Data Mining for Water Resource Management

Part 1 – Answering Contemporary Questions with Historical Databases

Paul A. Conrads[1] and Edwin A. Roehl, Jr.[2]

AUTHORS: [1]Surface Water Specialist, U.S. Geological Survey South Carolina Water Science Center, Gracern Rd, Suite 129, Columbia, SC 29210; 803-750-6141; pconrads@usgs.gov, [2]Chief Technical Officer, Advanced Data Mining Intl, 3620 Pelham Rd., PMB 351, Greenville, SC 29615; 864-201-8679; ed.roehl@advdmi.com;

REFERENCE: *Proceedings of the 2010 South Carolina Water Resources Conference,* held October 13-14, 2010, at the Columbia Metropolitan Convention Center.

Abstract. This paper, part 1 of a 2-part paper, gives a broad overview of data mining and describes several applications where it has been applied. Data mining is an emerging field that addresses the issue of converting large databases into knowledge. Data mining methods come from different technical fields such as signal processing, statistics, and artificial intelligence. Data mining employs methods for maximizing the information content of data, determining which variables have the strongest relations to problems of interest, and developing models that predict future outcomes. This paper describes the results of three case studies where data mining, including artificial neural network models, has been applied to large-scale environmental issues in South Carolina and Georgia. For the Beaufort River, South Carolina, dissolved-oxygen models were developed and used for determining Total Maximum Daily Load of allowable point-source effluent loading to the Beaufort River. For the Savannah River estuary in Georgia, models were developed to simulate pore-water salinity and used to determine the potential impacts of deepening the Savannah Harbor on upstream freshwater tidal marshes. For the Pee Dee River in South Carolina, models were developed to estimate the minimum streamflow required to protect municipal intakes from seawater inundation along the Grand Strand of South Carolina. For the Congaree River, the effects of the flow release from the Lake Murray Dam on the river and groundwater level at Congaree National Park were analyzed. In the four studies, the models were able to convincingly reproduce historical behaviors and generate alternative scenarios of interest. To make the results of the studies directly available to all stakeholders, user-friendly decision support systems were developed as a spreadsheet application that integrates the historical database, models, user controls, streaming graphics, and simulation output.

Part 2 of this paper describes technical approaches and methods, and how they are used to solve different types of water resource problems.

INTRODUCTION

Natural-resource managers face the difficult problem of controlling the interactions between hydrologic and man-made systems in ways that preserve resources while optimally meeting the needs of disparate stakeholders. Finding success depends on obtaining and employing detailed scientific knowledge about the cause-effect relations that govern the physics of these hydrologic systems. This knowledge is most credible when derived from large field-based datasets that encompass the wide range of variability in the parameters of interest. The means of converting data into knowledge of the hydrologic system often involves developing computer models that predict the consequences of alternative management practices to guide resource managers towards the best path forward. Complex hydrologic systems are typically modeled using computer programs that implement traditional, generalized, physical equations, which are calibrated to match the field data as closely as possible. This type of model commonly is limited in terms of demonstrable predictive accuracy, development time, and cost.

The science of data mining presents a powerful complement to physics-based models. It provides a structure to convert large databases into knowledge, and is uniquely able to accommodate the high variability exhibited by the real-time, multivariate data typically collected for a hydrologic system. In a number of side-by-side comparisons to state-of-the-art physics-based models, Data-Mining solutions have been significantly more accurate, less time consuming to develop, and embeddable into spreadsheets and sophisticated Decision Support Systems, making them easy to use by broad communities of stakeholders and regulators (Conrads and Roehl, 1999).

An important part of the U.S. Geological Survey (USGS) mission is to provide scientific information for the effective water-resources management of the Nation. To assess the quantity and quality of the Nation's surface water, the USGS collects hydrologic and water-quality data from rivers, lakes, and estuaries by using standardized methods, and maintains the data from these stations in a national database. New technologies in environmental monitoring have made it cost effective to acquire tremendous amounts of hydrologic and water-quality data. Although these data are a valuable resource for understanding environmental systems, often these databases are under-utilized and not well interpreted for addressing contemporary hydrologic issues. The four data-mining applications presented in this paper (part 1 of 2) demonstrate how valuable information from the USGS database can be used to assist local, State, and Federal agencies to address contemporary water-resource management issues. The second paper of describes technical approaches and methods, and how they are used to solve different types of water resource problems.

