Environmental Science

http://www.aimspress.com/journal/environmental

AIMS Environmental Science, 8(3): 268–282
DOI: 10.3934/environsci.2021018
Received: 01 February 2021
Accepted: 25 May 20211
Published: 07 July 2021

Research article

Gaussian process regression for predicting water quality index: A case study on Ping River basin, Thailand

Kamonrat Suphawan*and Kuntalee Chaisee

Data Science Research Center, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand

* **Correspondence:** Email: kamonrat.s@cmu.ac.th.

Abstract: The water quality index (WQI) is an aggregated indicator used to represent the overall quality of water for any intended use. It is typically calculated from several biological, chemical, and physical parameters. Assessment of factors that affect the WQI is then essential. Climate change is expected to impact a wide range of water quality issues; hence, climate variables are likely to be significant factors to evaluate the WQI. We propose three statistical models; multiple linear regression (MLR), artificial neuron network (ANN), and Gaussian process regression (GPR) to assess the WQI using the climate variables. The data is the WQI of Ping River, which flows through the provinces in the north of Thailand. The climate variables are temperature, humidity, total rainfall, and evaporation. A comparison between these models is determined by model prediction accuracy scores. The results show that the total rainfall is the most significant variable to predict the WQI for the Ping River. Although these three methods can predict the WQI relatively good, overall, the GPR model performs better than the MLR and the ANN. Besides, the GPR is more flexible as it can relax some restrictions and assumptions. Therefore, the GPR is appropriate to assess the WQI under the climate variables for the Ping River.

Keywords: water quality index; multiple linear regression; artificial neural networks; Gaussian process regression; forecasting

1. Introduction

River water is a vital surface water resource for households, agriculture, and industry activities. It also plays an important role in health and environmental issues. Thus, assessing and monitoring the quality and quantity of the water carried in the River is essential. A water quality index (WQI) is a single number that provides information about overall water quality. It is calculated using a parametric expression from several biological, physical, and chemical parameters measured from a water sample

at a location and time. The parametric expression for WQI calculation involves the use of relative weights per involved parameter. According to the Water Quality Management Bureau, Pollution Control Department, Thailand, five water parameters including dissolved oxygen (DO), biological oxygen demand (BOD), total coliform bacteria (TCB), fecal coliform bacteria (FCB), and ammonia nitrogen (NH_3-N) are employed for the estimation of WQI. Moreover, the WQI is often represented as five levels of water quality; very good (91-100), good (71-90), average (61-70), poor (31-60), and very poor (0-30).

Many researchers have been interested in examining factors that affect water quality, both man-made and environmental. However, changes in climate conditions such as the increase of water temperature, precipitation and evaporation patterns, and heavy rainfall and flooding mainly affect the biological and chemical properties of water quality. As a result, the relationship between climate factors and WQI has been studied, as well as finding an appropriate statistical model to represent the relationship. The WQI parameters were estimated for water streams in Finland with air temperature, rainfall runoff, and precipitation and characteristics of catchment areas as independent variables using artificial neural network (ANN) model [1]. The relationship between the WQI and climate variables on the Euphrates River within Karbala city, Iraq, has been investigated [2]. The non-linear regression and ANN models were employed to forecast the relationship between the WQI and temperature, relative humidity, rainfall depth, and sunshine duration. The non-linear regression model predicted the WQI better than the ANN models. The multiple linear regression (MLR) is used to study the relations between climate variables: temperature, relative humidity, and selected water quality parameters in Lake Manzala, Egypt [3]. They found a positive relation between studied variables. Recently, [4] used MLR and ANN models for predicting the stream water quality parameters on Green River watershed, Kentucky, USA, with independent variables precipitation, temperature, and land-use data.

Although MLR and ANN are widely used to model the relationship between the WQI parameters and climate variables, choosing the suitable regression model and dealing with restrictions in underlying assumptions in MLR modeling could be challenging. One of the difficulties in ANN modeling is choosing an appropriate number of hidden nodes that can affect its performance. The Gaussian process (GP) is a Bayesian machine learning method which has gained attention due to its flexibility in modeling. A GP can be applied in regression analysis, called Gaussian process regression (GPR). It has been applied and compared to estimating and forecasting models in many fields. For example, in health science, [5] used the GPR to estimate the child mortality rate in Iran in 1990 - 2013. They found that the GPR can efficiently estimate the mortality rate with relatively high fitting precision and flexibility. In earth science, [6] used the capability of GPR to predict the porosity and permeability of the southern basin of the South Yellow Sea. They compared the performance of the GPR to the back propagation neural network (BPNN), generalized regression neural network (GRNN), and radial basis function neural network (RBFNN). In food science, [7] applied the GPR to model the drying time of mosambi peel. The study showed that the GPR could be used as an alternative method because it provided a better estimation than the ANN and the response surface method. These findings mentioned above revealed that the GPR is one of the most powerful techniques and can be used in diverse fields.

In this work, we want to use the GPR, MLR, and ANN to predict the WQI using climate variables. The data used in the study is the WQI of the Ping River, the major River in the north of Thailand. The climate-related variables are temperature, humidity, rainfall, and evaporation. We also assess the performance of the prediction ability of the models.

2. Materials and methods

2.1. Study area and data

The Ping River is located in the central part of northern Thailand (19°48′45”N 98°50′20”E). The Ping River originates in Chiang Dao district, Chiang Mai Province. It then flows through the provinces of Lamphun, Tak, Kamphaeng Phet, and Nakhon Sawan. Its estimated length is 658 km, with catchment areas of around 34,885 km^2. The average discharge is about 265 m^3/s. The Wang River is its main tributary. The Ping river joins the combined Nan and Yom rivers to form the Chao Phraya River, the major River in Thailand, formed in the center of the country and flows through Bangkok and then into the Gulf of Thailand. The Ping river basin is in the area, where is mainly influenced by the Southwest and Northeast monsoon between mid-May or June to mid-October.

