Open Access Scientific Reports

Empirical Algorithms for Estimating Total Phosphorus 1 During Restoration of Urbanized Polluted Rivers Using 2 Bacterial Technology

Amos T Kabo-bah[1*], Xie Yuebo[1] and Song Yajing[1]

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

Abstract

The expensive nature of water quality field campaigns calls for mathematical models to support forecasting of indicators. This is particularly important for bacterial technology for urban rivers where data is usually rare but rapid reporting of preliminary results during such project implementation is essential. This research developed mathematical models to estimate total phosphorus given a set of water quality indicators (NH3-N, pH & COD) are available. The models were tested with adjusted R2, corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), PRESS and RMSE. All models indicated a prediction error (RMSE < 20%) except TPM 3 (RMSE = 20.2%). This prediction error confirms with measurements error of 15-20% for most water quality variables. The TPM 1 mathematical model is recommended as the best option for forecasting purposes in water quality modelling. The research can contribute significantly towards numerical water quality modelling in China and maybe in other countries having similar water quality problems as discussed in this paper.

Keywords: Total phosphorus; Empirical models; Water quality

Introduction

The water quality management is a major issue in China as the economy grows fast to take up the lead in the world. Urbanized streams and rivers are exposed to high increases in nutrient loads. This influences their aesthetics. High-polluted rivers possibly reduce their capacity to withstand incoming storms and floods or excessive precipitation. The study of nutrient dynamics has been conducted in several parts of the world [1]. The restoration of urban rivers can possibly contribute towards flood prevention and effective control. Large cities are growing bigger in terms of population and industrial activity yearly. Due to the availability of jobs, advanced technology and other social opportunities, most people prefer to live in the cities. As noted in a recent research, urban growth is changing and populations have increased from 15 to 50% globally [2]. In urban ecological sustainability, nutrient recycling has been recommended as the best alternative for wastewater management. Urban rivers are subjected to sewage disposal and other waste as a result of poor practices from the community. Urban water and wastewater planners in China are struggling to meet the growing water and sanitation demands while attaining sustainable urban water system [3]. River restoration campaigns have been identified to support Agenda 21 and recent COPs 36 discussion on Climate Change [4,5].

Several techniques have been employed over the years towards stream and river restoration. In general, the techniques involved re-aeration using weirs shift in effluent discharge location, use of oxygenator to pump air into water body and use of engineered constructed wetland. Regarding, the various research into the cost and feasibilities of all these methods, wetlands were found to be the most efficient in treating streams or rivers. Wetlands have been found to remove significant amounts of nutrients (i.e. over 70%), and are far better for long-term maintenance and management. The bacterial technology has been used for many years in many countries for industrial and domestic wastewater treatment. However, the use of such a method for the restoration of urbanised rivers in China is relatively new. For instance, the method has been successfully used in the treatment of lakes, wastewater treatment plants and urban streams in Shenzhen, Rui'an and Wuxi of China. The method has received

great concern because of the innovative and fast way it lowers the concentrations of the BOD and COD concentrations [6]. These bacterial technologies have been seen as necessary protocols towards sustainable management of urban water systems. However, there is a lack of models for forecasting and monitoring of the nutrient concentrations during such treatment campaigns. Predictive modelling of nutrient storage in rivers is an important aid to monitoring exercises of urban rivers. For instance, mathematical expressions derived by Vollenweider were used to estimate the fate of phosphorus in water bodies [7]. The challenge was that measurements such as lake's areal surface, areal water load, lake water's residence time and inflow phosphorus concentration were used to estimate total phosphorus. In a typical river restoration program with bacterial technology, these measurements are irrelevant. In another study, statistical models were successfully used to investigate the Escherichia coli concentrations in beaches in Lake Michigan [8]. Also, the fate of faecal indicators were modelled in the UK using digital land use maps [9]. Further, another study on lakes in the U.S used regression models to monitor nutrient and bacteria concentrations in real time [10]. These models were typically applied to lake conditions, and parameters used were far different from those usually concerned with in bacterial technology programs for rivers. Therefore, mathematical models to predict total phosphorus dynamics to support monitoring, planning and management of urban rivers during bacterial technology programs remain a major setback. The estimation of total phosphorus presence in urban rivers is an important indicator for understanding eutrophication problems [11].