BEAUFORT RIVER DISSOLVED-OXYGEN MODEL

The Beaufort River is a complex estuarine river system that supports a variety of uses including shellfish grounds, fisheries nursery habitats, shipping access to Port Royal, receiving waters for wastewater effluent, and an 18-mile reach of the Intracoastal Waterway. From 1996-2008, the river was on the Section 303(d) list of impaired waters of South Carolina for low dissolved-oxygen concentrations. The Clean Water Act stipulates that a Total Maximum Daily Load must be determined for impaired waters. An empirical model was developed to simulate the impact of point-source discharges and rainfall on dissolved-oxygen concentrations in the Beaufort River (Conrads and others, 2003). The model used water level, specific conductance, temperature, and dissolved-oxygen data collected at 15-minute intervals from seven real-time gaging stations and effluent point-source data collected on a weekly basis for a 33-month simulation period.

The empirical model was developed using data-mining techniques, including artificial neural network (ANN) models, to quantify the relations between the time series of three wastewater point-source discharges and the dissolved-oxygen concentrations recorded at seven real-time gages distributed throughout the system. The data mining produced a water-quality model that can predict the impacts that point and non-point source loads have on the dissolved-oxygen concentration throughout the river system. The analysis included environmental factors such as tides, specific conductance, water temperature, and rainfall. The model is comprised of numerous sub-models based on ANN models.

The data analyses and model provided unique ways to evaluate complex tidal dissolved-oxygen effects from point-source discharges and rainfall. The model executes non-iteratively, making it amenable to long-term simulation runs of 33 months. The model also included a non-linear, constraint-based numerical optimizer to determine the maximum allowable daily effluent loading without violating the State's water-quality standard. Insights were garnered from this technical approach that leveraged the full historical record in which assimilative capacity was found to be constantly changing. For example, critical conditions for effluent impacts on dissolved-oxygen concentrations occur during neap tides due to the streamflow characteristics and limited flushing of the system. The predictive model allowed for a variety of wastewater treatment plant operating scenarios and regulatory options to be evaluated quickly. Several 33-month time series of daily loadings were simulated utilizing the predictive model. Frequency distributions of the allowable loading were subsequently generated from the time series of optimal loading. Water- resource managers can use the frequency distribution to help predict the percentage of time water-quality standards may be violated. Model dissemination is facilitated by incorporating the ANN sub-models and point-source optimizers into an Excel spreadsheet application.

Prior to this study, it was understood that the net streamflow of the Beaufort River was to the south (Conrads and others, 2003). Analysis of the tidal streamflow determined that the net streamflow of the system was to the north. The tidal flow dynamics were confirmed by the long residence times seen in the specific conductance data and the correlation analysis between the waste-water effluents and dissolved-oxygen response. The knowledge of the direction of net streamflow of the system had far ranging consequences from determining critical dissolved-oxygen conditions to calculating the assimilative capacity (the amount of effluent that can be discharged without violating the State water-quality standard) of the system.

SAVANNAH RIVER MODEL-TO-MARSH

The Savannah Harbor is one of the busiest ports on the East Coast of the USA and is located just downstream from one of the largest freshwater tidal marshes of the Savannah National Wildlife Refuge (SNWR). To evaluate the environmental impacts of a potential deepening of the Savannah Harbor, a three-dimensional (3D) hydrodynamic model (Tetra Tech, 2005) and a marsh succession model (MSM; Welch and Kitchens, 2007) were developed for the system. The 3D model predicts changes in water levels and salinity in the system in response to potential harbor geometry changes. The MSM predicts plant distribution in the tidal marshes in response to changes in the water-level and salinity conditions in the marsh.

The Model-to-Marsh Decision Support System (M2MDSS; Conrads and others, 2006) was developed using data-mining techniques to integrate the riverine predictions from the 3D model to the MSM. Artificial neural network models used in the M2MDSS were developed to simulate riverine and marsh water levels and salinity in the vicinity of the SNWR for the full range of historical conditions using over 11 years of data from the riverine and marsh gaging networks.

