A Machine Learning Tool for Weighted Regressions in Time, Discharge, and Season

Alexander Maestre, Ph.D. P.E.
Dept. of Civil, Construction and Environmental Engineering
The University of Alabama Tuscaloosa,
Alabama. USA

Eman El-Sheikh, Ph.D.
Department of Computer Science
University of West Florida Pensacola,
Florida. USA

Derek Williamson, Ph.D.
Dept. of Civil, Construction and Environmental Engineering
The University of Alabama Tuscaloosa,
Alabama. USA

Amelia Ward, Ph.D.
Center for Freshwater Studies
The University of Alabama Tuscaloosa,
Alabama. USA

***Abstract*— A new machine learning tool has been developed to classify water stations with similar water quality trends. The tool is based on the statistical method, Weighted Regressions in Time, Discharge, and Season (WRTDS), developed by the United States Geological Survey (USGS) to estimate daily concentrations of water constituents in rivers and streams based on continuous daily discharge data and discrete water quality samples collected at the same or nearby locations. WRTDS is based on parametric survival regressions using a jack-knife cross validation procedure that generates unbiased estimates of the prediction errors. One of the disadvantages of WRTDS is that it needs a large number of samples (n > 200) collected during at least two decades. In this article, the tool is used to evaluate the use of Boosted Regression Trees (BRT) as an alternative to the parametric survival regressions for water quality stations with a small number of samples. We describe the development of the machine learning tool as well as an evaluation comparison of the two methods, WRTDS and BRT. The purpose of the tool is to evaluate the reduction in variability of the estimates by clustering data from nearby stations with similar concentration and discharge characteristics. The results indicate that, using clustering, the predicted concentrations using BRT are in general higher than the observed concentrations. In addition, it appears that BRT generates higher sum of square residuals than the parametric survival regressions.**

***Keywords*—Machine Learning; Boosted Regression Trees; Survival Parametric Regression; Water Quality Modeling; Weighted Regressions in Time; Discharge; and Season**

I. INTRODUCTION

The United States Geological Survey (USGS) has developed linear models for predicting daily concentration of water constituents in rivers and streams using physical and temporal explanatory variables. The majority of these models are based on regressions that evaluate the correlation between observed concentrations and other variables including water discharge and time. Recently, a new model has been developed by the USGS to estimate daily concentrations using Weighted Regressions in Time, Discharge, and Season (WRTDS) [1]. Two main advantages of WRTDS include the possibility of conducting regressions with censored information (non-detects) using parametric survival regressions. In addition, WRTDS uses a jack-knife cross validation approach that evaluates the importance of each survival regression by selecting subsets of the complete dataset. The cross validation approach is also used to identify trends of the constituent concentration in time.

WRTDS has been created by a series of routines written in R, a free package for statistical computing and graphics [2]. The statistical method estimates the concentration using two libraries: dataRetrieval and EGRET. The first library, dataRetrieval [3], automatically downloads existing records of water discharge and water constituent concentrations from a dedicated server. Approximately 14,500 parameters are available for download using the dataRetrieval tool. The list of parameters available in the server includes nutrients, pesticides, organics, and physical properties among others. The second library, EGRET [4], was created to explore and generate graphics associated with river concentration trends. EGRET conducts the parametric survival regressions and estimates daily concentrations in those periods when samples were not collected.

WRTDS has been tested in more than two dozen stations in the U.S. [1][5-11]. The use of this technique has become popular in recent years because it uses locally weighted regressions to estimate daily concentrations. During the regression process, WRTDS establishes the regression coefficients using only observed concentrations with similar discharge, season, and time to the day that is being estimated [9]. However, one of the restrictions of this method is that requires a large number of samples (minimum 200) collected at the specific station with daily water discharge records collected for at least 20 years without major gaps [1].

There are approximately 26,000 USGS stations installed throughout the U.S. A large percentage of these stations have long historical records of daily water discharge but only a few have more than the required 200 water quality samples. Fortunately, other agencies (including state and local environmental agencies) have been collecting additional water quality samples for several decades. The information collected by these agencies has been motivated by cities, non-profit organizations and communities to assess and manage the quality of rivers and streams.

Support from NSF EPSCoR EPS-1010607 and the 2013 USGS 104(G) Competitive Grants Program/Alabama Water Resources Research Institute (Project 2013AL156B).

Improvements on these water quality stations include the installation of real time stations. Currently there are approximately 1,700 stations in the U.S. that collect water discharge with a frequency of 15 minutes or less. The information collected by these stations can be downloaded automatically via Internet.

In this paper, we analyze the possibility of generating nitrate + nitrite-N concentration estimates in stations that have few samples. To achieve that, we generate a training data set with samples collected in stations with similar concentration and discharge characteristics and generate a Boosted Regression Tree (BRT). BRT is a non-parametric technique that successfully identifies the influence of the predictors in the response when the interaction occurs in a complex and non-linear way [12]. It has been used to investigate high variance traffic crash data in Taiwan [13], predict fishing effort distributions [14], and the identification of processes that drive the richness, composition, and occurrence of plants species in northwest Finland [15]. Machine learning methods using BRT have also been used to determine the best set of automatic methods for fine-tuning the code executed on the graphics processing unit (GPU) in different computer architectures [16]. Although BRT has been used for a variety of problems, no literature was identified on its use for water quality modeling. Our work reported in this paper on applying BRT to this problem is thus novel.

