Research on evaluating water pollution determinants using multiple logistic regression

Ruoyan Shi[1*]

[1]Faulty of Arts and Science, University of Toronto, Toronto, M5S 1A1, Canada

Abstract. Water pollution is a pivotal challenge, underpinning urgent conversations around environmental sustainability, public health, and ecosystem viability. This research aims to assess the degree of water pollution, dissect and understand the myriad factors contributing to it, and pave the way for formulating effective mitigation strategies and policies to preserve the integrity of water bodies worldwide. It highlights that rapid industrialization, population growth, and agriculture cause pollution. Industrial activities release pollutants like heavy metals, while agriculture contributes through runoff. Urbanization also exacerbates the problem. The study uses a dataset from Kaggle and selects variables like aluminium, ammonia, etc. A multiple logistic regression model analyses factors affecting water potability. Results show that aluminium, chloramine, and ammonia positively correlate with potability, while uranium and barium have negative ones. Interaction terms added to the model improve its fit. The study emphasizes understanding individual contaminants and their interactions for effective water management strategies. Accounting for these interactions enables a more comprehensive understanding of the factors affecting water safety. These insights are crucial for developing targeted and effective water management strategies that ensure safe drinking water and support public health.

1 Introduction

Water pollution is a pivotal challenge, underpinning urgent conversations around environmental sustainability, public health, and ecosystem viability. This pervasive issue, rooted in a nexus of anthropogenic activities, prompts a critical examination of its multifaceted causes and far-reaching consequences [1]. This research aims to assess the degree of water pollution, dissect and understand the myriad factors contributing to it, and pave the way for formulating effective mitigation strategies and policies to preserve the integrity of water bodies worldwide.

At the outset, it is imperative to recognize that water pollution epitomizes a significant 21st-century dilemma exacerbated by rapid industrialization, burgeoning population growth, and extensive agricultural practices. These elements collectively strain existing water resources and overwhelm the natural regenerative capacity of water bodies [2]. For instance, the unchecked discharge of industrial effluents and agricultural runoff introduces a spectrum

* Corresponding author: Ruoyan.shi@mail.utoronto.ca

of pollutants-from heavy metals to synthetic chemicals-into aquatic systems, compromising water quality and posing a grave threat to biological diversity and public health [3].

Industrial activities release pollutants such as heavy metals, organic compounds, and endocrine-disrupting chemicals into waterways, elevating the risk of acidification and eutrophication [3]. Acidification, a byproduct of sulfur and nitrogen emissions from fossil fuel combustion, has been shown to alter the ecological dynamics of aquatic environments, leading to the decline of fish populations and water quality degradation [4]. Similarly, eutrophication, fuelled by the excessive input of nutrients like nitrogen and phosphorus, triggers algal blooms that reduce oxygen levels, harm aquatic life, and degrade water bodies' aesthetic and recreational value.

Agriculture, another principal actor in this environmental drama, contributes to water pollution through the runoff of pesticides, herbicides, and fertilizers. These chemicals poison aquatic life and percolate through the ecosystem, affecting the health of terrestrial species and contaminating the human food chain [5]. Furthermore, the expansion of agricultural land often leads to soil erosion, another source of water pollution, which diminishes the natural filtration ability of the soil and increases the sediment load in rivers and lakes.

Urbanization and population growth also play critical roles in exacerbating water pollution. The proliferation of impervious surfaces in urban areas prevents natural water absorption by the soil, leading to increased runoff and pollutant loads in urban waterways [6]. Moreover, inadequate or outdated wastewater treatment facilities in many growing cities must cope with the increased sewage volumes, resulting in the direct discharge of untreated or partially treated effluents into nearby water bodies [7].

On the health front, the repercussions of water pollution are profound and far-reaching. Contaminated water sources lead to an array of diseases, including diarrhoea, skin conditions, and more severe illnesses such as cancer and neurological disorders [8]. Communities near polluted water bodies, such as those around the Turag River in Bangladesh, frequently suffer from exacerbated health issues, underscoring the urgent need for improved water management and sanitation infrastructure.

Water pollution is an environmental problem and a multifaceted challenge affecting all aspects of life. Recognizing the complexity of water pollution requires a holistic approach to its research and management that combines scientific, regulatory and community perspectives [9]. This study aims to delve into these interactions using relevant theories, models, and empirical data to assess the extent of water pollution, identify and prioritize pollutants and affected ecosystems, and support the development of effective mitigation strategies.

