GIS-BASED PREDICTION OF WATER QUALITY INDEX USING MACHINE LEARNING: A CASE STUDY OF THE THA CHIN RIVER, THAILAND

Pasin PROMHAN [1]*, ***Sitang PILAILAR*** [2]

DOI: 10.21163/GT_2025.202.19

ABSTRACT

This study integrates field data, machine learning (ML), and Geographic Information System (GIS) techniques to compute and predict the Water Quality Index (WQI) along the Tha Chin River Basin, Thailand. Five key water quality parameters—DO, BOD, COD, pH, and Salinity—were collected from monitoring stations between 2019 and 2022. These were normalized into Q-values and used to compute WQI through weighted aggregation. Six supervised ML algorithms were tested, with the Random Forest model yielding the highest predictive performance ($R^2 = 0.664$, MAE = 0.855) at a 10-hour lead time. Incorporating spatial indicators such as urban land cover and population density significantly enhanced model accuracy. Computed WQI values were visualized using Inverse Distance Weighted (IDW) interpolation to assess spatial and temporal trends. Results indicated a consistent decline in water quality from upstream to downstream, with all zones classified as "moderately degraded" or "degraded." The lowest WQI values were observed in Banglen District, ranging from 39 to 42, linked to high urban density and pollutant accumulation. In contrast, upstream areas such as Mueang Suphan Buri slightly improved over time. The study confirms that urban expansion is a major contributor to river water degradation. The proposed ML-GIS framework supports proactive monitoring, spatial prioritization, and evidence-based water resource management in rapidly urbanizing river basins.

Key-words: *Tha Chin River; Water Quality Index (WQI); Machine Learning; GIS; Spatial Interpolation. Environmental Monitoring; Random Forest.*

1.INTRODUCTION

Water quality deterioration is a growing concern in many regions across the globe, particularly in areas undergoing rapid development (Abbasi & Abbasi, 2012; Horton R.K, 1965) Increased agricultural production, unregulated industrial discharge, and expanding urban settlements contribute to elevated pollution loads in river systems, which serve as primary conduits for surface runoff and point-source wastewater. Monitoring and managing these dynamic water systems is critical for protecting public health, maintaining biodiversity, and ensuring water security.

The Water Quality Index (WQI) has long served as a composite metric that condenses multiple water quality parameters—such as Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), chemical oxygen demand (COD) (Belete & Huchaiah, 2021), total suspended solids (TSS), Salinity, and pH—into a single value representing overall water quality status (Lumb, Halliwell, & Sharma, 2006). While effective for summarizing complex data, conventional WQI assessments are limited by their retrospective nature. To support real-time decision-making (Liu, Wang, Sangaiah, Xie, & Yin, 2019), predictive modeling approaches are increasingly required.

However, in practice, environmental datasets frequently contain gaps and noise due to missing measurements, equipment malfunctions, or human error (Magyari-Sáska, Haidu, & Magyari-Sáska, 2025). Machine learning (ML) techniques are gaining traction for their ability to capture non-linear relationships and generate accurate forecasts from complex datasets (Chen et al., 2020; Najah Ahmed et al., 2019).

[1] *Department of Water Resources Engineering, Kasetsart University, 10900 Bangkok, Thailand, (PP)* pasin.pro@ku.th *,(SP) fengstpl@ku.ac.th

For instance, (Ahmed et al., 2019) introduced a Wavelet De-noised Adaptive Neuro-Fuzzy Inference System (WDT-ANFIS) for predicting key water quality parameters, significantly outperforming traditional models through effective preprocessing and input optimization. Similarly, (Magyari-Sáska et al., 2025) demonstrated self-imputation techniques for handling incomplete hydrological time series, improving model robustness under data scarcity.

In addition, (Kouadri, Elbeltagi, Islam, & Kateb, 2021) compared multiple ML algorithms—including multilayer regression (MLR), random forest (RF), and support vector regression (SVR)—for WQI estimation in data-scarce regions, highlighting the benefits of sensitivity analyses in reducing input uncertainty while maintaining high predictive accuracy.

When combined with Geographic Information Systems (GIS), which add spatial dimensions such as land use, population density, and urbanization patterns (Ly et al., 2021). ML models can deliver more context-aware predictions. Integrating GIS-derived variables into ML workflows enhances spatial interpretability, enabling forecasts of site-specific WQI, mapping of pollution hotspots, and visualization of basin-wide trends (Kouadri et al., 2021; Magyari-Sáska et al., 2025; Wang, Kim, & Li, 2021).

The Tha Chin River basin in central Thailand presents an ideal setting to apply such an integrated ML–GIS framework. Stretching over 300 kilometers through diverse agricultural and urban landscapes, the basin exhibits a pronounced gradient of anthropogenic pressures affecting water quality. This study aims to: (i) evaluate the performance of six machine learning algorithms in forecasting WQI, (ii) incorporate GIS-derived spatial indicators into model development, and (iii) produce spatial visualizations of predicted WQI to support targeted water resource management.

2. STUDY AREA

The Tha Chin River is a major distributary of the Chao Phraya River, originating in Chainat Province in central Thailand. It flows southward for approximately 325 kilometers through Suphan Buri, Nakhon Pathom, and Samut Sakhon Provinces before discharging into the Gulf of Thailand. The river traverses a diverse range of landscapes, including agricultural zones, urban centers, and industrial areas, making it a representative basin for examining the interplay between land use and water quality dynamics (Pollution Control Department. PCD, 2020).

The upper and middle reaches of the basin are dominated by rice cultivation, aquaculture, and livestock farming, which contribute significant non-point source pollution through runoff containing nutrients and organic waste. In contrast, the lower basin−particularly around Nakhon Pathom and Samut Sakhon−is highly urbanized and industrialized, with dense residential populations and a concentration of agro-industrial facilities. These characteristics result in a complex pollution profile influenced by diffuse and point sources.

Water quality in the Tha Chin River has shown persistent degradation over the past two decades. According to the Pollution Control Department (PCD), downstream segments frequently fall into the "poor" or "very poor" categories of the national WQI classification, with elevated levels of BOD, COD, and reduced DO. Seasonal variation further complicates the basin's condition, as pollutant concentrations typically rise during the dry season due to reduced flow and dilution capacity (Pollution Control Department PCD, 2022).

