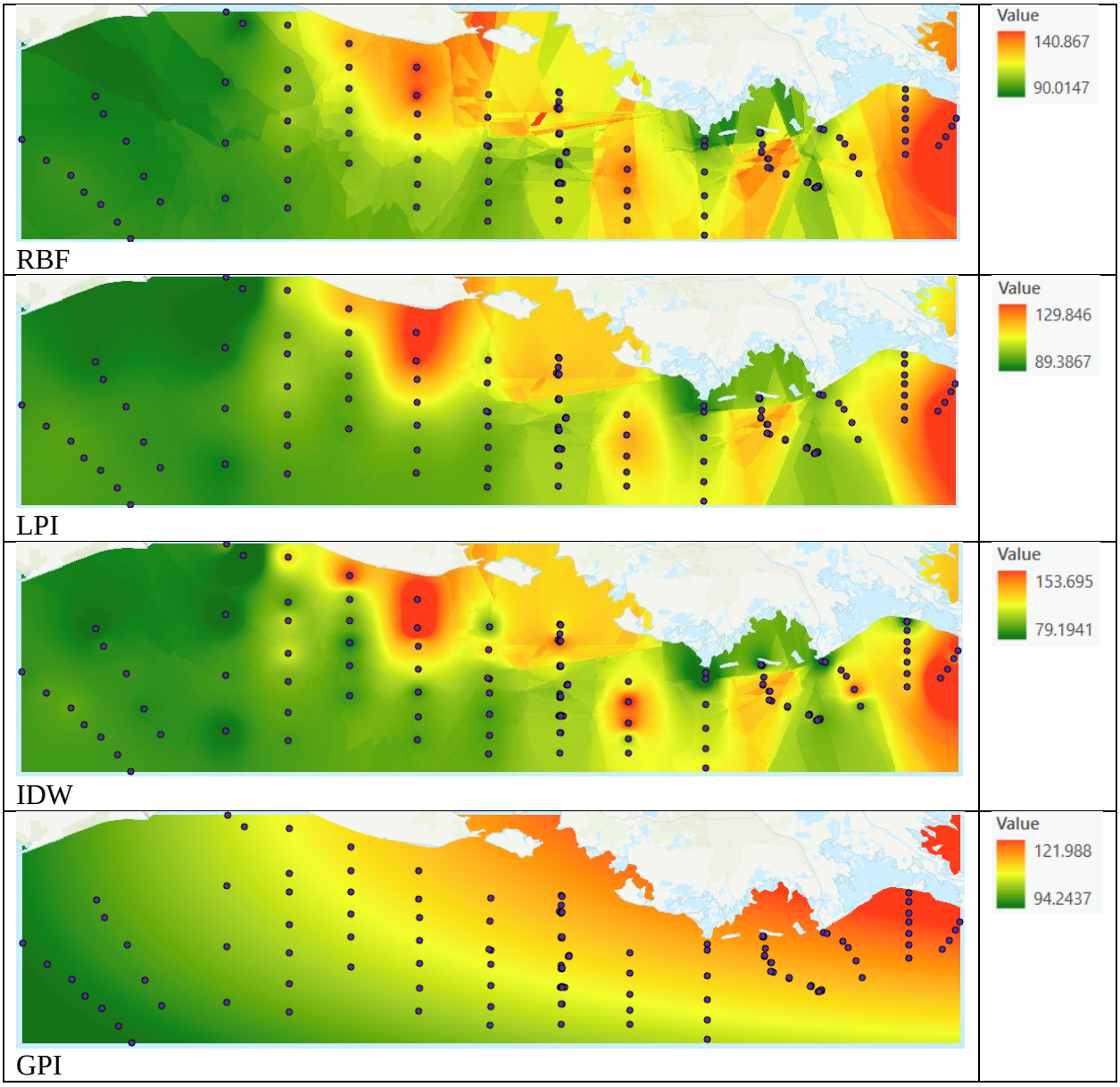
Table 6. Dissolved oxygen (mg/L) interpolations





