

EFFECTS OF NONPOINT SOURCE MARSH LOADING ON COMPLEX ESTUARIES

Edwin A. Roehl, Jr.
Advanced Data Mining Intl
Greenville, SC
ed.roehl@advdmi.com
John B. Cook, PE
Advanced Data Mining Intl
Greenville, SC
john.cook@advdmi.com

ABSTRACT

A foundational element in determining Total Maximum Daily Loads (TMDLs) is the determination of the relative impacts of point and nonpoint source impacts on the dissolved oxygen concentration (DO) of an impaired stream. It is not uncommon that the ultimate oxygen demand during a rain event can be greater than the effects of the fully permitted point source loading to the stream. Traditionally, the loading of oxygen consuming constituents is often estimated with a watershed model which is then coupled with a mechanistic model of the receiving stream. Calibrating coastal system applications of watershed models and mechanistic models to match the behavioral variability observed in actual field data is particularly difficult due to low watershed gradients, poorly defined drainage areas, chaotic forcing functions, and insufficient understanding of watershed and marsh process physics and chemistry.

Data mining offers an alternative approach for analyzing and modeling tidal DO signals to quantify their responses to point and nonpoint source loadings. Data mining can be used to extract DO signal components whose forcing is by nonpoint loading caused by rainfall events and tidally entrained organics from marshes and mudflats. It has been successfully applied by the authors to South Carolina's Cooper, Pee Dee, Savannah and Beaufort Rivers. At one gaging station on the Cooper River, rainfall was found to decrease DO concentrations at a rate of approximately 0.25 milligrams per liter (mg/L) per inch of rainfall. Similarly, it was found that specific conductance, water level, and tidal range, which indicate tidal forcing, modulate DO in the range of 3.1 mg/L. As confirmation of the process approach, rainfall impacts on the Beaufort River were found to be as high as 0.8 mg/L, whereas point source loads had an effect as high as 0.4 mg/L. This paper examines the approach of using data mining results to improve and expedite the calibration of mechanistic models for coastal applications and more accurately quantify the effects of nonpoint source load estimates.

INTRODUCTION

Much research and development over the last 20 years has focused on improving mechanistic watershed models and their coupling with riverine and estuarine models. Applications in coastal areas have remained particularly difficult due to low watershed gradients, poorly defined drainage areas, tidal complexities, and a lack of understanding of watershed and marsh processes. Nonetheless, ideally, good water-resource management requires an accurate accounting of the effect that nonpoint sources are having on impaired waters.

Many coastal streams along the Southeast coast are naturally low in DO due to loading of organic material from tidal marshes and mudflats. With changes in land-use in coastal areas, the

contributions of anthropogenic loads are compounding the natural loading to receiving streams during rainfall events. Often, the ultimate oxygen demand of a load pulse during an event can be greater than the fully permitted point source loading.

Since the beginning of water quality modeling, mechanistic models have been the state of the practice for regulatory evaluations of point source and nonpoint source impacts. However, the emergence of advanced data mining technologies has created new opportunities to develop models that are built directly from extracting the information contained in the data, and, the authors have found, represent natural system behavior much more accurately in estuarine settings. Data mining is defined as “the extraction of information from massive databases” (Weiss and Indurkha, 1998). It is comprised of several important technologies that include signal processing, multivariate statistics, multi-dimensional visualization, Chaos Theory, Information Theory and machine learning from the field of Artificial Intelligence (AI).

Previous studies by the authors and others have employed data mining, including a form of machine learning called artificial neural networks (ANN), to predict hydrodynamic and water-quality behaviors in the Cooper, Pee Dee, Beaufort and Savannah River estuaries (Conrads and Roehl, 1999; Roehl and others, 2000; Conrads and others, 2002a; Conrads and others, 2002b; Conrads and others, 2003; Daamen and others, 2005), and stream temperatures in western Oregon (Risley and others, 2002). These studies have demonstrated excellent performance by data mining in predicting water level (WL), dissolved oxygen concentration (DO), and specific conductance (SC). Data mining has also been used to assess the impacts of reservoir releases and point and nonpoint sources on receiving streams.

The concurrent growth of coastal communities and stormwater regulation and permitting has created an immediate need to better account for nonpoint source loading from coastal watersheds, and to better assess the responses of receiving streams. This paper describes how data mining has been applied to quantify the effects on a complex estuary of nonpoint source loading on DO and confirmed by application to a second estuary. It also presents an approach to integrating results from data mining to enhance the calibration and accuracy of mechanistic water quality models of receiving streams.

