Monte Carlo Uncertainty Analysis and Regression Models on Water Quality Estimation Strait of Tuba, Langkawi

Hasmida Muhamad[1], Ernieza Suhana Mokhtar[1], Muhammad Akmal Roslani[2], Mohammed Oludare Idrees[3], Azlan Abdul Aziz[4], Noraini Nasirun[5]*
[1]Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, MALAYSIA
[2]Faculty of Applied Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, MALAYSIA
[3]Department of Surveying & Geoinformatics, Faculty of Environmental Sciences. University of Ilorin PMB 1515, Ilorin. Kwara State
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, MALAYSIA
[5]Faculty of Business and Management, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, MALAYSIA
**Corresponding author email: ernieza@uitm.edu.my*

ABSTRACT

Received: 31 Mar, 2022
Reviewed 12 June, 2022
Accepted: 1 July, 2022

Preserving and maintaining water quality is essential in providing sustainable development for a nation, particularly for the protection of marine life and human health. In acquiring primary data for water quality parameters (WQP), it is indisputable how prominent the in-situ sampling technique is as opposed to other methods (e.g., satellite imagery, drones). However, the unpredictable condition of the environment and other uncertainties that may exist in the acquired data from in-situ sampling needs to be assessed to ensure its consistency and credibility. Hence, this research aims to identify the best linear regression model for estimating WQP in the Strait of Tuba, Langkawi by i) reducing the existing error in sampling data using Monte Carlo (MC) analysis and ii) determining the best spatial interpolation method to interpolate the WQP. About 71 sampling points were collected in December 2021 situated in Selat River, Langkawi. An uncertainty MC analysis was applied to the sampling data and the mean distribution from the MC analysis result has less error compared to the observed data. The result for the standard deviation of turbidity, salinity, temperature, pH, and dissolved oxygen (DO) are 0.2333, 0.6695, 0.4711, 0.2671, and 1.3230 respectively. For the spatial interpolation model comparison, Inverse Distance Weightage (IDW) spatial interpolation model outperforms kriging in terms of RMSE, Mean Absolute Error (MAE), and standard deviation. Finally, nine linear regression equations were established to predict the WQP. This study will help decision-makers in predicting the WQ by reducing the ground sampling data collection with the proposed linear regressions.

Keywords: *Water Quality, Geospatial Analysis, Spatial Interpolation, uncertainty analysis, regression model*

INTRODUCTION

Water bodies serve as habitats for a wide range of organisms and their response to the stressors may vary among the producers and consumers. Not to mention that it is a key substance in sustaining vital activities of humans such as nutrition, respiration, circulation, excretion, and reproduction (Kılıç, 2020). Over the years, as human populations increase, industrial and agricultural activities expand, and climate change threatens to disrupt the hydrological cycle that thereafter declining water quality has become a global issue of concern. With that, a discussion and measures to address water quality status and other regulating factors, which may affect the animal community in either way have become the main concern

since it could lead to the general loss of biodiversity including loss of fish production. Not only that, a wide range of aquaculture sustainability complications never fail to attract the attention of legislators and environmentalists in pursuing a sustainable life for the sake of future generations (Jayanthi et al., 2021). Hence, this study aims to identify the appropriate linear regression model for estimating WQP in the Strait of Tuba, Langkawi to i) reduce the existing error in sampling data by using MC uncertainty analysis and ii) determine the appropriate spatial interpolation method to interpolate the WQP. The outcome of this study will help the authorities such as the Fisheries Development Authority of Malaysia and even fishermen to estimate the WQP for Selat River as well as be able to indicate the water quality conditions.