The WQI of the rivers in Thailand is monitored by the Pollution Control Department, Water Quality Management Bureau, The Ministry of Natural Resources and Environment of Thailand. There are 14 water quality monitoring stations located in many areas along the basin, shown in Table 1. The spatial plot of the stations and areas of study is illustrated in Figure 1. The WQI is measured in three-month intervals, mostly in February, May, August, and November.

The climate data are daily provided by the Northern Meteorological Center, Meteorological Department, Ministry of Digital Economy and Society of Thailand. To related the WQI and climate data, we then use the monthly average temperature ($^\circ C$), monthly average humidity (%), monthly total rainfall (mm.), and monthly average evaporation (mm.) as climate variables in this work.

The WQI and climate data originate from different sources that unfortunately are collected in different locations and timespans. Therefore, data were filtered and only those collected at the same locations and timespans (January 2010-December 2019) were selected for this analysis. Consequently, the selected data used in modeling originate from eight water monitoring stations located in four areas as follows: (1) Mueang-KamphengPhet (PI04 and PI05), (2) Mueng-Tak (PI06, PI07, and PI08) (3) SamNgao-Tak (PI09), and (4) Mueang-Chiang Mai (PI12 and PI13). In each station, we expect 40 WQI observations, however, there are some missing values. Therefore, the selected data consist of 284 observations from eight stations.

Table 1. List of 14 stations on the Ping River basin.

Station	Station name	District	Province	latitude	longitude
PI01	Phitsanulok Bridge	Mueang	Nakhon Sawan	15.7114	100.146
PI02	Tong Kung bridge	Banphot Phisai	Nakhon Sawan	15.9354	99.9768
PI03	Saen To bridge	Khanu Woralaksaburi	Kamphaeng Phet	16.0643	99.8602
PI04	Wang Yang bridge	Mueang	Kamphaeng Phet	16.3767	99.5691
PI05	Lan DokMai bridge	Mueang	Kamphaeng Phet	16.6278	99.4320
PI06	Ta Ta-Kraw bridge	Mueang	Tak	16.7931	99.1707
PI07	Kittikachorn Bridge	Mueang	Tak	16.8563	99.1247
PI08	WangMoung bridge	Mueang	Tak	16.9604	99.1138
PI09	Ban Tak Bridge	Sam Ngao	Tak	17.0417	99.0677
PI10	Kong Hin bridge	Hod	Chiang Mai	18.1787	98.6303
PI11	Had Nak bridge	Jomthong	Chiang Mai	18.3482	98.6973
PI12	Police Station bridge	Mueang	Chiang Mai	18.7599	98.9972
PI13	Wang Sing Khum bridge	Mueang	Chiang Mai	18.8103	99.0032
PI14	Cho lae bridge	Mae Taeng	Chiang Mai	19.1459	99.0074

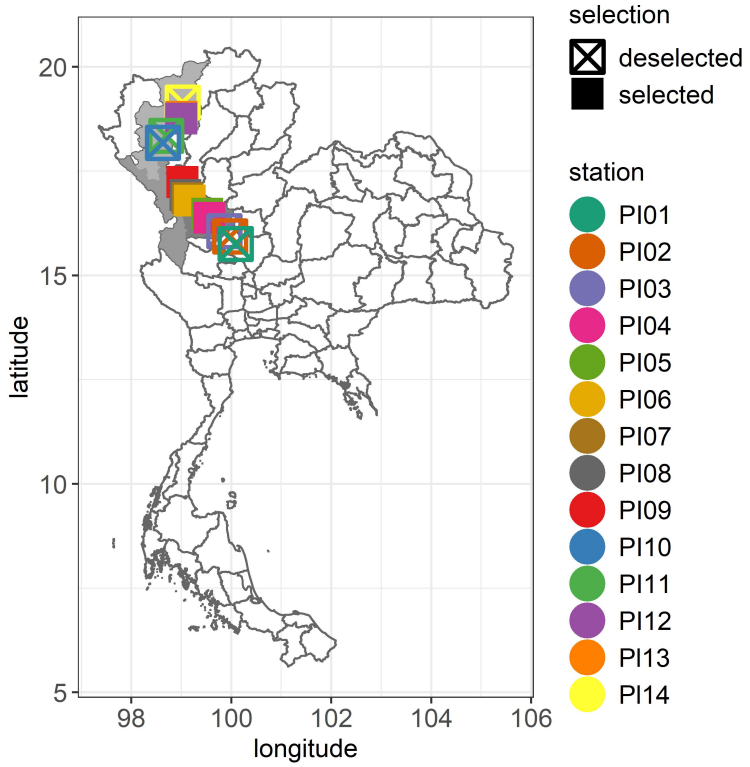

Figure 1. Location of stations.

2.2. Multiple linear regression (MLR)

Multiple linear regression (MLR) attempts to predict a dependent or response variable, y, on the basis of an assumed linear relationship with several independent or explanatory variables, $x_1, x_2, ..., x_d$. The MLR model can be expressed as

$$y = f(x_1, ..., x_d) + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d + \epsilon, \tag{1}$$

where $f(.)$ is a transition function, mapping the relationship between the response and independent variables. The ϵ is a random error assumed to be independent and identical normally distributed with zero mean and constant variance, $\epsilon \sim N(0, \sigma_\epsilon^2)$. For n observations, the model can be expressed in the form of vector and matrix as

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y}$ is $n \times 1$ vector of observed values of the response variable, $\mathbf{X}$ is a $n \times (d + 1)$ matrix of independent variables, $\boldsymbol{\beta}$ is $(d + 1) \times 1$ vector of regression coefficient parameters, and $\boldsymbol{\epsilon}$ is $n \times 1$ vector of random errors. The parameter estimation for $\boldsymbol{\beta}$ and σ_ϵ^2 can be performed using several approaches such as the least square and maximum likelihood estimation. After model fitting, the model adequacy and validation are investigated using some properties of the residuals, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, where $\mathbf{y}$ is the observed responses vector and $\hat{\mathbf{y}}$ is the predicted response obtained from the fitted model. The assumptions of the MLR require that residuals are normal distributed with no systematic pattern, constant variance, and no outliers.