ESTIMATING NONPOINT SOURCE IMPACTS USING DATA MINING

Typically, nonpoint source impacts on DO are estimated by inputting a computed watershed load of oxygen consuming constituents into a mechanistic model of the receiving stream. Figure 1 shows a simplified conceptual model of the watershed process. In some studies, there is a large amount of high quality time series data for rainfall, DO, and other parameters. However, due to sampling logistics and analytical costs, water chemistry data are often inadequate for calculating watershed loads or calibrating and confirming watershed loading models. In contrast, data mining assumes that that a system’s physics and chemistry are manifest as information in its behavior, which is described by its response time-series. Rather than computing load inputs to the system, data mining quantifies the sensitivity of DO response to variables such as meteorology, tides,

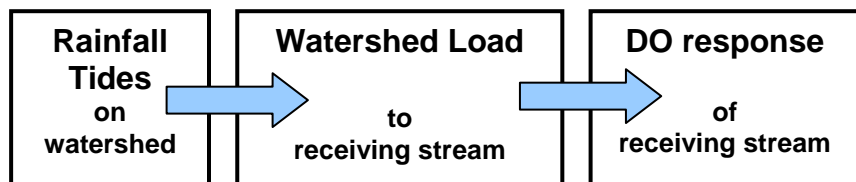


Figure 1. Simplified conceptual model of watershed process for non-point source loading to a receiving stream.

and point source loading.

Data mining has been used to estimate rainfall and tidal effects on DO in South Carolina’s Cooper and Beaufort River estuaries. Both are highly complex tidal systems with semi-diurnal tides exceeding 6 feet, expanses of tidal marshes, and multiple point source dischargers. Cooper River stream flows are affected by releases from a hydroelectric plant, and the Beaufort River is affected by connections to large tidal bays.

Both studies used all available point source and rainfall data in combination with multi-year time series of hydrodynamic and water-quality parameters, such as DO, WL, SC, flow (Q) and water temperature (WT) at multiple gaging stations. Calculated variables included tidal range (XWL) and DO deficit (DOD), which is DO normalized for WT and salinity (U.S. Geological Survey, 1981). The effect on DO of decaying organics transpires over several days. This effect can be difficult to discern when coupled with high frequency diurnal and semi-diurnal influence caused by WT and tides. Therefore, the hourly time series were spectrally filtered using fast Fourier transforms (Press and others, 1993) to segregate diurnal and semi-diurnal signal components. Rainfall data was collected from multiple National Weather Service and U.S. Geological Survey (USGS) meteorological stations in each watershed. These were spatially and temporally averaged to diminish the “spotty” nature of rainfall by averaging stations together and then applying a 2-day moving window average (MWA). This calculated rainfall is referred to as an average of an average rainfall, or RAINAA.

Sensitivity analysis quantifies the relationships between a response variable of interest and an explanatory variable (for example, DO is known to be dependant on WT and rainfall). Computing sensitivities requires defining the relationships between variables through modeling. Empirical modeling adapts generalized mathematical functions to optimally fit a line or surface through data from two or more variables. The modeling approach used here employed “multi-layer perceptron” (MLP) ANNs (Hinton, 1992) trained using back-propagation and conjugate gradient training algorithms (Rummelhart and others, 1986). MLP ANNs can synthesize functions to fit high-dimension, non-linear multivariate data, and have been used extensively in numerous industrial and environmental applications (Jensen, 1994; Devine and others, 2003).

MLP ANNs (hereafter ANNs) were used to “learn” how DOD at each station is affected by RAINAA and WT; indicators of tidal forcing WL, XWL, and SC; and point source discharges. In the Cooper River study (Conrads and others, 2002a), the sensitivity of DOD to RAINAA was quantified. In the Beaufort River study (Conrads and others, 2003), rainfall impacts were simulated for a three year period and compared to point source dischargers. These studies are summarized below.

