Parameter Estimation of Nonlinear Nitrate Prediction Model Using Genetic Algorithm

Rui Wu* Jose T. Painumkal* John M. Volk^ Siming Liu*
Sushil J. Louis* Scott Tyler^ Sergiu M. Dascalu* Frederick C. Harris, Jr*

*Department of Computer Science and Engineering
University of Nevada, Reno
Reno, NV, USA

^Department of Geological Sciences & Engineering
University of Nevada, Reno
Reno, NV, USA

*{rui, sushil, dascalus, fred.harris}@cse.unr.edu *{josepainumkal, liusiming}@nevada.unr.edu ^{jmvolk, styler}@unr.edu

***Abstract*—We attack the problem of predicting nitrate concentrations in a stream by using a genetic algorithm to minimize the difference between observed and predicted concentrations on hydrologic nitrate concentration model based on a US Geological Survey collected data set. Nitrate plays a significant role in maintaining ecological balance in aquatic ecosystems and any advances in nitrate prediction accuracy will improve our understanding of the non-linear interplay between the factors that impact aquatic ecosystem health. We compare the genetic algorithm tuned model against the LOADEST estimation tool in current use by hydrologists, and against a random forest, generalized linear regression, decision tree, and gradient booted tree and show that the genetic algorithm does statistically significantly better. These results indicate that genetic algorithms are a viable approach to tuning such non-linear, hydrologic models.**

Keywords—genetic algorithm; nitrate; model; prediction;

I. INTRODUCTION

Nitrogen is a major nutrient that is essential for plant and animal growth. Although nitrogen is often a limiting nutrient, an abundance of inorganic species such as nitrate (NO_3) causes excessive growth among primary producers that often results in low levels of dissolved oxygen, fish kills, toxic algal blooms, and toxicity to aquatic organisms. [1][2][3]. Nutrient enrichment from nonpoint sources such as fertilizer runoff was identified as one of the largest impairments to surface water quality in the United States [4]. Daniel et al. suggest 70% of the fertilizers and feed applied to farms in the US is either lost to soil storage or transported to surface or groundwater [5]. Additionally, sewage effluent, burning of fossil fuels (emits NOx and N_2O), energy production, and industrial activities also can lead to increased nitrate in the environment [6][7][8]. Nitrate is mobile in groundwater, and drinking nitrate contaminated water has been linked to infant methemoglobinemia (MetHb), among other human health issues [9][10]. For these reasons, it is critical to measure and predict nitrate loads in rivers to better inform governmental and nongovernmental agencies such as policy makers, environmental groups, and water suppliers.

There are different models utilized by hydrologists to determine nitrate content in water. These prediction models use other constituents present in water to predict the NO_3 content. The models mainly differ from each other in the number of constituents they need to make the predictions. In this paper, an improved nonlinear prediction model is used to predict the NO_3 content in water. The proposed model uses six constituents - organic nitrogen, orthophosphate, pH level, dissolved oxygen, temperature, and discharge to make the predictions on NO3 content. The model contains 12 parameters, which need to be calibrated effectively to improve the accuracy of the predictions. Due to the nonlinearity of the model, the calibration of the model parameters is highly complex. In this paper, a genetic algorithm based approach for parameter estimation of nitrate prediction model is proposed.

Genetic algorithms (GA) are a powerful adaptive search technique that use the concepts of natural selection to mimic the process of biological evolution to efficiently solve optimization problems [12]. On searching over a large multidimensional state space, GAs can outperform other conventional search techniques due to its simplicity, effectiveness, versatility and robustness [13]. For the calibration of the proposed model, a genetic algorithm was found to be a feasible approach due to the following reasons. 1) the nonlinearity of the model 2) presence of many parameters and 3) vast search space and 4) higher possibility of convergence towards optimal values for the parameters.

