Impact of Water Conditions on Dissolved Oxygen Level*

An Analysis of U.S. Water Quality Data From 1989 to 2019

Wentao Sun

April 19, 2024

This study conducts a thorough analysis of various water quality parameters including dissolved oxygen content, water temperature, salinity, and pH levels within aquatic environments. The research primarily investigates the impacts of temperature, salinity, and pH on the dissolved oxygen levels, key indicators of water quality, to discern patterns in environmental changes. Findings reveal that both temperature and salinity exert a negative influence on dissolved oxygen levels, with increases in these parameters correlating with reductions in oxygen availability. The comprehensive water quality analysis deployed in this research elucidates the intricate interactions among critical determinants of aquatic ecosystem health. Ultimately, this study enhances our understanding of how different environmental factors influence dissolved oxygen content, offering valuable insights for the management of aquatic ecosystems and advancing environmental sustainability efforts.

Table of contents

*Code and data are available at: https://github.com/TonySun1107/Water_Quality.git

1 Introduction

As we moved into the 21st century, the growth of industry and technology brought environmental sustainability into focus. Water resources of desired quality and quantity are the foundation for human survival and sustainable development (Huang et al. 2021). Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs (Brundtland 1987). The concept of sustainable development promotes accurate quantitative analysis and review of the human living environment, culminating in a burgeoning field of water quality research. This body of work is crucial in comprehending the nuanced impacts of modern society on natural water systems and in shaping policies for the preservation of these vital resources.

The dissolved oxygen (DO) concentration is a critical parameter for evaluating the ecological health of an aquatic environment (Huang et al. 2021). This concentration results from an equilibrium between oxygen-producing (e.g., photosynthesis and air diffusing) and oxygen-consuming (e.g., aerobic respiration, nitrification, and chemical oxidation) processes in an aquatic environment (Olyaie, Abyaneh, and Mehr 2017). In conditions of low dissolved oxygen concentrations prevalent in water bodies receiving heavy sewage pollution (Aston 1973). Combining quantitative data analysis and water quality research, this study investigated dissolved oxygen levels in a bay, a fishing pond, and three other pools to assess water quality in a variety of aquatic environments from 1989 to 2019.Data provided by the U.S. Department of the Interior were selected for this study to understand the various measurements of a given body of water, including salinity, ph, and water temperature. including salinity, ph, and water temperature. The dissolved oxygen of the waters were analyzed in detail using the Corrected Total Oxygen Reading (CTOR).The CTOR value can be used as a quantitative measure of the change in oxygen.The higher the CTOR, the higher the dissolved oxygen content of the waters and the healthier the water quality.The CTOR values are helpful in evaluating the impacts of the environmental policies and the natural variations on the aquatic health.The CTOR values are used to assess the impacts of the environmental policies and natural variations on the aquatic health.

The paper is structured to facilitate a comprehensive understanding of the study and its implications. Following Section 1, Section 2 presents the data, detailing the data sources, analytical

techniques, and the rationale behind the chosen methods. Section 4 discusses the results of the study, describing the trends summarized in the watershed quality data. Section 5 provides an in-depth discussion of these results, exploring potential factors influencing measurements, linking them to broader social policy and environmental issues, and providing recommendations for future research in this area.

1.1 Estimand

The focus of this study was to estimate the causal effect of dissolved oxygen content of waters on water quality, with the salinity and pH of the waters being equally worthy of consideration. By studying the data obtained from various measurements in five different waters, the aim is to understand the impact of environmental policies and natural changes on aquatic health.

2 Data

This section offers an overview of the datasets utilized in the analysis, focusing specifically on water quality indicators. The data, sourced from the open-access US data.gov database, encompasses measurements from a bay, a fishing pond, and three watersheds, covering the period from 1989 to 2019. The dataset includes detailed records of oxygen content, pH, salinity, and water temperature, each noted with specific dates. This comprehensive dataset enables a detailed examination of the temporal changes in water quality over the thirty-year period. Additionally, the analysis explores the relationships among oxygen content, pH, salinity, and overall water quality, providing insights into the environmental dynamics affecting these aquatic systems.

2.1 Source and Methodology

The data collection process for this database was meticulously executed by the U.S. Department of the Interior (DOI), employing a diverse array of methodologies such as sampling, naturalistic observation, and experimental testing. This multi-faceted approach guaranteed a comprehensive and reliable dataset, which facilitates a thorough analysis and understanding of the environmental variables at play. Specifically, water quality data for the Refuge, collected bi-weekly by volunteers, includes measures of turbidity, pH, dissolved oxygen (DO), salinity, and temperature. Sampling is conducted at designated locations across several water bodies, including the Bay, D-Pool (fishing pond), and C-Pool, B-Pool, and A-Pool. This structured and regular data collection strategy provides valuable insights into the temporal dynamics of water quality within these aquatic ecosystems.