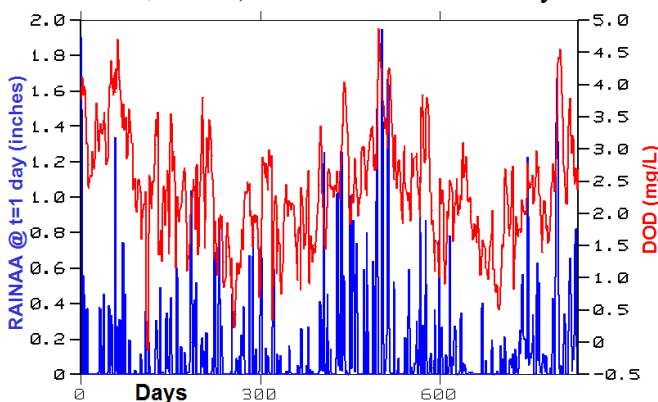


Figure 2: Average rainfall (RAINAA) with one day lag ($\tau=1$) and dissolved-oxygen deficit (DOD) versus time in Days (from Conrads and others, 2002a).

Effects of Rainfall on the Cooper River Estuary

The variability of DO in the Cooper River is a result of many factors including the quality of the water from Lake Moultrie and Charleston Harbor; loading of oxygen-consuming matter from tidal marshes, abandoned rice fields, and other nonpoint sources; point source effluent; and physical characteristics of streamflow, tidal range, salinity, and temperature. Figure 2 compares RAINAA and DOD at the station 02172037 located in a relatively pristine reach of the East Branch of the Cooper River. These parameters vary seasonally and large rainfall events coincide with spiking DOD. As shown in figure 3, an ANN with inputs RAINAA at multiple time delays (τ) predicts a significant fraction of the DOD variability ($R^2 = 0.28$). This suggests that of all the factors affecting DO at 02172037, rainfall accounts for approximately 28 percent of the variability.

ANNs can be used to generate response surfaces that show the interaction of two explanatory variables on a response variable. Figure 4 shows the ANN model's predicted response for DOD versus RAINAA at $\tau = 1$ and 3 days. Also shown are the actual data projected onto the surface. Note that the surface is nearly linear, and that the sensitivity of DOD to RAINAA at $\tau = 1$ is less than at $\tau = 3$ days.

The overall impact of rainfall can be estimated from figure 4 as follows. The total increase in DOD ≈ 2 mg/L (approximately 1.9 to 3.9 mg/L on the Z-axis). This occurs when the RAINAA for $\tau = 1$ and 3 days are both ≈ 2 inches of rain (simply, ≈ 0.8 mg/L at $\tau = 1$ plus ≈ 1.2 mg/L at $\tau = 3$ days). Because RAINAA is a 2-day MWA, a value of RAINAA = 2.0 inches is equivalent to 4 inches of rainfall over 2 days, or 8 inches over 4 days. The sensitivity of DOD to rainfall can be characterized as $\Delta DOD/inch \approx 2$ mg/L / 8 inches of rainfall over 2 days, or, $= 0.25$ mg/L per inch of rainfall.

The effect of tidally-entrained organics from marshes and mudflats can also be analyzed. The response surfaces shown in figure 5 were generated by the same model. In addition to the shown inputs XWL' and SC' , the model also had inputs for "unshown" variables RAINAA, WL' , and WT' ¹. Low and high WL' values were used to generate the left and middle surfaces respectively, while other unshown inputs were held constant. Below, the data used to train the model has been projected onto the low WL' surface. The range of DOD varies from 1.4 to 4.5 mg/L at low WL' ($\Delta DOD = 3.1$), and from 1.1 to 2.8 mg/L at high WL' ($\Delta DOD = 1.7$). DOD varies little on the rare occasions when SC' exceeds

