TRANSFORMING LARGE DATABASES INTO CRITICAL KNOWLEDGE USING DATA MINING– THREE CASE STUDIES IN SOUTH CAROLINA AND GEORGIA

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

[1]USGS SC Water Science Center
Stephenson Center, Suite 129
Columbia, SC 29036

[2]Advanced Data Mining, Greer, SC

ABSTRACT

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 relationships to problems of interest, and developing models that predict future outcomes. Data mining is used extensively in financial services, banking, advertising, manufacturing, and e-commerce to classify the behaviors of organizations and individuals, and 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, models were developed to simulated 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 determine the minimum streamflow required to protect municipal intakes from seawater inundation along the Grand Strand of South Carolina. In the three 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.

KEYWORDS

Data mining, artificial neural network models, decision support systems, dissolved oxygen, salinity

INTRODUCTION

While environmental monitoring technologies have made it cost effective to acquire tremendous amounts of real-time hydrologic and water-quality data, there is greater demand to transform these data into the essential knowledge needed by State and local water-resource managers. It is imperative that new technologies be developed and adopted that facilitate faster and more accurate data analysis, modeling, and regulatory tool development. Data mining is an emerging field that addresses the issue of converting large databases into knowledge to solve complex problems due to the large numbers of variables. Data mining methods come from different technical fields such as signal processing, statistics, and artificial intelligence and are used extensively in financial services, banking, advertising, manufacturing, and e-commerce to classify the behaviors. Data mining employs methods for maximizing the information content of data, determining which variables have the strongest relationships to problems of interest, and developing models that predict future outcomes. This knowledge encompasses both understanding of cause-effect relations and predicting the consequences of alternative actions.

There are many environmental systems where tremendous historical databases exist. Generally these databases are under interpreted and under utilized. Data mining offers an approach to transform these data into information and, ultimately, knowledge of functionality of the environmental systems. This paper describes the technical approach and results of three case studies where data mining was applied to large-scale environmental issues in South Carolina and Georgia to assist decision makers in the issuance of long-term permits. The three studies used existing databases for analysis and development of empirical models to understand how the system works and to address the salient concerns of decision makers. The large databases were transformed into information that provided new knowledge on how the systems function. The three case studies are (1) the determination of allowable point-source effluent loading to the Beaufort River for the design of a Regional Water Reclamation Facility; (2) the determination of the potential impacts of deepening the Savannah Harbor on upstream freshwater tidal marshes for the Environmental Impact Statement; and (3) the determination of minimum releases from North Carolina reservoirs required to protect municipal intakes from seawater inundation along the Grand Strand of South Carolina for a 50-year Federal Energy Regulatory Commission (FERC) license. The three case studies share several characteristics:

- Utilized large existing historical databases;

- Developed empirical models of complex tidal systems using Artificial Neural Network (ANN) models[1];

- Developed Decision Support Systems (DSS’s) that integrated databases, models, model simulation controls, streaming graphics, and model outputs in a easily disseminated spreadsheet application; and,

- Results from the studies were or are currently (2006) being used for water-resource management with large long-term environmental, economic, and societal consequences.

CASE STUDY 1: SIMULATING POINT-SOURCE EFFLUENT LOADING IMPACTS TO THE BEAUFORT RIVER

Beaufort and Jasper Counties are two rapidly growing coastal counties in South Carolina. According to the 2000 census, the region grew by 40 percent during the preceding 10 years. The population growth has increased the quantity of wastewater in the area. The principle receiving stream for treated effluent is the Beaufort River. The river is a complex tidal river system that is home to shellfish grounds and fisheries nursery habitats in addition to receiving treated wastewater from four civilian and military water reclamation facilities (WRFs, fig.1). Although not uncommon for coastal areas, the river is on the South Carolina 303(d) list of impaired waters for low dissolved oxygen (DO) (SCDHEC, 1998). The Clean Water Act stipulates that Total Maximum Daily Loads (TMDL) must be determined for all waters on the 303(d) list. Critical to the development of a defensible TMDL is the linkage between the impairment and the source of the impairment. The linkage is typically performed using a prediction model.