To capture this variability, the study utilizes water quality monitoring data from five key PCD stations located along the river: Chainat (upper reach), Suphan Buri (mid-upper reach), Song Phi Nong (midstream), Nakhon Chai Si (mid-lower reach), and Mahachai (river mouth), as shown in **Fig. 1.** These sites provide consistent seasonal data across multiple years, forming the empirical foundation for model training and validation. Supplementary spatial datasets−including land use classifications, population density, and administrative boundaries−were integrated to enhance the spatial representation of anthropogenic pressures within the modeling framework.

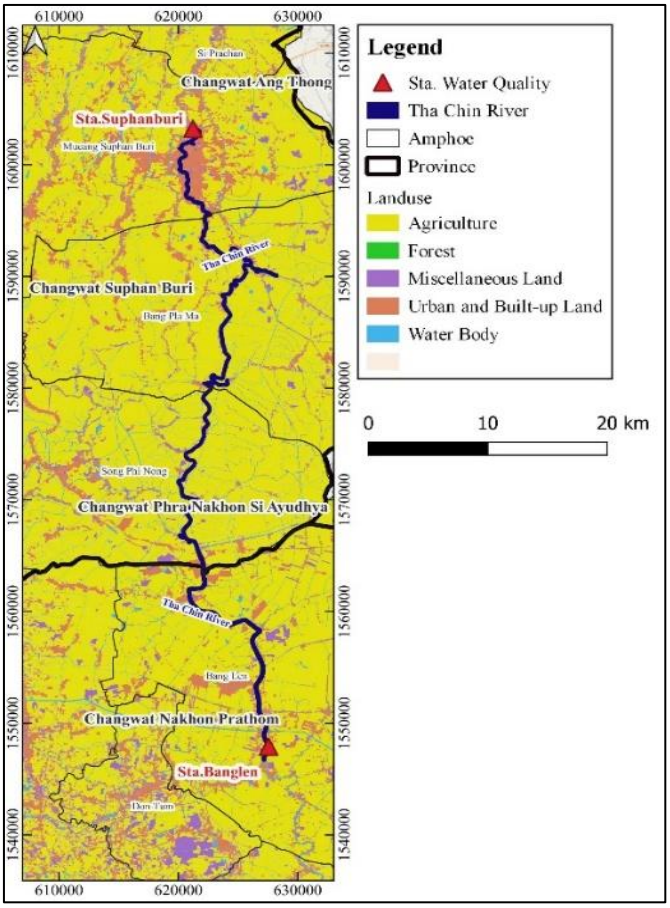

Fig. 1. The Chin River Basin and locations of water quality monitoring stations used in the study.

3. METHODOLOGY

This study employed an integrated approach combining water quality assessment, machine learning (ML) prediction, and Geographic Information System (GIS)-based spatial analysis. The methodology consisted of four main steps: (i) data collection and preprocessing, (ii) Water Quality Index (WQI) computation, (iii) machine learning model development, and (iv) spatial visualization of predicted WQI values, as shown in **Fig. 2.**

3.1. Data Collection and Preprocessing

In this study, five key parameters were selected in accordance with the Pollution Control Department (PCD) of Thailand's WQI framework: BOD, DO, COD, pH, and Salinity. These parameters represent both organic and inorganic pollution pressures, and their inclusion follows national reporting standards for surface water quality. The hourly water quality data from five PCD monitoring stations along the Tha Chin River were collected for 2019–2022. To ensure data reliability, outliers were removed using the interquartile range (IQR) method (Gazzaz, Yusoff, Aris, Juahir, & Ramli, 2012), and missing values were handled using linear interpolation. Z-score normalization was applied to standardize the dataset (Al-Faiz, Ibrahim, & Hadi, 2019).

Additional spatial data were compiled to represent anthropogenic influences within a 5 km buffer around each station. These included (i) population density, derived from official census records (2019–2022), and (ii) land use classifications from the Land Development Department (LDD) for the years 2019 and 2021. Land cover types were extracted using supervised classification in QGIS.

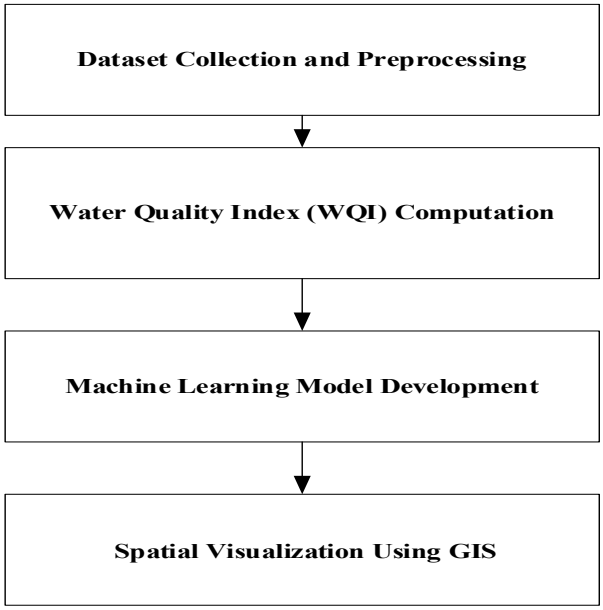

Fig. 2. Steps of work.

3.2. Water Quality Index (WQI) Computation

The Water Quality Index (WQI) was computed to integrate multiple physicochemical parameters into a single composite score that reflects the overall status of water quality in the Tha Chin River. This approach enables a consistent and comparable assessment across different sites and seasons (Abbasi & Abbasi, 2012; Ahmed et al., 2019). The computation of WQI involved two main stages:

(i) Q-value normalization, where raw parameter values were transformed into a dimensionless 0–100 scale based on PCD's rating curves; and

(ii) weighting factor application, where each parameter's Q-value was multiplied by its assigned weight to reflect its relative importance in the overall index. The final WQI score was obtained by summing the weighted Q-values and dividing by the sum of the weights.

The parameter ranges, classification thresholds, and corresponding weighting factors used in this study are presented in **Table 1**. Detailed descriptions of the normalization process and weighting scheme are provided in Sections 3.2.1 and 3.2.2.