GPI

Table 7. Oxygen saturation (%) interpolations



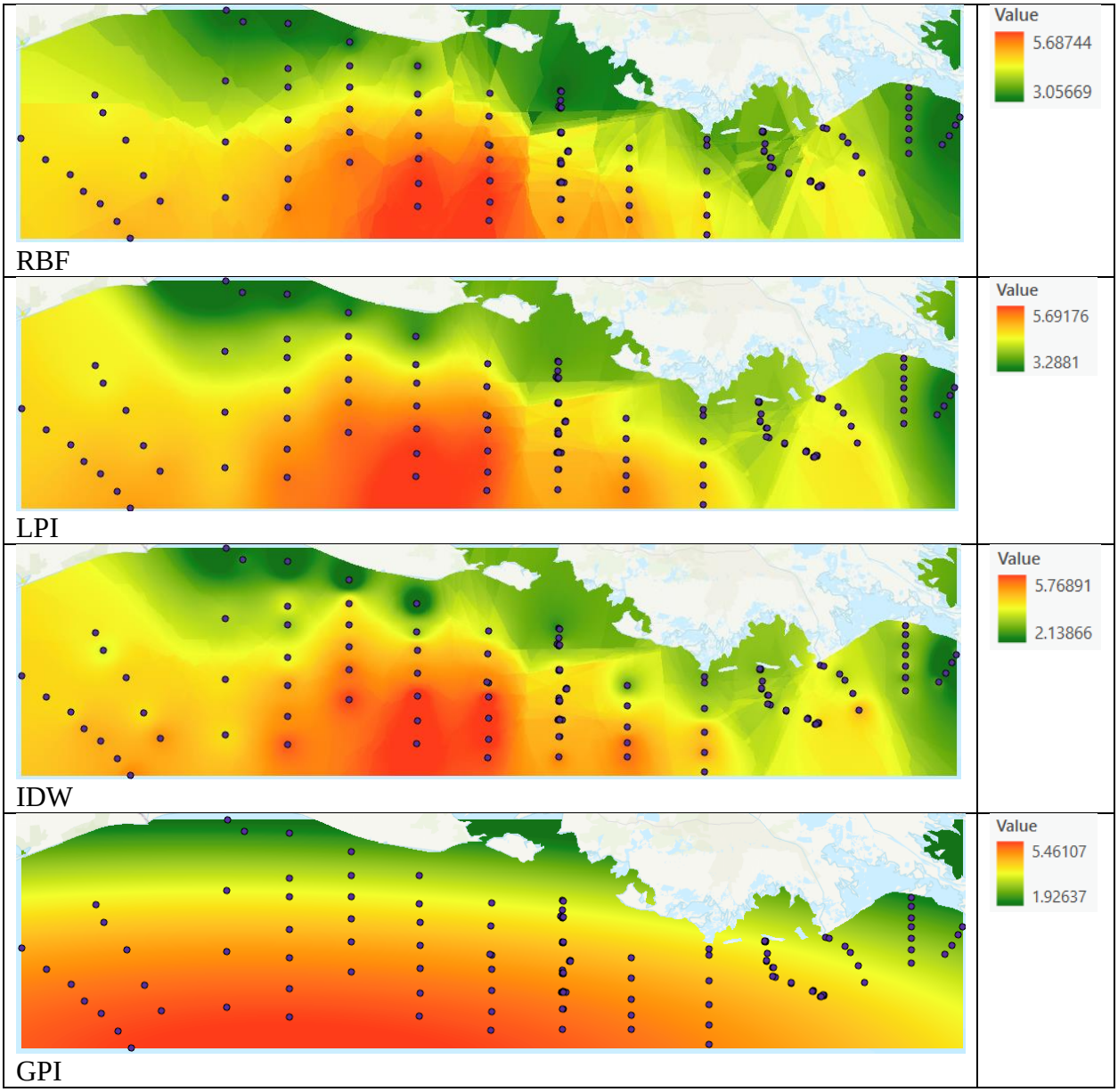
RBF

LPI

IDW

GPI

Table 8. Conductivity (S/m) interpolations



Overall, RBF, LGI order 0, and IDW produce similar models for oxygen and conductivity, while GPI can be interpreted as the outlier interpolator yielding different surface models. The oxygen levels (both concentration and saturation) are high close to the coastal areas and the delta of the Mississippi river where the Deepwater Horizon well is located. In contrast, conductivity is low in these areas and high in open sea.

Results Verification

The one-value cross-validation method only retrieves the root mean square error (RMSE) for the interpolators selected. This metric indicates how closely the interpolators predict the actual values:

the smaller the RMSE, the better. Table 9 provides the RMSE for each interpolator for each variable.

Table 9. RMSE. Green is the lowest RMSE for each variable, while red is the highest.

Variable	Interpolator	RMSE	RMSE Range*
Dissolved oxygen (mg/L)	RBF	1.191	0.142
	LPI	1.114	
	IDW	1.256	
	GPI	1.125	
Oxygen saturation (%)	RBF	17.285	1.848
	LPI	16.320	
	IDW	18.168	
	GPI	16.514	
Conductivity (S/m)	RBF	0.490	0.035
	LPI	0.477	
	IDW	0.512	
	GPI	0.484	

*IDW minus LPI

In all 3 cases, LPI order 0 yields the lowest RMSE, whereas IDW yields the highest. Conductivity and concentration of dissolved oxygen are the variables with the lowest RMSE. In contrast, oxygen saturation has a high RMSE which makes the models not reliable. It should be noted that the results within each variable are not statistically significant as the RMSE ranges (IDW minus LPI) are small, which means that the interpolators have roughly the same accuracy.

Discussion and Conclusion