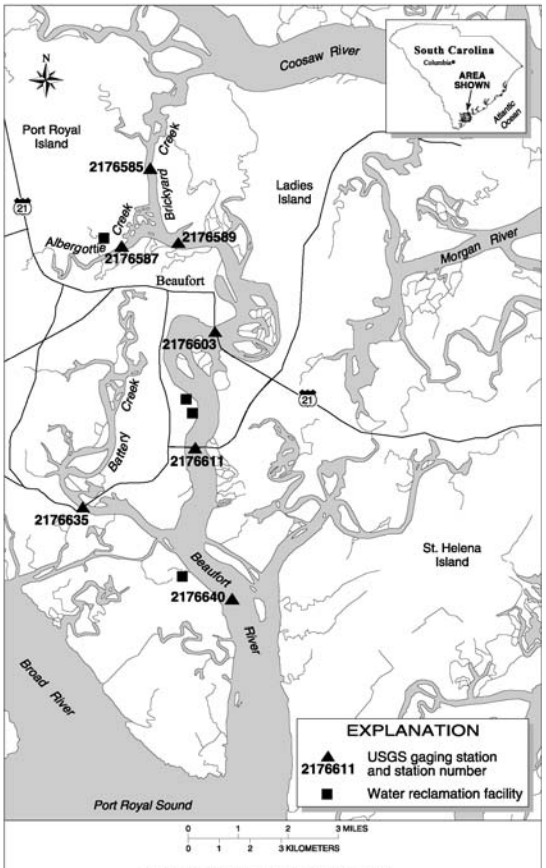

Figure 1: Beaufort River, S.C.

The Beaufort-Jasper Water and Sewer Authority (BJWSA) operate two WRFs on the Beaufort River. The facilities are operating at 70 percent of capacity and must be replaced in 2006 to handle increased wastewater flows for the growing coastal community. The Director of BJWSA envisioned a plan to build a regional WRF and to consolidate the civilian and military wastewater discharges into a single, high-quality effluent at the location of one of the current outfalls. The ambitious plan required an expedited permitting effort that included developing a predictive DO model of the Beaufort River to evaluate the effect of existing and future WRFs. Two previous modeling and permitting efforts along the SC coast (Myrtle Beach and Charleston) were lengthy processes taking more than 10 years from the initiation of data collection to issuance of permits. To meet the schedule for a new permit and the construction of a new WRF, a new approach to developing a predictive DO model was required. BJWSA assembled a team consisting of the U.S. Geological Survey (USGS), Advanced Data Mining, LLC (ADM), and Jordan Jones and Goulding (JJ&G) to analyze existing data, build an empirical DO model, and coordinate the permitting process with the South Carolina Department of Environmental Control (SCDHEC).

The team was successful in developing an accurate predictive DO model of the Beaufort River, disseminating the study results in a user-friendly DSS and obtaining the required permits to initiate construction of a new regional WRF. This study used the new modeling approach and was able to reduce the time from the initiation of data collection to the issuance of permits by 50 percent.

Approach

The variability of DO in the Beaufort River is a result of many factors including the quality of the water from Port Royal Sound to the south and the Coosaw River to the north, the loading of oxygen-consuming constituents from tidal marshes and other non-point sources, effluent from four permitted point sources, and the physical characteristics of streamflow, tidal range, salinity, and temperature. The following discussion is a brief summary of the data sets, data preparation, and ANN modeling. More detail descriptions of these technical aspects of the study can be found in Conrads and others (2003).

Data Sets and Data Preparation. The data used for analysis and modeling consisted of continuous (1-hour interval) tidal and water-quality data, daily total precipitation data, and weekly effluent data. In 1999, BJWSA, in cooperation with the USGS, established a network of seven gaging stations (fig. 1) on the Beaufort River that monitor water level (WL), water temperature (WT), specific conductance (SC), and DO. Three of the stations also record tidal streamflow. Precipitation data were obtained from the National Weather Service and two of the WRFs. Effluent data (sampled once a week) consisting of 5-day biochemical oxygen demand (BOD_5) and ammonia (NH_3) also were obtained from the WRFs.

Two calculated variables were derived — tidal range (XWL) and DO deficit (DOD). Tidal range is an important variable for determining the flushing dynamics of the tidal rivers. Tidal range, calculated from water level, is defined as the water level at high tide minus the water level at low tide for each semi-diurnal tidal cycle. The DOD is the measure of the difference between actual DO measurement and DO for fully saturated conditions. The DOD was computed using an algorithm that assumes a constant barometric pressure over the data collection period (USGS, 1981). The DOD was adjusted for salinity.

Tidal systems are highly dynamic and exhibit complex behaviors that occur over a range of time scales. The semi-diurnal tide is dominated by the lunar cycle, which is more influential than the 24-hour solar cycle; thus, a 24-hour average is inappropriate to use to reduce tidal data to daily mean values. For analysis and model development, the USGS hourly data were digitally filtered using a low-pass filter (Press and others, 1993) to remove semi-diurnal and diurnal variability (filtered variables are denoted by an “F” prefix, for example, FDO). The resulting filtered time series were then averaged over a 24-hour period to represent the daily mean for each parameter.