The analysis of this paper makes use of the `R` programming language (R Core Team 2020) for statistical computations and visualizing data. The `tidyverse` package (Wickham et al. 2019)

is installed to gain access to other importantR packages, including the `dplyr` package (Wickham et al. 2023) used to manipulate and clean data, the `readr` package (Wickham, Hester, and Bryan 2023) to read and import data, the `here` package (Müller 2020)to create a path to specific saved files, the `ggplot2` package (Wickham 2016) to create the data visualizations. Additionally, the `lubridate` package (Grolemund and Wickham 2011) facilitated date and time manipulation, the `kableExtra` package (Zhu 2021) enhanced the creation of complex tables and the `GGally` package (Schloerke et al. 2024) provided extended visualization capabilities, particularly for pairwise data exploration. The `modelsummary` package (Arel-Bundock 2022) was utilized for generating comprehensive summary tables. These tools collectively underpin the robust analytical framework that supports the findings presented herein.

2.2 Variables

In order to better understand the data and the research process, we have selected the first ten rows of data to detail the research methodology used, explain its relevance and how it contributes to our understanding of the topic. Our focus is on salinity, oxygen, ph and air temperature in the selected waters at different times, which gives a comprehensive understanding of how water quality indicators change when different air temperatures are present.

Table 1: First Ten Rows of Water Quality from 1990 to 2019

Year	Month	Site ID	Salinity(ppt)	Oxygen(mg/L)	PH	Air.Temp(°F)	Water.Temp(°C)
1990	1	Bay	1.0	9.7	7.5	53.6	10.0
1990	2	Bay	3.4	10.8	7.5	50.0	10.0
1990	2	Bay	3.2	10.6	7.0	48.2	10.0
1990	2	Bay	4.2	11.6	7.5	41.0	5.0
1990	2	Bay	2.8	13.8	7.0	33.8	2.5
1990	3	Bay	2.9	11.6	7.0	47.3	7.0
1990	3	Bay	2.9	9.0	7.0	69.8	17.0
1990	3	Bay	2.8	7.8	7.0	57.2	15.0
1990	3	Bay	3.0	10.8	7.0	48.2	9.5
1990	4	Bay	1.8	9.6	8.0	59.0	5.0

table1, built with `kableExtra` (Zhu 2021), display the first ten rows for salinity, oxyge, PH, and air temperature of the selected waters. This is a more concise table after removing the vacant data, and it shows the statistics of each indicator for waters of different years and months, providing a streamlined view for subsequent analysis and processing.