The selection of training set stations is obtained from an arbitrary set classified by a clustering algorithm. Once the subset of similar stations is identified, the tree model is created using the stations located within a cluster. To evaluate the estimates, lack of fit of predicted and observed concentrations are compared for both WRTDS and BRT.

It is hypothesized that the use of Boosted Regression Trees could improve the concentration estimates in stations with less than 200 samples. A machine learning approach appears to be an ideal solution for such situations. As the model is analyzing new stations, a routine or program could identify patterns, similarities, and differences with previous runs and decide which combination of stations produces the best estimates.

II. Weighted Regressions in Time, Discharge, and Season (WRTDS) Method

Weighted Regressions in Time, Discharge, and Season (WRTDS) is one of the most recent methods developed by the United States Geological Survey (USGS) with the purpose of analyzing long-term, water quality data sets. One of the strengths of the method is that parameters of the mathematical model adjust to changes that occur with time. In addition, it has the capability of downloading data and metadata automatically from the National Water Information System (NWIS). It also includes multiple routines that allow the user to conduct preprocessing of the original data sets and identify the presence of outliers and influential observations that may cause bias in the estimated concentrations.

Equation (1) shows the mathematical equation that serves as the foundation of the WRTDS method:

$$\ln(c) = \beta_0 + \beta_1 t + \beta_2 \ln(Q) + \beta_3 \sin(2\pi t) + \beta_4 \cos(2\pi t) + \varepsilon \qquad (1)$$

Where c is the concentration, the β terms are the unknown regression coefficients, Q is the discharge, t is the time, and ε are the independent random errors.

In a regular regression, the fitted coefficients are constant for the entire data set. In WRTDS, each observed concentration is recalculated using a jack-knife cross validation procedure in which a subset is extracted based on windows that involve ranges in time, discharge, and season. The parametric survival regression conducted by the method has the advantage of accepting the presence of censored information.

Due to the generation of subsets, the number of samples and the period of data collected must be large in order to identify trends. Stations with few collected samples cause the method to calculate poor fitted coefficients.

III. Boosted Regression Trees (BRT)

Classification trees are an alternative to regression models to predict the concentration using the same terms included in equation (1). Classification trees have several advantages: (1) trees are very flexible and can accept broad types of responses including categorical, numerical, and survival data; (2) trees are invariant to monotonic transformations of the independent variables; (3) trees are easy to construct; and (4) trees are easy to interpret [17]. At the same time, trees have the disadvantage that they create poor predictors and in the case of large trees they are difficult to interpret [18].

When the response variable is numeric, the tree is considered a regression tree. On the other hand, when the response is categorical, the tree is called a classification tree. One advantage of classification trees is that they can be represented in a figure with branches and leaves representing the different homogeneous groups.

The tree is constructed by repeatedly breaking the data into exclusive subsets of homogeneous data to the extent possible. The splitting process continues until an overlarge tree is created, and then the tree is pruned to the desired size. In order to select the size of the tree that accurately predicts the prediction error, the method uses a procedure called cross validation. During cross validation, a portion of the observations is deleted and recalculated using the remaining observations. The recalculated values are compared with the original observations to calculate the prediction error.

Boosting appeared as a method to improve the poor prediction capabilities of classification trees [18-19]. Boosting is based on the idea that it is easier to find and average many weak classifiers than trying to find a single highly accurate prediction rule. The advantage of this method is that it is sequential. At each step the model is fitted iteratively to the training data by the current sequence of trees, and these classifications are used as weights to the next step. Incorrect classifications will have higher weights in the next step than cases that were hard to classify, increasing their chance to be correctly classified.

IV. Model Based Clustering

Preliminary analysis of the BRT method indicated that the station is one of the parameters with the highest influence

during the generation of the tree. The initial step during the generation of the BRT model is the selection of a training set for the model. Nitrate + nitrite-N concentration in rivers and streams varied greatly due to land use practices, location, and fluctuations in discharge [20-21]. The concentration of nitrate + nitrite-N at the test station could be estimated by selecting nearby stations with similar discharge and concentration distribution. For this reason, it was proposed to create a large database with nitrate + nitrite-N concentrations and discharge values for multiple stations located throughout the U.S. and, using a clustering method, select those stations with concentration distribution similar to the distribution observed at the test site.

The R package mclust was chosen to select the nitrate + nitrite-N concentration and discharge values from those stations similar to the test station [22]. The package mclust implements a Gaussian hierarchical clustering algorithms and the expectation–maximization (EM) algorithm for a parameterized mixture of models with the possible addition of a Poisson noise term [23]. One of the advantages of mclust is that it automatically selects among 10 different combinations of the parameterizations of the covariance matrix finding the clusters with the best Bayesian Information Criterion (BIC).

V. Research Methods

A Python program was combined with an R script to select information from desired stations and evaluate if there was an improvement in the estimation of nitrate + nitrite-N concentration using the BRT model. The program and script perform four steps during the process: (1) generation of a master dataset; (2) identification of stations with similar characteristics; (3) generation of BRT model; and (4) comparison between WRTDS and BRT models for stations along the Sipsey River (located near Tuscaloosa, Alabama).