LITERATURE REVIEW

Although the in-situ sampling and laboratory testing approach for water quality assessment has many downsides, such as being labour intensive and costly, it provides more accurate data as opposed to other methods (e.g., satellite imagery and drone images). However, it is close to impossible to individually gather the data in every part of a water body, especially in a large-scale area (Gholizadeh et al., 2016). Prior study has used several spatial interpolation models such as Spline, Inverse Distance Weighting (IDW), and Kriging to interpolate topography data in generating landform (Ikechukwu et al., 2017). Ordinary Kriging and IDW interpolation model were used in several previous studies to predict the geographical parameters of the study area to compare the performance of both interpolation methods (Ikechukwu et al., 2017; Borges et al., 2015; Bay et al., 2014). However, there is no consistency in determining which of these two models is better in estimating water quality data.

An in-situ sampling of water quality parameters in large-scale locations imposes the integration of several processes occurring on various time and space scales. This process might cause a few uncertainties (Moreno-Rodenas et al., 2019). This inevitable uncertainty may be present in prediction models and even during data acquisition. Hence, the implementation of uncertainty analysis and the interpretation of the generated results, the definition, recognition, and treatment of these uncertainties are consequently critical (Pappenberger and Beven, 2006; Schellart et al., 2010). Prior study has to use Monte Carlo uncertainty analysis to generate random data combinations in predicting inundated areas and flood depth (Mokhtar et al., 2018)

Regression analysis was done by quite a few prior studies to establish estimation models for predicting spatial information (Wang et al., 2017; Mara et al., 2020; Ghazali et al., 2020). In establishing a regression model, the relationship between each variable must be first identified through correlation analysis of the acquired data. In a previous study, the researchers uses regression models to estimate the concentration of phosphorus around the area with missing observation (Ravichandran and Ramakrishnan, 2007). Although the estimation models for water quality estimation have already been created by many researchers from around the world, some remote areas like the Strait of Tuba, Langkawi were often overlooked.

STUDY AREA

The study area was conducted around the Strait of Tuba, Langkawi Malaysia and is located south of the infamous Langkawi Island in Peninsular (see Figure 1). This river is a habitat for a multi-variation of aquatic animals including fishes, crabs, and phytoplankton which plays a crucial role in maintaining the biological clock of the ecosystem around the study area (Yusuf, 2020). The reason for the selection of this study area is that there is a lack of studies on water quality mapping at Selat River.