*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, kabobah@hhu.edu.cn 7

So, this research developed mathematical algorithms to model the fate of the total phosphorus in urbanised biologically treated rivers. The developed algorithms were tested with independent data sets from other restoration experiments. The research envisages enormous contribution towards recent works in numerical water quality modelling, management of wastewater treatment plants and possible support towards approximation of nutrients in modelling packages such ASM1, ASM2, ASM2D or ASM3 [12].

Methods

Study area

The research data was obtained from a biological treatment project in the Rui'an city in the Zhejiang Province of China. This project was undertaken in the months of November and December 2010. The results of this data included daily measurements of *Total Phosphorus (TP), Ammonia (NH3-N), Chemical Oxygen Demand (COD) and pH*. The project was implemented in two riverine systems: the Fenghu-Songyong River. Six sampling points were observed and all measurements used for this research (Figure 1). The source of Fenghu River is Liangmian river and the outlet is Wenruitang river. The Songyong River flows directly into the Fenghu River. The study area is characterized by relatively warm weather conditions, short duration of winter and longer summer. The precipitation period for this area is divided into three parts: May to April, May to June and August to September. This is partly because the study is located in the typhoon zone of China. The Songyong River is about 280m long. The breadth of the Songyong River averages 5-18m and the water depth is about 1-3m. The Fenghu River is about 740m to the outlet. This river has a breadth averaging between 6-15m and the water depth of about 1m.

The independent datasets were obtained from the Xuxi River. The Xuxi River is located in the Chang Nan District of the Wuxi City, Jiangsu province. The total length of Xuxi River is 1.36 km. It has an average surface width of 4.50 m and water depth of 1.40 m. The poor environmental sanitation around this river contributes towards the high sewage loads in the river. It is estimated that about 10,000m³ of sewage is discharged daily into the river partly because of the non-existence of the sewage treatment facility [6]. The river is generally characterised by dark brown sediment and floating algae causing the river to appear dark green.

Bacterial technology

The purpose of this paper is not to reinvent the wheel by narrating all the processes involved in this particular method. The emphasis is more on the ability to used monitored data for the derivation of the algorithms. In this particular method, the selected bacteria and microbial accelerator were directly injected into the river to activate the native bacteria. In all, thirty-four buckets each containing 150kg of bacteria were used for restoring the quality of the Fenghu-Songyong River. For convenience of the method, bacteria were cultured nearby the river and injected into the six sections (Figure 1).

Data

The derivation and testing of the models were done using two independent datasets – the bacterial technology experiments in Fenghu and Songyong Rivers (FSR) in the Rui'an city, Zhejiang province and Xuxi River in Jiangsu province. The data from FSRs was obtained in December 2010 (Table 3A in appendix). This FSRs data was used to derive the models. The validation of the models was done using the independent dataset from Xuxi River sampled in October 2009 (Table 3B in appendix). The datasets recorded daily measurements of Total Phosphorus (TP), Ammonia (NH3-N), Chemical Oxygen Demand (COD) and pH. Due to the scarce nature of data from such projects, it was assumed that derived model's ability to predict effectively the variables from an independent dataset implied that, the derived model was functional for practical purposes. This is because the climatic and weather conditions of the two datasets were different and if the derived model predictive power was good, this implied its appropriateness for future practical use. The Ministry of Environmental Protection 2009 report on surface water quality reported that about 42.7% of the rivers in China varies from Grade IV to V, which represents a fairly poor water quality status. This implies that about 42.7% of most rivers in China are not suitable for aquatic life and domestic uses. Therefore, we selected the Fenghu- Songyong Rivers as a representative case for majority of the rivers which are in the inferior category of Class V in China (Table 3A in appendix).