On the one hand, it was expected to see a dramatic decrease in oxygen levels, especially in those areas close to the Deepwater Horizon well due to microbes utilizing oxygen to degrade the heavy carbon load from the oil spill. Furthermore, the Louisiana continental shelf to the west of the Mississippi river is characterized by the presence of 'dead zones' due to hypoxia caused by the introduction of nutrient loads from the Mississippi and Atchafalaya rivers. The values of dissolved oxygen can drop up to 2 mg/L and the most severe effects are seen between June and August every year (Rabalais, Turner, & Wiseman, 2001). On the other hand, a report presented by NOAA, EPA (Environmental Protection Agency), and OSTP (Office of Science and Technology Policy) in September 2010 stated that there were no dead zones observed or expected as part of the BP oil spill accident. Scientists found that oxygen levels dropped by 20% from the average but there was no continued decreased tendency over time (NOAA, 2010). Hence, dissolved oxygen and oxygen saturation were not hugely impacted by oil; however, if a dropping trend had been found, it would

have been uncertain how much could be attributable to oil since there was already a hypoxia condition.

Regarding conductivity, Zheng et al. (2018) found that the standard values at the sea surface are around 5.5 S/m. Besides, oil is an organic compound that does not conduct electrical current very well and thus, lowers conductivity when present in water (EPA, 2022). The interpolations obtained here effectively show low conductivity values close to the well which corresponds to what is expected. However, the conductivity is also low close to other coastal areas. The latter may happen due to the introduction of freshwater from rivers into the ocean which has low conductivity in comparison to seawater. Indeed, the low conductivity close to the well can be caused by the freshwater coming from the Mississippi river delta too. That is, both freshwater and oil pollution may be responsible for low water conductivity close to the well and there is no more information in the water variable dataset to know what percentage is attributable specifically to the oil spill.

