[*]

Wentao Sun

April 17, 2024

Table of contents

[*] Code and data are available at:

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.

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

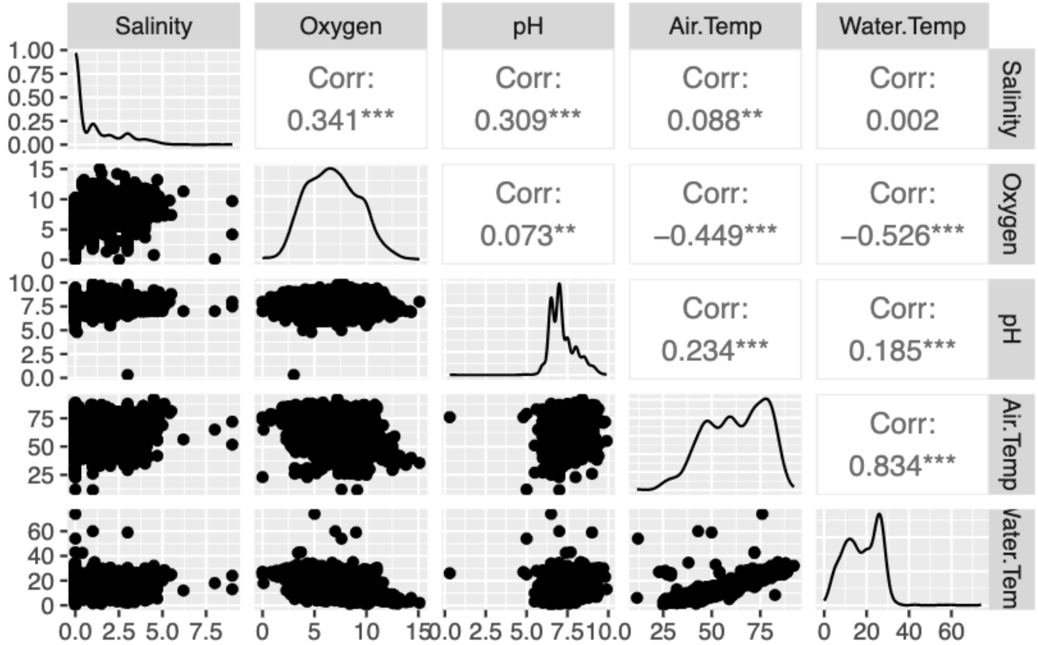

Figure 1: Correlation Matrix of Water Quality Indicators from 1989 to 2019

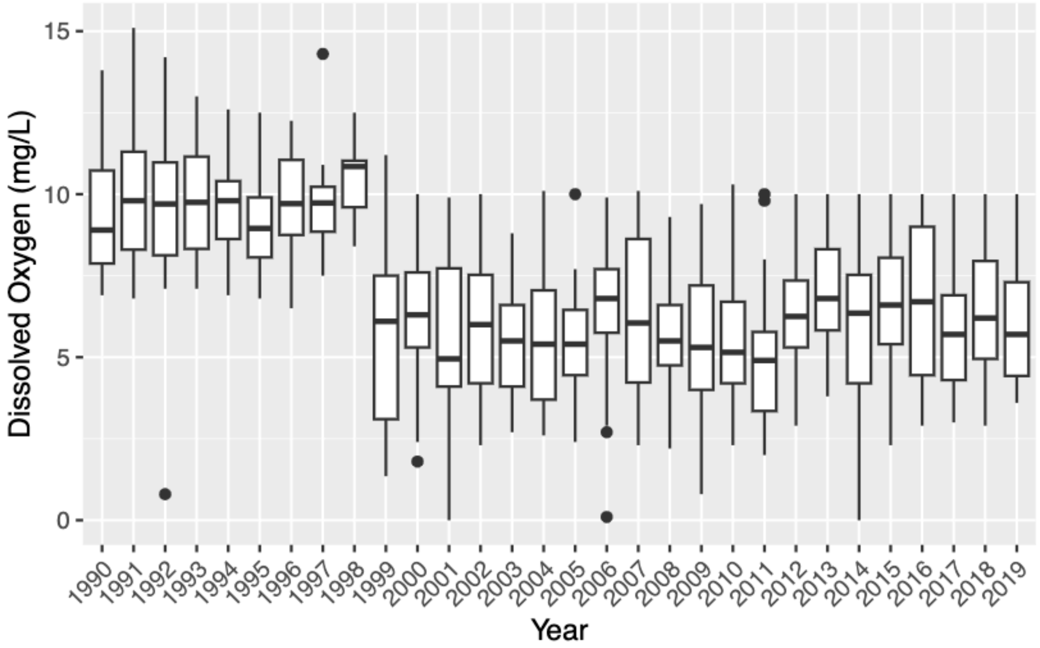