The Virtual Beach (VB) MLR model

The VB Multiple Linear Regression (MLR) Tool by USGS was used to support the work of this study. The study employed the statistical capability of the package to handle multiple linear regressions. This was used to develop the model for predicting the fate of TP against a given set of water quality variables. The MLR model [13] is given by

$$P = \alpha_0 + \sum_{\substack{i=n \\ j=n}}^{n} \alpha_i X_j + \varepsilon$$

(1) Where P is the predicted total phosphorus α0 is the intercept

αj is the slope for the ith explanatory variable

ε is the remaining unexplained noise in the data – error

The MLR analysis is dependent on least squares method to fit models. It is therefore subject to many considerations i.e. variable interactions, multicollinearity and model selection [14]. The VB uses 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 [14-16]. The VB has a function to perform data transformations. By default,
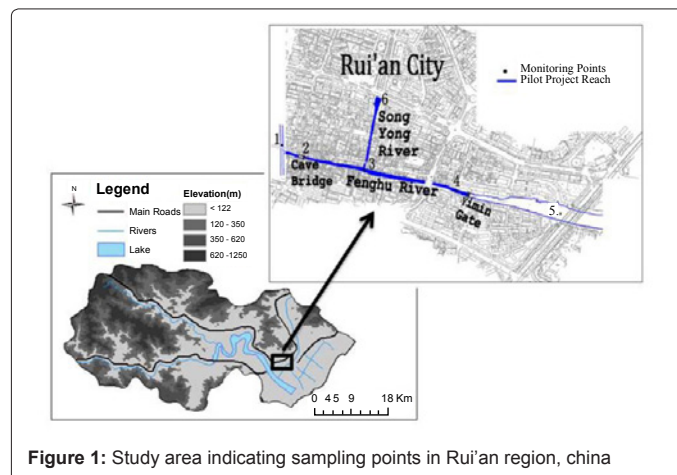

Figure 1: Study area indicating sampling points in Rui'an region, china

MLR equations 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.

Model evaluation

VB offers several options for checking the fits of the selected models and the forecasts. In general, goodness of fit and predicative capacity are important to describe a model's ability to predict. In this particular paper, emphasis was laid on the use of adjusted R2 (R_a^2), Prediction Sum of Squares (PRESS), Corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC) and Root Mean Square Error (RMSE). The ( ), AICc and BIC were applied to choose the most suitable model for description of the fate of total phosphorus. The PRESS and RMSE were used to determine the predictive power and accuracy, respectively, of each model.

Generally, R2 tend to increase as we add more variables and hence picking the biggest R2 will not necessary select the best model. This mostly results in over-fitting. Therefore R2 is simply perfect for situations in which models have the same number of variables. For the purposes of this research, the adjusted R2 was used. This was to penalize overly complicated models as in the equation below:

$$\overline{R}_a^2 = 1 \left(\frac{n-1}{n-p} \right) \left(1 - R_p^2 \right) \text{ (4)}$$ Where p is the number of models.

The adjusted R2 has been exhaustively used to illustrate model performance in several literatures [17,14,13]. The Akaike Information Criterion (AIC) [18] evaluate the goodness of fit of a selected model. In the most general form, the AIC is given by

AIC= 2i-2In(j) (5) Where i is the number of the variables in the model, and

j is the maximized value of the likelihood function for the estimated model.

For a given selected models for a particular data, the preferred model is the one with the minimum AIC value. AIC shows significantly the degree of goodness of fit of the model and also ensures the degree of penalising 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. [19] in their research found that it was strongly recommended using AICc (Equation 6). The AICc was a refinement by [20] to cater for the bias in regression for smaller sample sizes. In this research, the AICc was therefore adopted. It was also realised that the use of AIC presented better results compared to BIC. Despite this, the BIC were 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. The generic form of BIC is shown in equation 7. BIC is commonly known as the Schwarz criterion [21].

$$AIC_c = AIC + \frac{2i(1+1)}{n-i-1} \text{ (6)}$$ Where i is the number of parameters

$$BIC = -2 * \log L + (\log N) * d \text{ (7)}$$ Where N is the number of data points in the observed data

K is the number of parameters to be estimated

L is the maximised value of the likelihood function for the estimated model

The PRESS is the sum of the squared external residuals [22]. See PRESS in equation (2, 3).