To evaluate the performance of the proposed approach, we compared results from the GA tuned model with the prevalent environmental tool, LOADEST applied for predicting nitrate loads in Hellbranch Run, a protected stream in central Ohio [14]. LOADEST is a software tool offered by the United States Geological Survey (USGS), and is widely used by hydrologists to estimate the constituent loads [M/T] in water channels [15]. Besides LOADEST, the results of the genetic algorithm were compared with four other regression methods such as generalized linear regression, gradient boosted tree regression, random forest regression, and decision tree regression.

The rest of this paper is organized as follows: Section II describes the prior work. Section III introduces the methodology and various aspects of the proposed approach. Section IV presents the results and its interpretation. Section V concludes our ideas and introduces our future work.

978-1-5090-4601-0/17/$31.00c©2017 IEEE

II. PRIOR WORK

Much researches has been done to study how effectively nitrate content in water can be predicted. Almasri et al. proposed the use of Modular Neural Networks (MNN) to predict the nitrate distribution in water [16]. The MNN-based approach was simple and economical. Although it could efficiently predict the distribution of nitrate concentration in water, its performance deteriorated drastically with noisy data due to high sensitiveness to errors in the input data. Yesilnacar et al. used Artificial Neural Networks (ANN) based approach to predict the nitrate concentration in 24 observation wells in the Harran Plain, located in Turkey [17]. The developed model was cost-effective and gave a satisfactory fit to the experimentally obtained nitrate data. Poor et al. proposed the use of tree analysis to improve the predictions of low-flow nitrate in Willamette River[18]. Although regression tree analysis greatly improved the predictability compared to multiple linear regression, the results show that this approach was highly inaccurate with smaller datasets and shows an inconsistent relationship between nitrate and some other parameters. Arabgol et al. proposed the use of Support Vector Machine (SVMs) Models in predicting the nitrate concentration in ground water resources [19]. SVM models were fast, reliable and cost-effective. The prediction accuracy of SVM was better than ANN. However, the prediction accuracy of SVM models with noisy data has not yet been proven. To acquire more accurate results, we proposed a numerical equation and tune the parameters with GA in this paper.

The four objectives of our study are: 1) Use genetic algorithm to optimize the parameters of the nonlinear NO_3 prediction model 2) Evaluate the performance of GA with the results from LOADEST software. 3) Compare the performance of GA approach with four other machine learning techniques such as gradient boosted tree regression, random forest regression, decision tree regression and generalized linear regression. 4) Deal with missing fields in the dataset and evaluate how it affects the prediction capability of the model.

Our results show that our GA-based approach produced nitrate level predictions that were closer to the observed values than LOADEST and statistically significantly (t-test, $p=8.19*10^{-47}$) different from LOADEST predications. Furthermore, the GA tuned model performed better than the four other estimation methods described earlier. Therefore, using our proposed approach hydrologists can make more accurate predictions on nitrate content in water.

III. METHODOLOGY

We used a GA with rank based selection. Two-point crossover was used as the crossover technique, and bit-wise mutation was used as the mutation strategy. We also compared the performance of rank based selection strategy with other prominent selection techniques used in the genetic algorithm. The results of the comparison between different selection strategies are given in section IV.

The proposed NO_3 prediction model uses six constituents present in water to make predictions on the NO_3 level. Organic nitrogen, orthophosphate, pH level, dissolved oxygen, temperature, and discharge are the six constituents required by the model. The model maintains a nonlinear quadratic relationship with the various constituents and contains 12 parameters whose value lies in [-10.24,10.24]. The proposed model is represented as below:

$$\begin{aligned}\Psi = {} & a0 + a1 * \mathrm{Ln}\,Q + a2 * (\mathrm{Ln}\,Q)2 + a3 \\ & * \sin(2\pi * \mathrm{dtime}) + a4 \\ & * \cos(2\pi * \mathrm{dtime}) + a5 * \mathrm{dtime} + a6 \\ & * \mathrm{dtime2} + a7 * \mathrm{DO} + a8 * \mathrm{T} + a9 * \mathrm{ON} \\ & + a10 * \mathrm{OP} + a11 * \mathrm{TP}\end{aligned}$$