The M2MDSS integrates the 3D hydrodynamic and marsh-succession modeling studies by combining the riverine and marsh databases, the riverine and marsh ANN models, model control, 3D hydrodynamic model input, and simulation output into a spreadsheet application that is readily disseminated. The application allows users to utilize the predicted water-level and salinity impacts from the 3D model as input to the M2MDSS to predict the impacts in the tidal marshes. Output from the application to be used as input to the marsh succession model is a grid of the marsh system (either 10 or 100 meter). The salinity values and grid parameters, for example cell size and corner coordinates, can be exported as an ASCII file for input into other mapping packages.

WACCAMAW RIVER AND ATLANTIC INTRACOASTAL WATER SALINITY-INTRUSION MODEL

Six reservoirs in North Carolina discharge into the Pee Dee River, which flows 160 miles through South Carolina to the coastal communities near Myrtle Beach, South Carolina. During the Southeast's record-breaking drought from 1998 to 2003, salinity intrusions overwhelmed a coastal municipal freshwater intake, limiting water supplies. To evaluate the effects of regulated flows of the Pee Dee River on salinity intrusion in the Waccamaw River and Atlantic Intracoastal Waterway, the South Carolina Department of Natural Resources and a consortium of stakeholders entered into a cooperative agreement with the USGS to apply data-mining techniques to the long-term time series to analyze and simulate salinity dynamics near the freshwater intakes along the Grand Strand of South Carolina (Conrads and Roehl, 2007). Salinity intrusion in tidal rivers results from the interaction of three principal forces - streamflow, mean tidal water levels, and tidal range. To analyze, model, and simulate hydrodynamic behaviors at critical coastal gages, data-mining techniques were used to evaluate over 20 years of hourly streamflow, coastal water-quality, and water-level data. Artificial neural network models were trained to learn the variable interactions that cause salinity intrusions. Streamflow data from the 18,300-square- mile basin were input to the model as time-delayed variables and accumulated tributary inflows. Tidal inputs to the models were obtained by decomposing tidal water-level data into a "periodic" signal of tidal range and a "chaotic" signal of mean water levels. The ANN models convincingly reproduced historical behaviors and generated alternative scenarios of interest.

To make the models directly available to all stakeholders along the Pee Dee and Waccamaw Rivers and Atlantic Intracoastal Waterway, an easy-to-use decision support system (DSS) was developed as a spreadsheet application that integrates the historical database, ANN models, model controls, streaming graphics, and model output. An additional feature includes a built-in optimizer that dynamically calculates the amount of flow needed to suppress salinity intrusions as tidal ranges and water levels vary over days and months. This DSS greatly reduced the number of long-term simulations needed for stakeholders to estimate the minimum flow required to adequately protect the freshwater intakes.

CONGAREE RIVER STAGE AND GROUNDWATER MODEL

The Congaree National Park was established "…to preserve and protect for the education, inspiration, and enjoyment of present and future generations an outstanding example of a near-virgin, southern hardwood forest situated in the Congaree River flood plain in Richland County, South Carolina" (Public Law 94-545). The resource managers at Congaree National Park are concerned about the timing, frequency, magnitude, and duration of flood-plain inundation of the Congaree River. The dynamics of the Congaree River directly affect ground-water levels in the flood plain, and the delivery of sediments and nutrients is constrained by the duration, extent, and frequency of flooding from the Congaree River. The Congaree River is the southern boundary of the Congaree National Park and is formed by the convergence of the Saluda and Broad Rivers 24 river miles upstream from the park. The streamflow of the Saluda River has been regulated since 1929 by the operation of the Saluda Dam at Lake Murray. The USGS, in cooperation with the National Park Service, Congaree National Park, studied the interaction between surface water in the Congaree River and ground water in the flood plain to determine the effect Saluda Dam operations have on water levels in the Congaree National Park flood plain (Conrads and others, 2008).