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \text{ (2)}$$

Where,

$$e_{(i)} = Y_i - \overline{Y}_{(i)} \text{ (3)}$$

The external residual for the ith observation is equal to the calculated external predicted value Ȳ(i) without the use of the ith observation. The external predicted values and external residuals are independent of Ȳ(i)and Ȳ(i) is not used in fitting the regression model [22]. The best regression model will have the smallest predictive sum of squares.

The VB tool uses Genetic Algorithm (GA) to effectively and efficiently search for the best possible MLR model [23]. GA uses a class of stochastic search procedures called evolutionary algorithms. These algorithms use computational models of natural processes to develop computer-based problem solving systems [24]. The process mimics the natural biological phenomena where organisms produce successive generations. The application of GAs in hydrological and water quality modelling is receiving attention in recent times [25-28].

Results and Discussion

Model development

The models were derived and their evaluation statistics computed using the VB tool. The TP was modelled as the dependent variable and all other variables (NH3-N, COD and pH) were modelled as independent. These produced various possible models that could be used to practically estimate the values of TP. For a complete understanding of the permutations used to derive this, [23]. Seven (7) models (TPM 1, TPM 2, TPM 3, TPM 4, TPM 5, TPM 6 and TPM 7) with an adjusted R2 greater 75% were considered to be appropriate for practical purposes. The models are described in Table 1 below with the relevant computed statistics for , AICc and BIC for each of the model. Ranking the models in terms of goodness of fit (R_a^2), the TPM 1 ranked first followed by TPM 3, TPM 7, TPM 2, TPM 5, TPM 6 and TPM 4 respectively. The AICc further checks the goodness of fit of the model by penalising the addition of independent variables. Using this criterion, TPM 1 model has the lowest AICc value and indicates the most appropriate. Again, other models such as TPM 3, TPM 7, TPM 2, TPM 5, TPM 6 and TPM 4 ranked after TPM 1 using this criterion. The BIC statistics give the same rankings for the models. In this particular instance, all the statistical tests provide the same rankings for all the models. It is fairly correct to say that each of these selection criteria give a reasonable way for model selection. This confirms with previous research that AICc,

(R_a^2) and BIC are appropriate techniques for the selection of regression models [29]. Therefore, it can be observed that based on the available data, each of these models is equally appropriate for practical use. However, if all measurements of pH, NH3-N and COD are available, TPM 1 is the most appropriate (for use).

These derived models provide a basis for estimating total phosphorus with few measurements. For instance, TPM 7 and TPM 5 can be easily estimated with measurements of pH and NH3-N. This is practically economical for water resources management and wastewater treatment monitoring. Also, the presence of pH, NH3-N and COD show a good estimation of TP as shown in TPM 1. For general

restoration project involve in the determining TP, measurements of these three parameters may be adequate.

Evaluation

The models derived (Table 1) were tested with measurements from another bacterial technological experiments in Xuxi River, Zhejiang Province of China. The correlation of the plots of the observed data (TP) and the predicted values from (TPM 1, TPM 2, TPM 3, TPM 4, TPM 5, TPM 6 & TPM 7) are shown in figure 1 & 2. These illustrations indicate the linearity behaviour of the observed and predicted measurements. All plots show a relatively high correlation between the observed and predicted values.