In this study, we use a stepwise method to select the independent variables. The variance inflation factor (VIF) is used to investigate the multicollinearity issue among independent variables. The residual assumptions are diagnosed by Breusch-Pagan (BP) test for testing the constant variance assumption, Shapiro-Wilk (SW) test for normality, and the Durbin-Watson (DW) test for autocorrelation. The Box-Cox transformation is employed to find the best transformation if the assumption(s) is violated.

2.3. Artificial neural network (ANN)

An artificial neural network (ANN) are multi-layer fully-connected neural nets that consist of three-layer processing units; an input layer, multiple hidden h layers, and an output layer. The model is inspired by the human brain that tries to find data structures and algorithms for learning and classifying data. The relation between the output y and the input $(x_1, x_2, ..., x_d)$ can be written as follow

$$y = w_0 + \sum_{j=1}^{h} w_j \cdot g\left(w_{0,j} + \sum_{k=1}^{d} w_{k,h} x_k\right) + \epsilon, \tag{3}$$

where $w_{k,h}$ with $k = 1, ..., p; j = 1, 2, ..., h$ is the connection weight, d and h are number of input vectors and number of hidden nodes, respectively. Moreover, g is a sigmoid transfer function, w_0 and $w_{0,j}$ are weights from the bias terms, and ϵ is the error term. Each of the inputs is multiplied by a connection weight or synapse. A given node takes the weighted sum of its inputs and passes it through a non-linear activation function. The output of the node then becomes the input of another node in the next layer. The number of nodes of the input layer corresponds to the number of variables describing the attributes being tested. The weight parameters of the ANN model are estimated by optimization solution to minimize the sum of squared of residual. The package `neuralnet` in R programming [8] ,

by training of neural networks using backpropagation network, is used in this study for obtaining the weights estimates and fitted responses. More details about ANN modeling can be found in [9].

2.4. Gaussian process regression (GPR)

A Gaussian process (GP) is a stochastic process involving random variables, any finite number of which have a joint Gaussian distribution. It is considered to be a non-parametric method for modeling data, as the model structure is determined from the data rather than through a parametric model. In the same way that a Gaussian random variable is characterized by its mean and variance, a GP is completely characterized by its mean function and covariance function, which are functions of the input vector. We define the input vector $\mathbf{x}$ as a collection of inputs, where $\mathbf{x} = (x_1, ..., x_n)^{'}$. We use the following notation to denote that $f(.)$ is a Gaussian process:

$$\begin{aligned} f(.) &\sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^{'})), \\ \text{where} \quad m(\mathbf{x}) &= \mathrm{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}^{'}) &= \mathrm{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^{'}) - m(\mathbf{x}^{'}))] = \mathrm{Cov}[f(\mathbf{x}), f(\mathbf{x}^{'})]. \end{aligned} \tag{4}$$

The $m(\mathbf{x})$ is the mean function and $k(\mathbf{x}, \mathbf{x}^{'})$ is the covariance function. The mean function has the effect of shifting a zero mean GP by the amount of $m(\mathbf{x})$, often set to be zero. It is the covariance function, which largely determines the properties of samples from the GP model. In this study, we employ the squared exponential (SE) covariance, which is the most commonly used in GP modeling. If we consider the input $\mathbf{X}$ with d-dimensional, the covariance function has the form of a direct sum as follows

$$k_{SE}(x, x^{'}) = \sigma_{gp}^2 \exp\left[\sum_{k=1}^{d} -l_k \left(x - x^{'}\right)^2\right]. \tag{5}$$

The SE covariance function is often controlled by parameters, called hyper-parameters: σ_{gp}^2 controls the amplitude of variation of the sample functions and the correlation parameters $l_k, k = 1, ..., d$, controls the smoothness of the samples. We define θ to be a set of hyper-parameters for a given mean and covariance function, $\theta = (\sigma_{gp}^2, \{l_k\}_{k=1}^{d})$. More details of the GP modeling can be found in [10].

In regression modeling, we can instead model the transition function $f(x)$ using a GP as

$$f(\mathbf{X}) \sim GP(m(\mathbf{X}), k(\mathbf{X}, \mathbf{X}^{'})).$$

Given a collection of inputs $\mathbf{X}$, the vector $\mathbf{f} = f(\mathbf{X})$ has a multivariate Gaussian distribution,

$$\mathbf{f}|\mathbf{X} \sim N(m(\mathbf{X}), K),$$

where $m(\mathbf{X})$ is the prior mean vector, and $K = k(\mathbf{X}, \mathbf{X}^{'})$ is the covariance matrix or Gram matrix.

Consider regression model $\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is the additive Gaussian noise $N(0, \sigma_{\epsilon}^2)$, the distribution of the output $\mathbf{y}$ is then $\mathbf{y}|\mathbf{f} \sim N(\mathbf{f}, \sigma_{\epsilon}^2 I)$, and $\mathbf{y}|\mathbf{x} \sim N(m(\mathbf{X}), K + \sigma_{\epsilon}^2 I)$, where I is the identity matrix.

Given a set of training input vectors and output, we are often interested in predicting test output $\mathbf{y}^* = \left(y_1^*, ..., y_m^*\right)$ and latent function variables $\mathbf{f}^* = f(\mathbf{X}^*)$, at test input $\mathbf{X}^*$, where $\mathbf{X}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_d^*)$.

