Hydrology
Current Research
ISSN: 2157-7587

Kabo-bah et al., Hydrol Current Res 2012, 3:3
DOI: 10.4172/2157-7587.1000135

Review Article Open Access

Regression Models for Determining the Fate of BOD_5 under Biological Treatment Method in Polluted Rivers

Amos T. Kabo-bah*, Xie Yuebo and Song Yajing

State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

Abstract

The estimation and prediction of BOD_5 is an important criterion for wastewater management and biological treatment of polluted rivers. The biological treatment method has been identified as the most optimal and technologically efficient technology to treat polluted urban rivers and streams. This practice has seen tremendous acceptability and applicability in most parts of China. However, the high cost of measurements, laboratory tests and sampling uncertainties associated with water quality variables make monitoring and prediction of desired water quality variables during biological treatment campaigns difficult. This paper has developed empirical models to predict the fate of BOD_5 during a biological treatment method. The developed ten models were evaluated using ten-stratified cross validation technique. The results indicate high R^2 relationships between observed and computed values. Prediction accuracy of the models were also assessed and showed errors in the range of ± 26% ~ ± 37%. These errors seem acceptable according to previous work on BOD_5 measurements and forecasting. It is presumed that the unexplained nature for empirical formulae to integrate all the natural processes underpinning BOD_5 processes might have been the cause. This notwithstanding, the results show plausible application for prediction and management of biological treatment projects and replicable for wastewater treatment systems.

Keywords: Prediction; Models; Ten-fold cross validation; Water quality

Introduction

Urban river pollution is on increase in developing economies as a result of economic boom, rapid population growth, and urban development and expansion. Such economies are also plagued by large, small and medium scale industrial growth. The discharge of effluents from these industries directly into river bodies further accumulates the water quality stress. Under such an event, support capacity of such riverine systems to accommodate aquatic life is limited. This situation is particularly true for most developing countries across the world. In water quality monitoring and assessment, the measurement of the level of organic matter amounts, activity and content is imperative to assess the degree of deterioration of the riverine system. One important measure is to determine the quantity of oxygen required to stabilise a certain amount of organic matter. Over the past years, several techniques that include manual and automatic approaches have been used to measure the quantity of organic matter in a polluted river system. Regardless of these available techniques, careful in-situ data collection and strict adherence to standard laboratory procedures are required to produce accurate results for assessment and monitoring. Typical parameters measured during water quality assessment of wastewater (i.e. including polluted rivers) are biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total organic carbon (TOC) [1]. Others include total phosphorus (TP), total nitrogen (TN) and ammonia-nitrogen (NH3-N). Ammonia-nitrogen is primarily considered during biological treatment method campaigns.

Biological treatment campaigns have become the most acceptable way to restore polluted rivers in urban settlements naturally. This method, for instance, in China has received laudable recognition among city authorities and research institutions. This is partly because this technology is efficient, cost-effective and sustainable. This technology also suggests benefits for current discussions and efforts for Agenda 21 and COPs discussions on Climate Change [2,3]. The emerging Biological Treatment Method (BTM) is not new but is gaining grounds recently for restoration of heavily polluted river systems. For instance, the BTM has been successfully used in many countries worldwide including China for domestic and industrial wastewater treatment. The technology has been used in urban rivers in Shenzen, Rui'an and Wuxi of China. This method has been known to rapidly reduce the concentrations of effluent BOD and COD [4]. The spatio-temporal measurement of water quality variables involved in BTM are time consuming and require an efficient workforce to be able to carry out the treatment process.

The measurements of water quality variables during BTM by regular sampling method also require enormous work. Moreover, BTM is not a regular operation but necessary for populated riverine systems. A further challenge is that, a review of existing research in water quality monitoring, prediction and management indicates that there are hardly any universal correlation algorithms that exist for general wastewater and polluted river systems [1]. This is partly because the fluxes of water quality variables (e.g. BOD, COD, TN, TP and NH3-N) are dependent on diverse processes in the river watershed or environment. This non-linear dynamics among the water quality variables make it difficult to universally generalize.