Explanatory variables for a particular response variable are often themselves correlated. It is difficult, if not impossible, to identify the individual effects of these variables (sometimes known as confounded or correlated variables), on a response variable. Empirical models have no notion of process physics or the nature of interrelations among input variables. To be able to clearly analyze the effects of confounded variables, the unique informational content of each variable must be determined by “de-correlating” the confounded variables. The precipitation, XWL, WL, and SC were systematically, non-linearly decorrelated using ANN models.

Because of the limited number of data points of the effluent data (BOD_5 and NH_3) as compared to the river gaging data, a subset of the dataset was excised that included only the daily digitally filtered data. In addition, time derivatives (the 1-day change in a variable) of the hydrodynamic (WL and XWL) and water-quality variables were computed and added to the dataset. The derivatives of filtered variables are denoted by an E prefix, for example, ESC.

Simulating Dissolved-Oxygen Deficit

The goal of the model was to predict the impact of the point- and non-point sources on DO. Had the goal only been to predict DO and not the effects of the WRFs, this could have been done easily and accurately using only WT owing to their strong inverse relation. Linear regression produces a coefficient of determination (R^2) of 0.88, indicating that approximately 88 percent of the variability of FDO is explained by FWT alone (fig. 2), and that only approximately 12 percent of the variability is caused by other factors.

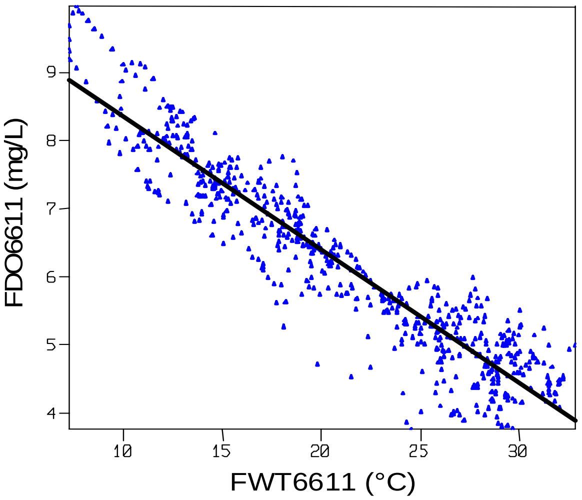

Figure 2: Scatter plot of filtered DO (FDO) and filtered water temperature (FWT) and least-squares regression line (R^2=0.88) for Station 02176611.

The real goal of a regulatory model is to predict how much of the variability in DO is attributable to point-source discharges. The use of DOD rather than DO as the response variable normalizes the DO signal with respect to WT to emphasize the effects of external loadings on DO. The response of DOD to BOD_5, NH_3, rainfall, and the other explanatory variables was predicted using ANN models that were trained using the back-propagation and conjugate-gradient algorithms.

Visual inspection of the BOD_5 loading data from the WRF and the 1-day change in DOD (EDOD) at station 02176611 (fig. 3) suggests a potential relation (note that the EDOD scale has been inverted so decreases in daily DO rise on the scale). The number of coincident peaks, for example observations 6, 31, 35, 39, and 58, indicate that BOD_5 loading may account for a significant part of the remaining 12 percent of the variability in DO.

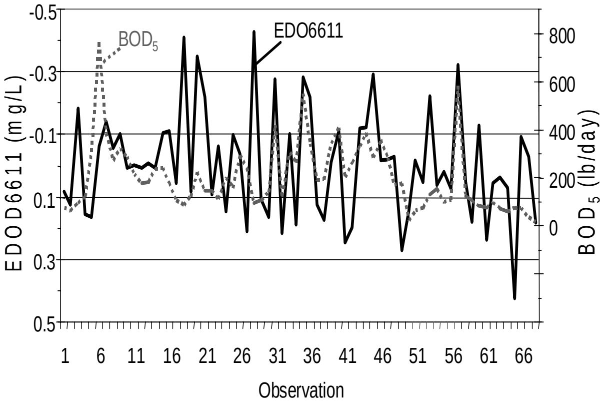

Figure 3: One-day change in DO (EDO) and BOD_5 (at a 1 day time delay).