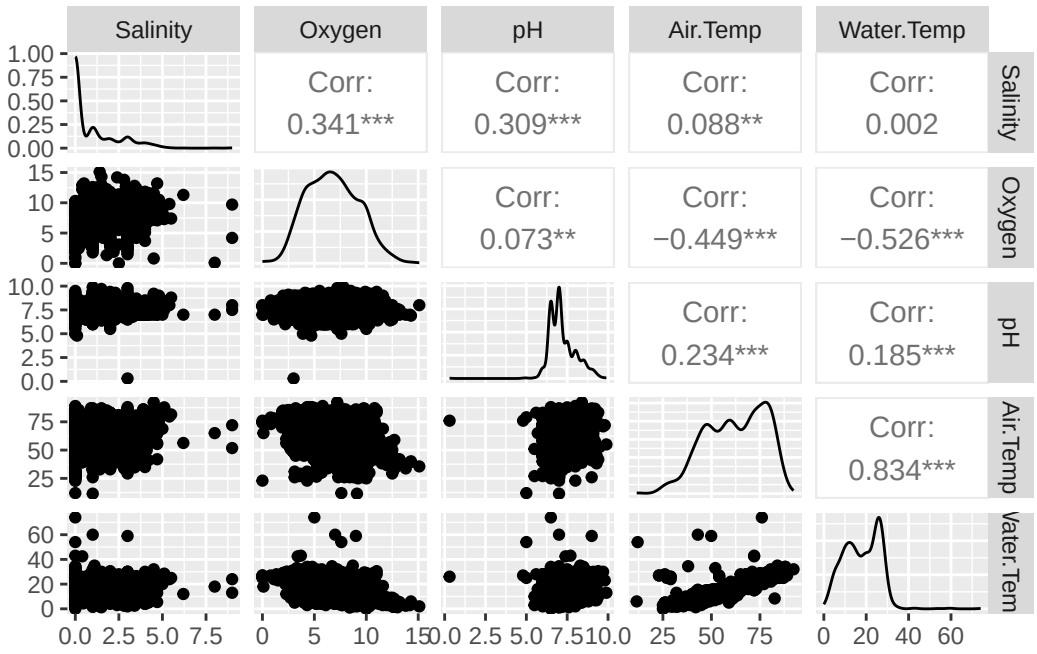

Figure 1: Correlation Matrix of Water Quality Indicators from 1989 to 2019

2.3 Preliminary Analysis

Figure 1, bulit with `GGally` (Schloerke et al. 2024), presents a comprehensive correlation matrix of water quality indicators, encompassing salinity, dissolved oxygen content, pH levels, and air temperature, collated from 1989 to 2019. Each plot within the matrix elucidates the pairwise relationships between these critical environmental parameters. The diagonal histograms reveal the distribution of individual indicators, where the dissolved oxygen content exhibits a normal distribution, while pH levels display a bimodal distribution. The scatter plots below the diagonal illustrate the interaction between variables, with the correlation coefficient denoted for each. Noteworthy is the strong negative correlation between air temperature and dissolved oxygen, suggesting a marked decrease in oxygen levels as temperatures rise. On the other hand, salinity and pH exhibit a moderately positive correlation.

Figure 2 displays a visual exploration of dissolved oxygen trends from 1989 to 2019. The boxplots demonstrate relatively stable median oxygen levels over the years, with interquartile ranges suggesting consistent variability. However, a distinct dip in levels between 1998 and 1999 indicates a significant disruption in oxygen balance, without noticeable recovery in the following years. This pattern may reflect specific environmental events impacting the water body's oxygenation.

Figure 3 displays a monthly distribution of dissolved oxygen levels across an aquatic environment, showcasing the seasonal fluctuations within a calendar year. The boxplots are arranged to represent each month, revealing both median oxygen concentrations and the spread of values

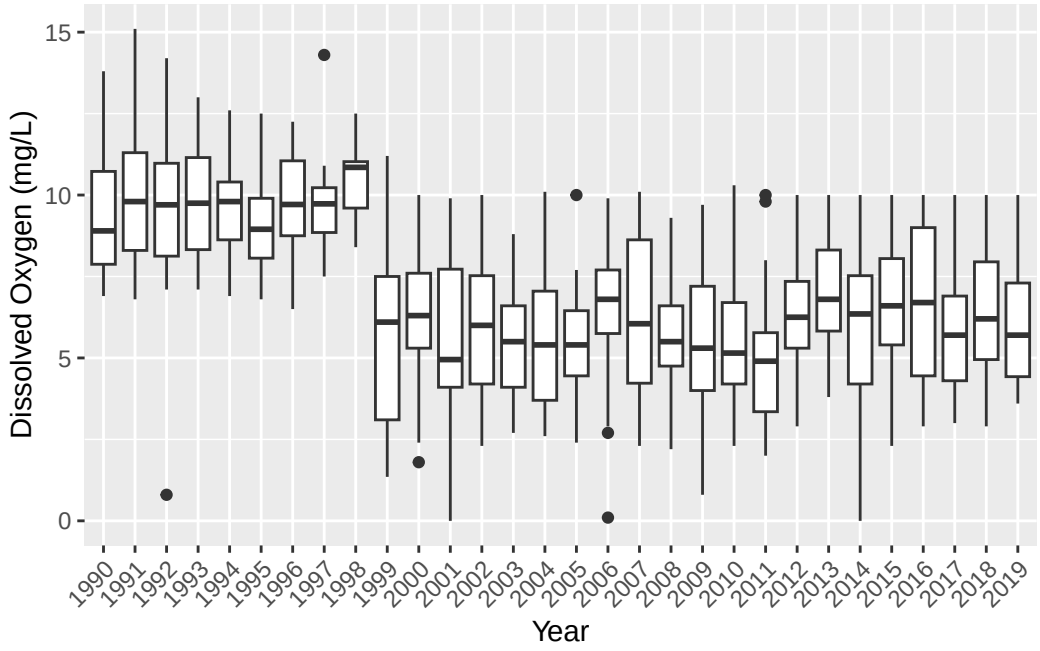

Figure 2: Annual Variability of Dissolved Oxygen Concentrations from 1989 to 2019