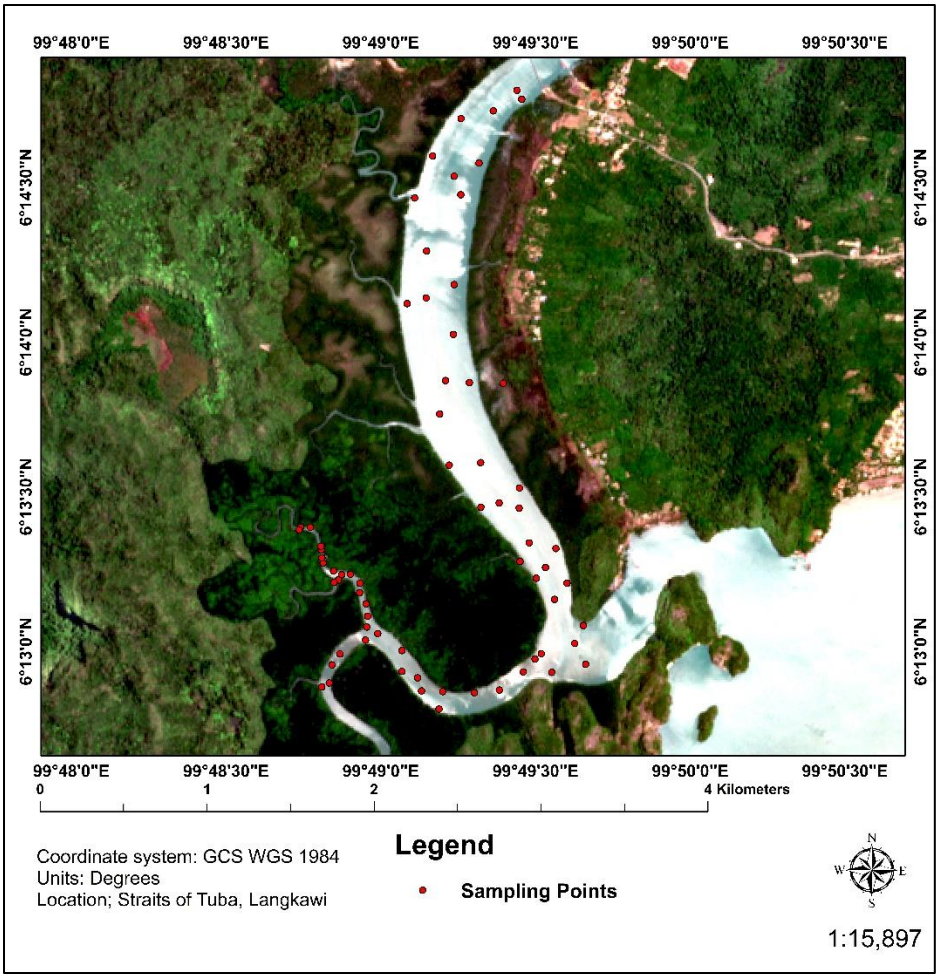

Figure 1: Map of Study Area Located in Strait of Tuba, Langkawi

MATERIALS AND METHODOLOGY

Data Acquisition

In-situ sampling has been carried out to obtain WQP samples abundance at random points along the Selat River in December 2021 during high tides. Water quality parameters such as pH, salinity, temperature, dissolved oxygen (DO), and turbidity were measured by using a pH meter, refractometer, thermometer, DO meter, and Secchi disk respectively (see Figure 2). Each of the WQP readings was taken two times to get the average value. A total of 71 sampling stations (see Figure 1) were collected in December 2021 where the high tides commonly occur around Peninsular Malaysia.

Figure 2: Apparatus and instruments used for data collection (from left Secchi disk, DO meter, and pH meter equipped with temperature detector)

Monte Carlo Uncertainty Analysis

The water quality parameter data has been sorted and analyzed in SPSS for the removal of any outliers and to conduct the descriptive analysis. In this process, three (3) sampling stations were detected as outliers and were removed and sorted out before proceeds to the next step which is now left with 68 sampling data. At this stage, all 68 sampling data have gone through MC uncertainty analysis where its purpose is to eliminate most of the uncertainties that may be present caused by the data collection process, environmental effects, and other possible factors (Farrance and Frenkel, 2016). The WQP data were fed into a Monte Carlo simulation excel, which produced findings in the form of a probability distribution centred on a mean value (Sonnemann et al., 2003). The uncertainty in the water quality datasets was analyzed by generating 8100 randomized values between -20% to 20% of the observed WQP as such was done by previous research (Mokhtar et al., 2018). Then, the mean of its WQP value from this process was recorded as a new dataset for MC. The performance of MC analysis was measured by statistical analysis which can be explained by the standard deviation value (see Table 1).

Spatial Interpolation

Despite having several drawbacks, including being labour-intensive and costly, there is no doubt that in-situ sampling and laboratory testing methodology yield more reliable data than alternative approaches. However, it is virtually impossible to collect the data by hand in every location of a water body, especially in a vast area. Hence, in this paper, spatial interpolation is applied as an alternative approach to avoid labour and cost problems. In the spatial interpolation process, two (2) models of spatial interpolation were tested which are kriging and IDW by using ArcGIS Pro software. These two interpolation models were often used to estimate the spatial parameters near sampling points (Wang et al., 2017; Bay et al., 2014). The majority of sampling points were used for spatial interpolation analysis with a ratio of 7:3 out of 68 sampling points. Considering the ratio, 48 sampling points were used for spatial interpolation while 20 sampling points were used for verification purposes. At the end of this process, the RMSE, MAE, and standard deviation of each dataset were compared side by side (see Table 2).