The derivation of the regression models in the field of bacterial technology of streams is relatively new. As a result it is very difficult

to conduct comparative analysis of findings here with related research. Notwithstanding, a recent study on real-time water-quality monitoring using regression analysis to estimate nutrient and bacteria concentrations in Kansas streams, [10] found that the derived mathematical models produced R2 values of (0.509~0.964) for total phosphorus prediction. This estimation was based on independent variables i.e. turbidity, water temperature and specific conductance. In this research, the independent variables used are different. Also, the applied , which is a far better criterion than R2 [13] was used. The R2 values show a range of (0.793~0.858). Comparatively, the results presented in this research show better prediction capabilities.

In order to assess the predictive power of each model, additional statistics were introduced the PRESS and RMSE. It is imperative to test the predictive power of the model to test its reliability and practical
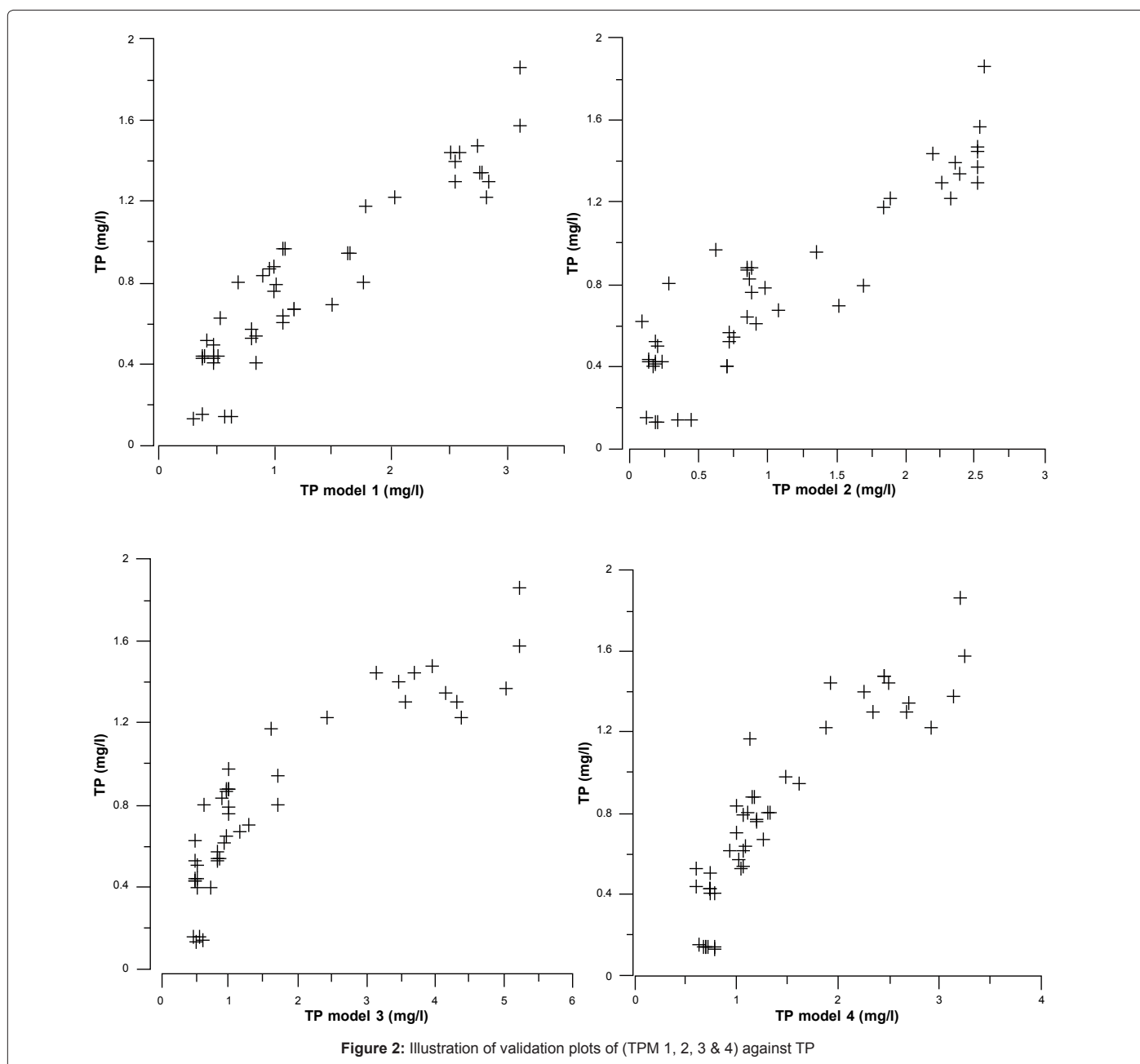