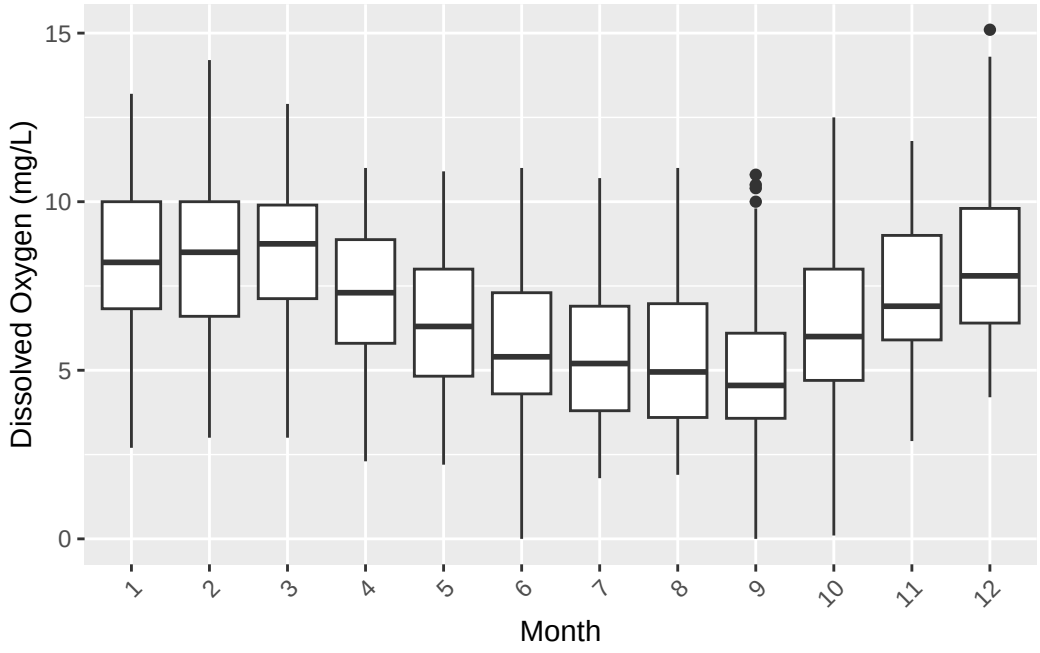

Figure 3: Seasonal Variation in Dissolved Oxygen Concentration by Month from 1989 to 2019

above and below the median—reflected by the interquartile range. It is notable that the later months exhibit greater variability, as evidenced by the length of the whiskers and the presence of outliers, which are indicative of sporadic events affecting water oxygenation. This visualization captures the cyclical nature of oxygen levels, potentially correlating with temperature changes and biological cycles within the ecosystem.

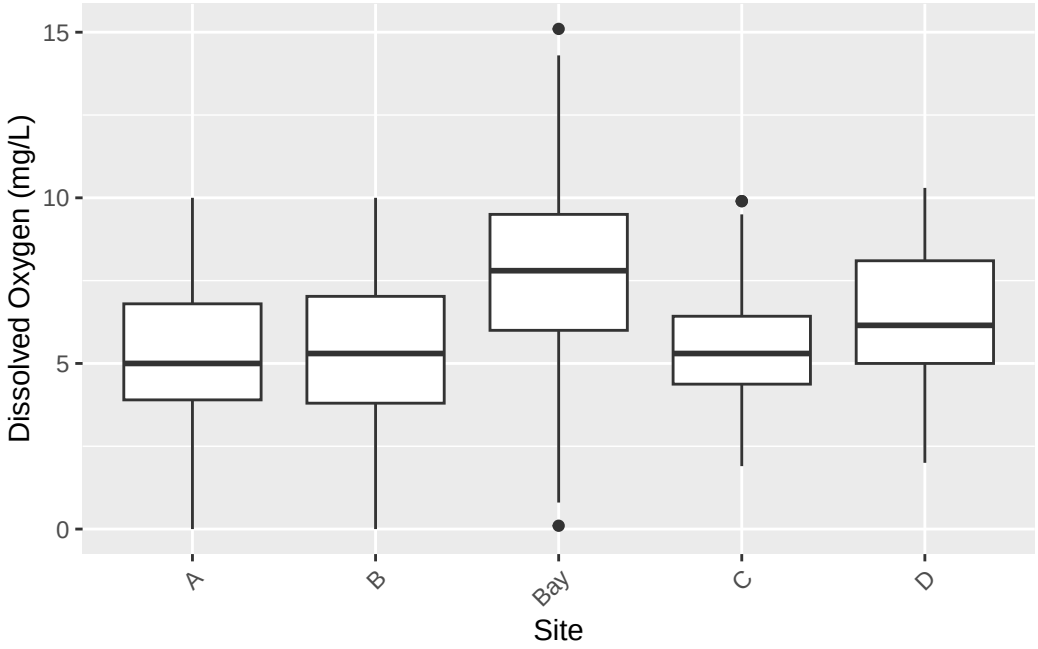

Figure 4: Distribution of Dissolved Oxygen Levels Across Sampling Sites from 1989 to 2019