For each station, an ANN model of the EDOD, having BOD_5, rainfall, and decorrelated filtered WL, XWL, SC and WT as inputs was generated to provide a more comprehensive assessment of the relation between the BOD_5 and the DO. Figure 4 show that the model fits most of the higher peaks in the EDOD. The R^2_{ANN} = 0.57, indicating that approximately 57 percent of the variability in the EDOD is accounted for by variability in the input variables. A similar approach was used for modeling the impact of NH_3 on DOD.

The Beaufort River DO Model is a super-model composed of 118 cascaded sub-models. Separate sub-models were constructed for each combination of river gage location, discharge type (BOD_5 or NH_3), and relative time delay. The impacts of BOD_5 and NH_3 were computed by sub-models for each river gage that used decorrelated XWL, WL, SC, and WT and their 2-day time derivatives as inputs. Also included are inputs for each WRFs BOD_5 and NH_3 at appropriate time delays. The output of each sub-model is a prediction of the 1-day EDOD.

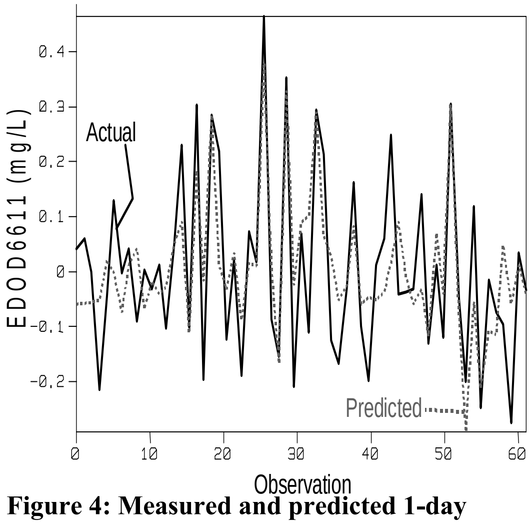

Figure 4: Measured and predicted 1-day change in DO deficit (EDOD). ANN used BOD_5 as an input at a time delay of 1 day.

Development of the Decision Support System

Commonly, a DSS is often a software package built around a model, making the model the DSS's most important component because ostensibly it can correctly predict, *"What will happen if we do A instead of B?"* Models are often developed at considerable expense; therefore, the packaging is done only to maximize the usefulness of the model to the broadest possible community of users. Complex mathematical (mechanistic) models based on first principles physical equations are often developed and operated by senior scientists or engineers; however, the interests and computer skills of the actual decision makers and other stakeholders are quite varied. A DSS was developed to meet the needs of the technically diverse group of stakeholders for equal access by all to the body of scientific knowledge needed to make the best possible decisions.

The DSS for the Beaufort River was developed as Microsoft®Excel/Visual Basic for Applications (VBA)[2] programs. This allowed the DSS to be prototyped, easily modified and also allowed the DSS to be distributed in a familiar form. The DSS is operated through a graphical user interface (GUI) composed of menus and controls that require no typing. This makes the DSS easy to use and eliminates the need to trap user input errors. The DSS incorporates a database of measured and calculated time-series variables for running long-term simulations. Under user control, a VBA program loops through database records, assembles input vectors, executes super-model instructions, outputs model predictions, and drives graphics. The DSS incorporated the following in addition to the database and ANN sub-models:

- *Simulation Controls:* Model controls, including start/stop dates, user-defined settings, and optimizations run (discussed below), are set with the point and click GUI. The DSS executes multiple model simulations simultaneously such as the no-discharge load, actual (historical) loading, user-defined, and optimized conditions.
- *Spatial Interpolation:* The DO predictions are spatially interpolated by the modeling application using a "natural cubic spline" algorithm (Burden and others, 1981). The longitudinal

DO response in the system for the no load, actual load, and optimized load is dynamically displayed during the model simulation.

- *Volumetric Averaging:* Regulatory analysis of impacts to a system typically is done on river segments over a specified period rather than on an “any time and any place” basis. The DSS allows the user to subdivide the estuary into a maximum of four segments for volumetric averaging of historical and computed hydrodynamic and water-quality variables.
- *Constrained Optimization:* A constrained optimizer was configured to represent South Carolina state law that governed the maximum allowable impact that nutrient loads from the WRFs could have on riverine DO. The optimizer modulates controllable inputs during simulations to obtain predictions that match user-set points. Users can allocate the TMDL load among the BOD and NH_3 discharges from each WRF. At each time step, the optimizer iterates load inputs as assimilative capacity changes. Output from the optimizer is time series of allowable loadings for WRFs that meet the water-quality standard.