Figure 2: Illustration of validation plots of (TPM 1, 2, 3 & 4) against TP
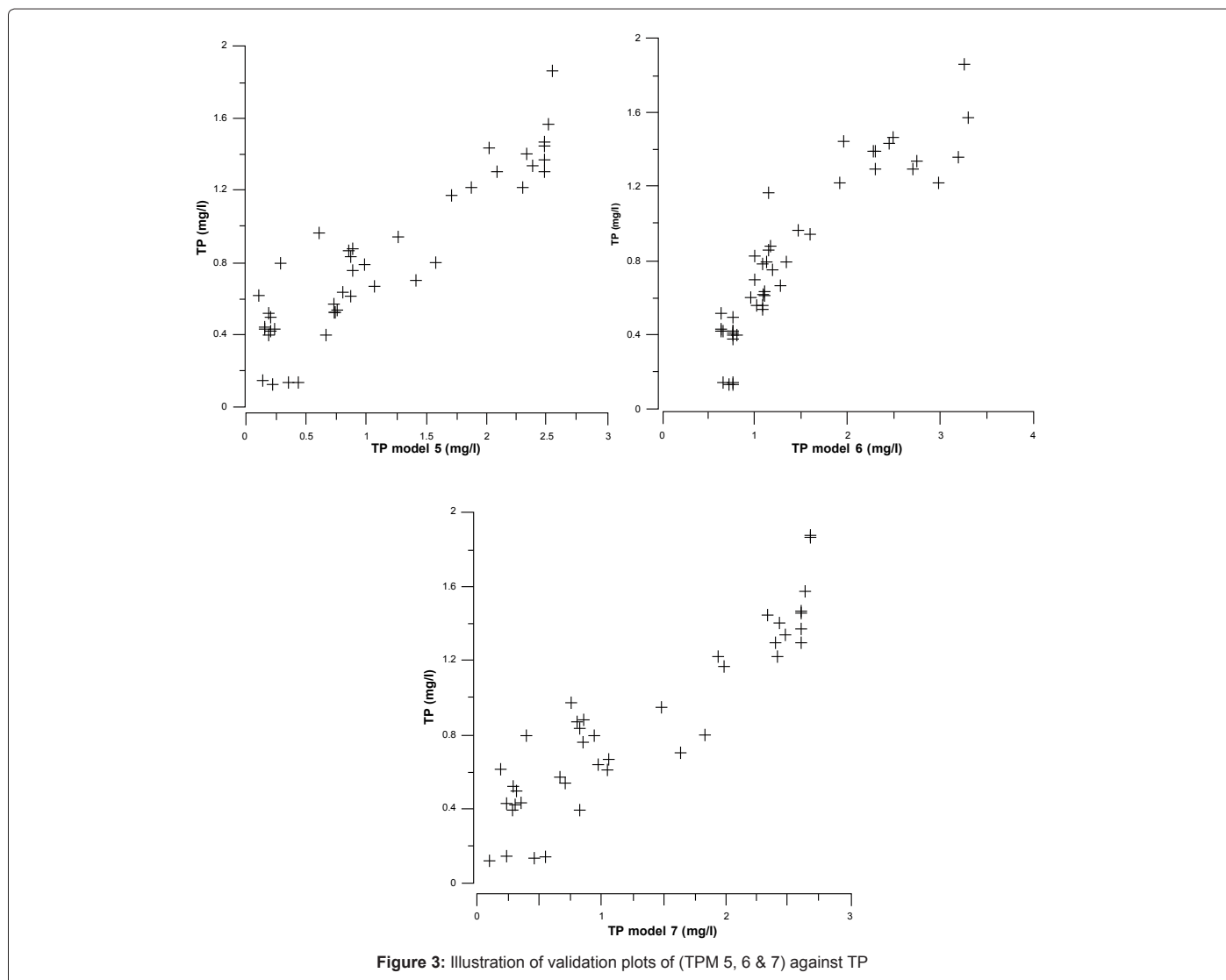
Figure 3: Illustration of validation plots of (TPM 5, 6 & 7) against TP