3.2.1. Q-Value Normalization

Raw measurements of each water quality parameter were transformed into a standardized 0–100 scale, referred to as the Q-value, using parameter-specific rating curves developed by the PCD of Thailand. The Q-value represents the degree to which the measured parameter meets the desired water quality condition, with higher values indicating better quality. For each parameter, the transformation was carried out by matching the observed value to its corresponding score on the PCD's rating curve. These curves are non-linear and account for the unique response of each parameter to environmental conditions. For example, DO values close to saturation receive higher Q-values, while elevated BOD and COD—indicating greater organic pollution—receive lower scores. Similarly, pH values outside the optimal range (6.5–8.5) are penalized, and salinity values above freshwater tolerance levels result in reduced Q-values. This method aligns with previous applications of parameter normalization for WQI computation (Marzieh Mokarram & Zarei, 2021)and ensures comparability between parameters with different units and environmental implications. Alternative normalization methods, such as z-score or min–max scaling, were not assessed in this study in order to maintain consistency with the national water quality assessment framework.

3.2.2 Weighting Factors

In the composite WQI calculation, each normalized Q-value was multiplied by a weighting factor that reflects its relative importance in determining overall water quality. The weighting factors for DO, BOD, COD, pH, and Salinity were adapted from the PCD guidelines (PCD, 2021) and are supported by prior WQI studies in Thailand (PCD, 2017). The weighting factors applied in this study were: DO (0.28), BOD (0.24), COD (0.20), pH (0.14), and Salinity (0.14). These values sum to unity, ensuring that the resulting index remains bounded within the 0–100 scale. The overall WQI for each sampling site was computed using the weighted arithmetic mean method as expressed in Eq. (1):

$$\mathrm{WQI} = \sum(W_i \times Q_i) \tag{1}$$

where

Q_i = the water quality parameter s' index i

W_i = the associated weighing factor for each parameter

No sensitivity or uncertainty analysis was conducted in this study to test the impact of alternative weighting schemes, although such assessments have been recommended in recent literature. The chosen factors align with those used in previous WQI studies in Southeast Asia (Najah Ahmed et al., 2019), supporting methodological consistency and comparability with earlier work.

Table 1.

Equations for calculating Q-values for selected water quality parameters (adapted from PCD standards).

Variables	Range	Sub-index Function
DO	0.0≤x≤4.0	y=15.25x+1.667
	4.1≤x≤6.0	y=5x+41
	6.1≤x≤8.4	y=12.083x-1.5
	8.5≤x≤8.9	y=-78x+755.2
	9.0≤x≤11.2	y=-13.043x+177.09
	x≥11.3	y=7.561x+115.68
BOD	0.0≤x≤1.5	y=-19.333x+100
	1.6≤x≤2.0	y=-20x+101
	2.1≤x≤4.0	y=-15x+91
	x≥4.1	y=-6.4583x+56.833
Temperature	x<0	y=100
	0≤x≤30	$y=-0.0325x^2-1.8851x+96.236$
	x>30	y=0
COD	x≤20	-1.33x+99.1
	x>20	$103e^{-0.537x}$-0.04x
pH	x<2	y=0
	2≤x<4	$y=1.13x^2-3.66x+5.04$
	4.1≤x≤7.5	$y=-1.7479x^2+44.845x-144.25$
	7.5<x<10	$y=-3.1167x^2+23.279x+92.915$
	10≤x≤13	$y=1.157x^2-31.088x+210.06$
	x>13	y=0
Salinity		$y=14.341x^2-69.274x+86.549$

3.3. Machine Learning Model Development

Six supervised ML algorithms were applied to predict WQI values: Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), and AdaBoost (AB). Input features included the five water quality parameters and spatial variables (population and land use). The dataset was split into training (80%) and testing (20%) subsets. Features were normalized using Z-score normalization (Jha et al., 2015) (Eq. (2)):

$$Z = \frac{(x-\mu)}{\sigma} \tag{2}$$

where:

x = the raw value

μ = the mean of the feature

σ = the standard deviation

A 10-fold cross-validation was applied to each model to assess generalizability. The performance of each model was evaluated using the Coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE) (Fathi et al., 2022).

3.4. Spatial Visualization Using GIS

The predicted WQI values were spatially visualized using the Inverse Distance Weighting (IDW) interpolation method in QGIS 666(Mueller et al., 2004). This method estimates WQI at unsampled locations based on the spatial proximity and magnitude of known values. The interpolation was conducted separately for wet and dry seasons to capture seasonal variations in water quality(Magesh, Krishnakumar, Chandrasekar, & Soundranayagam, 2012).

$$Z(x_0) = \frac{\Sigma(\frac{Z_i}{d_i^p})}{(\frac{1}{d_i^p})} \tag{3}$$

where:

$Z(x_0)$ = the estimated WQI

Zi = the known value

di = the distance between points

p = the power parameter (set 2)

The output maps provided spatial insights into water quality patterns and supported the identification of potential pollution hotspots.

4. RESULT AND DISCUSSION

This section presents the results of the water quality assessment, machine learning model predictions, and spatial analysis, followed by a discussion of key findings. The analysis is based on data from five monitoring stations along the Tha Chin River, spanning dry and wet seasons from 2019 to 2022. Results are structured to highlight the transformation and classification of raw water quality data, the predictive performance of machine learning algorithms, and the spatial patterns of WQI concerning land use and population pressures.

Integrating Q-value normalization, spatial indicators, and machine learning allowed for a comprehensive understanding of how physicochemical and anthropogenic factors interact to influence river water quality. The following subsections discuss (i) Q-value normalization and WQI classification, (ii) predictive performance of ML models, (iii) influence of spatial indicators, and (iv) spatial interpolation of WQI for decision support.

4.1. Field Survey Results and Spatial Correlation with Anthropogenic Activities

From the total of 55 Pollution Control Department (PCD) monitoring locations within the Tha Chin River Basin, 26 field survey sites were selected through a stratified, criteria-based approach to ensure representativeness and practicality.

Fig. 3. Spatial distribution of EC, pH, DO, BOD, and TDS values across 26 sampling stations within the Tha Chin River study area.

The selection process involved three key steps:

1. Stratification by river reach and hydrological conditions – The river was divided into five longitudinal zones: upper, mid-upper, midstream, mid-lower, and lower reaches. This ensured that each section's distinct hydrological regime was captured, including variations in flow velocity, tidal influence, and upstream–downstream pollution dynamics.

2. Classification by dominant land use and pollution sources – Using recent land-use GIS layers, historical PCD WQI records (2013–2022), and point-source pollution data, candidate sites were categorized according to dominant influences: intensive agriculture, livestock farming, aquaculture, mixed urban–rural settlements, and industrial zones. Priority was given to sites exhibiting contrasting pollution profiles (e.g., "good" vs. "poor" historical WQI ratings) to enhance model sensitivity to varying water quality conditions.