Figure 4 illustrates the distribution of dissolved oxygen levels across various sites, as represented by boxplots. Each boxplot corresponds to a unique site ID, summarizing the central tendency and variability of oxygen measurements at each location. The median line within each box denotes the median dissolved oxygen concentration, while the hinges of the box reflect the interquartile range. Notably, certain sites exhibit a wider range of values, indicating more variability in oxygen levels, while others display a tighter interquartile range, suggesting more stable conditions. Outliers are depicted as individual points outside the whiskers, indicating observations that stand out from the general distribution and may warrant further investigation. The sites are labeled along the x-axis, which has been angled for improved readability. This figure provides a comparative view of oxygen content variability, a crucial indicator of water quality and aquatic health, among the sampled sites.

2.4 Measurements

In this study, water quality in a variety of aquatic environments was assessed by analyzing dissolved oxygen levels, a key indicator of ecosystem health. The dataset used in this study

consists of readings collected from multiple sites over a thirty-year period from 1989 through 2019, with measurements recorded every two weeks to capture seasonal and annual trends.

The data were sourced from recognized online databases, ensuring the reliability and accuracy of the records. We employed the Corrected Total Oxygen Reading (CTOR). The CTOR is derived by dividing the unique oxygen measurements (types) by the square root of twice the total number of all measurements (tokens). This approach offers a normalized value, facilitating a valid comparison of oxygen levels across different sites and times. It effectively minimizes the influence of data quantity, allowing for a direct assessment of water quality. The CTOR values thus serve as a quantitative measure of oxygen variability and are instrumental in evaluating the impacts of environmental policies and natural changes on aquatic health.

3 Model

Since the dependent variable is continuous, a linear model is considered instead of a logistic model. As part of the multiple linear regression modeling, we will include as explanatory variables the salinity of the test waters, the ph of the water, the water temperature, and the time of day when the watershed metrics were measured. For the analysis in the rest of this section, the dataset is split 80%/20% for training/testing purposes. The statistical programming language R (R Core Team 2020) is used to run this model and present the findings from it.

Equation (1) represents the model used in this paper

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \tag{1}$$

In Equation (1):

- Y the dependent variable, is the dissolved oxygen per liter of water.
- β_0 is the coefficient for intercept.
- X_1 is the covariate of salinity of water.
- X_2 is the covariate of PH value of water.
- X_3 is the covariate of temperature of the water.
- X_4 is the covariate of the year after 1999.
- $\beta_1\beta_2, \beta_3, \beta_4$ are the coefficients of the variables.

The general assumption behind this modeling is that, holding other variables constant, the average oxygen content has decreased on average after 1999 compared to before. In addition, Salinity has a negative effect on oxygen content, with an increase in Salinity leading to a decrease in oxygen content, holding other variables constant. This may be due to problems such as the destruction of the natural environment such as biological invasions that occurred in 1999, or it may be the result of illegal emissions from anthropogenic chemical plants. There

may also be a link between changes in temperature and dissolved oxygen content, given that the measurement of dissolved oxygen content is based on the temperature of the water. The model therefore attempts to explore whether there is indeed a statistically significant link between the dissolved oxygen content of the waters and the above factors.