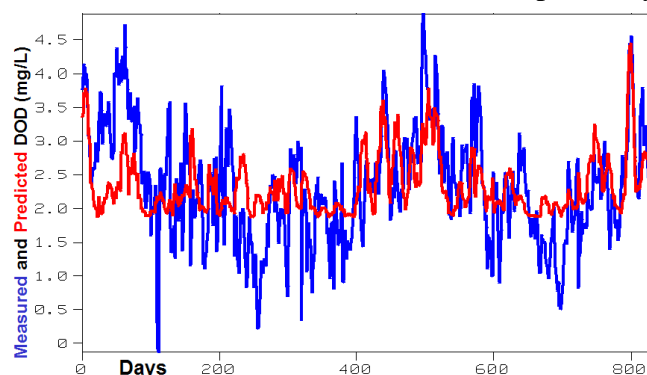


Figure 3: Measured and ANN prediction of DOD (from

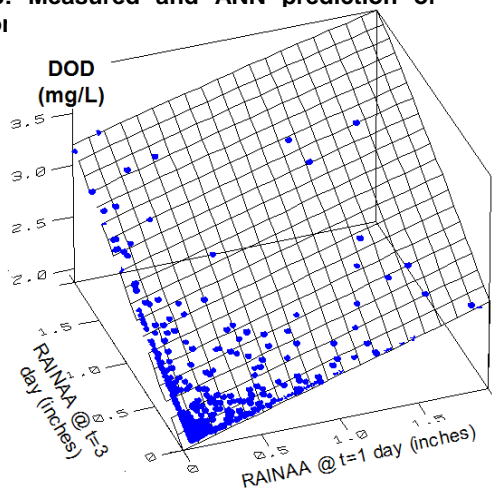


Figure 4: DOD versus RAINAA @ $\tau = 1$ and 3 days (from Conrads and others, 2002a).

¹ Primes denote decorrelated variables. The decorrelation method used is described by Conrads and others, 2003. Decorrelation can shift a variable's scale to (+/-) values.

50 μ -siemens/cm. Strictly speaking, determining sensitivities from non-linear, multivariate functions involves computing partial derivatives; however, response surfaces can also be visually inspected to provide some idea of what the ANN has learned about a process' physics. At low SC', ranging XWL' varies DOD from 3.0 to 4.5 mg/L at low WL (Δ DOD=1.5), and from 1.5 to 2.8 mg/L at high WL (Δ DOD =1.3). At low SC', and XWL', DOD falls from 4.5 to 2.8 mg/L (Δ DOD=1.7) when WL is increased. At low SC', and high XWL', DOD falls from 3.0 to 1.5 mg/L (Δ DOD=1.5) when WL is increased. The model clearly indicates that DOD is highest when XWL and WL are low, and falls as they are increased, bringing dilution and higher quality water. This is not surprising, but it is reassuring to see that the ANN's functional relationships follow expectations and are now quantified. This assessment also indicates that drawing generalizations about sensitivities among non-linear, multivariate relationships is inherently problematic, even though people tend to seek such paradigms.

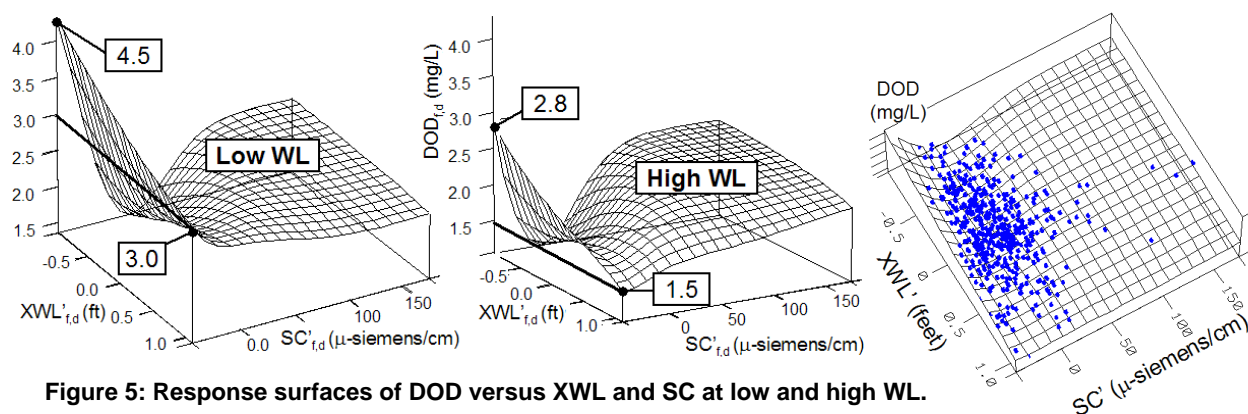


Figure 5: Response surfaces of DOD versus XWL and SC at low and high WL.

Simulation of Rainfall Effects on the Beaufort River Estuary

The Beaufort River study (Conrads and others, 2003) expanded the approach taken on the Cooper River study by developing a simulator in MS ExcelTM that integrates a large historical database, 118 ANN models, graphics, and a graphical user-interface (GUI). The simulator was used to determine permit limits for all three wastewater treatment facilities which discharge to the Beaufort River. The simulator predicts the effect of rainfall on DOD at seven gages on the river and its tributaries. Rainfall and point source inputs to the simulator can be modified as a percent (0 to 150%) of the historical values in the database. To evaluate the impact of rainfall on DO, the simulator was run setting the rainfall inputs to zero (and point source loads to the actual condition) and comparing the results to simulations using actual rainfall. Figure 6 shows results for stations on the river's two tributaries, Battery (2176635) and Albergottie (2176587) Creeks, and two stations on the Beaufort River (2176603 and 2176589).