We can write the joint distribution of training output $\mathbf{y}$ and latent function variables $\mathbf{f}^*$, and training output $\mathbf{y}$ and test output $\mathbf{y}^*$ as follow

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix} \sim N\left(\begin{bmatrix}m(\mathbf{X})\\ m(\mathbf{X}^*)\end{bmatrix}, \begin{bmatrix}K+\sigma_\epsilon^2 I & k(\mathbf{X}^*,\mathbf{X})\\ k(\mathbf{X},\mathbf{X}^*) & k(\mathbf{X}^*,\mathbf{X}^*)\end{bmatrix}\right),$$
$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}^*\end{bmatrix} \sim N\left(\begin{bmatrix}m(\mathbf{X})\\ m(\mathbf{X}^*)\end{bmatrix}, \begin{bmatrix}K+\sigma_\epsilon^2 I & k(\mathbf{X}^*,\mathbf{X})\\ k(\mathbf{X},\mathbf{X}^*) & k(\mathbf{X}^*,\mathbf{X}^*)+\sigma_\epsilon^2 I\end{bmatrix}\right).$$

Bayes's rule is then applied to obtain the joint posterior distribution of training and test latent variables given the training outputs

$$p(\mathbf{f},\mathbf{f}^*|\mathbf{y}) = \frac{p(\mathbf{f},\mathbf{f}^*)p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y})}.$$

The predictive distribution of $\mathbf{f}^*$ at given test locations can be derived using conditional distribution of two Gaussian random variables, yields

$$\begin{aligned}\mathbf{f}^*|\mathbf{X}^*,\mathbf{y},\theta &\sim N(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*),\\ \text{where}\quad \boldsymbol{\mu}^* &= m(\mathbf{X}^*)+k(\mathbf{X}^*,\mathbf{X})(K+\sigma_\epsilon^2 I)^{-1}(\mathbf{y}-m(\mathbf{X})),\\ \boldsymbol{\Sigma}^* &= k(\mathbf{X}^*,\mathbf{X}^*)-k(\mathbf{X}^*,\mathbf{X})(K+\sigma_\epsilon^2 I)^{-1}k(\mathbf{X},\mathbf{X}^*).\end{aligned} \tag{6}$$

The predictive distribution of the target outputs, $\mathbf{y}^*$ is

$$\mathbf{y}^*|\mathbf{X}^*,\mathbf{y},\theta \sim N(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*+\sigma_\epsilon^2\mathbf{I}). \tag{7}$$

The predictions are the mean vector $\boldsymbol{\mu}^*$, and variances can be obtained from the diagonal of the covariance matrix $\boldsymbol{\Sigma}^*+\sigma_\epsilon^2\mathbf{I}$. The package `mlegp` in R programming [11] is used to obtain the hyper-parameter estimates and the GPR predictive mean and variances, as shown in Equation (7).

2.5. Evaluation of prediction accuracy

The scores used to compare the predictive performance of the models are the root mean squared error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE), given by

$$\begin{aligned}\text{RMSE} &= \sqrt{\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{n}},\\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^n |y_i-\hat{y}_i|,\\ \text{MAPE} &= \frac{1}{n}\sum_{i=1}^n \frac{|y_i-\hat{y}_i|}{y_i},\end{aligned}$$

where y_i and $\hat{y}_i$ are the i^{th} observed and the predicted response for $i = 1,...,n$ and n is the number of observations. These scores are negatively-oriented: lower values are better. Moreover, the plot between the observed and the predicted is illustrated along with the diagonal line. A perfect alignment with the line indicates no difference between observed and predicted, yields perfect prediction. The coefficient of determination (r^2) is computed to measure how well the model predicts the data, given by

$$r^2 = 1-\frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{\sum_{i=1}^n (y_i-\bar{y})^2}.$$

3. Results

3.1. Exploratory data analysis

The descriptive statistics of the WQI and climate data of the selected stations and areas are shown in Table 2. The time series plots and boxplots of the WQI from 8 stations and the climate data from four areas are displayed in Figure 2 and 3, respectively.

The means and medians of the WQI are between 58-70. The minimum is 41, and the maximum is 89 at station PI13 and PI04, respectively. The standard deviations are similar in many stations, but at station PI04 and PI09, they are slightly higher than the others. The boxplots show that there are some outliers at stations PI04, PI05, PI06, and PI08. According to Thai Meteorological Department, Thailand has three seasons: summer (February-May), rainy season (June – September), winter (October – January). The statistics of WQI in the seasons are as follows: the means (standard deviation) are 66 (8.46), 62.7 (7.93), and 66.73 (8.39). This means that the WQI in the rainy season is, on average, lower than summer and winter season. However, there is no clear patterns on the variability of WQI by area, as seen in Figure 2.

For climate data, the monthly average temperature (AvgTemp), average humidity (AvgHumid), and average evaporation (AvgEvapor) have similar means and medians in every area. The means of the monthly total rainfall (TotalRainfall), particularly in KamphengPhet and Chiang Mai, are relatively high, whereas the medians are not much different. This indicates the extreme values in the total rainfall data. Moreover, we provide the scatterplot matrix of all data in Figure 4 to show the relationship among variables. The small values of the correlation coefficients indicate a weak relationship between the WQI and the climate variables. Although the plots suggest that the climate variables are moderately correlated, the VIFs indicate no multicollinearity issue.