Discussion of Case Study I

Prior to this study, it was understood that the net streamflow of the Beaufort River was to the south through Port Royal Sound. Analysis of the tidal streamflow determined that the net streamflow of the system was to the north through Brickyard Creek (and other tidal creek connections) to the Coosaw River. The tidal flow dynamics were confirmed by the long residence times seen in the SC data and the correlation analysis between the WRF effluents and DO response. The knowledge of the net streamflow of the system had far ranging consequences from determining critical DO conditions to calculating the assimilative capacity (the amount of effluent that can be discharged without violating the State water-quality standard) of the system.

The Beaufort River DSS enabled stakeholders and regulators to use new approaches to analyzing critical conditions and allowable loading to coastal systems (Conrads and others, 2003). The assimilative capacity of a system is a dynamic phenomenon that changes with the changing hydrologic and water-quality conditions. For regulatory purposes, the assimilative capacity is a fixed quantity representing the allowable loading as determined by the critical conditions for the system. For the regulator, the question becomes one of selecting the steady-state load that the WRFs will be permitted.

For the Beaufort River, the critical condition occurs during neap tides and has a recurrence interval of 14 days (Conrads and others, 2003). Rather than select one neap tide to use as a critical condition, the allowable loading was computed for the full 2½-year period of record (March 1999 to September 2001). A frequency distribution of the allowable loading (ultimate-oxygen demand, in pounds per day) was generated to better understand the range and occurrences of the predicted loads. Figure 5 shows the load frequency distribution and the cumulative percentile plot.

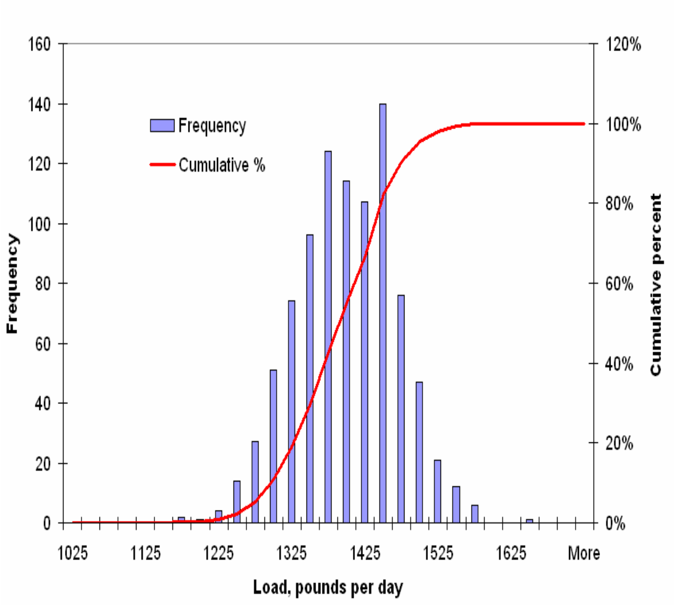

Figure 5. Frequency distribution for allowable loading (pounds per day) for WRF near station 02176611 for March 1999 to September 2001.

Using the percentile plots, regulators could select a constant allowable loading based on a frequency of occurrence. Once selected, the allowable load was simulated in the model as a constant load and the frequency of meeting the water-quality standard was evaluated. For the Beaufort River, a 90-percent reduction in loading was required to obtain the new permit for the regional WRF plant for the Beaufort River.

The successful role played by the Beaufort River ANN model in developing a regional WRF demonstrates that an innovative modeling approach using data mining techniques can be undertaken if the river system is well characterized by continuous data and there is a cooperative relationship between the permitted dischargers and the regulatory agencies involved. Advances in environmental monitoring over the past 20 years have made it cost effective to acquire tremendous amounts of hydrologic and water-quality data and large databases exist for many riverine and estuarine systems. Empirical models of complex river systems can be developed directly from the data. These models often can be developed faster than traditional modeling methods and easily disseminated to meet the needs of a broad range of stakeholders.

CASE STUDY 2: INTEGRATING THREE-DIMENSIONAL HYDRODYNAMIC TRANSPORT AND ECOLOGICAL PLANT MODELS OF THE SAVANNAH RIVER ESTUARY

Under sponsorship from the U.S. Army Corps of Engineers (USCOE) and the Georgia Ports Authority (GPA), the Lower Savannah River estuary and the surrounding freshwater tidal marshes of the Savannah National Wildlife Refuge (SNWR) have been studied for years by a variety of governmental agencies, water users, universities, and consultants. Their interests are in controlling water quality and predicting the potential impacts on the estuary and tidal wetlands of a proposed harbor deepening. Two major initiatives for the development of the Environmental Impact Statement were the application of a three-dimensional hydrodynamic model (3DM) by a team of hydrologists, and the development of a marsh succession model (MSM) by a team of plant ecologists. The 3DM predicts changes in riverine water levels and salinity in the river system in response to potential harbor changes. The MSM predicts plant distribution in the tidal marshes in response to changes in the water-level and pore-water salinity conditions in the marsh. A mechanism for linking riverine and marsh behaviors was needed. This case study describes the integration of these models and their respective databases.