Through comprehensive research and concerted global efforts, it is possible to manage and mitigate the effects of water pollution. This research is a step towards understanding the key factors contributing to water pollution. It aims to support the development of strategies to ensure the planet's and its inhabitants' health for generations to come. This study will provide valuable insights into the development of strong evidence-based policies that can mitigate water pollution and promote a sustainable future.

This study is of great significance for understanding and dealing with water pollution. First, by deeply exploring the multiple influencing factors of water pollution, this paper can better identify the main pollution sources and key influencing processes and provide a scientific basis for formulating more effective water quality management strategies and pollution control measures [10]. Given water pollution's complexity and regional nature, this comprehensive study will help policymakers and environmentalists tailor locality - and goal-specific solutions to address specific problems in different Settings. Secondly, this study enhances the understanding of the relationship between environmental degradation and human health by analyzing the impact of water pollution on public health. Water pollution is directly related to human health and quality of life. Therefore, improving water quality and

ensuring the safety of drinking water is an important issue in the field of public health. This study will provide important theoretical support for improving water resources management and protecting public health. In addition, with the intensification of global climate change and population growth, the pressure on water resources is increasing, so it is more urgent to study the causes and consequences of water pollution. The study can contribute to a better understanding the global water crisis and promote concerted international action on sustainable water resources management and conservation.

2 Methods

2.1 Data sources

The dataset used in this paper is fetched from the Kaggle website (Water Quality- Dataset for water quality classification). The original dataset contains water quality metrics for 8,000 different water bodies. Through data cleaning, 7,997 samples will be used for analysis in the paper. The dataset remained in .csv format.

2.2 Variable selection

The data in this research contains 7,997 different water samples collected to evaluate water quality and its suitability for human consumption.

Table 1. List of Variables

Variable	Logogram	Meaning
aluminium	x1	Level od aluminium in water per liter
ammonia	x2	Level od ammonia in water per liter
arsenic	x3	Level od arsenic in water per liter
barium	x4	Level od barium in water per liter
cadmium	x5	Level od cadmium in water per liter
chloramine	x6	Level od chloramine in water per liter
chromium	x7	Level od chromium in water per liter
copper	x8	Level od copper in water per liter
flouride	x9	Level od flouride in water per liter
bacteria	x10	Level od uranium in water per liter
viruses	x11	Level od viruses in water per liter
lead	x12	Level od lead in water per liter
nitrates	x13	Level od nitrates in water per liter
nitrites	x14	Level od nitrites in water per liter
mercury	x15	Level od mercury in water per liter
perchlorate	x16	Level od perchlorate in water per liter
radium	x17	Level od radium in water per liter
selenium	x18	Level od selenium in water per liter
silver	x19	Level od silver in water per liter
uranium	x20	Level od uranium in water per liter
potability	Y	Indicate if water is safe for human consumption

Among the variables included are the water's physical, chemical, and biological characteristics. The data consists of 20 variables (aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites,

mercury, perchlorate, radium, selenium, silver, uranium), each representing a different aspect of water quality and one dependent variable (potability). These variables were chosen for their relevance in determining the potability of water, which is crucial for public health. The specific description of this dataset is shown in Table 1.

2.3 Modal selection

This paper uses a multiple logistic regression model to analyze the factors affecting water potability. Logistic regression is chosen due to the binary nature of the dependent variable, which indicates whether the water is potable (1) or not potable (0). The purpose of this section is to determine the significance of the various independent variables, including aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium, on the potability of water. This paper aims to identify the key determinants contributing to water safety by doing so.

The multiple logistic regression model is a statistical model that estimates the probability of a binary outcome based on multiple explanatory variables. It uses the maximum likelihood estimation (MLE) technique to estimate the model's parameters, ensuring that the predicted probabilities are between 0 and 1. Logistic regression is particularly effective in evaluating the relationship between a binary response variable and multiple predictors, making it well-suited for this study.

By utilizing logistic regression, the objective is to understand how changes in each independent variable influence the likelihood of water being classified as potable. The model results will be analyzed to assess the strength and direction of the relationship between each variable and potability. This analysis will provide valuable insights into the significant factors affecting water quality and support the development of effective water management strategies.

3 Results and discussion

3.1 Logistic regression analysis

The analysis in this paper the relevance and significance of the independent variables, including aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium, in predicting water safety.