relevance. The smallest PRESS value indicates the best predictive model. The results shown (Table 2) show that TPM 1 is the best choice. It also has the lowest RMSE of 0.167. In general, all other models show a relative RMSE of less than 20%. 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 [30,31]. In this case, the prediction error of less 20% justifies the probable use of all models for forecasting except TPM 3 (RMSE slightly >20%).

The TPM 1 shows a better predictive power when checked against the AICc and BIC statistics (Table 2). Therefore, it is strongly recommended for predictive works under bacterial technology in urban streams and related studies. It is envisaged that for rapid reporting of preliminary results under the bacterial technology to stakeholders involve in such a project, these algorithms will be useful to achieve this goal. However, the difference between the performance of TPM 1 and the others is not significant in practical terms. As a result, depending on the work required or the available measured data for predictive works, each of these models has a good chance of providing reliable results as discussed.

Conclusions

Bacterial technology for restoration the water quality of urbanised polluted rivers promises to have appropriate use for the majority of polluted rivers in China. The past decade has seen some experimental and pilot projects through the use of this technology in parts of China. Related data on nutrient concentrations have been collated and collected during these projects. There is limited or non-existent work on mathematical models for forecasting and planning of biological treatment programs in urban streams or rivers. The models have shown a good performance against AICc, BIC and RMSE statistics. The recorded RMSE for all models was less than 20% and represent a comparative range with measurement errors (15-20%) associated with water quality variables. The TPM 1 shows the best choice of all the seven (7) models. However, the significant difference between the models in terms of practical relevance is small and hence, all models are equally good enough depending on the type and use of research. The derived models in this research envisage a contribution towards water quality monitoring and management of urban rivers and streams in China. It is also believed that, with urban rivers and streams in other

Model Number	Description of Model	Evaluation Statistics		
		R_a^2	AICc	BIC
TPM 1	TP=0.273-0.377*[pH]+(93.04Ee-04)*([pH]*[NH3-N])+(11.86e-04)*([pH]*[COD])	0.897	-2.44.33	-3.10.14
TPM 2	TP=(28.025e-04)+(93.879e-03)*[NH3-N]	0.836	-213.14	-282.94
TPM 3	TP=0.453+(11.845-04)*([NH3-N]*[COD])	0.874	-232.09	-301.89
TPM 4	TP=-0.955+(20.326e-03)*[COD]+0.154*[pH]	0.753	-182.51	-250.
TPM-5	TP=(31.474e-03)+(12.226e-03)*[pH]*[NH3-N]	0.826	-209.01	-278.81
TPM 6	TP=0.171+(27.797e-04)*[pH]*[COD]	0.759	-185.67	-255.47
TPM 7	TP=3.348-0.447*[pH]+(13.552e-03)*[pH]*[NH3-N]	0.849	-218.14	-285.91

Table 1: Derived Models and Evaluation Statictics of Each Model

Model No.	Evaluation Statistics				
	R_a^2	AICc	PRESS	BIC	RMSE
TPM 1	0.858	-103.670	1.235	-144.830	0.167
TPM 2	0.830	-96.104	1.475	-137.260	0.183
TPM 3	0.793	-87.772	1.802	-128.930	0.202
TPM 4	0.819	-93.296	1.599	-134.450	0.189
TPM 5	0.834	-96.976	1.447	-138.130	0.181
TPM 6	0.812	-91.876	1.657	-133.030	0.192
TPM 7	0.827	-95.300	1.498	-136.460	0.185

Table 2: Predictive capacities and evaluation of different models

parts of the world with similar conditions like the Class V standard grouping of the Chinese National Board, might be applicable to use the algorithms developed here.