For specific research regions across the world, various methods have been used to develop empirical methods for monitoring general water quality variables in rivers and lakes based on a given set of river discharge and land use practices in the watershed [5-7]. The measurement of five-day BOD is a tedious process and encompasses

***Corresponding author:** Amos T. Kabo-bah, State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China, E-mail: kabo-bah@greenwaterhut.org

Received March 08, 2012; **Accepted** July 07, 2012; **Published** July 11, 2012

Citation: Kabo-bah AT, Yuebo X, Yajing S (2012) Regression Models for Determining the Fate of BOD_5 under Biological Treatment Method in Polluted Rivers. Hydrol Current Res 3:135. doi:10.4172/2157-7587.1000135

the possibility of introduction of errors if care and precaution measures are not followed. This is because, accurate measurement of BOD_5 requires a precise measurement of dissolved oxygen (DO) [8]. For the purposes of BTM, the overall monitoring of the BOD_5 is very relevant for deciding other purification campaigns in the river system. For an example, the effluent conditions of the BOD_5 concentrations of a treated river give information about the next strategy for follow-up treatment campaigns. In view of this, the ability to predict the flux of BOD_5 becomes important and relevant for monitoring and management of biological treatments. This research used statistical regression methods to derive models based on a set of water quality variables (i.e. pH, water temperature, COD, TN, TP, NH3-N) to predict the amounts of BOD_5. This research seeks to support undergoing efforts in water quality monitoring, environmental modeling and water resources planning for urban cities especially river quality restoration programmes.

Methods

Study area

The Xuxi River is derived from the mouth of Jing-Hang and flows into the ancient canal towards the end of its lower reach. It is located in the Chang Nan District of Wu Xi city of China. The total length of the river is 1.36 km with an upstream surface width of 4.5 m and a depth of about 1.4 m. The river is characterized by muddy sediments that can be up to 1.6 m. This river is located in a north sub-tropical humid zone. This zone is affected by the monsoon circulation phenomenon and thus, has four distinct seasons. The recorded annual precipitation is generally higher than the annual evaporation. The hydrodynamics of the river are generally poor. For instance, natural river siltation, indiscriminate rubbish dumping and unstable slopes of the river make its hydrodynamic conditions poor and inhabitable for aquatic life. It was estimated in an independent research that about 10,000 m^3 of sewage was discharged daily into the river [4]. This was partly due to the non-existence of a common wastewater treatment facility in the city.

Data

The models were developed using datasets conducted in October 2009. Due to the limited nature of BTM projects to undertake BOD_5 measurements, it was difficult to obtain an independent dataset for use during this research. This notwithstanding, the data collection points were significantly varied spatially. This data were collected in four days (i.e. 17th–22nd October). The consistency in the time was to allow for comparability of the results. The variables measured included DO, TP, pH, NH3-N, COD and BOD_5. These were sampled and analysed by the Environmental Protection Agency (EPA) of the Wuxi City Council. A total of 8 different locations were sampled along the river at 3.00 pm each day. The measurements were reasonably spaced out evenly to cover the whole length of the river under study.

The virtual beach (VB)

VB is a software package used to construct specific site Multi -Linear Regression (MLR) models for the prediction of pathogens indicator levels at recreational beaches. The MLR approach has been proven to work better for beaches where conditions of hydrology, weather, human and animal activities are high and changes significantly daily [9]. The VB version 2.0 is a spread-like grid surface that facilitates the imports and processing of data using the MLR model (see equation 1 below).

$$BOD_5 = \alpha_0 + \sum_{k=n}^{j=n} \alpha_i X_j + \varepsilon \quad (1)$$

Where BOD_5 is the predicted total

BOD_5

α_0 is the intercept

α_j is the slope for the *j*th explanatory variable

ε is the remaining unexplained noise in the data – error

The VB MLR model is dependent on least squares method to fit derived models. It considers many variable interactions, multi-collinearity and model selection [9]. The VB model uses the backward elimination method to help the user select the best appropriate model with the specified explanatory variables. The VB model facilitates model development and offers better chances for developing good models with limited datasets. The VB has been successfully used to develop models for the fate of biological contaminants in beaches [9-11]. The VB has a function that performs data transformations. By default, MLR models are linear and this has the tendency to limit value of explanatory variables. VB offers a number of transformation methods such as square root and square, inverse and polynomial functions. Another important aspect of the VB is that, it uses a parsimonious model. This allows the user to identify the best explanatory variables from the mix of variables for fitting. This is because each variable tends to increase the estimated variance. Practically, it is preferable to use a smaller set of variables from which a parsimonious model is finally selected [9]. In this work, the number of models recommended for the parsimonious model was between 2 and 5. For the purpose of uniformity, five variables were retained.