Correlation Analysis and Establishment of Regressions

By using the value of correlation between WQP, regression models can be established to estimate the value of other WQP by using another WQP. For example, regression has been established to estimate the pH of water using ground sample data of Chlorophyll-a and dissolved oxygen (Zang et al., 2011). In this study, the correlation between pH, salinity, turbidity, temperature, and dissolved oxygen is calculated to establish a regression of WQP estimation. For the data that has a normal distribution, non-parametric (Pearson's) coefficients are recommended (Valentini, Borges, and Muller 2021). Hence, Pearson's

correlation analysis was conducted by using SPSS v20 software to obtain the relationship between WQPs. From that, a total of 9 linear and multivariate linear regressions were established to predict the WQP (see Table 4).

RESULTS AND DISCUSSIONS

A comparison for std dev value was done in previous studies where it compares the standard deviation of the original data and the MC data (Roberto and Couto, 2013; Farrance and Frenkel, 2016). Table 1 shows a descriptive analysis to determine the difference in mean value and standard deviation (std. dev) between observed and MC data. The highest std. dev value for observed data is DO (1.3268 Mg/L) which signifies that as the most uncertain value while the lowest std. dev is turbidity: 0.2333 m which signifies that the data has few uncertainties. The temperature, salinity, and pH of observed data showed std. dev values of 0.4758℃, 0.6675 PSU and 0.2681 respectively. However, after the MC analysis, the std. dev value of the WQP for temperature, pH, and DO shows an improvement with a reduction of 0.047℃, 0.0009, and 0.0038 Mg/L respectively. The decrement in std. dev value shows that the data has become less dispersed concerning its mean value which consecutively reduces the uncertainty in data to produce a more stable prediction model. However, the std. dev value for salinity shows a slight increment from 0.6675 PSU to 0.6694 PSU after an MC analysis. According to National Water Quality Standard for Malaysia, the mean value of observed pH (7.8907) in this research was categorized as class II which is suitable for water supply, recreational use, and fishery (sensitive aquatic species only). As for the mean value for observed temperature (27.8821), it befalls in class III which is suitable for water supply, recreational use, water supply (extensive treatment needed), and fishery (common species) (Ministry of Natural Resources and Environment Malaysia, 2014).

Table 1: Observed Data VS Monte Carlo (MC)

	Units	Mean value		std. dev (σ)	
		observed	MC	observed	MC
Turbidity	m	0.8448	0.8524	0.2333	0.2333
Temp	℃	27.8821	27.8909	0.4758	0.4711
Salinity	PSU	29.6418	29.6454	0.6675	0.6694
pH	-	7.8907	7.8893	0.2681	0.2672
DO	Mg/L	5.5927	5.5970	1.3268	1.3230

The spatial interpolation performance as shown in Table 2 was measured by RMSE, MAE, and standard deviation. The pH value presents the lowest RMSE using the IDW technique with 0.0838 (RMSE), 0.0692 (MAE), and 0.0858 (std. dev). On the other hand, the salinity value presents the highest RMSE using the IDW technique with 0.6581 (RMSE), 0.5097 (MAE), and 0.6772 (std. dev) in comparison with other WQP. Overall, the IDW model outperforms the kriging model with RMSE and MAE values for DO, pH, temperature, and turbidity being significantly lower than the kriging model. In comparison, IDW models perform better than kriging as proven in the previous study (Gong et al., 2014).

Table 2: IDW VS Kriging

	RMSE		MAE		std. dev (σ)	
	IDW	*Kriging*	*IDW*	*Kriging*	*IDW*	*Kriging*
DO	0.5172	0.5681	0.4191	0.4629	0.5007	0.5604
pH	0.0838	0.0939	0.0692	0.0825	0.0858	0.0967
Salinity	0.6581	0.5864	0.5097	0.4799	0.6772	0.6037
Temp	0.4503	0.4581	0.3270	0.3191	0.4651	0.4731
Turbidity	0.1684	0.1829	0.1294	0.1420	0.1734	0.1887