Where Q denotes discharge; DO denotes dissolved oxygen; T denotes temperature; ON denotes organic nitrogen; TP denotes pH level; OP denotes orthophosphate; dtime (decimal time - center of decimal time); ψ denotes nitrate load at dtime. Decimal time (calculated as decimal years in LOADEST) is an important explanatory variable for load modeling. In the LOADEST model, the third and fourth terms represent a first-order Fourier series in dtime to capture seasonal variations and the fifth and sixth terms in dtime are meant to capture linear and quadratic temporal trends [3]. Decimal time is the decimal equivalent of the date and time. To convert the date and time to its decimal equivalent, one year is represented as one revolution around the unit circle. Therefore, the values within a year are converted to their respective values between 0 and 2π. Center of decimal time is the average of all the decimal equivalents for the entire time period.

The objective function in the genetic algorithm for the estimation of optimal parameters (a0 to a11) in the proposed nonlinear NO_3 prediction model is taken as minimizing the mean square root of sum of squares between the observed and predicted nitrate content in water and is given by

$$\min \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2}$$

where P_i and A_i represent the predicted and observed values of nitrate content respectively and n is the total number of observations. In genetic algorithms, the fitness function often defined for the canonical GA. To meet this requirement, the fitness function is represented as the reciprocal of the objective function. Therefore, the RMSE values will be minimized on maximizing the fitness function.

The string length of 11 was chosen to represent each variable and was encoded with the binary digits. This is because the value of each variable lies in [-10.24, 10.24] and the precision is 0.01, which means there are 2^{11} possible numbers in total. Since there were 12 variables, from a0 to a11, the total chromosome length of the individual was 132. The population size was chosen as 200 and the number of generations was fixed to 500. A mutation probability (P_m) of .01 and a crossover probability (P_c) of 0.9 were used in the genetic algorithm to estimate the optimal values of the variables in the proposed nonlinear nitrate prediction model. After many experiments were done, we found these chosen GA parameters guarantee good results for this problem. The genetic algorithm was run for 30 times with different random seeds. The best (minimum) of the 30 runs was selected as the solution to the problem.

There are some existing hydrological tools or libraries that can predict NO_3 content in water based on the available constituent details. However, none of them can guarantee accurate predictions. To evaluate the performance of the model, several quality metrics were used. Root Mean Square Error(RMSE), Percent bias (PBIAS) and Nash-Sutcliffe efficiency (NSE) are some of the popular quantitative statistics used to perform the statistical evaluation of model accuracy. The statistical parameters are defined by the following:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{\sum_{i=1}^{n}(A_i - A^*)^2}$$

$$\text{PBIAS} = \frac{\sum_{i=1}^{n}(A_i - P_i)\ 100}{\sum_{i=1}^{n} A_i}$$

where P*i* and A*i* represent the predicted and observed values respectively.

RMSE is a widely used error metric which indicates how close the predictions to the actual values [20]. Since RMSE is the difference between the actual observed values and the values predicted by the model, for an efficient model, the RMSE value should be lower.

PBIAS is a performance metric which shows the behavior of the predictions made by the model. It indicates whether the predictions from the model overestimate or underestimate the actual observations. Positive PBIAS values denote overestimates, whereas negative values denote underestimates. For an effective model, PBIAS values should be close to zero.

NSE is a commonly used criterion in hydrology to evaluate the quality of the predictions made by the model. NSE is a normalized statistic which determines the ability of the model to make predictions that fit 1:1 line with the observed values. NSE value ranges between -infinity and 1. The higher the NSE value, the more accurate the model. To consider a model with an acceptable level of performance, NSE values should be close to 1.