Figure 1 shows the relevance analysis between the dependent and independent variables. Aluminium shows the highest positive correlation with potability, while Chloramine and Ammonia also show moderate positive correlations. On the other hand, Uranium and Barium have negative correlations with potability, indicating that higher levels of these variables may reduce water safety. These correlations suggest that aluminium, chloramine, and ammonia are more significant in determining water potability than other variables.

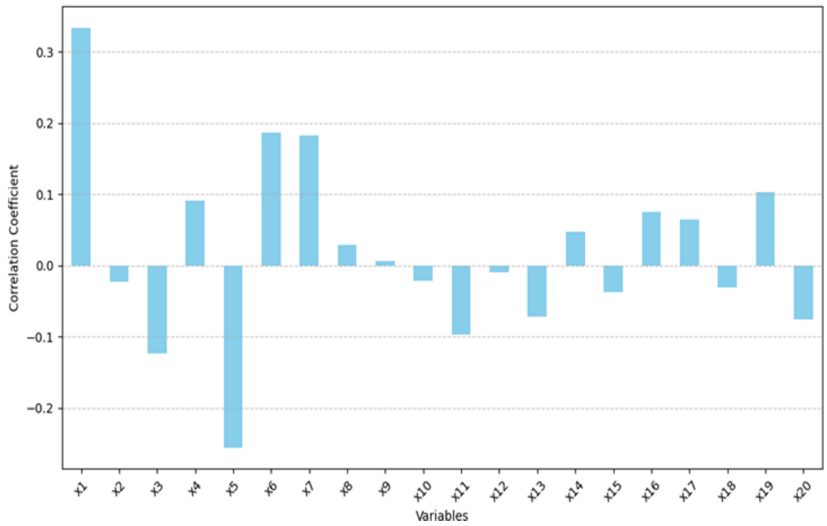

Fig. 1. Relevance Analysis Between Dependent and Independent Variables

Table 2. Regression coefficient table

	Coefficient	SE	P-value	Odds Ratio	VIF
const	0.660	0.234	0.005	1.936	31.575
x1	0.715	0.032	<0.001	2.044	1.351
x2	-0.025	0.005	<0.001	0.976	1.052
x3	-3.046	0.309	<0.001	0.048	1.618
x4	0.121	0.040	0.002	1.129	1.496
x5	-20.641	1.783	<0.001	<0.001	1.332
x6	0.179	0.020	<0.001	1.196	2.021
x7	1.226	0.179	<0.001	3.409	1.780
x8	-0.378	0.071	<0.001	0.686	1.111
x9	0.129	0.097	0.184	1.138	1.005
x10	0.772	0.214	<0.001	2.165	2.085
x11	-1.255	0.183	<0.001	0.285	1.871
x12	-1.605	0.746	0.031	0.201	1.036
x13	-0.050	0.008	<0.001	0.951	1.009
x14	-0.312	0.098	0.002	0.732	1.516
x15	-36.236	13.949	0.009	<0.001	1.008
x16	-0.026	0.003	<0.001	0.975	1.900
x17	-0.056	0.020	0.005	0.946	1.285
x18	-4.839	1.469	0.001	0.008	1.014
x19	-1.320	0.345	<0.001	0.267	1.715
x20	-12.926	1.605	<0.001	<0.001	1.007

Table 2 shows that the logistic regression model indicates several important findings regarding water potability. This step allows us to quantify the strength and direction of the relationships previously highlighted in Figure 1. The intercept has a coefficient of 0.6604, representing the baseline log odds when all predictor variables are zero, and its small p-value (0.0048) confirms its statistical significance.

Several predictors also exhibit statistically significant effects ($p < 0.05$) on water safety. Notable among these is aluminium, with a coefficient of 0.7148 and a very low p-value (2.2223e-108). This positive coefficient suggests that higher levels of aluminium are associated with an increased likelihood of water being safe, supported by an odds ratio of

2.043874, indicating approximately double the odds of safety. Cadmium and mercury exhibit large negative coefficients (-20.6412 and -36.2361, respectively) and significant p-values, suggesting a strong negative impact on water safety. Cadmium's extremely low odds ratio reinforces its adverse influence. Similarly, mercury's odds ratio is low, indicating a significantly negative effect on safety. In contrast, variables like fluoride (p-value: 0.1838) show higher p-values, meaning there is insufficient evidence to conclude that they significantly impact water potability.