Figure 2: Annual Variability of Dissolved Oxygen Concentrations from 1989 to 2019

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.

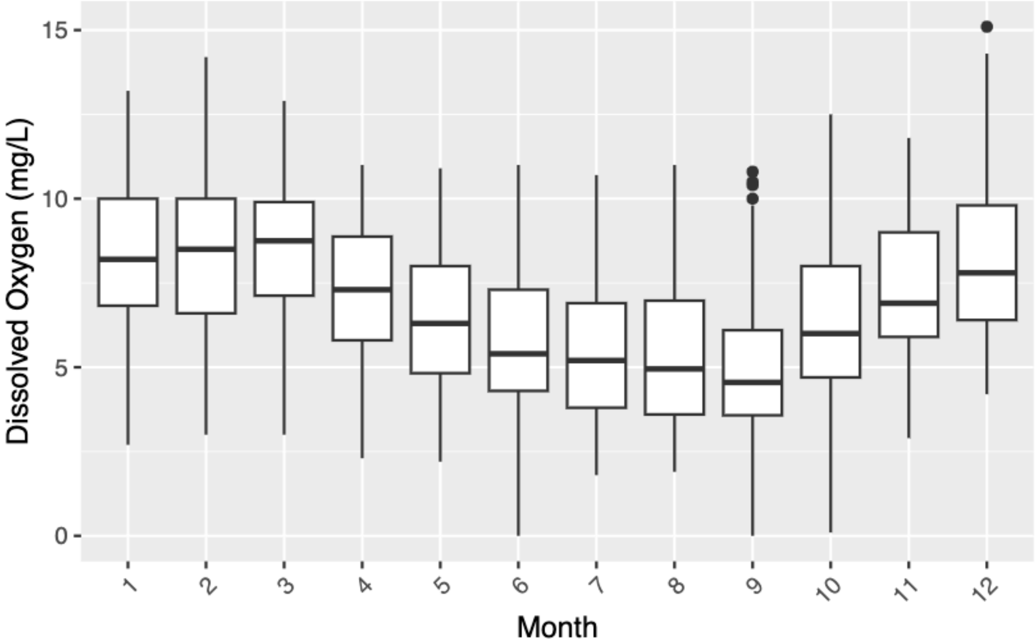

Figure 3: Seasonal Variation in Dissolved Oxygen Concentration by Month from 1989 to 2019

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 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 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.