A. Generation of Master Dataset

The first step in the process was to retrieve relevant information from two previous studies (Mississippi River [7,11] and the Chesapeake Bay [1] basins) and stations located near the Sipsey River. An interface tool was created to generate the training dataset. The user has the capability of either using the tool or creating a text file that includes the list of stations, the parameter to be analyzed, and the period of analysis. In the text file, each row corresponds to a station of the training dataset. The tool and interface are explained in section VI. Once all the stations have been entered into the system, the program will classify those stations with similar nitrate + nitrite-N concentration and discharge distributions.

The stations near the Sipsey River were selected from a recent analysis on the variability of nitrate + nitrite-N concentration and discharge completed for rivers of the Mobile Alabama River System (MARS) [8]. Table I shows a summary of the data for stations included in the comparison. Note that the station located in Sipsey River was not included in the training dataset (USGS station 02446500). The last column in the table indicates the assigned cluster that will be discussed in Section VII Results. The stations included in Table I were based on previous references, geographic proximity to the Sipsey River, similar drainage area size, and similar land uses in the basin. Indeed, the MARS stations were selected because they have similar climate conditions to those expected at the Sipsey River.

TABLE I. Data For Selected Stations

Basin	USGS Station Number	Nitrate + Nitrite Concentration (mg-N / L)		Logarithm of Discharge (m^3/s)		Cluster
		$\bar{x}$	σ	$\bar{x}$	σ	
CHESAPEAKE	01491000	1.28	0.405	1.288	1.432	3
	01578310	1.09	0.375	7.353	1.077	3
	01594440	1.17	0.406	2.451	1.018	3
	01646580	1.12	0.491	5.54	1.272	3
	01668000	0.49	0.291	3.644	1.581	3
	01673000	0.268	0.106	3.013	1.402	1
	01674500	0.155	0.088	2.74	1.615	1
	02035000	0.23	0.144	5.102	1.214	1
	02041650	0.171	0.139	3.133	1.587	1
MOBILE ALABAMA RIVER SYSTEM	02411000	0.135	0.088	5.082	0.949	1
	02419890	0.172	0.067	4.163	1.016	1
	02424000	0.315	0.081	3.504	0.824	1
	02429500	0.142	0.086	5.553	0.678	1
	02444160	0.046	0.145	4.355	0.867	1
	02446500 [a]	0.105	0.071	2.256	1.23	1
	02454055	0.046	0.191	1.567	1.343	1
	02462000	5.03	2.458	1.334	0.701	4
	02464000	5.03	2.458	-0.197	2.146	4
	02466031	0.229	0.155	4.215	0.956	1
	02469762	0.207	0.125	5.893	0.994	1
	02411000	1.11	0.4	9.034	0.795	3
MISSISSIPPI	05420500	1.67	0.954	7.377	0.608	2
	05465500	5.39	2.508	5.47	0.928	4
	05586100	4.17	1.719	6.743	0.828	4
	05587455	3.06	1.271	8.007	0.646	2
	06934500	1.345	0.738	7.789	0.637	2
	07022000	2.41	0.913	8.836	0.614	2
	07373420	1.38	0.508	9.698	0.537	2

[a] USGS Station Sipsey River near Elrod, not included in the training dataset

B. Identification of Stations with Similar Characteristics

In general, the distribution of water discharge follows either power law or lognormal distribution [24]. Stations with similar median logarithm of discharge and median logarithm of nitrate + nitrite-N concentration could originate from areas of similar land use, catchment area, or times of concentration. Clustering

analysis was conducted on stations that shared similar median and standard deviation of the natural logarithm of the discharge and nitrate + nitrite-N concentration.

It was hypothesized that, as the number of stations in the cluster increased, the results of the BRT would improve by increasing the number of observations in the training set. The statistical program R was selected to calculate the median and standard deviation of the natural logarithm of the nitrate + nitrite-N concentration and discharge of all the stations included in the analysis.

One of the assumptions behind the idea of clustering stations of similar characteristics was that all the stations in the clusters would be affected by the same phenomena that were regional or national in scope. For example, it was hypothesized that if a specific year was wet, all the instruments included in the cluster recorded large discharge values that year. These two conditions could impact the coefficients related to time and discharge in equation 1. On the other hand, it was also considered that clustering stations located in regions with different climate patterns (i.e., northern versus southern U.S.) may affect the seasonal terms of the equation. For this reason, it was also considered preferable to select stations located within the same region.

C. Generation of the BRT Model

In the previous step the function mclust identified four clusters. In this step, mclust identified which cluster was associated with the station located in the Sipsey River (in this case, Cluster 1). The stations within the same cluster of the Sipsey River were selected for the generation of the Boosted Regression Tree. The BRT model was created using the library gbm for the General Boosted Model [25].

The R function gbm.step was used to generate the General Boosted Model. This function determines the optimal tree size using the k-fold cross validation procedure [26]. The default option in gbm.step uses 10 folds and a bag fraction of 0.5, which indicates that 50 percent of the observations of the observed variables are selected to construct the model. As indicated previously, since the distribution of nitrate + nitrite-N concentration and discharge followed a lognormal distribution, it was assumed that the logarithm of these parameters should follow a normal distribution. The model requires the selection of a method to calculate the loss function. Because both discharge and nitrate + nitrite-N concentration are continuous variables, it was decided to use the Gaussian option to focus on minimizing the square error between the observed and predicted values.