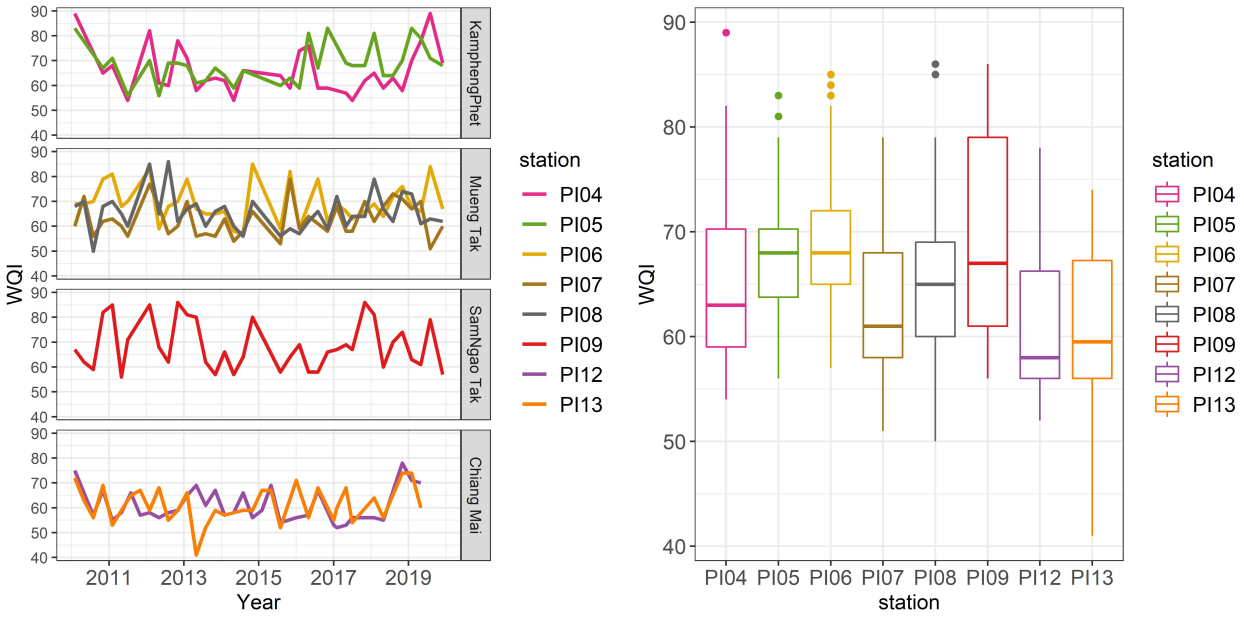

Figure 2. Time series plots and boxplots of the WQI data.

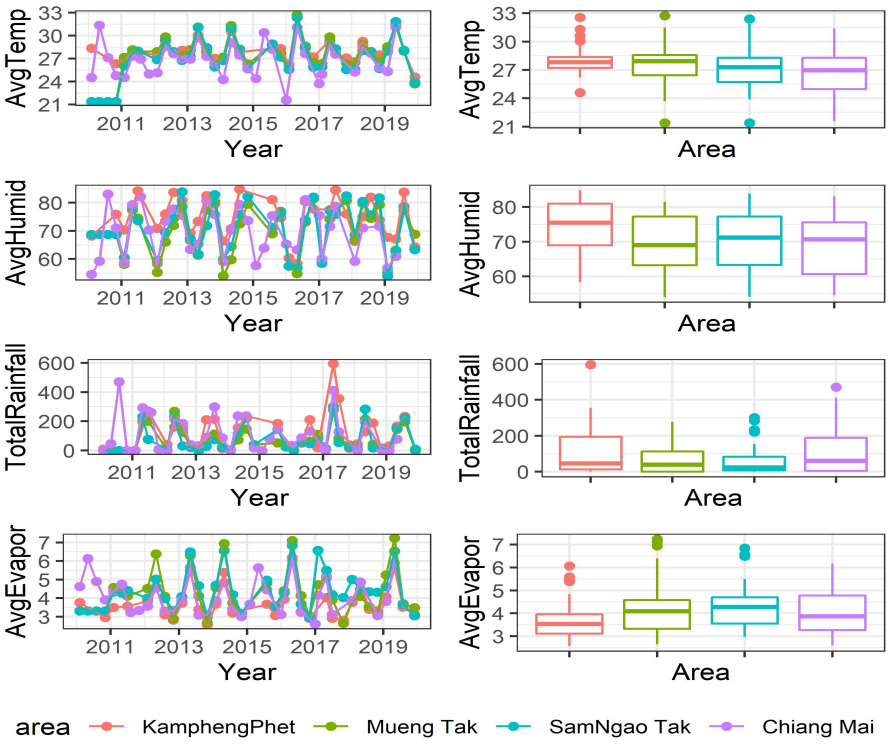

Figure 3. Time series plots and boxplots of the climate data.

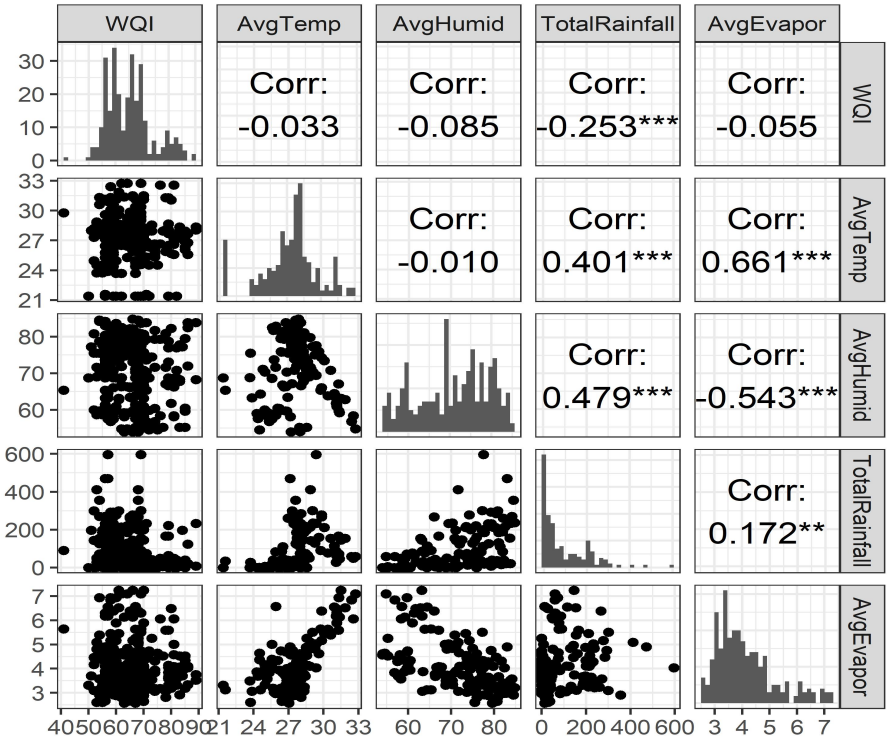

Figure 4. Scatterplot matrix of the data.

Table 2. Descriptive statistics of the data.