To support 3DM and MSM development, many disparate databases had been created that described the natural system's complexity and behaviors, but they had not been compiled and integrated into a usable form. Variables having particular relevance include those describing bathymetry, meteorology, WL, SC, WT, and DO concentration (specific conductance is a field measurement that is commonly used to compute salinity concentration). Most of the databases were composed of time series that varied by variable type, periods of record, measurement frequency, location, and reliability. It was recognized that data mining techniques, which include ANN models, could be used to link riverine and marsh behaviors and integrate the 3DM and the MSM.

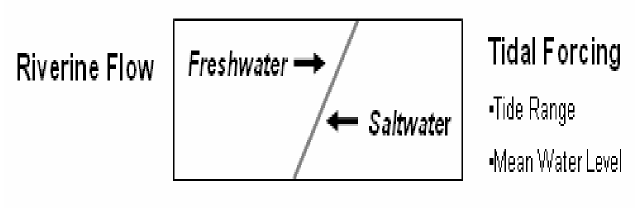

Figure 6. Diagram showing simplified conceptual model of the location of the freshwater/saltwater interface in estuarine rivers.

Approach

The estuarine portions of the Lower Saver River estuary are constantly integrating the changing streamflow, changing tidal conditions of the Atlantic Ocean, and changing meteorological conditions including wind direction and speed, rainfall, low and high pressure systems, and hurricanes. The location of the saltwater/freshwater interface is a balance between upstream river flows and downstream tidal forcing (fig. 6). During periods of high streamflow, it is difficult for salinity to intrude upstream and the saltwater/freshwater interface moves downstream towards the ocean. During periods of low streamflow, salinity is able to intrude upstream and the saltwater/freshwater interface moves upstream.

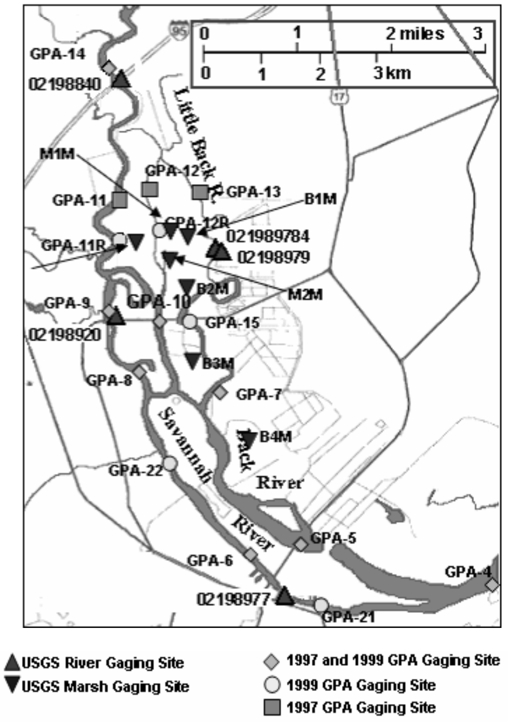

Figure 7. Gaging sites of multiple agencies in the study area.

Linking the riverine predictions of the 3DM to the MSM required that another model be developed, called the M2M for "model-to-marsh". The M2M simulates riverine and marsh water levels and salinity in the vicinity of the SNWR for the full range of historical conditions using data from the riverine and marsh gaging networks.

Data Sets and Data Preparation. The locations of the real-time gages are shown in Figure 7. The available data required extensive clean up for problems such as erroneous and missing values and phase shifts. The resulting database was composed of 11½ years of half-hourly data (200,000+ time stamps) for 110 variables. The original sources of data were:

- Q_{Clyo} and WL_{Harbor} : 11½ years of half-hourly WL signals in Savannah Harbor and river flows measured 30 miles upstream of the SNWR at Clyo, Georgia, by the USGS;

- USGS riverine WL and SC: 11½ years of half-hourly signals collected from four stations in the Lower Savannah River by the USGS;

- GPA riverine WL and SC: half-hourly signals collected on behalf of the GPA from 14 stations over three months each in 1997 and 1999. Some stations recorded both surface and bottom SC measurements (SC_{top}, SC_{bottom});

- USGS marsh WL and SC: 4½ years of half-hourly signals from 7 stations; and,

- GPA marsh WL and SC: 19 months of half hourly SC and WL from 10 stations;

Much of the field data were collected during a record setting 4½-year drought, raising concerns that it was not representative of "normal" hydrodynamic conditions. Figure 8 shows that the record low river flows during the drought led to unprecedented seawater intrusions far inland, even without a deepened harbor. It was expected that the ANNs could reasonably extrapolate from the field data by "learning" the full range of behaviors exhibited over 11½ years, which also included two El Niño events when flows were significantly above average, and presumably periods of normal conditions.