The last two parameters in the function gbm.step are the tree complexity and learning rate. The learning rate refers to how quickly the estimated value is calculated based on the previous estimated value plus a portion of the value obtained by the fitted regression model. On the other hand, tree complexity refers to the depth of the tree (also known as the interaction depth), which is a function of the number of terminal nodes in the tree. It has been recommended for the learning rate to be as small as possible and obtain the optimum number of iterations by cross validation [25]. It is important in BRT models to avoid a large number of iterations because that can cause overfitting [27]. Overfitting occurs when the model starts depicting the random error instead of the relation between the predictors and the response.

The authors conducted preliminary analyses using sites located in Alabama, varying the tree complexity between 2 and 20 and the learning rate between 0.0001 and 0.05. The results of these analyses indicated that, as the tree complexity increases, the number of trees decreases. The same pattern was observed between the learning rate and the number of trees. The lowest cross validation correlation standard error was observed when the tree complexity was 5 and the learning rate was 0.01.

D. Comparison between WRTDS and BRT Model

In WRTDS the estimates are based on the observations from the same station. On the other hand, BRT estimates are based on observations from other stations. The goal is to observe which method generates better estimates of nitrate + nitrite-N concentration for each of the observed concentrations. A perfect fit creates a straight line between the observed and predicted values. The sum of square errors (SSE) was selected as a measure of fitness between the WRTDS and BRT models. The model with the lowest SSE would produce the best estimates.

VI. System Description

The interface tool was developed in Python. The WRTDS model, BRT model, clustering analysis, and comparison between models were completed using the statistical program R. Figure 1 shows a flow diagram describing how the Python tool and the R script interact during the estimation of the nitrate + nitrite-N estimates.

The graphical interface tool was developed using the Tkinter/ttk package that provides dynamic interaction between the program and the routines executed by R. The interface performs two main tasks: (1) processes information about the stations and parameters included in both models; and (2) executes an R script that creates and compares the WRTDS and BRT models. Figure 2 shows the interface tool that runs the simulation. The user enters the information of each station by completing the fields available on the main screen. Among the parameters needed by the model are the station number, parameter to be analyzed, discharge information, and period of analysis. The interface allows the user to either download automatically the information from the USGS website or access it from a text file that follows the format required by WRTDS.

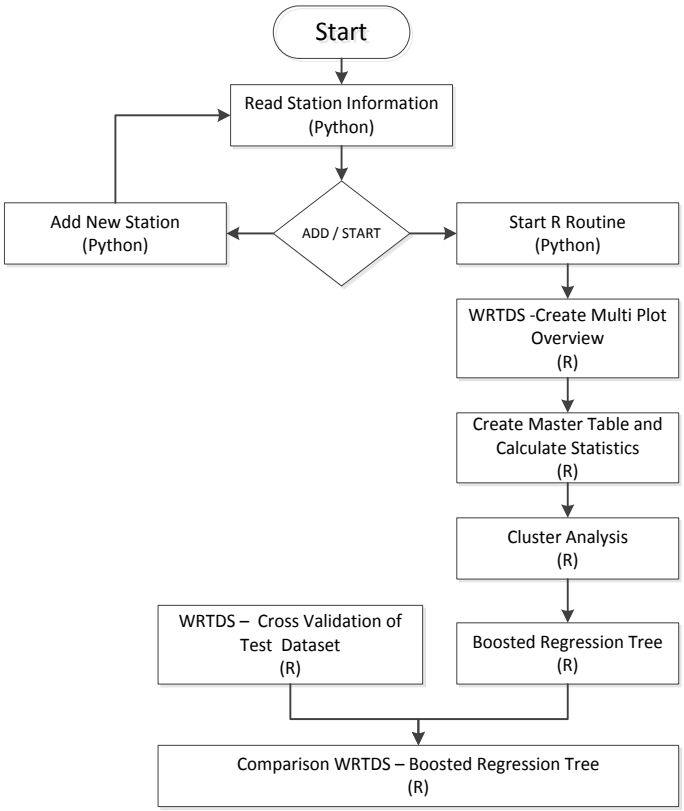

Fig. 1. Flow diagram of the steps involved in the development of the tool.

The "Add Station" button adds the information to a text file that will be read by the R script. The user adds as many stations as needed to run the model. The "Start" button initiates the R script program. In the background, R reads the information from the text file created by the interface tool and creates a data frame with all the records obtained from the selected stations. During this process, the tool automatically generates three figures for each station: concentration versus time, discharge versus time, and a multi-plot data overview (Figure 3). All the figures generated by the script are saved as images and pdf files in a separate folder.

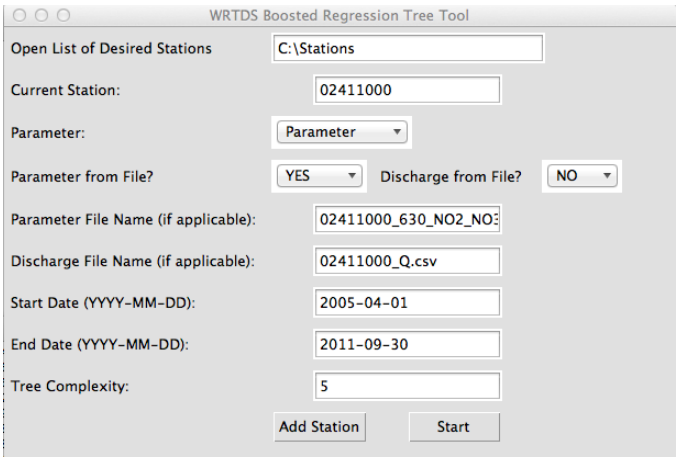

Fig. 2. Interface of the machine learning tool for comparison of WRTDS and BRT methods to estimate nitrate + nitrite-N concentration.