The dataset used in the study was collected from the United States Survey for Big Darby Creek Watershed in Ohio. The dataset comprised 435 water samples monitored during the period of 20 years between December 1, 1996 and August 25, 2016. There were two challenges faced with this dataset. 1) cleaning the data and 2) filling the missing fields in the dataset. The dataset contains many constituent details which are not relevant for the proposed model. Finding the required constituents information and removal of unwanted constituent details from this USGS dataset was the first challenge. Out of the 435 water samples, only 140 samples had the measurements for all the constituents required by the model. For the remaining 235 samples, at least one of the constituent readings were missing. Missing fields in dataset is a common issue with environment data and researchers have employed various strategies to deal with the problem. Artificial neural networks, support vector machines, interpolation or regression techniques, and Bayesian approaches, multiple imputations are few of them. However, for this experimental study, simple linear regression was used to fill the missing fields. Because when using a deterministic linear regression approach, if results go bad, it is easy to pinpoint where the issue lies, whether with the data or the GA approach. We did not only use the complete 140 samples because we want to test if linear regression technique is good enough for this problem and it is very common that there are some missing data in the real world environmental observations. These incomplete (with some missing data) samples are also very important to most studies.

The proposed approach was compared with the results from LOADEST, which is a prominent load estimation tool used by hydrologists. On specifying the input constituents, LOADEST performs its own calibration and estimation procedures using several statistical estimation methods and forms a regression model to predict the estimated constituent. Besides LOADEST, the results of the genetic algorithm were compared with four other machine learning techniques: Generalized linear regression, gradient boosted tree regression, decision tree regression, and random forest regression. Linear regression is a popular modeling technique used to estimate values for an unknown parameter [21]. The data for the known variables (features) is used to map a linear relationship with the parameter to be estimated. Linear regression is not suited for problems which maintain a nonlinear relationship between predicted parameter and features. Generalized linear regression is more accurate than linear regression, as it allows transform predictors and interactions [22]. Decision tree regression uses decision tree as the predictive model and is widely used in data classification research [23]. It breaks down data into smaller datasets, by incrementally developing an associated decision tree. Random forest regression is similar to decision tree regression, where random forest regression uses multiple decision trees to improve the regression results [24]. Gradient boosted tree regression is another machine learning technique which follows a stage-wise fashion to build an additive prediction model using the combination of other predictive models [25]. It is a popular technique which is used by Google and Yahoo for page ranking in search engine.

In the next section, we have done several experiments to compare the performances of the introduced six methods and analyze the results.

IV. Results and Analysis

We used three performance measurements (RMSE, PBIAS, and NSE) to compare six approaches introduced. The six prediction methods are a generalized linear regression, gradient boosted tree regression, random forest regression, decision tree regression, GA, and LOADEST. LOADEST contains many methods and we chose the best results to compare with other methods.

We tested these methods with 30-fold cross-validation and used 70% of the data for training and used 30% of the data for testing. Furthermore, the models have been run 30 times and the average values are used as the final result. Some interesting phenomena are found from these results.

TABLE I. RESULTS OF DIFFERENT TECHNIQUES

Name of the method	RMSE	PBIAS	NSE
Generalized linear regression	2.02044227	2.3559	0.27549
Gradient boosted tree regression	1.96091777	1.9744	0.72986
Random forest regression	1.89409494	1.8084	0.58736
Decision tree regression	1.95253682	2.2497	0.55093
Genetic Algorithm	2.03574731	-0.8012	0. 3761
LOADEST	2.89	59.958	-0.474

Table I displays our results from all six methods. From the table, it is clear that the random forest regression has the lowest RMSE, which means this regression method prediction is closer to the observed values. GA has the best PBIAS, which shows that GA does least overestimate or underestimate compared to other methods. The gradient boosted tree regression has the best NSE. This means the method is more efficient and its prediction fits 1:1 line with the observed values. The GA is not as good as other machine learning methods, but it is slightly better than LOADEST based on RMSE and much better based on PBIAS and NSE. However, this does not mean GA is less useful. The best result of the 30 GA model runs RMSE is only 1.896525565, which is better than most other methods. This means GA can obtain the good results but it is not very robust.