Model evaluation

VB provides several options for assessing the quality of the fits of the derived models. In general, goodness of fit and predicative capacity is important to describe a model's ability to predict. In this particular paper, emphasis was laid on the use of adjusted R^2 (R_a^2), Prediction Sum of Squares (PRESS), Corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC) and Root Mean Square Error (RMSE). The (R_a^2), AICc and BIC were applied to choose the most suitable model for description of the fate of BOD_5. The PRESS and RMSE were used to determine the predictive power and accuracy, respectively, of each model. The R^2 tend to increase as we add more variables and hence picking the biggest R^2 will not necessarily select the best model. This mostly results in over-fitting. Therefore, R^2 is simply perfect for situations in which models have the same number of variables. For the purposes of this research, the adjusted R^2 was used. The adjusted R^2 has been extensively used to illustrate model performance in several literatures [9,12,13]. The Akaike Information Criterion (AIC) [14] evaluates the goodness of fit of a selected model. For a given selected models for a particular data, the preferred model is the one with the minimum AIC value. In addition, AIC shows significantly the degree of goodness of fit of the model and also ensures the degree of penalisin g when increasing the function of the number of estimated variables. AICc is AIC with correction for finite sample sizes. In fact, AICc is AIC with more penalties for extra variables in the model [15] strongly recommended the use of AICc. The AICc was a refinement by [16] to cater for the bias in regression for smaller sample sizes. In this research, the AICc was therefore adopted. The BIC was included in the research to complement the works of the AICc estimates. The BIC penalises complex models most and gives preference to simpler models in selection.

Cross validation

In order to assess the quality and further assess the accuracy for derived models, cross-validation statistical method was applied. Cross validation has been widely used for model selection and stratified cross-validation is considered best for elimination of bias and variance [17]. According to Kohavi [16], the use of ten-fold stratification is best to indicate the best results. In this particular research, a stratum was considered to be the measurements taken on a particular day. In this case, 6 different days means 6 different strata for cross validation. Now, also, all the measurements were randomised in excel using RAND function (it returns a random number that is equal to or greater than 0 and less than 1). Further, the data was divided equally into four parts. In this case, a ten-fold strata was obtained (i.e. sub test A, sub test B, sub test C, sub test D, sub test E, sub test F, sub test G, sub test H, sub test I and sub test J). Sub tests A-F were measurements taken respectively on 17^{th}, 18^{th}, 19^{th}, 20^{th}, 21^{st} and 22^{nd}.

Sub test G-J were randomised data grouped (i.e. 1^{st} quarter–sub test G, 2^{nd} quarter–sub test H, 3^{rd} quarter– sub test I and last quarter–sub test J). A rapid assessment was made by comparing the observed values and predicted values by each model under this stratified cross validation. By performing this rigorous assessment, it helps to decide the level of applicability of the derived model for predicting BOD.

Results and Discussion

Model development

The models for the BOD_5 were derived based on a set of 4 parameters that include pH, ammonia-nitrogen (NH3-N), Dissolved Oxygen (DO) and Chemical Oxygen Demand (COD). A total of ten models were derived (see table 1 for more detail description of each model). These models were evaluated with standard statistics criteria including R_a^2, AICc and BIC. However, it is advised to refer to [18] for further details. All the derived models show very high R_a^2 of above 90%.The AICc examines the goodness of fit of the models by penalising out the addition of independent variables. On the other hand, the BIC penalises the effect of sophisticated models and hence give priority to simple models. In that case, the best model is the one, which has the lowest BIC and lowest AICc. According to AICc and BIC ranking, the models BODe2 is the best choice followed by BODe1, BODe4, BODe5, BODe6, BODe9, BODe3, BODe7, BODe8 and BODe10 respectively. However, the small difference between the R_a^2, for all models, shows that all models look plausible. Notwithstanding, it is worth mentioning that with the AICc and BIC, it is easier to evaluate in detail the performance of each model. In this case, the selection of the best model is possible as suggested in previous researches [19]. All the models were further verified with cross validation techniques. This was because no other auxiliary datasets were available.