The script also generates three text files that could be used for further analyses: (1) a summary table with all the observations from all the stations; (2) a table that includes the median, standard deviation, and first and third quartiles of concentration and discharge for each station; and (3) a table that indicates the cluster assigned to each station during the cluster analysis.

VII. Results

In this article, we present the results for the stations included in Table I. Figure 3 shows an example of one of the multi-plot data overview figures generated by WRTDS (station 02411000, Coosa River, near Wetumpka, AL). The chart allows the identification of gaps, outliers, as well as influential points, and provides a general idea of the number of samples collected by month. The figure has four panels. In the upper left panel is a scatterplot of concentration versus discharge. This plot shows extreme events and potential correlations between discharge and concentration. Notice that both axes are in log scale matching the terms included in Equation (1). In the upper right box there is a scatterplot of the concentration versus time. This plot shows gaps and major changes in concentration with time. In the lower left corner is the distribution of samples by month. This box plot confirms that samples were collected throughout the year with a relative similar frequency. The lack of samples during specific times of the year could have an impact on seasonal components. Finally, in the lower right corner there are two box plots that compare the distribution of the discharge records for the whole period of analysis and when samples were collected.

In this case, the results demonstrated that there was a positive correlation between discharge and nitrate + nitrite-N concentration, no significant gaps or censored observations, a good distribution of samples collected during the year except for January, February, and May, and that the discharge distribution of the sampling dates was similar to the distribution of the whole period of analysis.

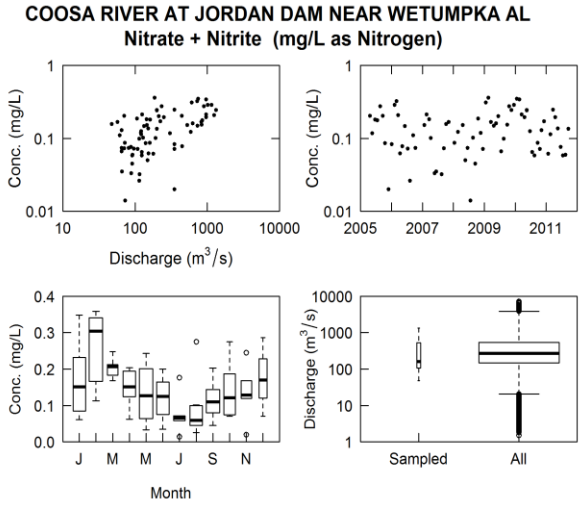

Fig. 3. Example of one of the multi-plot data overview figures generated by WRTDS.

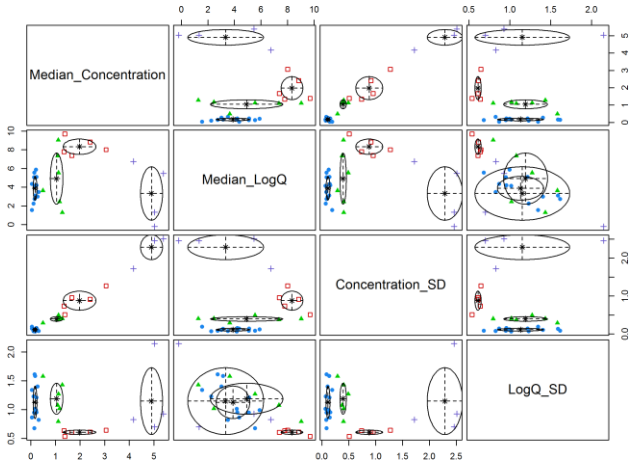

Fig. 4. Clustering example using the library mclust for USGS stations based on median and standard deviation of nitrate + nitrite-N concentration and logarithm of discharge. The four clusters are: Cluster 1 (blue circle), Cluster 2 (red square), Cluster 3 (green triangle), and Cluster 4 (purple plus symbol).

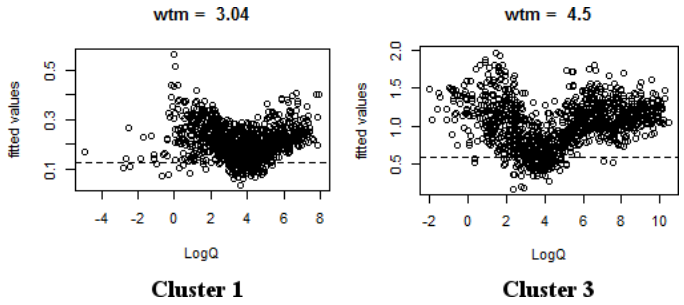

Fig. 5. Fitted values from the BRT models for Cluster 1 (low nitrate + nitrite-N concentration and intermediate discharges) and Cluster 3 (samples collected mainly in the northern part of the Chesapeake Basin).

Figure 4 and Table I show the results of the clustering analysis based on the concentration, discharge median and standard deviation of all the stations included in the analysis. The clusters are identified by colors: Cluster 1 (blue circle); Cluster 2 (red square); Cluster 3 (green triangle); and Cluster 4 (purple plus symbol). The analysis indicated that the best cluster model was the one that uses equal volume, equal shape, and variable orientation (EEV) with four clusters.