Different selection strategies were tried to improve the execution of the genetic algorithm. We compared the performance of rank based selection with other popular selection strategies such as truncation selection, fitness proportionate selection, tournament selection and elitism. For truncation selection, the candidate individuals were sorted in the decreasing order of their fitness value, and the individuals were picked from the first half of the population to generate offspring. To perform tournament selection, a set of 40 individuals were randomly selected from the population and the individual with the best fitness was chosen. To implement elitism, the fittest 25 individuals in the population were copied to the next generation and thus ensure that the best chromosomes are not being lost during the evolution process. Table II shows the results obtained with different selection strategies. Among the five selection strategies implemented, rank based selection performed the best, whereas truncation selection was the worst.

TABLE II. COMPARISON OF DIFFERENT SELECTION STRATEGIES

Selection strategy	RMSE (Average of 30 runs)
Fitness proportionate selection	2.077242
Truncation selection	2.925081
Tournament selection	2.190722
Rank based selection	2.035747
Elitism	2.100074

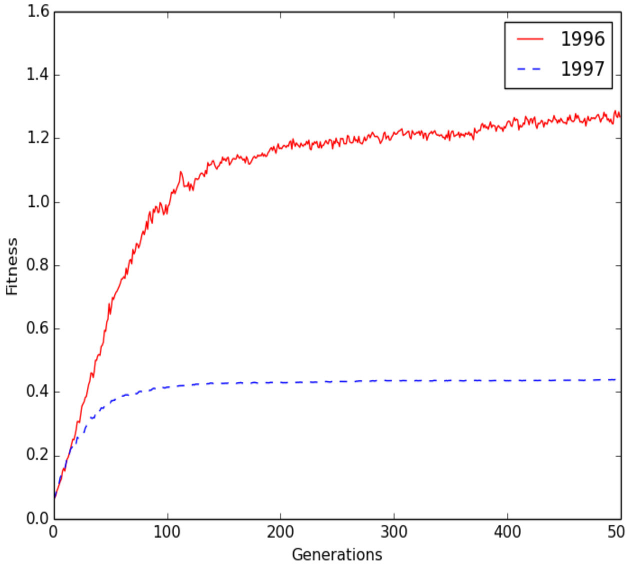

Figure 1. Comparison of the best (year 1996) and worst (year 1997) results

Different years have different characteristics. Some years are very dry (droughts) and some years are very wet (floods). Even though the year information is built in “dtime” in our fitness, the GA model performs different year by year. Figure 1 shows the best and worst results of year-wise comparison. This means that the year information is not well-built in the current fitness function. We have run the GA model with and without year information. The result shows that the year information can improve the results (with year RMSE is 2.035 and without year RMSE is 2.145).

To prove that GA method is significantly different from the LOADEST estimations, the one-tailed T-Test has been done and the p-value obtained was $8.19*10^{-47}$, which shows that the predictions of these two methods are significantly different. Also from Table III, it is clear that even though some methods, such as gradient boosted tree, have better RMSE and NSE, their T-Test values show that the predictions of these methods were not significantly different from GA predictions. Therefore, it is wrong to state that the performance of those techniques was superior to GA approach. Table IV contains more statistics results of the introduced methods.

Some other fitness functions were also tried. However, most of them did not perform very well. For example, one of the fitness function is created with the assumptions that the nitrate has non-linear quadratic relations with all the parameters. The results show that this assumption cannot guarantee good results for all the occasions and most cases it can lead to a worse result than the previously introduced fitness functions. This experiment shows that professional knowledge of the target problem is very necessary to build a good fitness function.

TABLE III. RESULTS OF T-TEST

Methods Names	P-value	T-value
GA vs LOADEST	8.19*10^{-47}	1.648
GA vs random forest	0.026	1.648
GA vs generalized linear regression	0.102	1.648
GA vs decision tree	0.274	1.648
GA vs gradient booted tree	0.486	1.648

TABLE IV. MEANS AND VARIANCE

Methods Names	Mean	Variance
GA	2.458	4.250
LOADEST	4.082	11.585
random forest	2.621	2.492
generalized linear regression	2.545	1.279
decision tree	2.515	3.536
gradient booted tree	2.461	4.317

For all the experiments mentioned above, we used the original observed dataset from USGS and filled the missing data gaps using linear regression. To test the accuracy of data filling technique, we ran the models with the original dataset and also with the modified dataset filled with linear regression values. Figure 2 shows the comparison of the results between the original dataset and the modified dataset. From Figure 2, it is evident that the performance of GA deteriorated with the filled data values. Thus, we could conclude that linear regression is not a reliable technique for filling missing data fields in environmental datasets and could be replaced with other efficient data filling techniques.