The odds ratios help quantify each predictor's effect; an odds ratio greater than 1 implies increased odds of safety, while an odds ratio less than 1 implies reduced odds. Notably, chromium has an odds ratio of 3.4089, suggesting a significant positive effect on water safety, while cadmium and mercury have very low odds ratios, aligning with their harmful impacts.

The Receiver Operating Characteristic (ROC) curve evaluates the logistic regression model's performance. After determining the key predictors and their effects in Table 2, the ROC curve is used to assess the model's ability to effectively classify water as potable or non-potable.

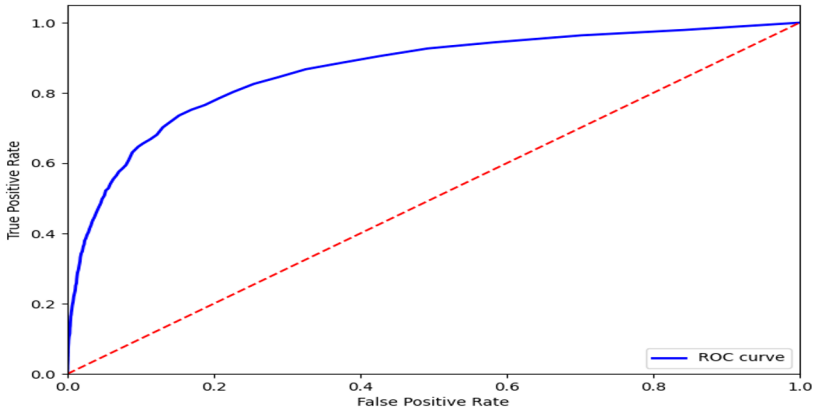

Fig. 2. Receiver Operating Characteristic (ROC) Curve

In Figure 2, the ROC curve indicates that the model has a good discriminatory ability, with the curve well above the diagonal line, suggesting a good balance between true positive rate and false positive rate. This implies that the logistic regression model can effectively distinguish between potable and non-potable water with this dataset.

3.2 Multiple linear regression with interaction terms

The analysis of the VIF values presented in Table 2 provides the foundation for exploring potential relationships among the variables. The VIF values show that all variables have VIF values below 10, indicating no severe multicollinearity in the data. The variables bacteria (VIF = 2.0852), chloramine (VIF = 2.0210), and perchlorate (VIF = 1.8998) have relatively higher VIF values, suggesting that there may be some relationships among these variables worth further exploration, such as potential interaction effects.

The next analysis phase involves adding interaction terms to the model, as shown in Table 3. The interaction terms (`x6x10`, `x6x16`, `x10x16`) represent the combined effects of chloramine and bacteria, bacteria and perchlorate, and chloramine and perchlorate, respectively.

Table 3. Multiple Linear Regression Model analysis results with interaction terms

	Coefficient	Standard Error	P-value
const	-2.521	0.120	<0.001
x_1	0.514	0.031	<0.001
x_6	0.312	0.033	<0.001
x_2	-0.029	0.005	<0.001
x_{10}	-1.213	0.284	<0.001
x_{16}	0.018	0.006	0.001
x_6x_{10}	0.190	0.057	0.001
x_6x_{16}	0.022	0.008	0.009
$x_{10}x_{16}$	-0.012	0.001	<0.001

From Table 3, the regression results show that most of the predictors, including the interaction terms (x6x10, x6x16, x10x16), have p-values less than 0.05, indicating statistical significance. It also indicates that interaction terms such as x6x10 (chloramine and bacteria) and x6x16 (bacteria and perchlorate) positively influence potability, suggesting a beneficial combined effect. Conversely, x10x16 (chloramine and perchlorate) has a negative coefficient, indicating a compounded adverse impact when both variables are high. These results highlight that specific combinations of pollutants may significantly alter water quality outcomes.

Table 4. Model Comparison

	AIC	BIC	Log-Likelihood
Original model	4894.7185	4936.6387	-2441.3593
Updated model	4720.7263	4783.6066	-2351.3632

After adding these interaction terms, the original and updated models are compared. The original model includes the variables aluminium, chloramine, ammonia, bacteria, and perchlorate, while the updated model includes these variables plus the interaction terms (x6x10, x6x16, x10x16). Table 4 highlights this comparison, with metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Log-Likelihood used to evaluate the models' effectiveness. The results show that the updated model, including the interaction terms, has a better fit than the original model. The lower AIC and BIC values and the higher Log-Likelihood suggest that including interaction terms provides a more accurate data representation, thereby improving the model's predictive power.