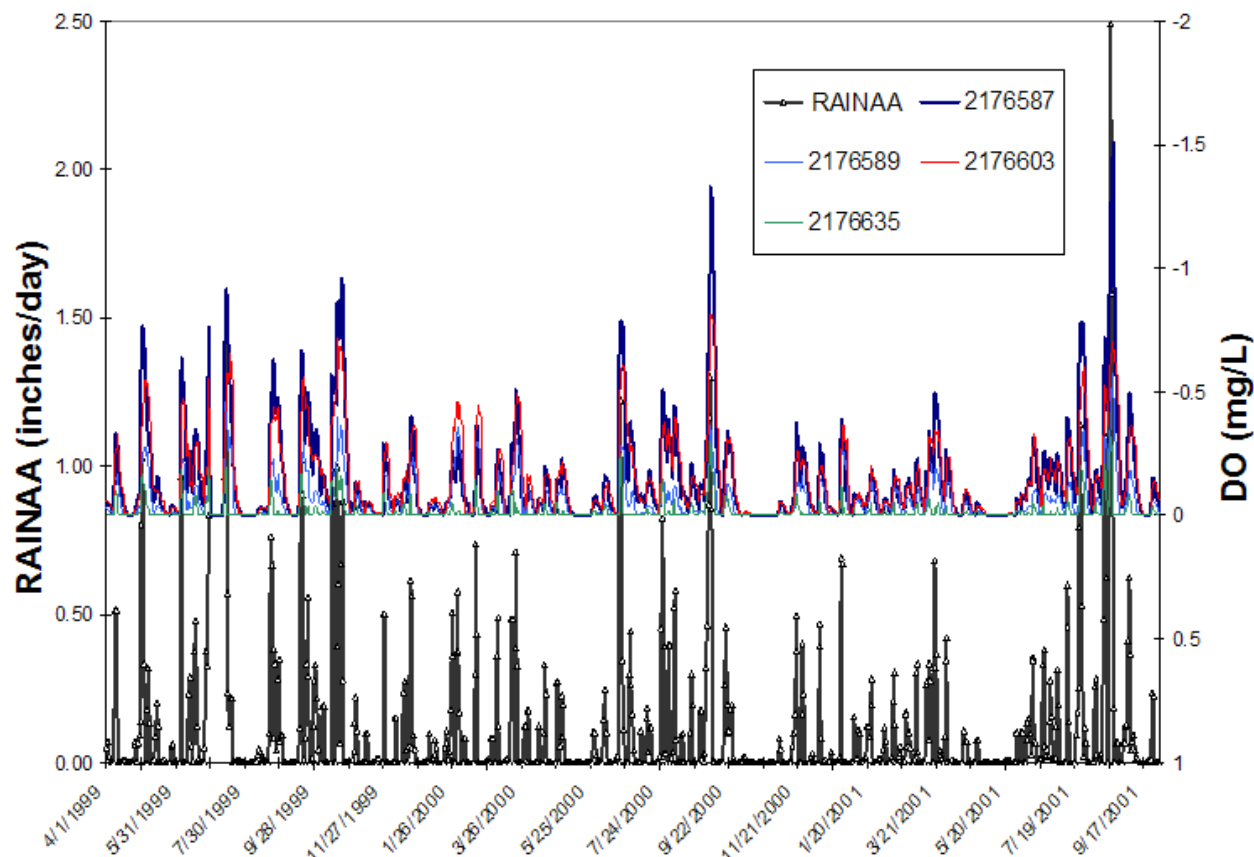


Figure 6. Predicted DO impact and RAINAA at four Beaufort River USGS stations. (From Conrads and others, 2003.)

The largest impact is in Albergottie Creek, where rainfall decreases DO by as much as 1.5 mg/L. The smallest impact is on Battery Creek where the DOD concentration increases by less than 0.5 mg/L. Although the riparian tidal marsh of Battery Creek drains the western side of the City of Beaufort, the low observed impact may result from the station's location on the lower reach of the creek where the channel is large and tidal exchange with the lower Beaufort River is also large. Of the two river stations, 2176603 had the larger decrease in DO; probably because 2176589 is connected to a high quality water source.