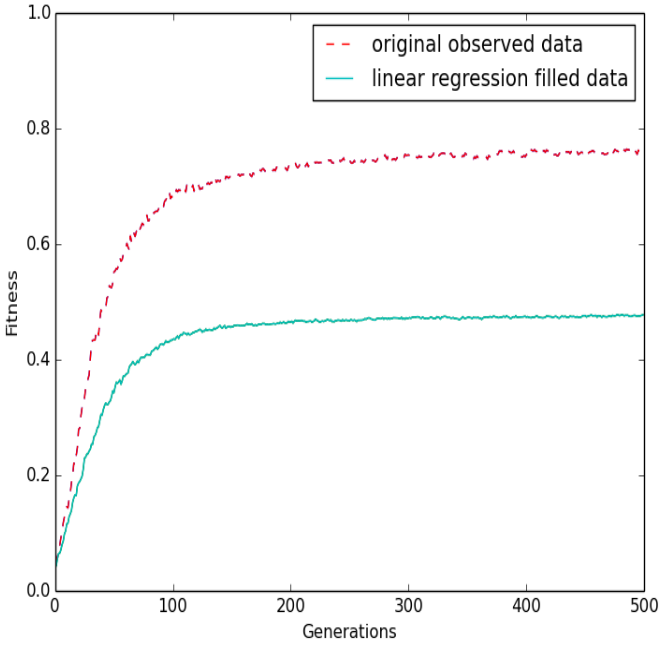

Figure 2. Comparison of the results using original and filled data

TABLE V. PARAMETER ESTIMATES OF BIG DARBY CREEK WATERSHED USGS DATASET

Parameter	Genetic Algorithm	LOADEST
a0	1.26	2.6248
a1	-6.06	0.2632
a2	3.15	-0.0030
a3	0.72	0.6947
a4	0.43	0.3325
a5	-0.17	-0.0359
a6	2.95	0.0039
a7	-0.02	-0.1511
a8	0	-0.1002
a9	1.28	-0.0344
a10	6.61	5.3633
a11	-5.55	-1.7135

Figure 3 compares the LOADEST predications, GA predictions, and the observed data. Only two methods results are shown in the graph because it is clearer than crowding all the results in a single graph. From the comparisons, we can tell that GA predictions are closer to the observed than the LOADEST predications. However, GA method generates some predictions below zero and NO_3 values cannot be negative. In the future, we plan to set some extra rules to make predications more accurate, such as turn negative value into zero to make the results more accurate.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a GA method to calibrate the parameters of an improved nonlinear hydrological nitration prediction model. From RMSE, PBIAS, and NSE, the GA method is better than LOADEST. GA predictions are significantly different from the LOADEST predictions based on T-test p-value (<0.001). We have also used some other popular machine learning techniques (generalized linear regression, gradient boosted tree regression, random forest regression, and decision tree regression) to predict nitrate content with the same dataset. The results show GA has the best PBIAS value than other methods. This means GA does least overestimates and underestimate compared with other six introduced methods.

GA's best results are as good as random forest regression predications based on RMSE.