Figure 3 shows that the predicted probability of potability increases with higher chloramine levels, and this effect is more pronounced when bacteria levels are low. This suggests that chloramine is effective at increasing potability, but its effectiveness diminishes in the presence of higher bacteria concentrations. Figure 4 shows that the predicted probability of potability increases with higher perchlorate levels, particularly when low bacteria levels. However, higher bacteria levels reduce the positive impact of perchlorate on potability, suggesting that interactions between these pollutants can significantly alter their individual effects on water quality. Figure 5 illustrates that chloramine has a stronger positive effect on potability when perchlorate levels are low. In contrast, the effect is much less pronounced when perchlorate levels are high, indicating that high perchlorate levels may diminish the positive impact of chloramine on water quality.

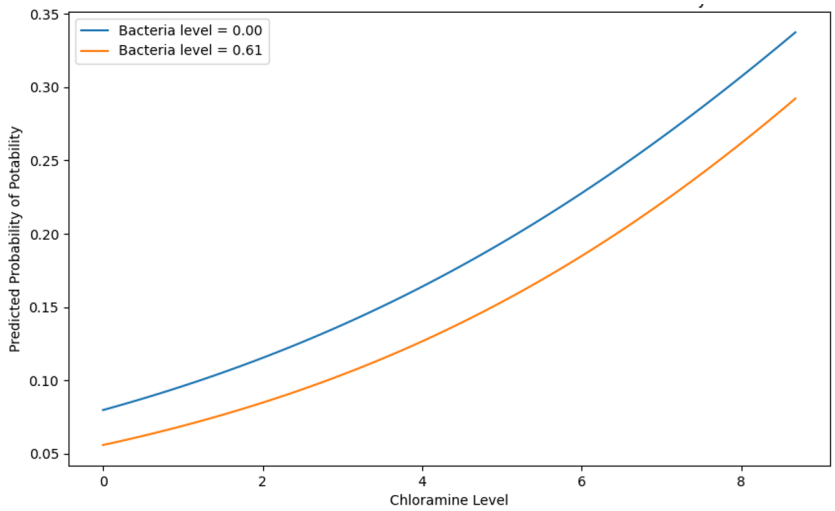

Fig. 3. Interaction plot: Effect of Chloramine and Bacteria on Potability

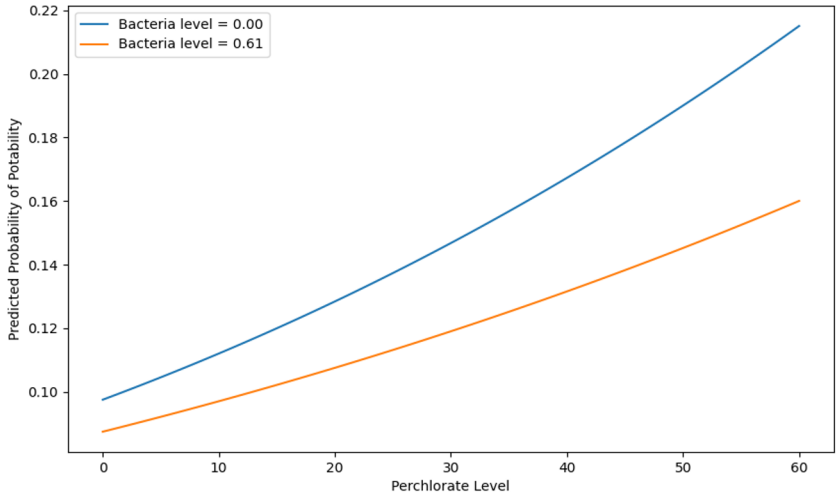

Fig. 4. Interaction Plot: Effect of Perchlorate and Bacteria on Potability

Together, these analyses, starting with VIF assessment, followed by the inclusion of interaction terms, and ending with a comparison of model fit and visualization of interactions, build a comprehensive understanding of the factors affecting water potability. By understanding the individual effects and the combined influences of different pollutants, the study supports the development of more targeted and effective water management strategies.