3. Evaluation of accessibility, safety, and logistical feasibility – Sites with restricted access, overlapping influence zones with other stations, or safety risks (e.g., hazardous navigation, unstable banks) were excluded. This ensured that selected sites could be sampled consistently under both dry and wet season conditions.

The final 26 sites thus provided balanced coverage of the basin's geographic extent, anthropogenic pressures, and hydrological diversity, while remaining feasible for repeated seasonal sampling. This strategic selection aimed to capture the full spatial variability of water quality conditions within the Tha Chin River Basin, thereby enhancing the robustness of the subsequent GIS-based machine learning analyses. Water quality measurements focused on five key parameters: Electrical Conductivity (EC), pH, DO, BOD, and Total Dissolved Solids (TDS). **Fig. 3** shows a clear spatial degradation pattern that emerges as the river flows downstream. BOD concentrations steadily increase, especially around Banglen District, indicating rising levels of organic pollution. Meanwhile, DO levels decline, reflecting oxygen depletion associated with microbial activity. EC and TDS values also fluctuate, with notable peaks near Mueang Suphan Buri, where rice mill effluents contribute high-ion discharges, and downstream communities with intensive land use. Occasional spikes in BOD at specific stations were linked to high-density zones, such as temples or clustered residential areas.

These pollution patterns are closely associated with increasing human settlement and land use changes. Between 2019 and 2021, the study area experienced a notable expansion in urban and built-up areas. According to land classification results, urban land increased by 1.54% in Chai Nat, 11.89% in Nakhon Pathom, and 14.91% in Suphan Buri. These changes are visualized in **Fig. 4.**

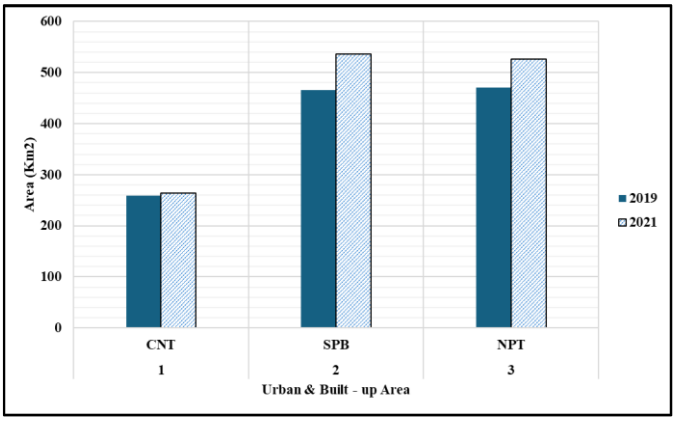

Fig. 4. Urban & Built-up Area expansion in Chai Nat (CNT), Nakhon Pathom (NPT), and Suphan Buri (SPB) from 2019 to 2021.

The correlation between these land use changes and the observed water quality deterioration underscores the significant impact of anthropogenic activities on river ecosystems. Increased urbanization, population density, and unregulated wastewater discharge contribute to higher pollutant loads, particularly in downstream segments. These field findings reinforce the rationale for integrating spatial indicators such as population density and urban land use into machine learning models discussed in later sections of this study.

4.2. Q-Value Normalization and WQI Classification

To enable standardized analysis across parameters with different units and scales, raw values of the five selected water quality indicators−DO, BOD, COD, pH, and Salinity−were first transformed into sub-indices using Q-value normalization. This approach scaled the parameters onto a standard 0‑100 range, facilitating the computation of the Water Quality Index (WQI) using the weighted arithmetic method (D. Satish Chandra, 2017). The transformation followed guidelines from the Pollution Control Department (PCD) of Thailand and aligned with WHO standards for water quality classification.

Fig. 5 provides an overview of Q-values across all parameters and stations, highlighting spatial and seasonal variation. Among the five indicators, DO and BOD showed the highest variability, consistent with seasonal hydrology and pollutant loading patterns. These indicators are susceptible to organic pollution and oxygen demand, making them reliable indicators of environmental stress.

A closer examination of Q-value behavior for DO and BOD is shown in **Fig. 6**.