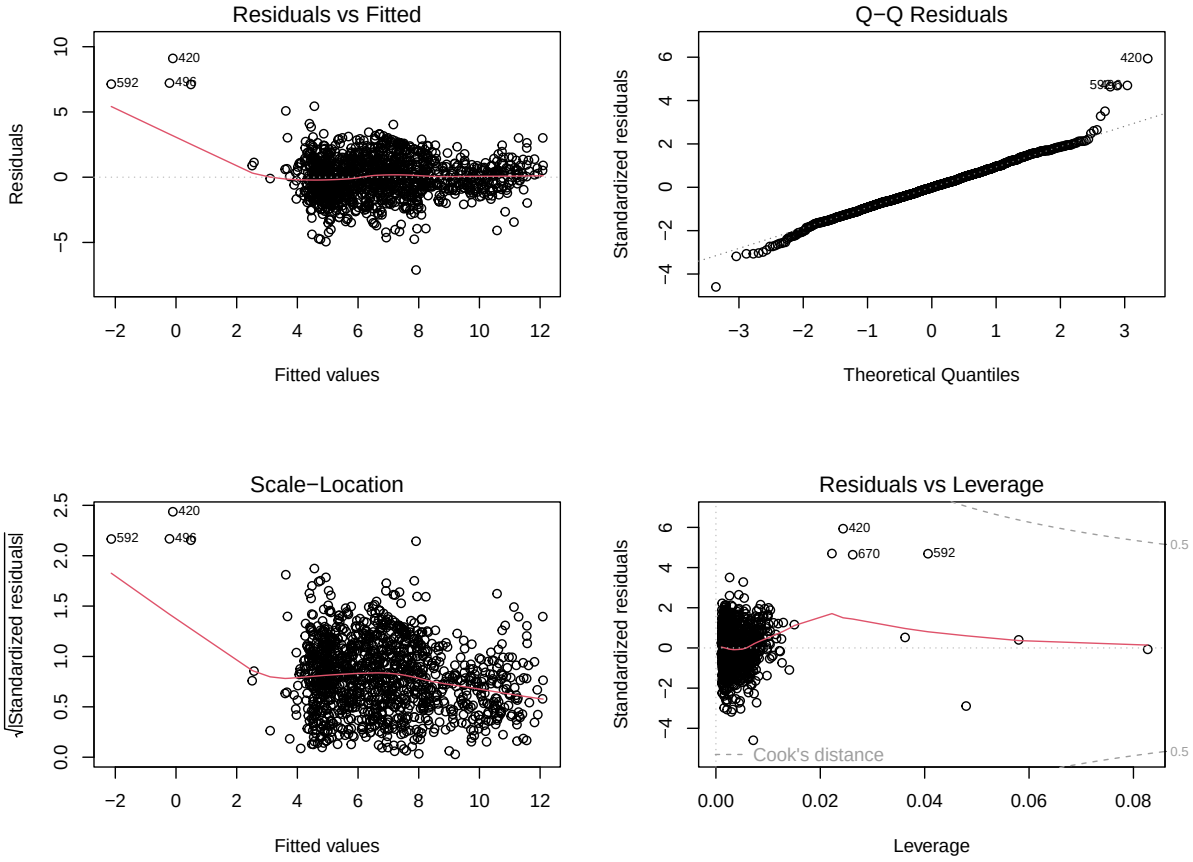

Figure 5: Model Diagnostic Plots

The diagnostic plots of the model are shown in Figure 5 The fitted values vs. residuals graph (top-left) shows a roughly horizontal line, which indicated a linear relationship between the dependent and independent variable. Ideally, the residuals should be scattered randomly around the horizontal line at zero, with no discernible pattern. A clear pattern, as we see with the red line curving, could indicate issues with linearity or equal variance.

Normal Q-Q (Quantile-Quantile) Plot: This plot checks the assumption that the residuals are normally distributed. The points should fall approximately along the 45-degree line. The figure shows that most of the points fall on the dashed line, but a significant number of points fall on the right tail of the dashed line, indicating that the residuals may not be perfectly normally distributed.

Scale-Location Plot: This plot, similar to the first, is used to check homoscedasticity. The

vertical spread of the points should be roughly the same across all levels of fitted values. If the spread increases or decreases with the fitted values, it suggests non-constant variance. The red line is not horizontal; a slope suggests issues with homoscedasticity.

Residuals vs Leverage: This plot helps to identify influential observations that might have a significant impact on the model's predictions. Points that stand out from the crowd, especially if they are outside the dashed Cook's distance lines, are worth checking as they might be unduly influencing the model. Note that in all four plots above, points 420, 496, and 592 are labeled as possible outliers, indicating that their inclusion would have a significant impact on the linear model.

4 Results

4.1 Model Summary

Table 2: Model Summary of Dissolved Oxygen Based on Salinity, PH, Water.Temp and After year 1999

term	estimate	std.error	statistic
(Intercept)	11.21	0.43	26.14
Salinity	−0.20	0.04	−4.47
pH	0.18	0.06	3.10
Water.Temp	−0.14	0.01	−26.91
year_1999	−3.81	0.14	−26.48

Table 2 shows the coefficients for the model predictor variables. We are concerned with the upper half of the table because it provides coefficient values representing the intercept and coefficient values representing the factors affecting the dissolved oxygen content of the water. The standard errors of the estimated coefficients are also included in parentheses. The degree of error in the predictions is shown in Figure 5. The time period explored in this model is from 1999 to 2019. It is worth noting that we specify the estimated relationship between the independent and dependent variables. Therefore, we combine the estimated relationship with the available data to estimate the coefficients of the model.