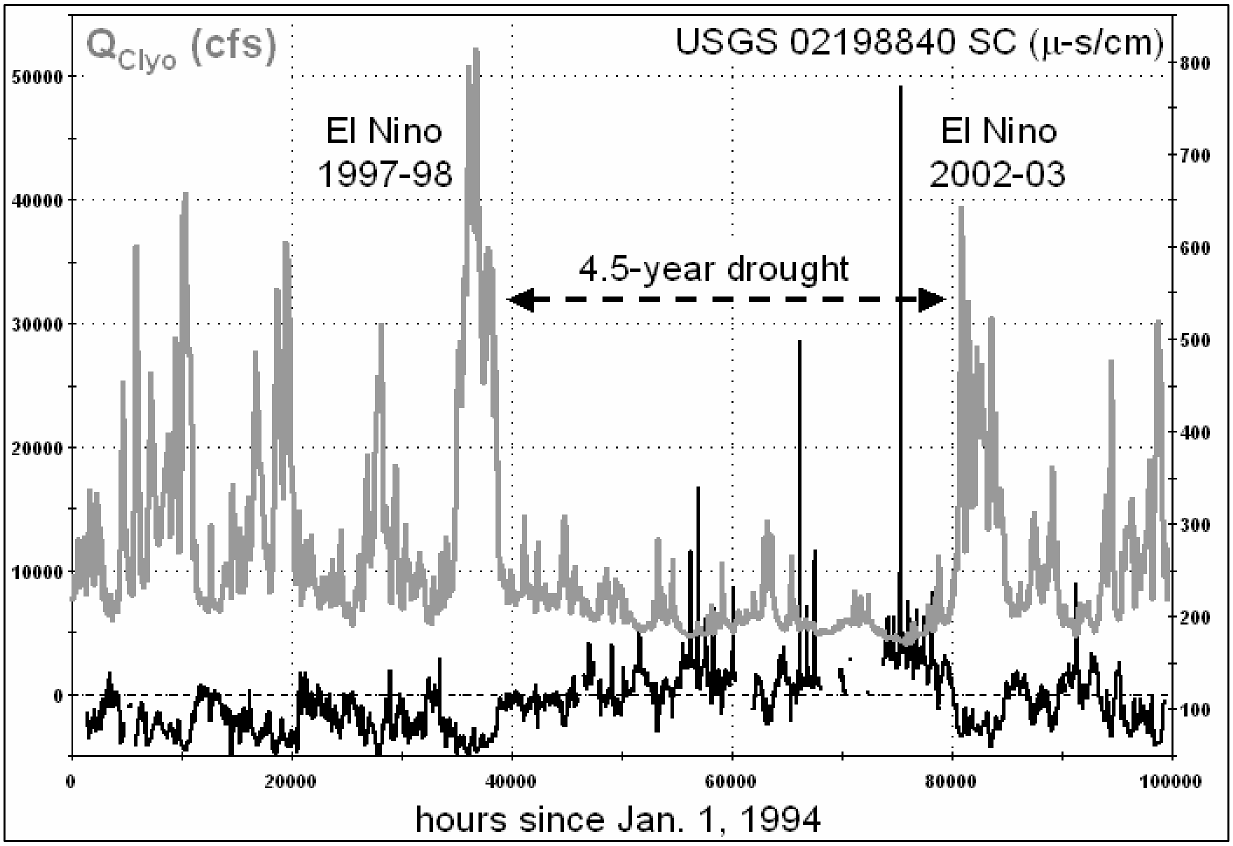

Figure 8. 11½ years of hourly Q_{Clyo} and SC at USGS 02198840, which was the farthest inland riverine gage. The SC spikes at center right occur at 28-day intervals, and are coincident peaking of the tidal range during the lowest flows of the drought.