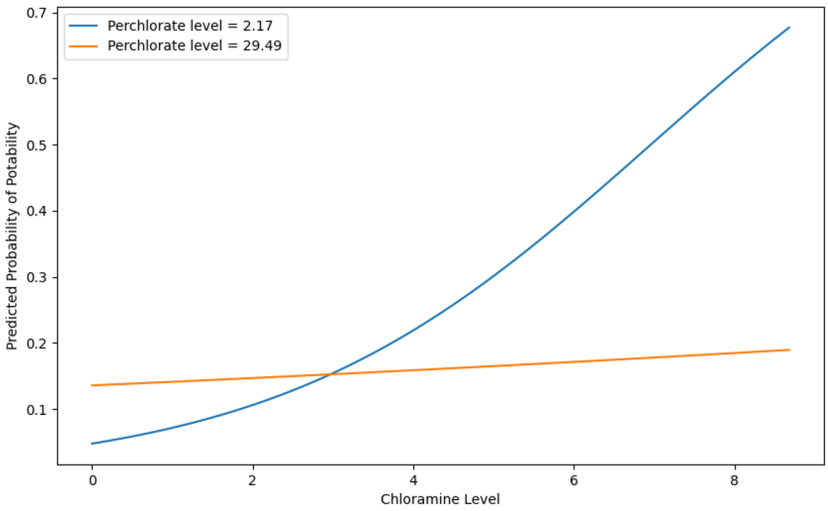

Fig. 5. Interaction Plot: Effect of Chloramine and Perchlorate on Potability

4 Conclusion

This study emphasizes the crucial importance of comprehending individual contaminants and how they interact to determine water potability. According to a multiple logistic regression model, elements like uranium, cadmium, and mercury harm portability. At the same time, certain factors like aluminium, chloramine, and ammonia strongly correlate with water safety. The model's accuracy was increased by including interaction terms, highlighting the significant impact that coupled pollution effects can have on water quality results.

These findings' wider ramifications highlight how urgently tailored water management measures are needed. Water pollution is still a major environmental problem threatening ecosystems, public health, and sustainable development. Diseases including cholera, dysentery, and hepatitis are spread via contaminated water supplies, which disproportionately afflict underprivileged groups. Hazardous chemicals that bioaccumulate in aquatic life can also upset food systems, putting human health and biodiversity at risk through ingestion. Future studies should build on these findings by enlarging the dataset to include more varied geographic locations and water sources to improve the model's generalizability.

In summary, better mitigation techniques are made possible by an awareness of the complex effects of contaminants and how they interact. Water quality protection initiatives can be better informed by considering these intricate relationships, guaranteeing access to clean drinking water and building stronger, healthier communities. An all-encompassing strategy is essential as we work to achieve sustainable water management for future generations and deal with mounting environmental concerns.

References

1. A. K. Dwivedi, Researches in water pollution: A review. Deen Dayal Upadhyay Gorakhpur University **1**, 16-23 (2017).
2. H. Du, X. Ji, X. Chuai, Spatial Differentiation and Influencing Factors of Water Pollution-Intensive Industries in the Yellow River Basin, China. International journal of environmental research and public health **19(1)**, 20-24 (2022).

3. L. Li, H. Yang, X. Xu, Effects of water pollution on human health and disease heterogeneity: A review. Review of Economy and Management **7**, 88-92 (2022).
4. F. N. Chaudhry, M. F. Malik, Factors affecting water pollution: A review. Review of Economy and Management **32**, 168-170 (2017).
5. J. N. Halder, M. N. Islam, Water pollution and its impact on the human health. Review of Economy and Management **16**, 120-126 (2015).
6. Y. Qian, W. Hongru, Regional Disparity and Influencing Factors of Water Pollution Emissions in the Yangtze River Economic Belt: 2004-2014. Review of Economy and Management **9(30)**, 140-145 (2024).
7. L. Håkanson, A. Bryhn, Water pollution - methods and criteria to rank, model and remediate chemical threats to aquatic ecosystems. Nature Water **6**, 148-150 (2008).
8. W. W. Zhang, Measuring the value of water quality improvements in Lake Tai, China. Journal of Zhejiang University-Science A (Applied Physics & Engineering) **9**, 710-719 (2011).
9. M. Shao, X. Xie, E. Chao, et al. Fine-resolution estimation for urban surface water pollution susceptibility with multi-modal earth observation data. International journal of environmental research and public health **6**, 168-170 (2024).
10. E. R. Jones, et al. Sub-Saharan Africa will increasingly become the dominant hotspot of surface water pollution. Nature Water **7**, 1 (2023).