Evaluation

It was necessary to evaluate the models against their predictive capacity and preciseness for practical use under the BTM projects or programmes. The analysis was performed and table 2 shows the model statistics for each of the model. Model BODe3 looks preferable in operations in which overall cumulative bias should be kept to the minimum. On the other hand, if the relative prediction error is most desired which is the key to water quality monitoring and forecasting, BODe2 is most preferred. However, the overall prediction error for all models is between 25~37%. This error relatively falls within the range obtained in other researches. It has been established that sampling and measurements in water quality have typical errors in the range of 15-20% for most water quality variables, and sometimes higher (30-40%) for BOD [20,21]. Therefore, all the derived models are all plausible for generic monitoring of BTM processes.. The other reason considered for choosing the models was that, each model equation required a set of either of these variables (TP, DO, NH3-N, pH, COD). This meant that if only two variables were available for estimating the total BOD_5, it was still practically feasible to do this. For instance, one can use only COD and NH3-N to estimate total BOD_5 using model equation 2. Also, visual inspection of the correlation between the observed measurements and the computed shows relatively good linear relationship between the two (Figures 2-5). This further indicates the goodness of prediction of each of the models.

In addition, the datasets were categorised according to 10-fold stratified cross validation stage. Cross validation helps to penalise out the effects of bias and variance of derived models on using the same datasets. The results of the cross validation are shown in Figure 1. Generally, the predicted R^2 correlation between the observed and predicted for all values is between 82% and 99%. Test 1-6 which represents the different station measurements show considerably high R^2 values except Test 5 and Test 6. Test 5 and Test 6 show relatively low R^2 values possibly due to fact that original measurements from these test stations may have had some higher sampling and laboratory errors compared to the others. In event of this, this might have also affected Test I which shows relatively low R^2. Notwithstanding, all cases

Model No.	Description of Model	Evaluation Statistics		
		R_a^2	AICc	BIC
BOD_eqn1	$BOD_5 = 93.1452e-01+12.6097e-03*[COD][NH_3-N]+65.7477e-01*TP$	0.922	135.80	108.06
BOD_eqn2	$BOD_5 = 12.5032e00+14.5457e-03*[COD][NH_3-N]$	0.921	134.71	106.38
BOD_eqn3	$BOD_5 = -93.9783e-01+12.5041e-03*[COD][NH_3-N]+61.9133e-01*TP+26.3608e-01*PH$	0.920	138.51	111.11
BOD_eqn4	$BOD_5 = 18.4417e00+14.1608e-03*[COD][NH_3-N]-65.6279e-03*[DO][COD]$	0.920	136.69	108.95
BOD_eqn5	$BOD_5 = 14.1712e00+14.6332e-03*[COD][NH_3-N]-12.361e-02*[DO][NH_3-N]$	0.920	136.71	108.96
BOD_eqn6	$BOD_5 = -12.6157e00+14.2552e-03*[COD][NH_3-N]+35.0377e-01*PH$	0.919	136.98	109.23
BOD_eqn7	$BOD_5 = -16.8157e00+14.2863e-03*[COD][NH_3-N]-14.4308e-02*[DO][NH_3-N]+43.6124e-01*PH$	0.918	138.96	111.56
BOD_eqn8	$BOD_5 = -72.9307e-01-66.8275e-03*[DO][COD]+13.8549e-03*[COD][NH_3-N]+36.0482e-01*PH$	0.918	139.16	111.76
BOD_eqn9	$BOD_5 = 11.4174e00+14.6564e-03*[COD][NH_3-N]+85.1954e-02*[DO][TP]$	0.918	137.34	109.59
BOD_eqn10	$BOD_5 = 18.2255e00-92.1819e-03*[DO][NH_3-N]-49.4919e-03*[DO][COD]+14.3207e-03*[COD][NH_3-N]$	0.917	139.30	111.90

All measurements are in mg/l except pH as integer values.

Table 1: Derived models and evaluation statistics of each model.

Model No.	*PRESS*	*RMSE*	*Prediction error (%)*
BODe1	946.35	6.420	34
BODe2	1151.10	7.053	18
BODe3	905.34	6.254	32
BODe4	1191.80	7.175	27
BODe5	1087.00	6.812	27
BODe6	1081.70	6.803	29
BODe7	1001.50	6.499	25
BODe8	1160.40	7.060	27
BODe9	1189.70	7.190	37
BODe10	1178.10	7.145	26

Table 2: Evaluation statistics for regression models.