For the nitrate model parameter calibrations, GA is not perfect. From the experiments, it is clear that the year information is not well-built in the model. In the future, we plan to modify our fitness function to improve the "dtime". Also, based on the results in the Result Section, it is clear that the data filling using linear regression can make the predictions less accurate. This means the linear regression method is not very good for this problem. We planned to replace it with other methods, such as neural networks and compare the performance. Last but not least, we want to add some extra limitations or customize some conditions to obtain better results. For example, the predictions cannot be negative.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under grant numbers IIA-1329469 and IIA-1301726.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Moore, R.B., Johnston, C.M., Robinson, K.W. and Deacon, J.R., 2004. Estimation of total nitrogen and phosphorus in New England streams using spatially referenced regression models. *US Geological Survey Scientific Investigations Report*, *5012*, pp.1-42.

[2] Redfield, A.C., 1958. The biological control of chemical factors in the environment. *American scientist*, *46*(3), pp.230A-221.

[3] Biggs, B.J., 2000. Eutrophication of streams and rivers: dissolved nutrient-chlorophyll relationships for benthic algae. *Journal of the North American Benthological Society*, *19*(1), pp.17-31.

[4] EPA, U., 2009. *National water quality inventory: report to Congress*. Tech. rep., Washington, DC: Environmental Protection Agency.

[5] Daniel, T.C., Sharpley, A.N., and Lemunyon, J.L., 1998, Agricultural phosphorus and eutrophication: A symposium overview: Journal of Environmental Quality, v. 27, p. 251-257.

[6] Galloway, J.N., Dentener, F.J., Capone, D.G., Boyer, E.W., Howarth, R.W., Seitzinger, S.P., Asner, G.P., Cleveland, C.C., Green, P.A., Holland, E.A. and Karl, D.M., 2004. Nitrogen cycles: past, present, and future. *Biogeochemistry*, *70*(2), pp.153-226.

[7] Arheimer, B. and Liden, R., 2000. Nitrogen and phosphorus concentrations from agricultural catchments—influence of spatial and temporal variables. *Journal of Hydrology*, *227*(1), pp.140-159.

[8] Driscoll, C.T., Lawrence, G.B., Bulger, A.J., Butler, T.J., Cronan, C.S., Eagar, C., Lambert, K.F., Likens, G.E., Stoddard, J.L. and Weathers, K.C., 2001. Acidic Deposition in the Northeastern United States: Sources and Inputs, Ecosystem Effects, and Management Strategies: The effects of acidic deposition in the northeastern United States include the acidification of soil and water, which stresses terrestrial and aquatic biota. *BioScience*, *51*(3), pp.180-198.

[9] Hudak, P.F., 2000. Regional trends in nitrate content of Texas groundwater. *Journal of hydrology*, *228*(1), pp.37-47.

[10] Knobeloch, L., Salna, B., Hogan, A., Postle, J. and Anderson, H., 2000. Blue babies and nitrate-contaminated well water. *Environmental Health Perspectives*, *108*(7), p.675.

[11] A. Fan and V. Steinberg, "Health Implications of Nitrate and Nitrite in Drinking Water: An Update on Methemoglobinemia Occurrence and Reproductive and Developmental Toxicity", *Regulatory Toxicology and Pharmacology*, vol. 23, no. 1, pp. 35-43, 1996.