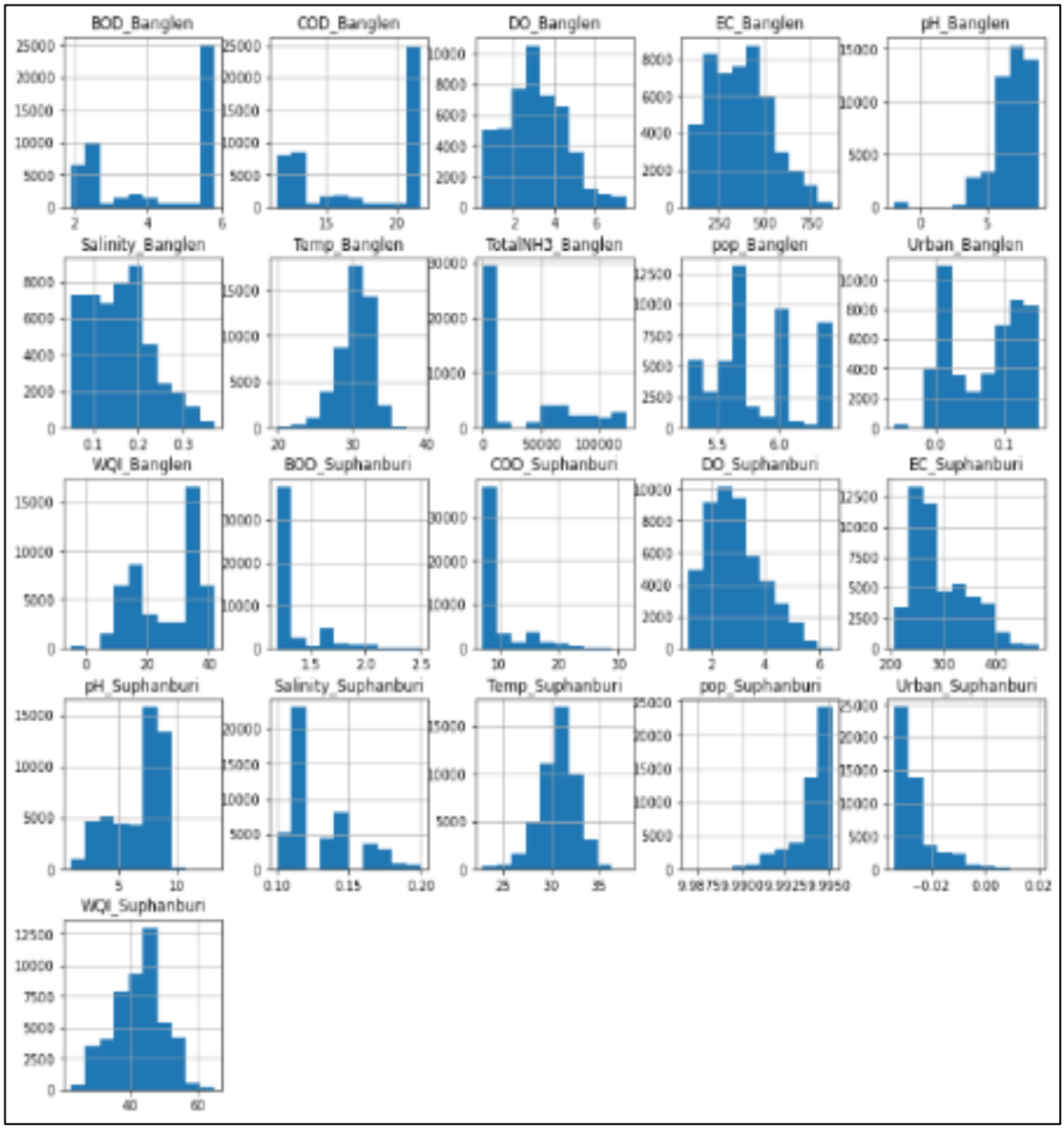

Fig. 5. Normalized Q-values of five water quality parameters across monitoring stations.

Fig. 6 demonstrates the successful normalization of these parameters across all monitoring stations and time periods. The transformation ensures that both extreme and moderate values are accurately represented in the index computation. Stations located in the downstream segment, particularly Mahachai and Nakhon Chai Si, displayed lower DO Q-values and elevated BOD Q-values, indicating higher organic pollution loads and reduced oxygen availability.

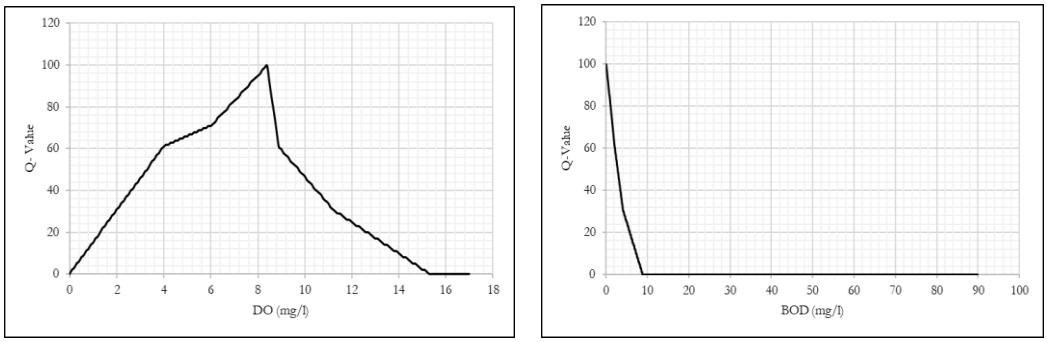

Fig. 6. Example of normalized Q-values for BOD and DO.

Following normalization, WQI scores were calculated by aggregating the Q-values using assigned weights. The resulting WQI values ranged from 23.5 to 81.6, covering classifications from "poor" to "good." A clear spatial trend emerged: upstream stations (e.g., Chainat, Suphan Buri) consistently showed higher WQI values, while downstream stations (e.g., Mahachai) exhibited significantly degraded conditions, especially during the dry season when river flow and dilution capacity were reduced. These results confirm that Q-value normalization not only standardizes data for WQI computation but also captures meaningful environmental gradients. The observed downstream decline aligns with land use pressures and pollution sources, supporting the case for integrating spatial indicators—such as population density and urban land use—into the modeling framework.

4.3. Correlation Analysis of Water Quality Parameters

To understand the interrelationships among water quality parameters and support the selection of input variables for predictive modeling, a Pearson correlation analysis (Senthilnathan, 2019) was conducted on the normalized dataset. This analysis revealed significant associations that reflect the underlying chemical and biological processes influencing river water quality, as shown in **Fig.7.**

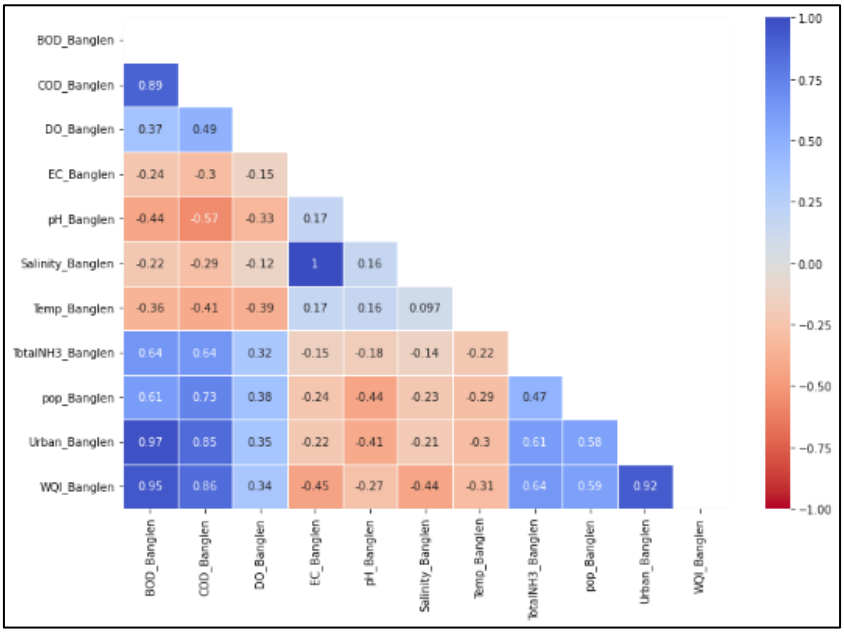

Fig. 7. Pearson correlation matrix of water quality parameters (DO, BOD, COD, pH, and Salinity) at Banglen Station.