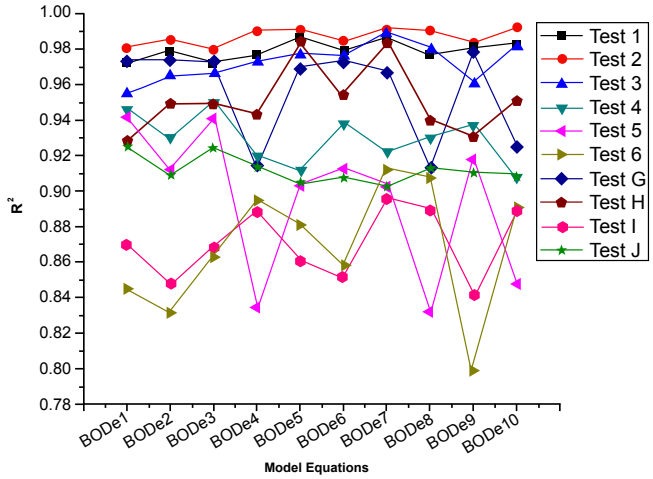

Figure 1: R^2 correlation coefficients for observed and computed values for each mathematically derived model.

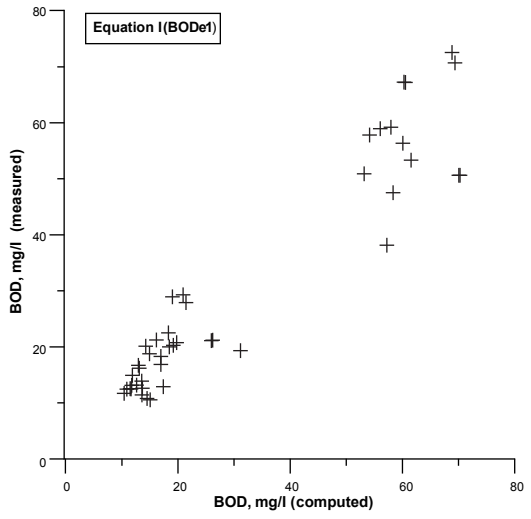

Figure 2: Relationship between BOD_5 estimated and computed using model equation 1.

indicate that all models are relevant for practical use. A comparison of the measured and computed BOD_5 using model equations 1, 3, 5 and 8 are presented in Figures 2 to 5. In these figures of 2 to 5, it can be seen that, the results are fragmented into two parts–(1), values between 10-30 mg/l and (2), values between 52–70 mg/l. The clustering of the

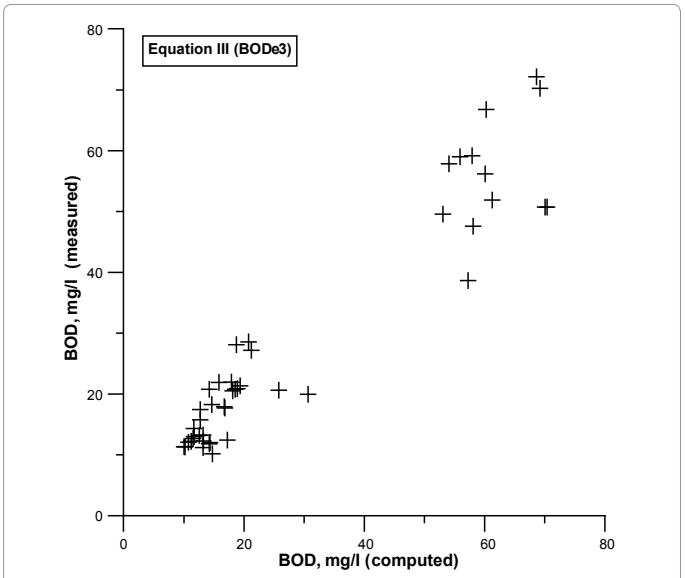

Figure 3: Relationship between BOD_5 estimated and computed using model equation 3.

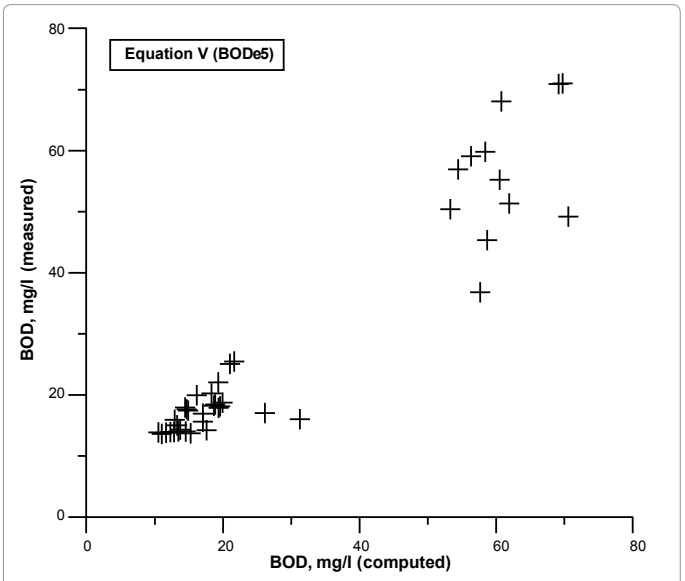

Figure 4: Relationship between BOD_5 estimated and computed using model equation 5.