Table 3 shows correlation values between each WQP with a significance level of P-value equal to 0.01. pH and DO have the highest correlation value of 0.948 which signifies that it has a very strong positive relationship. This means that every increment that may occur in DO or pH value significantly increases the value of one another and vice versa. This relationship between DO and pH can probably be explained by the effects of eutrophic waters' algae photosynthesis as per stated in a previous study (Zang et al., 2011). In contrast, salinity shows the weakest relationship with other WQPs ranging between -0.464 (turbidity) to 0.529 (DO) and has the weakest relationship with the temperature at only 0.368. However, turbidity shows a negative relationship with all of the other WQP; -0.418, -0.464, -0.724, and -0.713 for its relationship with temperature, salinity, pH and DO respectively as per studied in prior research (Gupta et al., 2005; Yap et al., 2011).

Table 3: Correlation between Water Quality Parameters

Parameter	*Turbidity*	*Temperature*	*Salinity*	*pH*	*DO*
Turbidity	1	-.418**	-.464**	-.724**	-.713**
Temperature	-.418**	1	.368**	.512**	.510**
Salinity	-.464**	.368**	1	.498**	.529**
pH	-.724**	.512**	.498**	1	.948**
DO	-.713**	.510**	.529**	.948**	1

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4 shows linear and multivariate regression to estimate pH, DO, and turbidity. The regressions were established by using WQP data from IDW interpolation analysis and were verified by using MC data. In estimating pH, equation 8 shows the lowest RMSE, MAE, and standard deviation compared to Equations 1 and 4. This is because the strength of a linear link between two variables is measured by correlation, whereas regression expresses the relationship as an equation. This can be explained by the correlation value of pH with turbidity and DO being at its highest compared to its correlation with other WQP as shown in Table 3. The establishment of a relationship and regression equation will be able to help in predicting the other WQP based on selected measured and predicted WQP. In addition, the outcomes of water quality monitoring are crucial for identifying Spatio-temporal trends in surface water and groundwater variations. The regression equation that has been developed in conjunction with suitable geographical interpolation can even be used as a tool for advocating for legislation to save the ecological system for future use.

Table 4: Regressions for WQP Estimation

No.	Regression	RMSE	MAE	Std. dev
1.	pH = 6.784+0.196(DO)	0.0845	0.0632	0.0847
2.	DO = -30.999+4.646(pH)	0.4195	0.3188	0.4186
3.	Turbidity = 5.279-0.563(pH)	0.1629	0.1377	0.1634
4.	pH = 8.621-0.857(turbidity)	0.1829	0.1544	0.1843
5.	DO = 9.112-4.046(turbidity)	0.9207	0.7279	0.9256
6.	Turbidity = 1.473-0.112(DO)	0.1648	0.1369	0.1659
7.	DO = -30.331+4.575(pH)-0.126(turbidity)	0.4176	0.3167	0.4166
8.	pH = 6.959-0.119(turbidity)+ 0.182(DO)	0.0834	0.0631	0.0829
9.	Turbidity = 4.660-0.020(DO)-0.470(pH)	0.1618	0.1370	0.1625

CONCLUSION

To conclude, MC uncertainty analysis has proven to reduce the uncertainty presence as compared to the original sampling data. IDW model serves as a better tool in the context of estimating WQP specifically around Selat River, Langkawi. From this research, it is also known that pH has the strongest relationship with DO while salinity has the weakest relationship with other WQP. This is however based on the limited number of WQP variations due to the limitation of instruments and time constraints. The result may differ when other WQP were included (e.g Ammonia, conductivity, nitrate chemicals, etc) or when conducted around the different study areas. Hence, the recommendation for further research is the exploration of other WQPs that has not been included in this research. This research will help the authorities, especially the Fisheries Development Authority of Malaysia and even fishermen to estimate the water quality parameters as well as able to indicate the water quality conditions for the Strait of Tuba, Langkawi Malaysia.