Visually represents the degree of correlation for each parameter. The color coding in **Fig. 7.** visually indicates the degree of correlation for each parameter, where a strong correlation is represented by a value of 1 (the blue shade). As the correlation value decreases towards 0 (the red shade), the color shade darkens, indicating weaker or no correlation. Darker shades also represent inverse correlations, showing a negative relationship between parameters. This visualization helps quickly identify which parameters are strongly correlated, both positively and negatively, which is crucial for understanding how these parameters interact and impact WQI.

The Pearson correlation analysis revealed the following key relationships:

- **At the Same Station:**
    - BOD negatively correlates with EC, Salinity, populations, and pH. It is loosely associated with Temperature WQI and negatively correlated with DO.
    - DO is weakly y correlated with all parameters.
    - EC is highly correlated with Salinity and loosely associated with BOD and DO. It is weakly correlated with pH and Temperature.
    - pH and Temperature are highly and positively correlated with Salinity. They are loosely associated with BOD and DO and negatively correlated with EC.
    - The population is highly negatively correlated with BOD, DO, and COD. It is weakly associated with pH, EC, Temperature, and Salinity.
    - Urban & Built-up Areas are highly positively correlated with BOD and DO. It is loosely associated with pH, EC, Temperature, and Salinity.
- **Between Banglen Station and Suphanburi Station:**
    - BOD at Banglen Station weakly correlates with BOD at Suphanburi Station. It is loosely associated with DO and Temperature and negatively correlated with EC, Salinity, and pH.
    - At Banglen Station, it is weakly correlated with DO at Suphanburi Station, BOD, and pH. It is loosely negatively correlated with EC, Salinity, and Temperature
    - pH, Salinity, and Temperature at Banglen Station are highly and positively correlated with pH, Salinity, and Temperature at Suphanburi Station. They are loosely negatively correlated with DO and BOD.
    - WQI at Banglen Station weakly correlates with DO, EC, pH, Salinity, and WQI at Suphanburi Station. It is lightly associated with BOD and Temperature.
    - The population and Urban & Built-up Area at both stations are highly negatively correlated with BOD, DO, and COD and negatively correlated with EC, Salinity, and pH.

These correlation patterns are critical for two reasons. First, they validate the inclusion of these parameters in the WQI formulation by highlighting their distinct yet interconnected roles. Second, they indicate which variables are likely to carry more predictive weight in the machine learning models discussed in the next section.

4.4. Correlation Analysis of Water Quality Parameters

To understand the interrelationships among water quality parameters and support the selection of input variables for predictive modeling, a Pearson correlation analysis was conducted on the normalized dataset. This analysis revealed significant associations that reflect the underlying chemical and biological processes influencing river water quality, as shown in **Fig. 7.**

Due to the limited availability of high-frequency water quality monitoring data, this study adopted a predictive modeling approach that utilizes upstream water quality information to forecast downstream conditions. The input dataset comprised five key physicochemical parameters−Temperature, BOD, DO, pH, and Salinity−along with spatial variables such as population density and urban built-up area coverage. The goal was to anticipate pollution severity downstream and provide timely data for management intervention.

To evaluate the robustness of model forecasts over different horizons, the Random Forest (RF) algorithm−identified as the best-performing model−was tested across three lead times: 5 hours, 10

hours, and 17 hours. As shown in **Table 2**, all lead times produced consistently high predictive performance. The 10-hour lead time yielded the best overall results with the highest R^2 and lowest MAE, suggesting it provides the most balanced and stable forecast window.

Table 2.

Performance of Random Forest model at different lead times.

Lead Time	R^2	RMSE	MAE
5 Hours	0.663	6.593	0.841
10 Hours	0.664	7.096	0.855
17 Hours	0.657	6.398	0.824

Further comparison of machine learning models at the 10-hour lead time revealed that Random Forest (RF) outperformed all other algorithms, including k-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and AdaBoost (AB). As shown in **Table 3**, RF achieved the highest R^2 value and the lowest prediction errors, confirming its suitability for modeling complex environmental data.

Table 3.

Performance comparison of machine learning models at 10-hour lead time.

Model	R^2	RMSE	MAE
Random Forest	0.664	7.096	0.855
KNN	0.593	11.532	1.197
Logistic Regression	0.588	14.685	1.353
Decision Tree	0.405	26.237	2.217
SVM	0.114	102.788	2.217
AdaBoost Classifier	0.158	107.419	6.327

Prediction error analysis showed that the RF model performed well in upstream locations such as Chainat and Suphan Buri, where water quality was more stable. In contrast, greater deviations occurred in Mahachai and Nakhon Chai Si, where urbanization and discharge variability introduced more complex conditions. Seasonally, WQI values were slightly overpredicted in the wet season (likely due to dilution from runoff) and underpredicted in the dry season, when pollutant concentrations peak due to low flow volumes. These insights highlight the importance of incorporating seasonal and site-specific variability into future model refinements.

4.5 Temporal and Spatial Change of WQI of the Tha Chin River

4.5.1. Relationship between WQI and accumulated Population along the River

To examine the long-term trends in surface water quality across the Tha Chin River Basin, the Water Quality Index (WQI), as illustrated in **Fig. 8**, was analyzed refering km.0 at Wat Sing District as the starting point. A comparative assessment of annual WQI values from 2020 to 2022 against the Pollution Control Department (PCD) 's standard classification for surface water quality revealed a consistent downward trend. The average WQI values were 39.9 in 2020, 39.0 in 2021, and further declined to 38.5 in 2022—each year, falling below the national standard.

Comparing these WQI values with the cumulative population along the river corridor from Wat Sing District downstream, a clear inverse relationship emerges. As the cumulative population increases, WQI values show a corresponding decline. This degradation is primarily attributed to elevated levels of BOD and COD, which rise with population density and accumulate downstream, contributing to deteriorating water quality.