Station	Area	Variable	Mean	SD	Median	Min	Max
PI04	KamphengPhet	WQI	65.88	9.49	63.00	54.00	89.00
PI05	KamphengPhet		68.31	7.70	68.00	56.00	83.00
PI06	Mueng Tak		69.41	7.64	68.00	57.00	85.00
PI07	Mueng Tak		62.73	6.86	61.00	51.00	79.00
PI08	Mueng Tak		65.54	7.51	65.00	50.00	86.00
PI09	SamNgao Tak		68.57	9.70	67.00	56.00	86.00
PI12	Chiang Mai		60.92	6.72	58.00	52.00	78.00
PI13	Chiang Mai		61.47	7.36	59.50	41.00	74.00
	KamphengPhet	AvgTemp	28.09	1.54	27.83	24.58	32.57
	Mueng Tak		27.32	2.69	27.93	21.39	32.73
	SamNgao Tak		27.00	2.69	27.29	21.39	32.40
	Chiang Mai		26.76	2.41	26.97	21.59	31.37
	KamphengPhet	AvgHumid	74.61	7.29	75.46	58.38	84.82
	Mueng Tak		69.68	8.47	69.01	53.98	81.43
	SamNgao Tak		70.77	8.60	71.20	54.04	83.85
	Chiang Mai		69.13	8.49	70.68	54.57	83.03
	KamphengPhet	TotalRainfall	117.4	132.8	46.65	0.00	595.9
	Mueng Tak		71.31	80.99	39.90	0.00	277.6
	SamNgao Tak		68.86	89.96	24.00	0.00	300.8
	Chiang Mai		107.2	123.9	61.45	0.00	470.6
	KamphengPhet	AvgEvapor	3.75	0.87	3.55	2.57	6.07
	Mueng Tak		4.26	1.19	4.10	2.65	7.24
	SamNgao Tak		4.38	1.08	4.28	2.97	6.85
	Chiang Mai		4.08	0.98	3.88	2.60	6.17

3.2. Modeling

The dataset is divided into two sets, training and test set, to validate the model performance for seen and unseen data. The output data is the WQI. The input data are the 4 climate variables: monthly average temperature, monthly average humidity, monthly total rainfall, and monthly average evaporation. All data in year 2010-2018 is selected to be the training set with 256 observations or 90.14% of the total number of observations. All data in the year 2019 is used as the test set with 28 observations or 9.86% of the total number of observations. Hence, the dimension of inputs for training and test set are 256 rows $\times$ 4 columns and 28 rows $\times$ 4 columns, respectively.

We normalize the data before modeling using the standardization technique. Therefore, each variable will be centered around zero and have approximately unit variance. The training data is used in modeling with three methods; MLR, ANN, and GPR. For the MRL, initially, we consider the full model with all four climate variables denoted by MLR1. After that, we use the stepwise method for model selection, and we found that only TotalRainfall is included in the model denoted by MLR2. In fact, the MLR1 and MLR2 do not meet the normality and independent assumptions. Then we use the Box-Cox transformation to MRL1 and MLR2, and it suggests a transformation with $\lambda = -0.5$ denoted by MLR3

and MLR4, respectively. As a result, the residuals assumptions for MLR3 and MRL4 are acceptable. The number of nodes in the hidden layer, h, is varied. We use the same set of climate variables used in the MLR1 and MLR2 denoted as ANN1 and ANN2. We found that the best number of hidden nodes, according to the least RMSE for ANN1 and ANN2, are 4 nodes and 1 node. The underlying GP is set to have zero mean and SE covariance function with unknown nugget variance. We also use the same set of climate variables in the ANN models denoted by GPR1 and GPR2. The fitted results from 3 approaches with 8 models are shown in Tables 3 - 5

Table 3. The coefficient estimates and the diagnostic tests of MLR models.

	Constant	AvgTemp	AvgHumid	TotalRainfall	AvgEvapor	p-value		
						BP test	SW test	DW test
MLR1	58.3911	0.373	0.016	-0.0254	-0.6087	0.1067	3.31e-05	0.0091
MLR2	66.9714			-0.0226		0.0059	7.58e-05	0.0091
MLR3	0.1311	3.60e-04	-1.61e-05	2.32e-05	6.52e-04	0.3220	0.0029	0.0109
MLR4	0.123			2.06e-05		0.0741	0.0049	0.0109

Table 4. The weights of ANN models associated to the input.

	AvgTemp	AvgHumid	TotalRainfall	AvgEvapor	w_0
ANN1	-0.8071	-3.8269	82.3872	-81.9345	-0.9049
ANN2			3.7453		0.3386

Table 5. The hyper-parameters estimates of GPR models.

	l_1	l_2	l_3	l_4	σ^2_{gp}	σ^2_ϵ
GPR1	5.6634	7.3907	1.35e-07	13.2	0.4708	0.6094
GPR2			2.32e+10		0.4206	0.6606

3.3. Comparison of models

To validate the prediction performance of the models, the predictive accuracy scores are presented in Table 6. The scores indicate that GPR models outperform the MLR and ANN models for both training and test sets, with the least values in RMSE, MAE, and MAPE. Besides, three scores from all models suggest that the methods perform better in the training set compared to the test set. The plots between the observed and predicted of the WQI of the training set (black dots) and the test set (red dots) are illustrated in Figure 5. We can observe that, for the training set, the results from the GPRs are more concentrated around the diagonal line than those from the MLRs and ANNs. This confirms that the GPRs give a better prediction, which corresponds to the highest values of r^2. Nevertheless, it is essential to point out that the ANN and GPR models perform well only when the WQI is in the range of 50 to 70. This can be explained by the density plot of WQI in the training and test set, shown in Figure 6. The WQI in the training set is mostly in range of 50 to 70, while in the test set, the range is from 50 to 90. The models are unable to capture the information and predict when the data are higher

than 70. Consequently, the accuracy of the prediction is relatively poor in the test set, particularly for the WQI over 70.

Table 6. Predictive accuracy scores.