ACKNOWLEDGEMENT

This research would not be able to be completed without permission and financial support from Universiti Teknologi MARA (600-RMC/LESTARI SDG-T 5/3 (104/2019) to our team. Also, many thanks to the personnel of our team and authors for being willing to spend their time, energy, and ideas in completing this research.

AUTHOR DECLARATION, CONFLICT OF INTEREST AND CONTRIBUTION

This research received financial support from Universiti Teknologi MARA (600-RMC/LESTARI SDG-T 5/3 (104/2019). The authors' contribution are as follows:

Hasmida Muhamad	Abstract, Introduction, Literature, Methodology, Result, Discussion and References
Ernieza Suhana Mokhtar	Abstract, Introduction, Literature, Methodology, Result, Discussion and References, Content Advisor
Muhammad Akmal Roslani,	Methodology, Result, Discussion
Mohammed Oludare Idrees	Introduction, Literature
Azlan Abdul Aziz	Methodology, Result, Discussion
Noraini Nasirun	Introduction, Literature

REFERENCES

Bay, Chesapeake, Rebecca R. Murphy, Frank C. Curriero, William P. Ball, and M. Asce. 2014. "Comparison of Spatial Interpolation Methods for Water Quality Evaluation in the Chesapeake Bay." (February 2010). doi: 10.1061/(ASCE)EE.1943-7870.0000121.

Borges, Pablo De Amorim, Johannes Franke, and Christian Bernhofer. 2015. "Comparison of Spatial Interpolation Methods for the Estimation of Precipitation Distribution in Distrito Federal , Brazil." 2012. doi: 10.1007/s00704-014-1359-9.

Farrance, Ian, and Robert Frenkel. 2016. "Uncertainty in Measurement : A Review of Monte Carlo Simulation Using Uncertainty in Measurement : A Review of Monte Carlo Simulation Using Microsoft Excel for the Calculation of Uncertainties Through Functional Relationships , Including Uncertainties In." (February 2014). doi: Clin Biochem Rev 35 (1) 201437.

Ghazali, Mochamad Firman, Ketut Wikantika, Agung Budi Harto, and Akihiko Kondoh. 2020. "Generating Soil Salinity, Soil Moisture, Soil PH from Satellite Imagery and Its Analysis." *Information Processing in Agriculture* 7(2):294–306. doi: 10.1016/j.inpa.2019.08.003.

Gholizadeh, Mohammad, Assefa Melesse, and Lakshmi Reddi. 2016. "A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques." *Sensors* 16(8):1298. doi: 10.3390/s16081298.

Gong, Gordon, Sravan Mattevada, and Sid E. O. Bryant. 2014. "Comparison of the Accuracy of Kriging and IDW Interpolations in Estimating Groundwater Arsenic Concentrations in Texas." *Environmental Research* 130:59–69. doi: 10.1016/j.envres.2013.12.005.

Gupta, A. K., S. K. Gupta, and Rashmi S. Patil. 2005. "Statistical Analyses of Coastal Water Quality for a Port and Harbour Region in India." *Environmental Monitoring and Assessment* 102(1–3):179–200. doi: 10.1007/s10661-005-6021-7.

Ikechukwu, Maduako Nnamdi, Elijah Ebinne, Ufot Idorenyin, and Ndukwu Ike Raphael. 2017. "Accuracy Assessment and Comparative Analysis of IDW , Spline and Kriging in Spatial Interpolation of Landform (Topography): An Experimental Study." 354–71. doi: 10.4236/jgis.2017.93022.

Jayanthi, M., S. Thirumurthy, M. Samynathan, P. Kumararaja, M. Muralidhar, and K. K. Vijayan. 2021. "Multi-Criteria Based Geospatial Assessment to Utilize Brackishwater Resources to Enhance Fish Production." *Aquaculture* 537(February):736528. doi: 10.1016/j.aquaculture.2021.736528.

Kılıç, Zeyneb. 2020. "The Importance of Water and Conscious Use of Water." *International Journal of Hydrology* 4(5):239–41. doi: 10.15406/ijh.2020.04.00250.