The coefficient estimates shown in Table Table 2 for the final model is equivalent to the following, shown in Equation (2):

$$\hat{Y}_i = 11.21 - 0.20X_{1i} + 0.18X_{2i} - 0.14X_{3i} - 3.81X_{4i} \tag{2}$$

Table 3: Summary of Corrected Total Oxygen Reading (CTOR) Across Various Sites and Years

Site ID	Year	Total Oxygen	Unique Oxygen Values	Tokens	CTOR
A	1999	11.8	3	3	1.2
B	1999	9.2	2	2	1.0
Bay	1990	298.4	21	32	2.6
C	1999	16.6	3	4	1.1
D	1999	28.4	5	5	1.6

Table 3 offers a detailed quantification of the Corrected Total Oxygen Readings (CTOR) for a select year, providing insights into the oxygen variability at different aquatic sites. For instance, in 1999, Site A recorded a total oxygen amount of 11.8 units, with 3 unique measurements (indicated by "Unique Oxygen Values"), out of 3 readings (indicated by "Tokens"), resulting in a CTOR value of 1.2. Contrastingly, the Bay area in 1990 exhibited a substantially higher oxygen content of 298.4 units, with 21 unique values from 32 readings, culminating in a CTOR of 2.6, indicating a greater diversity in oxygen measurements for that year and site.This could reflect a more dynamic ecosystem or varying water quality management practices.

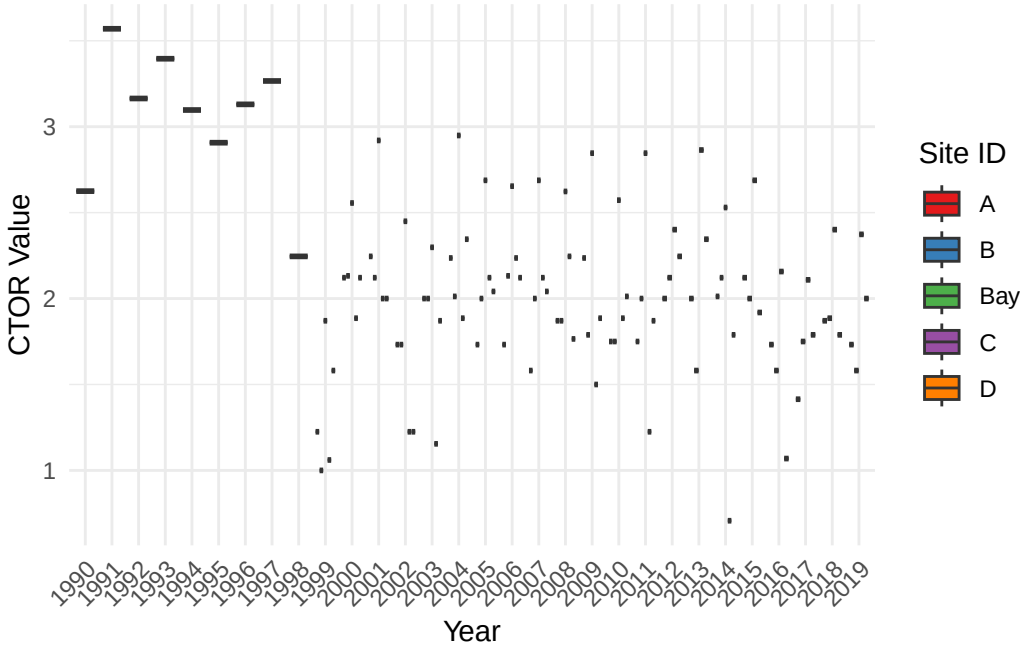

Figure 6: Corrected Total Oxygen Reading (CTOR) from 1989 to 2019

Figure 6 displays a scatter plot illustrating the distribution of Corrected Total Oxygen Readings (CTOR) from 1989 to 2019 across different sites. Each site is represented by a unique

color, and the plot points are dispersed along the y-axis to reflect the CTOR value for each year. The data points suggest variability in oxygen content over the years, with some sites showing wider fluctuations, evident from the vertical spread of points, while others exhibit a more consistent CTOR value across the observed timeframe. It revealing significant inter-annual variability and potential site-specific environmental influences. Sites such as the Bay exhibit peaks in oxygen diversity, which could be attributable to factors such as seasonal biological activity or watershed inputs. The temporal spread of the data points across the years underscores the importance of continual monitoring to understand the ecological health of these sites.

When synthesizing the data from Table 3 and Figure 6, the narrative of water quality over time and across different sites becomes apparent. The numerical data in Table 3, when viewed alongside the trends in Figure 6, suggest that oxygen levels, a proxy for water quality, are subject to fluctuations potentially linked to environmental changes or human interventions. The Bay area, with its higher CTOR values, might be indicative of a robust and varied aquatic environment, or it could reflect the impact of episodic events that warrant further investigation. The integration of these data sources provides a powerful tool for researchers to identify patterns and hypothesize about underlying causes, which can inform future studies and environmental policies.