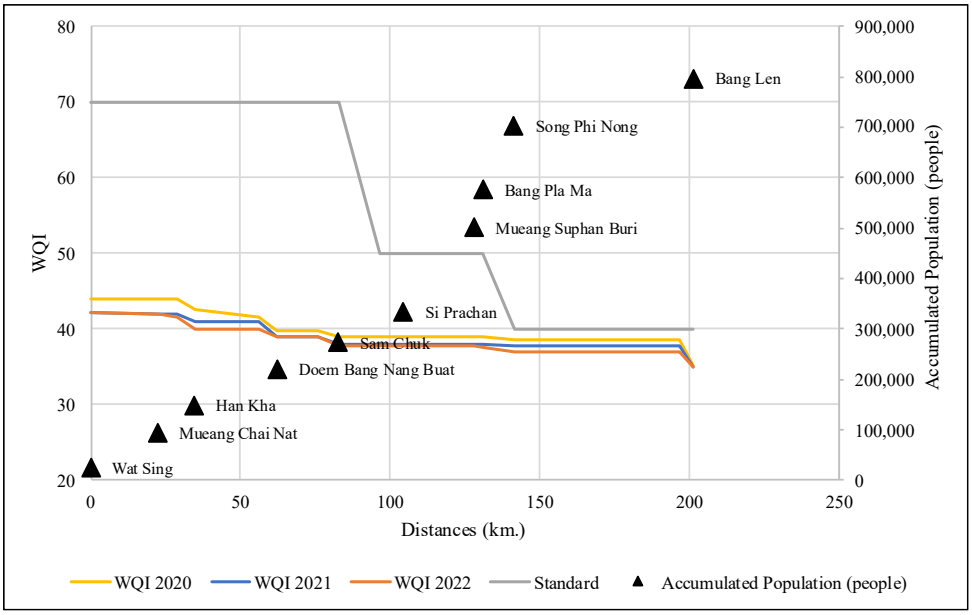

Fig. 8. The WQI values for the years 2020, 2021, and 2022 were compared against the national standard threshold established by the Pollution Control Department (PCD).

4.5.2. Spatial Visualization

To examine long-term patterns of surface water quality across the Tha Chin River Basin, Inverse Distance Weighted (IDW) interpolation was applied using QGIS 3.28 to generate annual spatial distribution maps of the Water Quality Index (WQI).

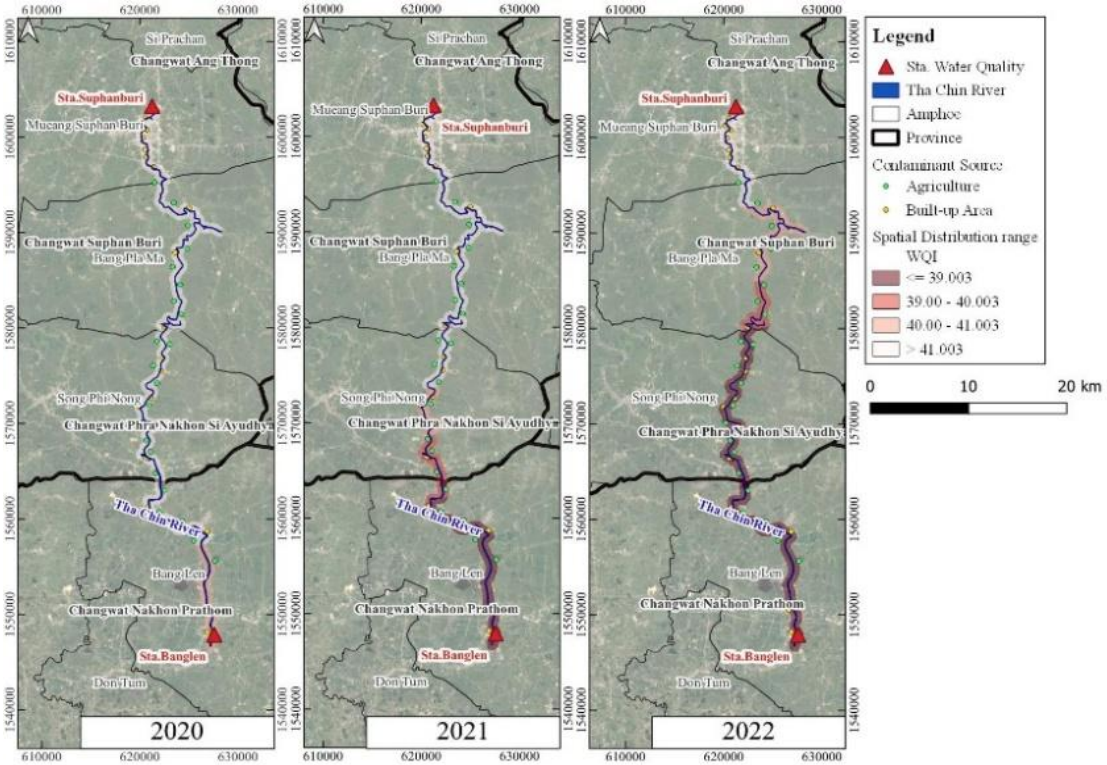

Fig. 9. Spatial distribution of WQI for 2020, 2021, and 2022, generated using IDW interpolation.

This geostatistical technique estimates unsampled values based on the weighted average of nearby measured points, offering a clear visual representation of spatial WQI variation (Tabios & Salas, 1985).

As shown in **Fig. 9**, the entire study area experienced persistent water quality degradation over the three years. All zones were categorized as either "moderately degraded" or "degraded," with no locations falling into the "good" or "excellent" categories. The deterioration trend is most pronounced in 2022, indicating a sustained decline in surface water quality across the basin.

When analyzed by district, Banglen consistently reported the lowest WQI values, ranging from 39 to 42, across all three years. This was attributed to high urban density and the accumulation of pollutants from upstream. Bang Pla Ma and Song Phi Nong displayed slightly higher and relatively stable WQI values within the "degraded" category. Mueang Suphan Buri District, located upstream, maintained moderately degraded water quality and showed a slight improvement trend from 2020 to 2022.

This decline in WQI aligns with spatial land use changes during the same period. Between 2020 and 2022, urban and built-up areas increased noticeably, particularly in Suphan Buri and Nakhon Pathom provinces. Although detailed agricultural land trends were not assessed, the documented expansion of the urbanized regions points to growing human pressure on the river system.

In summary, the IDW-interpolated maps provide strong visual and quantitative evidence of a long-term deterioration in water quality driven by increasing urban development. The spatial insight gained from this analysis highlights priority zones for policy intervention, wastewater control, and land use planning within the Tha Chin Basin.

5. CONCLUSION

This study demonstrates the effectiveness of combining field observations, machine learning (ML), and Geographic Information System (GIS) techniques to assess and predict water quality trends in the Tha Chin River Basin, Thailand. The approach analyzed key physicochemical parameters and spatial variables, including urban land expansion and population density, to generate a multidimensional understanding of river water quality between 2020 and 2022.