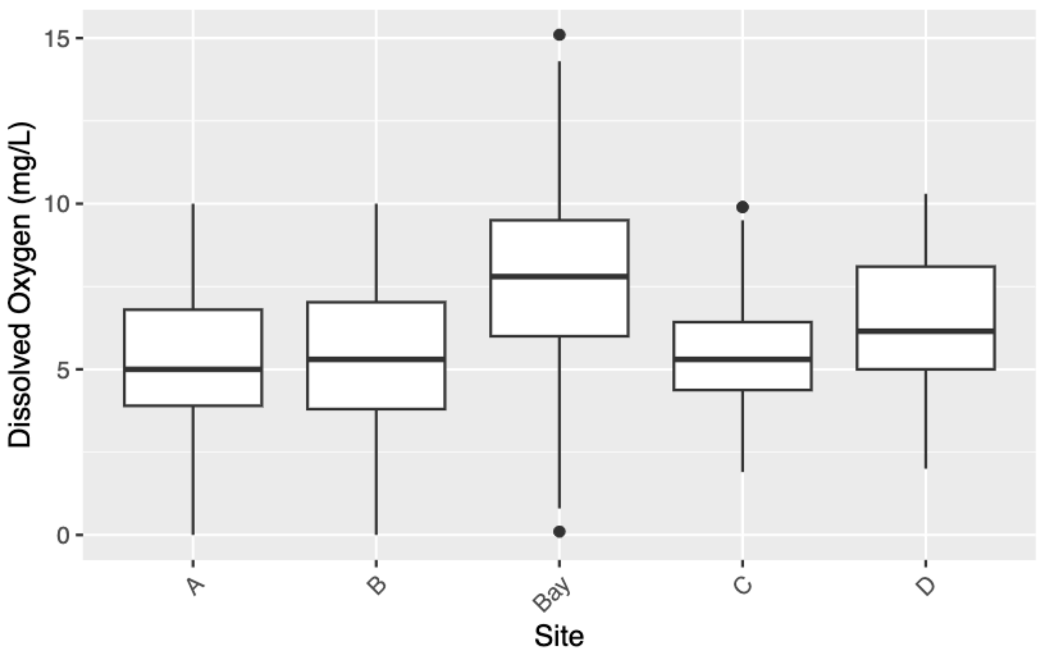

Figure 4: Distribution of Dissolved Oxygen Levels Across Sampling Sites from 1989 to 2019