	Training set				Test set			
	RMSE	MAE	MAPE	r^2	RMSE	MAE	MAPE	r^2
MLR1	7.8748	6.2603	0.0966	0.0985	9.9875	7.4044	0.1020	0.0353
MLR2	7.9071	6.2212	0.0960	0.0911	10.1451	7.4621	0.1028	0.0698
MLR3	7.8893	6.2234	0.0949	0.1036	10.2707	7.7134	0.1057	0.0292
MLR4	7.9275	6.2204	0.0949	0.0948	10.3962	7.7563	0.1063	0.0683
ANN1	6.3748	5.1402	0.0795	0.4092	10.2884	8.2624	0.1183	0.0035
ANN2	7.6810	6.0743	0.0937	0.1423	9.6650	7.6470	0.1099	0.0180
GPR1	5.7006	4.4945	0.0692	0.6056	8.8942	6.9560	0.1037	0.0008
GPR2	5.9578	4.6782	0.0722	0.5984	8.5610	6.6551	0.0978	NA

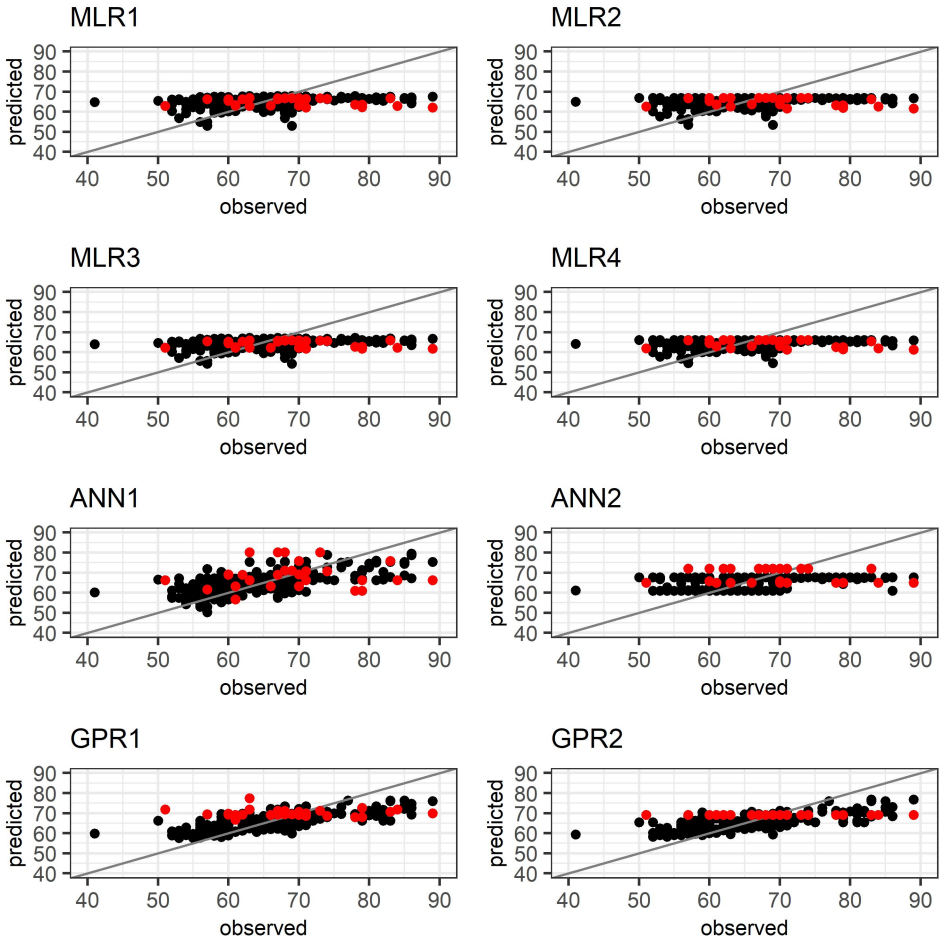

Figure 5. Plots between the predicted and observed WQI.

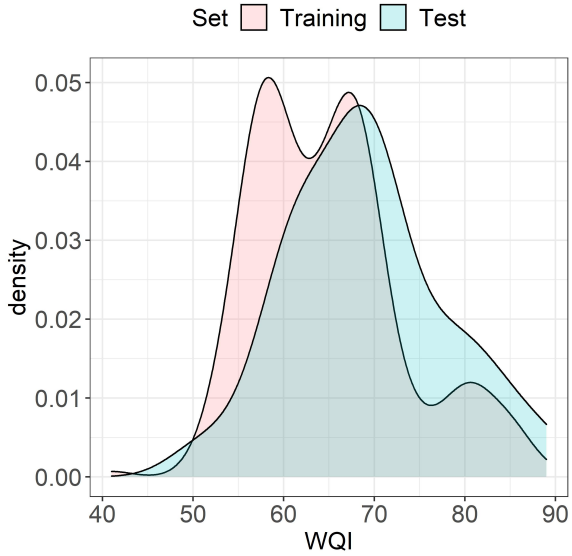

Figure 6. Density plot of WQI in training and test set.

We also investigate the distribution of residuals produced from all models using boxplots shown in Figure 7. For the training set, all models generate residuals with the center location of zero. The deviation of residuals is small in the case of GPRs compared to MLRs and ANNs. For the test set, the mean and median of residuals produced from MLRs are most deviated from zero, followed by ANNs and GPRs. The standard deviations of the residuals from the test set are not significantly different, but they are still larger than those from the training set.