measurements indicates that, model predictions might be more relevant within these limits. Since there was no other datasets to test for model performance for BOD_5 values lower 10 mg/l and higher than 70mg/l, it is strongly recommended for the purpose of using the models in this paper to consider the modelling limits. Also, one can observe from the graphs that, clusters of 10 mg/l-20 mg/l appear linear compared to the clusters of 52 –70 mg/l and hence, attention should be paid on this. These BOD_5 values are all far above the Class V ($BOD_5 \leq 10$ mg/l) of the GB 3838-2002 on surface water quality standard. In general, BOD_5 values of less than 1mg/l are considered for pristine rivers while moderately river systems may have BOD_5 values varying between 2-8 mg/l. However, it is worth to mention that, the exact definition of BOD classes for poor to pristine river systems depends on the country law on surface water quality. This implies that these values indicate the highly polluted river system. Therefore as described in the study area,

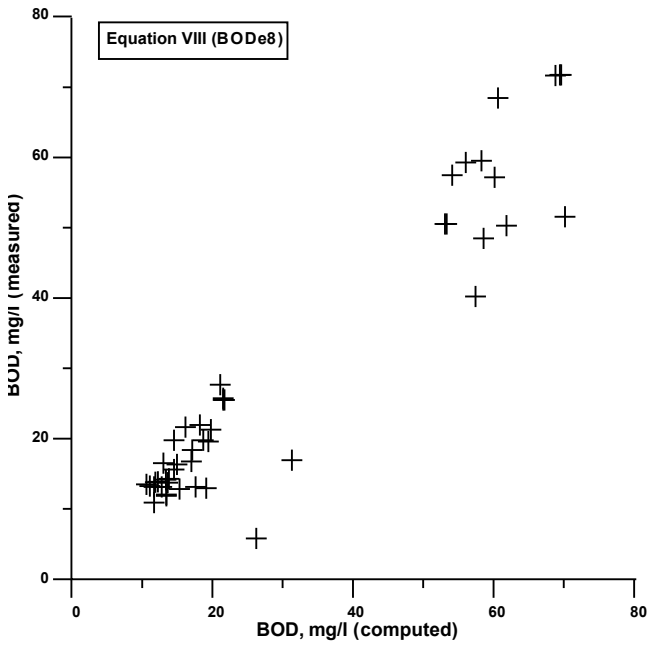

Figure 5: Relationship between BOD_5 estimated and computed using model equation 8.

this Xuxi River which is the case study is a highly polluted river system due to the direct sewage discharge into this river. Also, the sampling of the data within 4 days was not sufficient to inform the biological activity during this period. Generally, the biological activity is allowed to operate for a period of 3 months, then a further sampling and testing is conducted. In this particular instance, the measurements obtained were all within the early stages of the biological activity indicating at some places a low of 10 mg/l and a high of 70 mg/l. The Xuxi River typically represents a highly polluted river system and could be considered as the many cases in China and around the world in places where rivers have been heavily polluted by industrial activity. In such a particular instance, consideration of the use of the models developed here would be useful. It is noteworthy that scientific uncertainties in modelling models are impossible to avoid but provides a means for decision makers for selecting alternative measures and if possible consider other experimentation and observation [22] in water quality monitoring and management. Therefore, though these models derived are purely empirical, they still have relevance in providing information about the fluxes of BOD_5 especially during a biological treatment processes for polluted streams and rivers. However, the waiting period of five-day BOD measurements subject to large errors could be avoided with empirical approximation to support prediction and management of biological treatment campaigns.