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

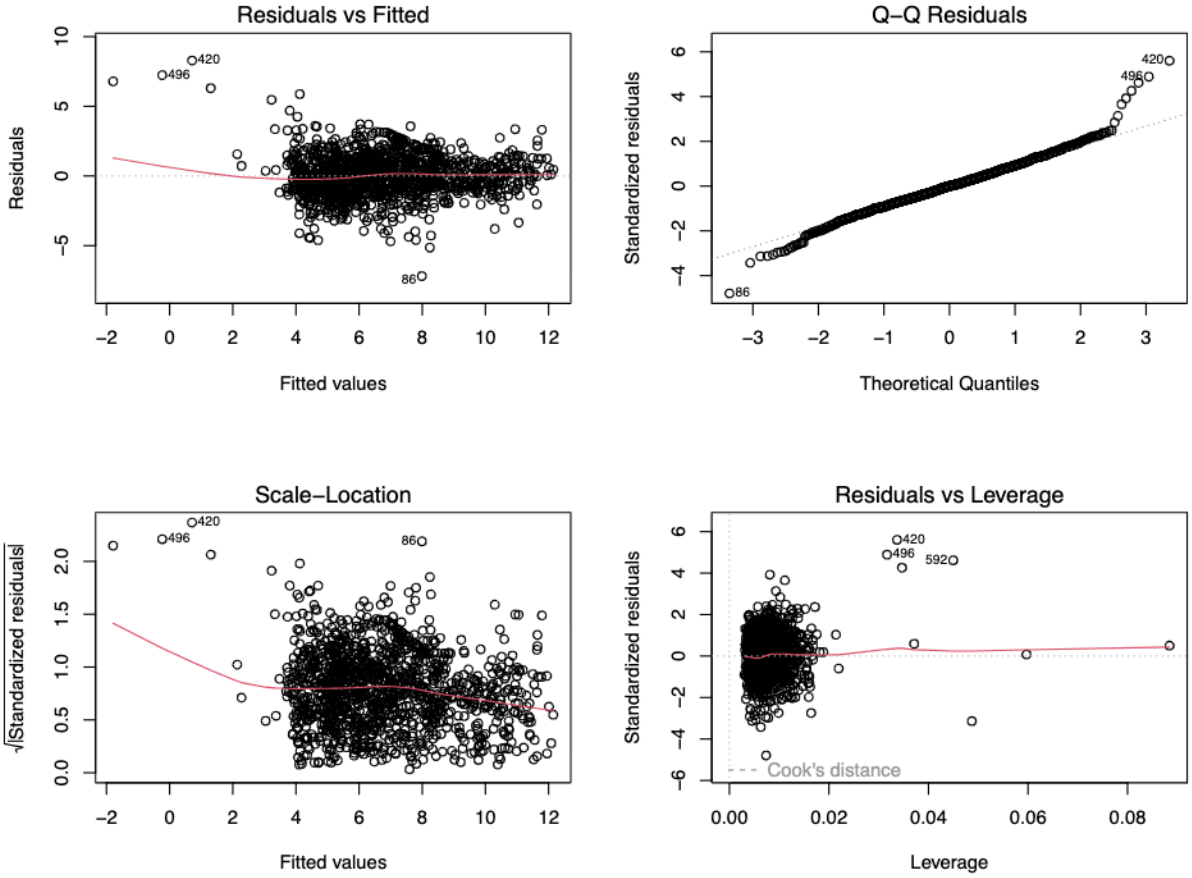

Figure 5: Model Diagnostic Plots

upload

4 Results

Table 2: 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

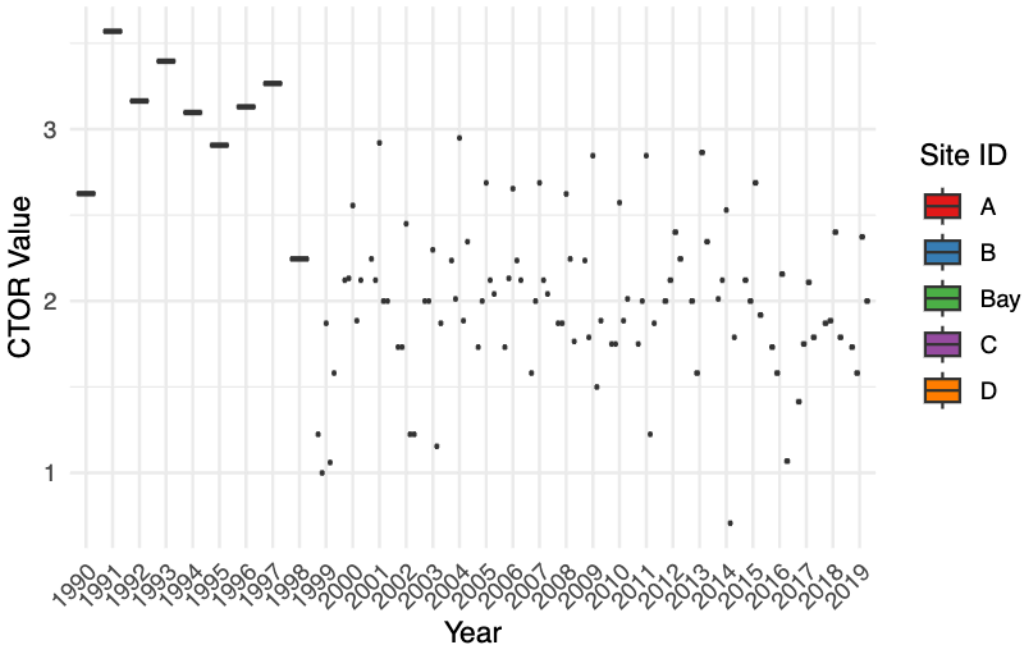

Figure 6: Corrected Total Oxygen Reading (CTOR) from 1989 to 2019

5 Discussion

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).

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.

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.