Simulations were run to compare rainfall impacts on DO to those of point source loads. This involved comparing a run using the actual historical point source loads with a run using a no-load condition. The results for station 2176603 are shown in figure 7. Point source loads decreased DO by as much as 0.4 mg/L. Rainfall decreased the DO concentrations by as much as 0.8 mg/L. Not surprisingly, the behaviors of the two types of loading are quite different. The point source impact is quite variable and sustained throughout varying hydrologic conditions; consistent with point source loading that varies but is never zero. Rainfall causes intermittent pulse loads whose impacts are often higher than those of the point sources, but are transient and disappear during periods of no rainfall. Although the maximum rainfall impact is twice that of all

the point sources, their average impacts were approximately equal over the simulation period.

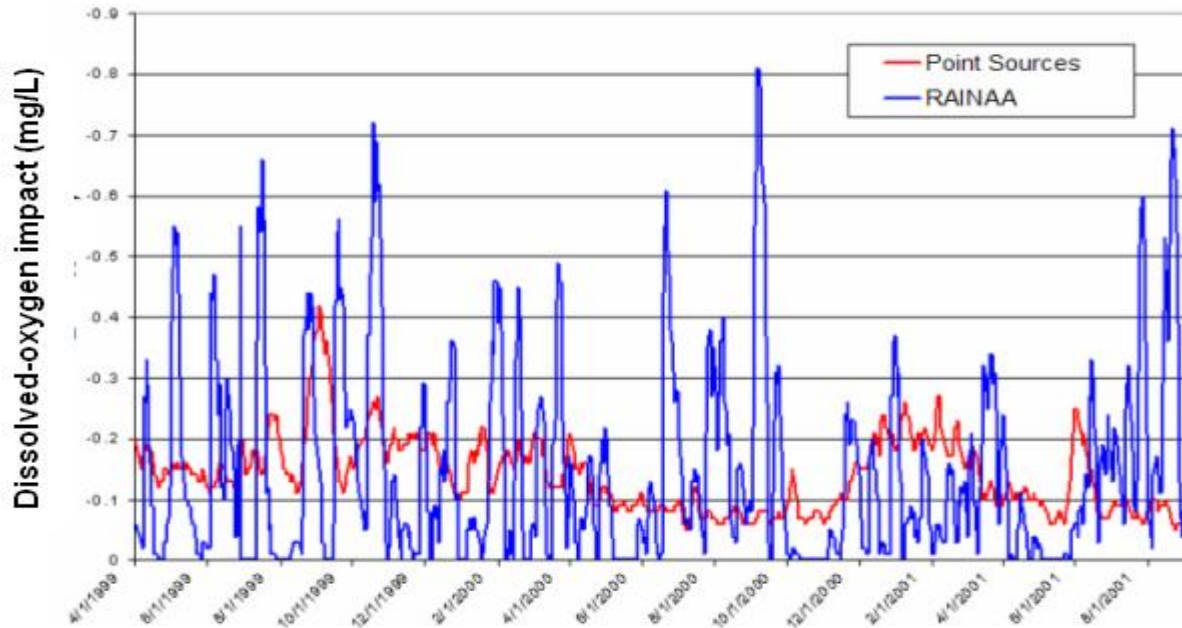


Figure 7. Predicted RAINAA and point-source load impacts on DO station 2176603. (From Conrads and others, 2003.)

INTEGRATING DATA MINING WITH MECHANISTIC MODELS

A critical element of TMDLs is determining the nonpoint source impacts on impaired streams. As seen in the above two examples, nonpoint sources contribute significantly to the low DO concentrations in the Cooper and Beaufort Rivers. Accounting for nonpoint source loads is often problematic in the application of mechanistic models of riverine and estuarine systems. As mentioned previously, there usually is inadequate data to quantify nonpoint source loadings. When there is adequate data, the loading equations typically used (such as Universal Soil Loss Equation or Modified Universal Soil Loss Equation) can significantly over- or under-estimate watershed loads.

Data mining provides an approach for calibrating and validating watershed load inputs to mechanistic models. Rather than estimating watershed loads, the above studies separated out the impacts of the nonpoint source loads in the DO time-series. Instead of computing loads, the response of the system to rainfall and tidal events was determined. In the Cooper River study, the sensitivity of DO to rainfall events was estimated as the change in DO per inch of rainfall. The sensitivity of DO to tidal forcing was estimated by varying indicator variables WL, XWL, and SC. In the Beaufort River study, a time-series of predicted rainfall impacts was computed using three years of changing meteorological and tidal conditions.