Spatial and temporal analysis of the Water Quality Index (WQI) revealed a clear degradation pattern throughout the study area. IDW-interpolated maps showed that all monitoring locations were classified as either "moderately degraded" or "degraded," with no sites reaching "good" or "excellent" quality. The temporal decline was most pronounced in 2022, with the poorest water quality consistently recorded in downstream districts such as Banglen. Conversely, Mueang Suphan Buri, located in the upstream portion of the basin, maintained a moderately degraded status and showed slight improvement over the three years.

In parallel, land use analysis confirmed a notable increase in urban and built-up areas between 2020 and 2022, especially in Suphan Buri and Nakhon Pathom provinces. This urban expansion corresponded with the declining WQI values in affected districts, reinforcing the impact of human-induced pressures on surface water quality.

To support predictive decision-making, six supervised ML algorithms were tested. The Random Forest model achieved the best performance ($R^2 = 0.664$, MAE = 0.855) at a 10-hour lead time and showed consistent accuracy across other forecast intervals. The inclusion of spatial indicators—such as population density and urban land cover—substantially enhanced model performance.

In conclusion, this study provides compelling evidence of spatially uneven and temporally worsening water quality conditions along the Tha Chin River. The ML-GIS integrated framework offers a scalable tool for identifying pollution hotspots, evaluating long-term degradation, and guiding targeted interventions in rapidly urbanizing river basins.

REFERENCES

Abbasi, T., & Abbasi, S. A. (2012). Water-Quality Indices. In *Water Quality Indices* (pp. 353-356).

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water, 11*(11). doi:10.3390/w11112210

Al-Faiz, M. Z., Ibrahim, A. A., & Hadi, S. M. (2019). The effect of Z-Score standardization (normalization) on binary input due the speed of learning in back-propagation neural network. *Iraqi Journal of Information & Communications Technology, 1*(3), 42-48. doi:10.31987/ijict.1.3.41

Belete, D. M., & Huchaiah, M. D. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications, 44*(9), 875-886. doi:10.1080/1206212x.2021.1974663

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., . . . Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res, 171*, 115454. doi:10.1016/j.watres.2019.115454

D. Satish Chandra, S. A., M.V.S. Raju. (2017). ESTIMATION OF WATER QUALITY INDEX BY WEIGHTED ARITHMETIC WATER QUALITY INDEX METHOD A MODEL STUDY. *International Journal of Civil Engineering and Technology (IJCIET) 8*(4), 8.

Fathi, P., Ebrahimi Dorche, E., Zare Shahraki, M., Stribling, J., Beyraghdar Kashkooli, O., Esmaeili Ofogh, A., & Bruder, A. (2022). Revised Iranian Water Quality Index (RIWQI): a tool for the assessment and management of water quality in Iran. *Environ Monit Assess, 194*(7), 504. doi:10.1007/s10661-022-10121-9

Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Mar Pollut Bull, 64*(11), 2409-2420. doi:10.1016/j.marpolbul.2012.08.005

Horton R.K. (1965). An index number for rating water quality. Journal of Water Pollution Control Federation. *37*(3), 300-306.

Jha, D. K., Devi, M. P., Vidyalakshmi, R., Brindha, B., Vinithkumar, N. V., & Kirubagaran, R. (2015). Water quality assessment using water quality index and geographical information system methods in the coastal waters of Andaman Sea, India. *Mar Pollut Bull, 100*(1), 555-561. doi:10.1016/j.marpolbul.2015.08.032

Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science, 11*(12). doi:10.1007/s13201-021-01528-9

Liu, P., Wang, J., Sangaiah, A., Xie, Y., & Yin, X. (2019). Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. *Sustainability, 11*(7). doi:10.3390/su11072058

Lumb, A., Halliwell, D., & Sharma, T. (2006). Application of CCME Water Quality Index to monitor water quality: a case study of the Mackenzie River Basin, Canada. *Environ Monit Assess, 113*(1-3), 411-429. doi:10.1007/s10661-005-9092-6

Ly, Q. V., Nguyen, X. C., Le, N. C., Truong, T. D., Hoang, T. T., Park, T. J., . . . Hur, J. (2021). Application of Machine Learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the Han River, South Korea. *Sci Total Environ, 797*, 149040. doi:10.1016/j.scitotenv.2021.149040

Magesh, N. S., Krishnakumar, S., Chandrasekar, N., & Soundranayagam, J. P. (2012). Groundwater quality assessment using WQI and GIS techniques, Dindigul district, Tamil Nadu, India. *Arabian Journal of Geosciences, 6*(11), 4179-4189. doi:10.1007/s12517-012-0673-8

Magyari-Sáska, Z., Haidu, I., & Magyari-Sáska, A. (2025). Experimental Comparative Study on Self-Imputation Methods and Their Quality Assessment for Monthly River Flow Data with Gaps: Case Study to Mures River. *Applied Sciences, 15*(3). doi:10.3390/app15031242

Marzieh Mokarram , & Zarei, A. R. (2021). Determining prone areas to gully erosion and the impact of land use change on it by using multiple-criteria decision-making algorithm in arid and semi-arid regions. *Geoderma, 403*.

Mueller, T. G., Pusuluri, N. B., Mathias, K. K., Cornelius, P. L., Barnhisel, R. I., & Shearer, S. A. (2004). Map Quality for Ordinary Kriging and Inverse Distance Weighted Interpolation. *soil science Society of America Journal, 68*(6).

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., . . . Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology, 578*. doi:10.1016/j.jhydrol.2019.124084

PCD. (2017). *Total Water Quality Score for 5 Parameters. (New WQI Calculation)*. Pollution Control Department: Pollution Control Department

PCD, P. C. D. (2020). *A Decade of Water Quality Monitoring in Thailand's FourMajor Rivers The Result and the Implications for Management*. Pollution Control Department

PCD, P. C. D. (2022). *Water quality report for the year 2022*. Pollution Control Department: Pollution Control Department

Senthilnathan, S. (2019). Usefulness of Correlation Analysis. *SSRN Electronic Journal*. doi:10.2139/ssrn.3416918

Tabios, G. Q., & Salas, J. D. (1985). Geographic Spatial Distribution Map. *American Water Resouces Association, 21*(3), 16.

Wang, R., Kim, J. H., & Li, M. H. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Sci Total Environ, 761*, 144057. doi:10.1016/j.scitotenv.2020.144057