Mara, Fernanda, Coelho Pizani, Philippe Maillard, and Camila C. Amorim. 2020. "Estimation of Water Quality in a Reservoir from Sentinel-2 MSI and Landsat-8 OLI Sensors." (August). doi: 10.5194/isprs-annals-V-3-2020-401-2020.

Ministry of Natural Resources and Environment Malaysia. 2014. "National Water Quality Standards For Malaysia- Annex." *National Water Quality Standards for Malaysia- Annex*.

Mokhtar, Ernieza Suhana, Biswajeet Pradhan, Abd Halim Ghazali, and Helmi Zulhaidi Mohd Shafri. 2018. "Assessing Flood Inundation Mapping through Estimated Discharge Using GIS and HEC-RAS Model." *Arabian Journal of Geosciences* 11(21). doi: 10.1007/s12517-018-4040-2.

Moreno-Rodenas, Antonio M., Franz Tscheikner-Gratl, Jeroen G. Langeveld, and Francois H. L. R. Clemens. 2019. "Uncertainty Analysis in a Large-Scale Water Quality Integrated Catchment Modelling Study." *Water Research* 158:46–60. doi: 10.1016/j.watres.2019.04.016.

Pappenberger, F., and K. J. Beven. 2006. "Ignorance Is Bliss : Or Seven Reasons Not to Use Uncertainty Analysis." 42(March):1–8. doi: 10.1029/2005WR004820.

Ravichandran, Y. Dominic, and K. Ramakrishnan. 2007. "Correlation and Regression Studies of Water Quality Parameters: A Case Study of Water from the Bhavani River." *Asian Journal of Chemistry* 19(4):2679–82.

Roberto, Paulo, and Guimarães Couto. 2013. "Monte Carlo Simulations Applied to Uncertainty in Measurement."

Schellart, A. N. A., S. J. Tait, and R. M. Ashley. 2010. "Towards Quantification of Uncertainty in Predicting Water Quality Failures in Integrated Catchment Model Studies." *Water Research* 44(13):3893–3904. doi: 10.1016/j.watres.2010.05.001.

Sonnemann, Guido W., Marta Schuhmacher, and Francesc Castells. 2003. "Uncertainty Assessment by a Monte Carlo Simulation in a Life Cycle Inventory of Electricity Produced by a Waste Incinerator." *Journal of Cleaner Production* 11(3):279–92. doi: 10.1016/S0959-6526(02)00028-8.

Valentini, Marlon, Gabriel Borges, and Bruno Muller. 2021. "Multiple Linear Regression Analysis (MLR) Applied for Modeling a New WQI Equation for Monitoring the Water Quality of Mirim Lagoon , in the State of Rio Grande Do Sul — Brazil." *SN Applied Sciences* 3(1):1–11. doi: 10.1007/s42452-020-04005-1.

Wang, Mengmeng, Guojin He, Zhaoming Zhang, Guizhou Wang, and Zhengjia Zhang. 2017. "Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China." 1–16. doi: 10.3390/rs9121278.

Yap, C. K., M. W. Chee, S. Shamarina, F. B. Edward, W. Chew, and S. G. Tan. 2011. "Assessment of Surface Water Quality in the Malaysian Coastal Waters by Using Multivariate Analyses." *Sains Malaysiana* 40(10):1053–64.

Yusuf. 2020. "Phytoplankton as Bioindicators of Water Quality in Nasarawa Reservoir , Katsina State Nigeria." 32.

Zang, Changjuan, Suiliang Huang, Min Wu, Shenglan Du, Miklas Scholz, Feng Gao, Chao Lin, Yong Guo, and Yu Dong. 2011. "Comparison of Relationships between PH, Dissolved Oxygen and Chlorophyll a for Aquaculture and Non-Aquaculture Waters." *Water, Air, and Soil Pollution* 219(1–4):157–74. doi: 10.1007/s11270-010-0695-3.