References

1. Verhoeven JT, Arheimer B, Yin C, Hefting MM (2006) Regional and global concerns over wetlands and water quality. Trends Ecol Evol 21: 96-103.

2. Deelstra T, Girardet H (2000) Urban agriculture and sustainable cities. 43-65.

3. Sharma SK, Vairavamoorthy K (2009) Urban water demand management: prospects and challenges for the developing countries. Water and Environment Journal 23: 210-218.

4. Protocol K (1997) United Nations framework convention on climate change. Kyoto Protocol.

5. Yamin F, Depledge J (2004) The international climate change regime: a guide to rules, institutions and procedures. Cambridge Univ Pr.

6. Yudianto D, Yuebo X (2011) Numerical modeling and practical experience of Xuxi River's natural restoration using biological treatment. Water Environ Res 83: 2087-2098.

7. Beck MB, Van Straten G (1983) Uncertainty and forecasting of water quality. International Institute for Applied Systems Analysis.

8. Nevers MB, Whitman RL (2005) Nowcast modeling of Escherichia coli concentrations at multiple urban beaches of southern Lake Michigan. Water Res 39: 5250-5260.

9. Kay D, Wyer M, Crowther J, Stapleton C, Bradford M, et al. (2005) Predicting faecal indicator fluxes using digital land use data in the UK's sentinel Water Framework Directive catchment: the Ribble study. Water Res 39: 3967-3981.

10. 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 Sci Technol 45: 205-211.

11. Gao L, Zhou JM, Yang H, Chen J (2005) Phosphorus fractions in sediment profiles and their potential contributions to eutrophication in Dianchi Lake. Environmental geology 48: 835-844.

12. Henze M (2000) Activated sludge models ASM1, ASM2, ASM2d and ASM3.

13. Helsel DR, Hirsch RM (2002) Statistical methods in water resources. Techniques of Water- Resources Investigations of the United States Geological Survey.

14. 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.

15. Ge Z, Frick WE (2007) Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations. Environ Res 103: 358-364.

16. 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. Environ Sci Technol 40: 5022-5028.

17. Francy DS, Darner RA, County C (2006) Procedures for developing models to predict exceedances of recreational water-quality standards at coastal beaches. US Dept. of the Interior, US Geological Survey.

18. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Springer Verlag 267-281.

19. Burnham K P, Anderson D R (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer Verlag.

20. Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76: 297-307.

21. Schwarz G (1978) Estimating the dimension of a model. The annals of statistics 6: 461-464.

22. Mediavilla FAM, Landrum F, Shah V (2008) A comparison of the coefficient of predictive power, the coefficient of determination and AIC for linear regression. Journal of Applied Business and Economics 1261-1266.

23. Wolfe K, Parmar R, Cyterski M, Galvin M (2011) Virtual Beach 2.0 User Guide. United States Environmental Protection Agency.

24. Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. Machine Learning 3: 95-99.

25. Chau KW (2006) A review on integration of artificial intelligence into water quality modelling. Mar Pollut Bull 52: 726-733.

26. Franchini M (1996) Use of a genetic algorithm combined with a local search method for the automatic calibration of conceptual rainfall-runoff models. Hydrological Sciences Journal 41: 21-39.

27. Ng AWM, Perera BJC (2003) Selection of genetic algorithm operators for river water quality model calibration. Engineering Applications of Artificial Intelligence 16: 529-541.

28. Rauch W, Harremoa P (1999) Genetic algorithms in real time control applied to minimize transient pollution from urban wastewater systems. Water Research 33: 1265-1277.

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

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

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