Conclusions

Rapid urbanization in cities across developing economies continues to be a challenge for water managers largely, because rivers are getting heavily polluted from household and industrial effluents. The biological treatment method has been identified as the most feasible way to treat and restore polluted rivers back to habitable forms for aquatic life and safe human health. However, the ability to monitor the processes under biological treatment method in terms of the changes in the water quality variables still remains unexplored. Traditional standard models are not able to explain the linkages and further, with this treatment method, specific water quality parameters are desired. As a result, empirical model were derived using regression analysis for BOD_5 estimation and prediction based on a set of given water quality variables (pH, ammonia-nitrogen, dissolved oxygen and chemical oxygen demand). The derived formulae are to support future planning, monitoring and management. The models showed high adjusted R^2 statistics (i.e. above 90%). The prediction errors for validation set of the data indicated ± 26% ~ ± 37%. This prediction error falls within related research conducted in water quality modelling. Also, ten-fold stratified cross validation method was applied and the R^2 coefficient of the observed and computed results indicated values of above 80%. Unaccounted for errors in this research are considered to be purely due to the large sampling and measurement errors and unexplained processes of the BOD_5 processes. Notwithstanding, the models are to serve as a guide for further planning, forecasting and management of biological treatment projects.

References

1. Rene ER, M Saidutta (2008) Prediction of BOD and COD of a refinery wastewater using multilayer artificial neural networks. Journal of Urban and Environmental Engineering 2:1-7.
2. Protocol K (1997) United Nations framework convention on climate change. Kyoto Protocol.
3. Yamin F, Depledge J (2004) The international climate change regime: a guide to rules, institutions and procedures. Cambridge Univ Press.
4. Yudianto D, Xie Y (2011) Numerical Modeling and Practical Experience of Xuxi River's Natural Restoration Using Biological Treatment. Water Environment Research 83: 2087-2098.
5. Christensen VG, Rasmussen PP, Ziegler AC (2002) Real-Time Water-Quality Monitoring and Regression Analysis to Estimate Nutrient and Bacteria Concentrations in Kansas Streams. Water Science & Technology 45: 205-219.
6. Nevers MB, Whitman RL, Frick WE, Ge Z (2007) Interaction and Influence of Two Creeks on Escherichia coli Concentrations of Nearby Beaches: Exploration of Predictability and Mechanisms. Journal of Environmental Quality 36: 1338-1345.
7. Delzer GC, Mckenzie SW (2003) Five-day Biochemical Oxygen Demand: U.S. Geological Survey Techniques of Water Resources Investigations 7: 1-21.
8. Frick WE, Ge Z, Zepp RG (2008) Nowcasting and forecasting concentrations of biological contaminants at beaches: a feasibility and case study. Environ sci technol 42: 4818-4824.
9. Liu L, Phanikumar MS, Molloy SL, Whitman RL, Shively DA et al. (2006) Modeling the Transport and Inactivation of E. coli and Enterococci in the Near-Shore Region of Lake Michigan. Environmental Science & Technology 40: 5022-5028.
10. Ge Z, Frick WE (2007) Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations. Environ res 103: 358-364.
11. Helsel DR, Hirsch RM (2002) Statistical Methods in Water Resources, U.S. Geological Survey, Techniques of Water- Resources Investigations.
12. Francy DS, Darner RA (2006) Procedures for Developing Models to Predict Exceedances of Recreational Water-Quality Standards at Coastal Beaches. U.S. Department of the Interior, U.S. Geological Survey.
13. Akaike H (1973) Information theory and an extension of the maximum likelihood principle Springer Verlag.
14. Burnham KP, Anderson DR (2002) Model selection and Multimodel Inference: A practical Information-theoretic approach. Springer Verlag.
15. Hurvich CM, Tsai CL (1988) Regression and time series model selection in small samples. Biometrika 76: 297-307.
16. Kohavi R (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

17. Wolfe K, Parmar R (2010) Virtual Beach 2.0 User Guide. United States Environmental Protection Agency.

18. Hastie T, Tibshirani R (2009) Model assessment and selection: The elements of statistical learning. 219-259.

19. Ahyerre MG, Chebbo G, Tassin B, Gaume E (1998) Storm water quality modelling, an ambitious objective? Water science and technology 37: 205-213.

20. Radwan M, Willems P, Berlamont J (2004) Sensitivity and uncertainty analysis for river quality modelling. Journal of Hydroinformatics 6: 83-99.

21. Reckhow KH (1994) Importance of Scientific Uncertainty in Decision Making. Environmental Management 18: 161-166.

22. Yudianto D, Xie Y (2011) Numerical Modeling and Practical Experience of Xuxi River's Natural Restoration Using Biological Treatment. Water Environment Research 83: 2087-2098.