Murphy et al. (2010) carried out a similar project where water quality was evaluated by interpolating salinity, water temperature, and dissolved oxygen data taken by cruises between 1985 and 1994 at multiple locations. They found that kriging methods outperformed IDW for all parameters. With regards to LPI, no easily accessible literature was found about its use in water quality. However, a study about the spatial distribution of precipitation showed that LPI was the optimal method over IDW, GPI, RBF, and even ordinary kriging, which is the same case as in this project. From Lab 3, it had been learned that i) GPI is not a good interpolator, and ii) IDW yields low accuracy, which both were once again confirmed in this project.

In conclusion, the water variables analyzed here were not appropriate to evaluate the impact of the oil spill accident on water quality due to the introduction of high uncertainty. In future studies, it is recommended to obtain datasets with readings on the same date so temporal variability can be removed and snapshot interpolations can be performed to analyze trends over time. Similarly, other variables sensitive mainly to oil pollution should be considered to avoid ambiguity since dissolved oxygen, oxygen saturation, and conductivity can be affected by multiple factors. Finally, kriging methods should be considered too as literature shows they are good at modeling water quality variables.

This project helped me sharpen my coding and troubleshooting skills by facing challenges never experienced before. First, I learned to clean text files populated with unwanted characters and whitespaces. Second, I dealt with data different from what I needed since I was looking for stations measuring variables at multiple locations on the same dates to carry out analyses over time. Nonetheless, after the cleaning process, I encountered a dataset with stations measuring variables at multiple locations but on different dates. Third, I gained an understanding of cross-validation and how to run it to verify and validate the models generated. Forth, I struggled to filter the DataFrame based on date, station ID, and minimum depth at the same time, but I finally could create code that, although it is not efficient, is effective to obtain the desired output. Last but not least, I developed a block of code to perform a spatiotemporal intersection.

References

Deepwater Horizon explosion. (2022, August 26). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Deepwater_Horizon_explosion

Deepwater Horizon oil spill. (2022, September 20). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill

EPA. (2022, July 11). *Indicators: Conductivity*. Retrieved from National Aquatic Resource Surveys: https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity

ESRI. (2021). *Classification trees of the interpolation methods offered in Geostatistical Analyst*. Retrieved from ArcGIS Desktop: https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/classification-trees-of-the-interpolation-methods-offered-in-geostatistical-analyst.htm

ESRI. (n.d.). *NOAA_Gulf_Oil_Spill (MapServer).* Retrieved from ArcGIS REST Services Directory: http://maps1.arcgisonline.com/arcgis/rest/services/NOAA_Gulf_Oil_Spill/MapServer

Murphy, R., Currier, F., & Ball, W. (2010). Comparison of spatial interpolation methods for water quality evaluation in the Chesapeake Bay. *Journal of Environmental Engineering, 136*(2), 160-171. doi:10.1061/ASCEEE.1943-7870.0000121

NOAA. (2010, September 13). No dead zones observed or expected as part of BP Deepwater Horizon oil spill. *ScienceDaily*. Retrieved from www.sciencedaily.com/releases/2010/09/100913165807.htm

NOAA. (2021). *Physical, chemical, and biological data collected in the Gulf of Mexico from 02 Feb 2010 to 28 Oct 2010 (NCEI Accession 0117436).* Retrieved from DATA.GOV: https://catalog.data.gov/dataset/physical-chemical-and-biological-data-collected-in-the-gulf-of-mexico-from-02-feb-2010-to-28-oc

Rabalais, N., Turner, R., & Wiseman, W. (2001). Hypoxia in the Gulf of Mexico. *Journal of environmental quality, 30*(2), 320-329. doi:10.2134/jeq2001.302320x

US Census Bureau. (2018). CB 2018 US Nation 5m. *Cartographic Boundary Files - Shapefile*. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

Zheng, Z., Fu, Y., Liu, K., Xiao, R., Wang, X., & Shi, H. (2018). Three-stage vertical distribution of seawater conductivity. *Scientific Reports, 8*(1), 1-10. doi:10.1038/s41598-018-27931-y

Self-score

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	24
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	28
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	100