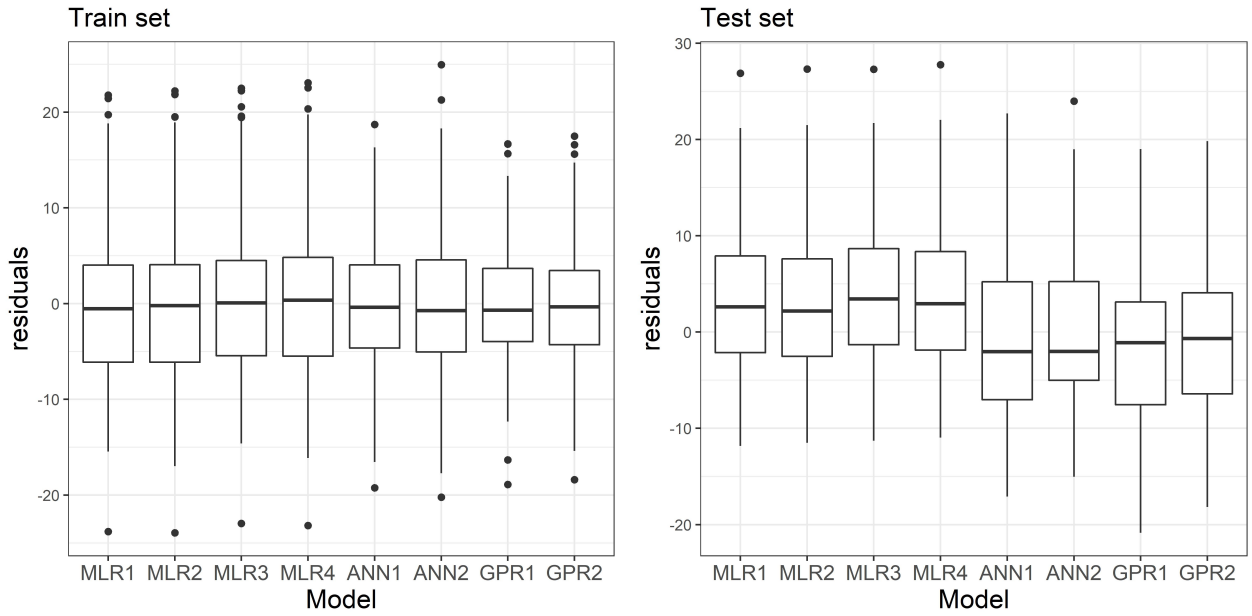

Figure 7. Boxplots of the residuals.

4. Discussion

Water quality is influenced by climate variabilities such as precipitation, temperature, rainfall level, and wind patterns. In this study, we assess the impact of climate change on the water quality of Ping River, Thailand. We focus on investigating the association of the climate data; temperature, humidity, total rainfall, and evaporation, and the water quality index using statistical modeling. An important result of this study is to develop a predictive model to assess how climate variables affect the WQI. The result suggested that the amount of total rainfall is the only significant climate variable that impacts WQI. The limitation of the available data allows us to analyze the WQI and climate data on the different locations. Moreover, the WQI is not collected in the same period as the climate data. These could be the main reasons that affect our results.

Three statistical models, MLR, ANN, and GPR, are used to analyze the relationship between WQI and climate data. The MLR is often used to describes the relationship between inputs and outputs by specifying a functional form mapping from inputs to outputs. We often set the function to be some specific form, such as combinations of linear, cubic, and higher-order polynomial terms with unknown parameters. However, choosing the correct function can be difficult. The ANN relies on the optimal number of nodes to obtain a better estimation. Unlike MLR and ANN, the GPR can learn the data well without specifying the function between inputs and outputs. It is considered a more flexible way to model the data. However, the GPR can be computationally expensive when the sample size is large. Our results have shown that the GPR is more efficient than the traditional MLR and the ANN as it can achieve better prediction accuracy. According to the results, the prediction accuracy from the full model (MLR1, ANN1, and GPR1), where all climate variables are included, is not much different from that of the models (MLR2, ANN2, and GPR2) consisting of only total rainfall variable.

5. Conclusions

This study demonstrated and compared the performance of the GPR, the MLR, and the ANN in the regression problem with the case study in predicting the WQI at Ping River basin, Thailand. The climate factors, temperature, humidity, total rainfall, and evaporation were used as the independent variables. We found that the total rainfall was the only significant variable to be included in the models. We also considered various models for each method with the same set of climate variables. In summary, the GPR models are superior to the MLR and the ANN models according to the prediction accuracy.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Faculty of Science, Chiang Mai University. We would also like to thank the Pollution Control Department, Water Quality Management Bureau, The Ministry of Natural Resources and Environment of Thailand and the Northern Meteorological Center, Meteorological Department, Ministry of Digital Economy and Society of Thailand for their assistance with the data.

Conflict of interest

The authors declare no conflict of interest.

References

1. Holmberg M, Forsius M, Starr M, et. al (2006) An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. *Ecol Model* 195: 51–60.
2. Majeed SAA, Saleh LAM, Aswed GK (2018) Modeling the water quality index and climate variables using an artificial neural network and non-linear regression. *Int J Eng Technol* 7: 1346–1350.
3. Sallam G, Elsayed EA (2018) Estimating relations between temperature, relative humidity as independed variables and selected water quality parameters in Lake Manzala, Egypt. *Ain Shams Eng J* 9: 1–14.
4. Anmala J, Venkateshwarlu T (2019) Statistical assessment and neural network modeling of stream water quality observations of Green River watershed, KY, USA. *Water Supply* 19: 1831–1840.
5. Mehdipour P, Navidi I, Parsaeian M, et. al (2014) Application of Gaussian Process Regression (GPR) in estimating under-five mortality levels and trends in Iran 1990-2013, study protocol. *Arch Iran Med* 17: 189–192.
6. Asante-Okyere S, Shen C, Ziggah YY, et. al (2018) Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability. *Energies* 11: 3261––3274.
7. Chaurasia P, Younis K, Qadri OS, et. al (2019) Comparison of Gaussian process regression, artificial neural network, and response surface methodology modeling approaches for predicting drying time of mosambi (Citrus limetta) peel. *J Food Process Eng* 42: e12966.
8. Fritsch S, Guenther F, Wright MN, et. al (2019) Training of Neural Networks. R package version 1.44.2
9. Shanmuganathan S (2016) Artificial neural network modelling: An introduction. Springer.
10. Rasmussen CE, Williams C (2006) Gaussian processes for machine learning. The MIT Press.
11. Dancik GM (2018) Maximum Likelihood Estimates of Gaussian Processes. R package version 3.1.7