Figure 4 shows that there is a linear relation between the nitrite + nitrate-N concentration median and its standard deviation. This relation appeared in all the clusters. Another panel that shows a potential correlation involves the median nitrate + nitrite-N concentration with the median logarithm of the discharge. The results show that clusters are mainly generated by the range of concentration and geographical location. For example, all the stations with elevated median nitrate + nitrite-N concentration (greater than 4 mg/L as N) were clustered together (Cluster 4). Two of the stations were located in the Mobile Alabama River System (MARS) and two in the Mississippi Basin. These four basins appeared to be associated with urban and agricultural activities.

Cluster 2 appeared to be associated with elevated discharge and nitrate + nitrite-N concentrations. All of the stations in Cluster 2 were located in the Mississippi Basin; Cluster 1 includes those sites with low nitrate + nitrite-N concentrations and intermediate flows. All of the Alabama sites and stations located in the Chesapeake Basin south of the Rappahannock River were included in this cluster. None of these sites have a median concentration greater than 0.5 mg/L as N. Finally, Cluster 3 shows similar discharge values to Cluster 1 but median nitrate + nitrite-N concentrations were between 0.5 and 1.5 mg/L as N. Based on these four clusters, four BRT models were created, one for each cluster.

Figure 4 also shows the clearly identified clusters in the plot of median nitrate + nitrite-N concentration and logarithm of discharge. It suggests that no large variation exists in the median concentration for a wide range of discharge variation except for Cluster 2. This could indicate that specific ranges of concentrations could be present for a wide range of discharges. In this panel, Clusters 1 and 3 show a similar range of discharge for small variations in nitrate + nitrite-N concentration.

Figure 5 shows the fitted values using the BRT models for these two clusters. There is a similar downward trend in concentration as the discharge increases until reaching a value of 50 m^3/s. From that point the nitrate + nitrite-N concentration increases again with the increase in discharge until reaching a plateau for values larger than 400 m^3/s. Figure 6 shows these patterns for the four clusters. Clusters 1 and 3 have a similar pattern. In these two clusters changes in discharge have a significant effect on concentrations with the initial decrease of concentration (dilution), increasing for values higher than 50 m^3/s (re-suspension) and the plateau after 400 m^3/s.

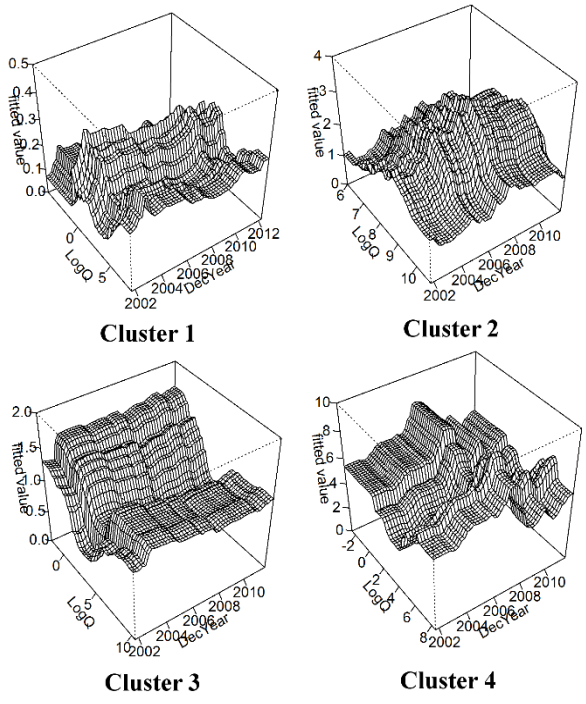

Fig. 6. Estimates of nitrate + nitrite-N concentration from the Boosted Regression Tree model at different conditions of time and discharge.

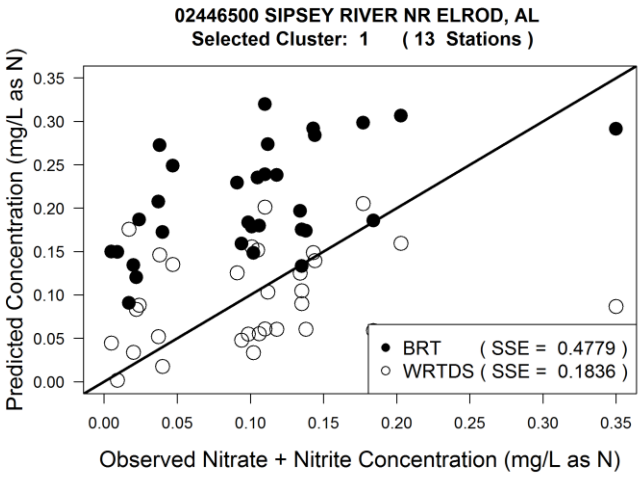

Fig. 7. Observed versus predicted nitrate + nitrite-N concentration values for the Sipsey River station at Elrod, AL. Predicted concentrations were obtained using WRTDS and BRT models.