Integration of data mining with mechanistic models would be straightforward. In the conceptual model shown in figure 1, the watershed load is the largest source of potential error. Data mining can be used to quantify the DO response of a system as a calibration time-series for the mechanistic model. To calibrate or confirm nonpoint watershed loads, the mechanistic model of the receiving stream would simulate a period with and without the nonpoint source watershed

load inputs. The difference between the DO time-series using the two simulations is the DO response using the computed watershed loads. The DO response using the computed watershed load is then compared with the response quantified by data mining, that is, the rainfall response time-series shown in figure 6 (blue trace). If the two response time series compare favorably, the data mining analysis confirms the computed watershed loads. If the watershed response is above or below the response computed by data mining, the watershed load is either over- or underestimated and would be adjusted.

SUMMARY

Coastal streams are naturally low in DO. Given continuation of current demographic trends, the stress to these systems will continue to increase. Better estimates of watershed loading and the impact of these loads need to be incorporated into the TMDL process in order to effectively manage coastal waters. Two studies were described that used data mining to quantify the DO response to nonpoint source loading in complex estuaries. The studies presented a novel approach to analyzing nonpoint source loading by modeling DO responses in long-term time series. This approach can be integrated with traditional mechanistic water quality models to significantly improve their accuracy and reduce development time. Using data mining technology, information can be extracted from data to provide an excellent means to understand highly complex and interacting behaviors in estuaries.

REFERENCES

- Conrads, P.A., Roehl, E.A., and Martello, W.P., 2003, "Development of an Empirical Model of a Complex, Tidally Affected River Using Artificial Neural Networks," Water Environment Federation TMDL 2003 Specialty Conference, Chicago, Illinois, November 2003
- Conrads, P.A., Roehl, E.A., and Cook, J.B., 2002a, "Estimation of Tidal Marsh Loading Effects in a Complex Estuary," American Water Resources Association Annual Conference, New Orleans, May 2002.
- Conrads, P.A.; Roehl, E.A., and Martello, W. P., 2002b, "Estimating Point-Source Impacts on the Beaufort River Using Neural Network Models," American Water Resources Association Annual Conference, New Orleans, May 2002.
- Conrads, P.A. and Roehl, E.A., 1999, "Comparing physics-based and neural network models for predicting salinity, water temperature, and dissolved oxygen concentration in a complex tidally affected river basin," South Carolina Environmental Conference, Myrtle Beach, March 15-16, 1999.
- Daamen, R.C., and Roehl, E.A., 2005, "Integrating Multiple Databases and Estuary Models Into A Comprehensive Software Tool for Regulatory Support," South Carolina Environmental Conference, Myrtle Beach, March 15-16, 2005.
- Devine, T.W., Roehl, E.A., and Busby, J.B., 2003, "Virtual Sensors - Cost Effective Monitoring," Air and Waste Management Association Annual Conference, June 2003
- Hinton, G.E., 1992, "How Neural Networks Learn from Experience," *Scientific American*, September 1992, p.145-151.
- Jensen, B.A., 1994, *Expert Systems - Neural Networks, Instrument Engineers' Handbook Third Edition*, Chilton, Radnor PA.
- Press, William H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., 1993, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1993.

- Risley, J.C., Roehl, E.A, and Conrads, P.A., 2002, "Estimating Water Temperatures in Small Streams in Western Oregon Using Neural Network Models,' U.S. Geological Survey Water Resources Investigations Report 02-4218, 2002.
- Roehl, E.A., Conrads, P.A., Roehl, T.A.S., 2000, "Real-Time Control of the Salt Front in a Complex Tidally Affected River Basin," *Smart Engineering System Design: Volume 10, Proceedings of the Artificial Neural Networks In Engineering Conference*, ASME Press, New York, pp. 947-954.
- Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, 318-362, Cambridge, MA: The MIT Press.
- U.S. Geological Survey, 1981, Technical Memorandum 81.11, Reston, Va. 1981.
- Weiss, S.M. and Indurkha, N., 1998, *Predictive Data Mining – A Practical Guide*, Morgan Kaufmann Publishers, Inc., San Francisco, p. 1.