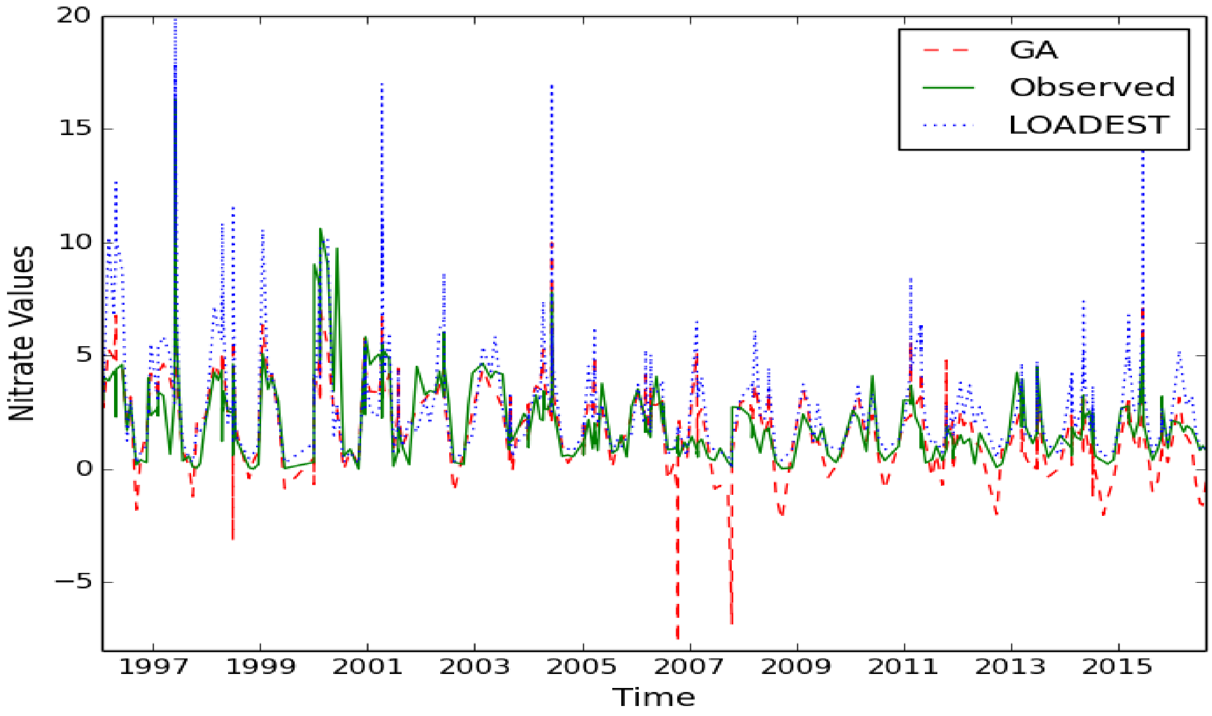

Figure 3. GA and LOADEST predictions with actual observed nitrate level.

[12] D. Goldberg and J. Holland, Genetic algorithms and machine learning, vol. 3, no. 23, pp. 95-99, 1988.

[13] Lucasius, C.B. and Kateman, G., 1993. Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and intelligent laboratory systems*, *19*(1), pp.1-33.

[14] Volk, J.M., 2011. Spatial and Temporal Variations of Water Quality in Hellbranch Run: A Historical Perspective. Thesis, The Ohio State University, Columbus, OH 69.

[15] Runkel, R. L., C. G. Crawford, and T. A. Cohn. Load estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers. Page 69. U.S. Geological Survey Techniques and Methods Book 4.

[16] M. Almasri and J. Kaluarachchi, "Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data", Environmental Modelling & Software, vol. 20, no. 7, pp. 851-871, 2005.

[17] M. Yesilnacar, E. Sahinkaya, M. Naz and B. Ozkaya, "Neural network prediction of nitrate in groundwater of Harran Plain, Turkey", *Environmental Geology*, vol. 56, no. 1, pp. 19-25, 2007.

[18] C. Poor and J. Ullman, "Using Regression Tree Analysis to Improve Predictions of Low-Flow Nitrate and Chloride in Willamette River Basin Watersheds", *Environmental Management*, vol. 46, no. 5, pp. 771-780, 2010.

[19] R. Arabgol, M. Sartaj and K. Asghari, "Predicting Nitrate Concentration and Its Spatial Distribution in Groundwater Resources Using Support Vector Machines (SVMs) Model", *Environmental Modeling & Assessment*, vol. 21, no. 1, pp. 71-82, 2015.

[20] Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), pp.679-688.

[21] Kutner, M.H., Nachtsheim, C. and Neter, J., 2004. Applied linear regression models. McGraw-Hill/Irwin.

[22] Nelder, J.A. and Baker, R.J., 1972. Generalized linear models. Encyclopedia of statistical sciences.

[23] Olaru, C. and Wehenkel, L., 2003. A complete fuzzy decision tree technique. Fuzzy sets and systems, 138(2), pp.221-254

[24] Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

[25] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.