Analysis of peak flows showed the reduction in peak flows after the construction of Lake Murray was more a result of climate variability and the absence of large floods after 1930 than the operation of the Lake Murray dam. Dam operations reduced the recurrence interval of the 2-year to 100-year peak flows by 6.1 to 17.6 percent, respectively. Analysis of the daily gage height of the Congaree River showed that the dam has had the effect of lowering high gage heights (95th percentile) in the first half of the year (December to May) and raising low gage heights (5th percentile) in the second half of the year (June to November). The dam has also had the effect of increasing the 1-, 3-, 7-, 30-, and 90-day minimum gage heights by as much as 23.9 percent and decreasing the 1-, 3-, 7-, 30-, and 90-day maximum gage heights by as much as 7.2 percent.

Analysis of the ground-water elevations in the Congaree National Park flood plain shows similar results as the gage-height analysis – the dam has had the effect of lowering high ground-water elevations and increasing low ground-water elevations. Overall, the operation of the dam has had a greater effect on the gage heights within the river banks than gage heights in the flood plain. This result may have a greater effect on the subsurface water levels of the surficial flood-plain aquifer than the frequency and magnitude of inundation of the flood plain.

DISCUSSION

The four case studies presented demonstrate how data-mining techniques can be applied to existing environmental databases to address concerns of long-term consequences. In each case, data were transformed into information, and ultimately, into knowledge. In the Beaufort River study, knowledge of the net flow to the north changed the understanding of the system and had long-term consequences for water-resource management of the river. The construction of the multi-million dollar water reclamation facility will meet the wastewater needs of the community for the next few decades. The Beaufort River was removed from the Section 303d list for dissolved-oxygen impairment.

In the Lower Savannah River estuary study, data mining was used to address various aspects of the Savannah Harbor Deepening Project. Data mining was used to develop models that estimate marsh pore-water salinity response to changing estuarine conditions, to integrate databases, and to integrate a hydrodynamic river model and ecologic marsh-secession models. By integrating the databases and models of various research groups, the M2M integrates the knowledge of river hydrologists and ecologists.

In the Pee Dee River study, data mining was used to understand the interaction between streamflow, tidal range, and mean tidal water level on salinity intrusion. With this understanding, stakeholders determined minimum streamflow needed to protect the intakes for a large range of hydrologic conditions but also realized that during extreme hydrologic conditions, the municipalities should have contingency plans to protect their intakes rather than required unrealistic flows from the reservoirs. The minimum streamflow will be used in the issuance of a 50-year permit by the Federal Energy Regulatory Commission for the operation of the North Carolina reservoirs.

Techniques and results from the Congaree River study show how long-term historical data bases can be utilized to answer contemporary questions of hydrologic concerns. Disparate data bases from collection periods of the 1920s and 1990s were integrated using data-mining techniques and the analysis provides valuable knowledge for sustainable water-resource management.

LITERATURE CITED

Conrads, P.A. and E.A. Roehl, 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 1999, 7p.

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., Daamen, R.C., and Kitchens,W.M., 2006, Simulation of water levels and salinity in the rivers and tidal marshes in the vicinity of the Savannah National Wildlife Refuge, Coastal South Carolina and Georgia: U.S. Geological Survey, Scientific Investigations Report 2006-5187, 134 p.

Conrads, P.A. and Roehl, E.A., Jr., 2007, Analysis of salinity intrusion in the Waccamaw River and the Atlantic Intracoastal Waterway near Myrtle Beach, South Carolina, 1995-2002: U.S. Geological Survey, Scientific Investigations Report 2007-5110, 41 p. 2 apps.

Conrads, P.A., Feaster, T.D., and Harrelson, L.G., 2008, The effects of the Saluda Dam on the surface-water and ground-water hydrology of the Congaree National Park flood plain, South Carolina: U.S. Geological Survey Scientific Investigations Report 2008-5170, 58 p.

Tetra Tech (2005), Development of the EFDC hydrodynamic model for the Savannah Harbor. prepared for the U.S. Army Corps of Engineers – Savannah District, Tetra Tech, Inc. Atlanta, Georgia.

Welch, Z.C., and Kitchens, W. M., 2007, Predicting freshwater and oligohaline tidal marsh vegetation communities in the vicinity of the savannah national wildlife refuge, *Proceedings of the 2007 Georgia Water Resources Conference*, held March 27–29, 2007, at the University of Georgia