Cluster 4 shows a similar pattern to the previous two clusters, but it is clear that there has been a strong decrease in nitrate + nitrite-N concentration with time in these four stations. This could be associated with changes in agricultural activities since 2008. In Cluster 2 the pattern is different than in the other clusters. It does not appear to have an initial dilution or change in concentration with time. Cluster 2 shows constant increase in nitrate + nitrite-N concentration until discharge values near 8,000 m3/s. A decrease in nitrate + nitrite-N concentration occurred in this cluster for larger discharges.

Finally, Figure 7 shows the observed versus predicted values at the Sipsey River station near Elrod, Alabama using both BRT and WRTDS models. It is expected that good estimates should all fall along a straight line with slope 1 that crosses the origin. This station was assigned to Cluster 1 with other 12 stations that have similar discharge and nitrate + nitrite-N concentration values.

The results of the WRTDS model are in white circles while the results of the BRT model are in black. Both models have similar results. Some of the predicted nitrate + nitrite-N concentrations using the BRT model fall directly on top of the straight line with slope 1, including the observation with the highest concentration. However, the sums of the square errors for both models indicated that the WRTDS model was better than the BRT. In fact, except for one estimate, all the estimates from the BRT model were higher than the observed nitrate + nitrite-N concentrations.

VIII. Discussion

In this paper, we presented an alternative method to WRTDS for the estimation of nitrate + nitrite-N concentrations in large rivers and streams. The results indicated that, even if the current estimates are not perfect or better than those obtained with WRTDS, the method has the potential of identifying stations with similar characteristics, correlations with several ranges of discharge, and trends with time. The method is promising because it can improve the estimates as data is collected and more stations are added to the system.

WRTDS is disadvantageous because it only uses data from the station that is being analyzed. For this reason, it requires a large number of samples collected during a period longer than 20 years. The combination of clustering and BRT allows the generation of large datasets with the goal of improving the accuracy of the estimates.

One of the advantages of combining clustering and BRT is the possibility of classifying streams and rivers based on the distribution of nitrate + nitrite-N concentration and discharge. Land uses and sources of nitrate + nitrite-N are in general different for each site included in the cluster. However, rain patterns, extreme events, and droughts can affect large areas of the country in a similar manner. The fact that many of the clusters were associated with geographical location suggest that changes in atmospheric deposition, human activity, and climate conditions will be observed in multiple stations at the same time.

Nitrate + nitrite-N concentration appeared to be highly correlated with water discharge in all the clusters. If stations located within the same cluster have the tendency to increase concentration as the discharge increases, it is expected that during a wet year the concentration estimates will be above average for all stations included in the cluster. It is important to continue research on methods or procedures that allow the extrapolation of trends and patterns observed in a group of stations to the station of interest.

IX. Conclusions

The use of a machine learning tool combined with cluster analysis offers great potential for the advancement of hydrological models. Clusters and the use of trees help identify trends and potential correlations between nitrate + nitrite-N concentration and discharge. In the past, these correlations were assumed to be linear. New methods like WRTDS enhance the capability of modeling non-stationary processes in rivers and streams.

The possibility of analyzing multiple stations with similar nitrate + nitrite-N concentration and discharge distributions opens the potential for developing simple models that effectively simulate dilution and re-suspension conditions. Boosted Regression Tree (BRT) models have the potential of simulating these processes as well as identifying trends with time.

The use of BRT and clustering did not appear to be a good alternative for the estimation of nitrate + nitrite-N concentration in sites with small number of samples. The combination of samples from multiple stations increases the variability of estimates. The machine learning tool could be improved if the influence of the multiple stations is removed during the process.

Recently, there has been an interest in the importance of representing the non-stationarity of the physical processes in future hydrological models [28]. The possibility of identifying regional trends by clustering stations with similar patterns could help future models rapidly identify changes caused for variations due to climate change, water infrastructure, and changes in land use and land cover. This approach makes the

use of machine learning techniques and algorithms a powerful tool for the next generations of hydrological models.

Future work in this area includes the design of systems that are able to identify patterns in the data collected in real time or from forecasting models. Such systems can identify variables that reduce the magnitude of errors in the estimates and potential correlations that reduce variability.

ACKNOWLEDGMENT

This research was conducted as part of the Northern Gulf Coastal Hazards Collaboratory (NGCHC), a consortium that focuses on advancing the science and engineering of coastal hazards across the states of Alabama, Mississippi, and Louisiana. The research was partly supported by the NSF EPSCoR EPS-1010607 grant, 2013 USGS 104(G) Competitive Grants Program and the Alabama Water Resources Research Institute (Project 2013AL156B).

REFERENCES

[1] R.M. Hirsch, D.L. Moyer, and S.A. Archfield, "Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay River inputs," JAWRA Journal of the American Water Resources Association, vol 46, pp. 857–880, October 2010. doi: 10.1111/j.1752-1688.2010.00482.x

[2] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[3] R.M. Hirsch and L. De Cicco, dataRetrieval: Retrieval functions for USGS data. 2012. R package version 1.2.1.

[4] R.M. Hirsch and L. De Cicco, Exploration and Graphics for River Trends (EGRET): An R-package for the analysis of Long term Changes in Water Quality and Streamflow, Including the Water Qauality Method Weighted Regressions on Time, Discharge, and Season (WRTDS) Draft Manual. 2013. R package version 1.2.3.