5 Discussion

This study provides a new perspective on water quality improvement and environmental sustainability. Quantitative analysis and integration of data confirmed that the amount of dissolved oxygen in water determines water quality. By analyzing the data obtained from the U.S. Department of the Interior, it provides help and new perspectives for environmental sustainability and social policy making.

5.1 Findings

The findings of this investigation corroborate the proposed hypothesis, delineating a significant association between the Corrected Total Oxygen Reading (CTOR) and aquatic health, notwithstanding some anomalies in the dataset. A comparative analysis of oxygen levels across various bodies of water revealed that the bay exhibited markedly higher oxygen content than its counterparts. Following closely, the oxygen levels in the fishing pond were also substantial, potentially suggesting an augmentation due to biotic activity. Seasonal trends observed over the three-decade span highlight a diminution in dissolved oxygen during the summer months, aligning with typical seasonal patterns. A point of particular concern was identified in 1999, where a pronounced decline in CTOR suggested a critical disruption in the water bodies, from which recovery was not observed until the close of the study period in 2019. These insights

are pivotal in understanding the temporal and fluctuations of water quality and underscore the need for continuous monitoring and preservation of aquatic ecosystems.

5.2 Sustainable Development and Environmental Impacts

Dissolved oxygen (DO) is a crucial indicator of water quality and a cornerstone of aquatic ecosystem health, integral to sustainable development and minimizing environmental impacts. Optimal DO levels support diverse aquatic life, aiding in achieving the sustainable management of water resources and the conservation of life below water, as advocated by the United Nations Sustainable Development Goals. Diminished DO levels, often caused by pollution from industrial, agricultural, and urban sources, can lead to ecological imbalances, disrupting food chains and impacting fisheries, which are vital for economic and community sustenance. Fisheries have rarely been 'sustainable' (Pauly et al. 2002). Efforts to sustain or improve DO concentrations are essential in maintaining the integrity of aquatic habitats and ensuring their resilience against environmental disturbances. The choice of pollution control instrument is a crucial environmental policy decision (Goulder and Parry 2008). Such initiatives include enforcing stringent waste disposal regulations, promoting green infrastructure, and preserving natural water purification systems like wetlands. These measures not only enhance DO levels but also contribute to broader environmental benefits, including improved terrestrial biodiversity and public health. Effective management of DO is, therefore, a key component of environmental policy, aligning human activities with ecological preservation to foster a sustainable and prosperous future.

5.3 Weaknesses and Future Research Directions

One limitation of our study is that there are unmeasured variables that can affect dissolved oxygen in waters, such as biological invasions or industrial discharges, that are not fully reflected in our analysis. Also, although the CTOR formula was used, it does not give a good indication of the potential factors that contribute to good or bad water quality. Pollution of waters from chemicals or radioactive substances cannot be well represented in the CTOR. These factors may include the fact that this study did not delve into the trace element content of the water and the abundance of organisms in the water.

Future research should aim to further unravel the complex interactions between dissolved oxygen levels and water quality, or alternatively, the data could be obtained by employing more fine-grained measurements for better processing and analysis. In addition, measurement of trace element content in water could provide a more detailed understanding of the factors influencing water quality. Understanding these factors is essential to fully analyze and understand water quality in the watershed.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Aston, RJ. 1973. "Tubificids and Water Quality: A Review." *Environmental Pollution (1970)* 5 (1): 1–10.

Brundtland, Gro Harlem. 1987. "What Is Sustainable Development." *Our Common Future* 8 (9).

Goulder, Lawrence H, and Ian WH Parry. 2008. "Instrument Choice in Environmental Policy." *Review of Environmental Economics and Policy*.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Huang, Ruixing, Chengxue Ma, Jun Ma, Xiaoliu Huangfu, and Qiang He. 2021. "Machine Learning in Natural and Engineered Water Systems." *Water Research* 205: 117666.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Olyaie, Ehsan, Hamid Zare Abyaneh, and Ali Danandeh Mehr. 2017. "A Comparative Analysis Among Computational Intelligence Techniques for Dissolved Oxygen Prediction in Delaware River." *Geoscience Frontiers* 8 (3): 517–27.

Pauly, Daniel, Villy Christensen, Sylvie Guénette, Tony J Pitcher, U Rashid Sumaila, Carl J Walters, Reg Watson, and Dirk Zeller. 2002. "Towards Sustainability in World Fisheries." *Nature* 418 (6898): 689–95.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. *GGally: Extension to 'Ggplot2'*. https://ggobi.github.io/ggally/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.