[5] R.M. Hirsch, 2012. Flux of nitrogen, phosphorus, and suspended sediment from the Susquehanna River Basin to the Chesapeake Bay during Tropical Storm Lee, September 2011, as an indicator of the effects of reservoir sedimentation on water quality: U.S. Geological Survey Scientific Investigations Report 2012–5185, 17 p. http://pubs.usgs.gov/sir/2012/5185/

[6] Hirsch, R.M., D.L. Moyer, and S. Phillips, 2013. Determining Nutrient and Sediment Loads and Trends in the Chesapeake Bay Watershed by Using an Enhanced Statistical Technique. U.S. Geological Survey Science Summary (January 2013).

[7] Sprague, L.A., R.M. Hirsch, and B.R. Aulenbach, 2011. Nitrate in the Mississippi River and Its Tributaries, 1980 to 2008: Are We Making Progress? Environmental Science & Technology 45(17):7209-7216. DOI:10.1021/es201221s

[8] Maestre, A. Williamson, D. Ward, A. 2012. "Nutrient Fluxes in Rivers of the Mobile – Alabama River System using WRTDS". 2012 Alabama Water Resources Conference. Orange Beach, Alabama.

[9] Moyer, D.L., Hirsch, R.M., and Hyer, K.E., 2012, Comparison of two regression-based approaches for determining nutrient and sediment fluxes and trends in the Chesapeake Bay watershed: U.S. Geological Survey Scientific Investigations Report 2012-5244, 118 p. http://pubs.usgs.gov/sir/2012/5244

[10] Markus, M., Demissie, M., Short, M., Verma, S., and Cooke, R. 2013. A Sensitivity Analysis of Annual Nitrate Loads and the Corresponding Trends in the Lower Illinois River. Journal of Hydrologic Engineering. doi:10.1061/(ASCE)HE.1943-5584.0000831

[11] Murphy, J.C., R.M. Hirsch, and L.A. Sprague, 2013. Nitrate in the Mississippi River and its tributaries, 1980 - 2010: An update. U.S. Geological Survey Scientific Investigations Report 2013–5169, 31p. http://pubs.usgs.gov/sir/2013/5169

[12] Lewin, W.-C., Mehner, T., Ritterbusch, D., and Bramick, U. (2014). The Influence of Anthropogenic Shoreline Changes on the Littoral Abundance of Fish Speces in German Lowland Lakes Varying in Depth as Determined by Boosted Regression Tress. Hydrobiologia, 724(1), 2014: 293-306. doi: 10.1007/s10750-013-1746-8

[13] Chung, Y. S. (2013). Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. Accident Analysis & Prevention, 61, 107-118. doi: 10.1016/j.aap.2012.08.015

[14] Soykan, C. U., Eguchi, T., Kohin, S., & Dewar, H. (2014). Prediction of fishing effort distributions using boosted regression trees. Ecological Applications, 24(1), 71-83. doi: 10.1890/12-0826.1

[15] le Roux, P. C., Luoto, M. (2014), Earth surface processes drive the richness, composition and occurrence of plant species in an arctic–alpine environment. Journal of Vegetation Science, 25: 45–54. doi: 10.1111/jvs.12059

[16] Bergstra, J., Pinto, N., & Cox, D. (2012). Machine learning for predictive auto-tuning with boosted regression trees. In: Innovative Parallel Computing, p. 1-9. IEEE. doi:10.1109/InPar.2012.6339587

[17] De Ath, G. and Fabricius, K. (2000). Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis. Ecology, 81 (11), 2000: 3178–3192. doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

[18] De Ath, G. (2007). Boosted Trees for Ecological Modeling and Prediction. Ecology, 88 (1), 2007: 243–251. doi: /10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2

[19] Elith, J., Leathwick, J.R., Hastie, T. (2008). A Working Guide to Boosted Regression Trees. Journal of Animal Ecology, 77 (4), 2008: 802-813. doi: 10.1111/j.1365-2656.2008.01390.x

[20] Williams, G.P. (1989). Sediment Concentration versus Water Discharge During Single Hydrologic Events in Rivers. Journal of Hydrology, v. 111, p. 89 – 106

[21] Johnes, P.J. (1996). Evaluation and Management of the Impact of Land Use on the Nitrogen and Phosphorous Load Delivered to Surface Waters: the Export Coefficient Modelling Approach. Journal of Hydrology, 183: 323 – 349

[22] Fraley, C., Raftery, A., Murphy, T., Scrucca, L. (2012). MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597. Department of Statistics, University of Washington. Seattle, USA.

[23] Fraley, C. and Raftery, A. (1999). MCLUST: Software for Model-Based Cluster Analysis. Journal of Classification, 16: 297-306

[24] Bowers, M. The Distributions of Seasonal Rivers Flows: Lognormal or Power-Law? Master of Science Thesis Purdue University, West Lafayette, Indiana. May 2012

[25] Ridgeway G. with contributions from others (2013). gbm: Generalized Boosted Regression Models. R package version 2.1. http://CRAN.R-project.org/package=gbm

[26] Hastie, T., Tibshirani, R., Friendman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer Science+Business Media. New York, NY.

[27] Jun, S. (2013). Boosted Regression Trees and Random Forests. Statistical Consulting Report for Michael Melnychuck. University of British Columbia.

[28] McCuen, R., 2012. Hydrologic Modeling in 2050: Knowledge Requirements in a Multi-Nonstationary Environment. In: Toward a Sustainable Water Future Visions for 2050. American Society of Civil Engineers.