The hydrodynamic and water quality behaviors observed in estuaries are superpositions of behaviors forced by periodic planetary motions and chaotic meteorological disturbances. Similar to the data preparation done for the Beaufort River Study, tidal signals were filtered to separate out the periodic and chaotic signals. Time derivatives of the tidal and streamflow signals were also computed. The primary chaotic inputs to the Lower Savannah River are the streamflow and the chaotic oceanic disturbances represented in the chaotic component of WL_{Harbor}.

The empirical representations of the dynamic behaviors that underlie periodic and chaotic signals are different. Multiply periodic signals are superpositions of individual periodic signals that are represented by three constants: phase, amplitude, and frequency. Abarbanel (1996) describes how chaotic univariate systems can be optimally represented by *dynamical invariants*: characteristic *time delays* and *dimensions*. Roehl and others (2000) describe an ANN model that predicted the salt-front location in the Cooper River, which incorporated signal decomposition and extended the univariate representation of chaotic behaviors to a multivariate system.

As shown in Figure 9, chaotic components were extracted from raw signals by applying a low pass spectral filter to remove high frequency (HF) diurnal and semi-diurnal variability. The important, multiply periodic tidal range XWL was computed from WL_{Harbor}. The chaotic component of Q_{Clyo} was further processed with moving window averages (MWA) of up to two weeks so that when input to an ANN with multiple time delays and time derivatives, flow histories of up to 44 days were represented.

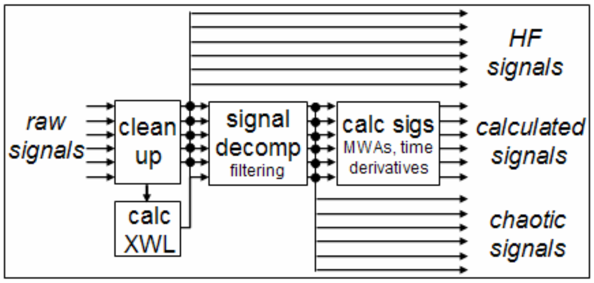

Figure 9. Signal processing and decomposition

Simulation of Marsh Pore-Water Salinity

The M2M super-model comprises 127 sub-models. Figure 10 shows that cascading sub-models predicted chaotic WL and SC signal components at riverine and marsh gaging sites. Using low pass filtered Q_{Clyo}, WL_{Harbor}, and XWL signal components for inputs, “chaotic sub-models” predicted chaotic WL and SC behaviors at four USGS gages in the main channel. These outputs were input to “HF sub-models” that also used HF WL_{Harbor} and XWL component inputs to obtain HF WL and SC predictions at the four gages.

The chaotic predictions at the main-channel gages were then transformed into calculated signals to decorrelate them and to represent dynamical behaviors that evolve over weeks. The calculated signals were used as inputs to model the historically shorter signals at the many remaining

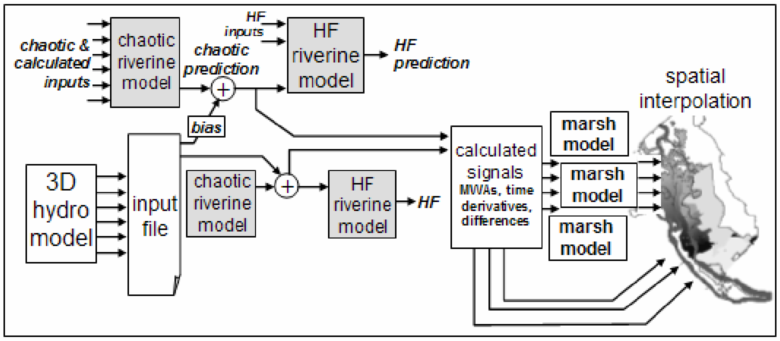

Figure 10. Data flow through the super-model decomposition. Separate sub-models were used for each WL and SC prediction.

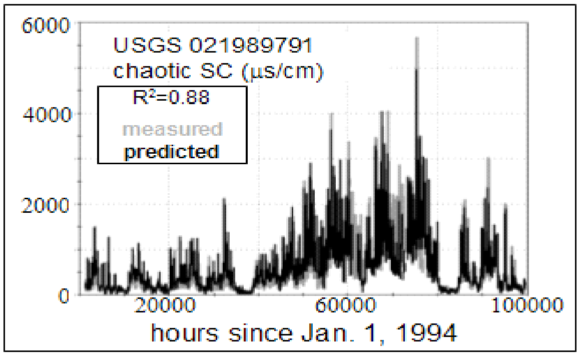

Figure 11. Measured and predicted chaotic riverine SC. Increased SC at center right occurred during the drought.

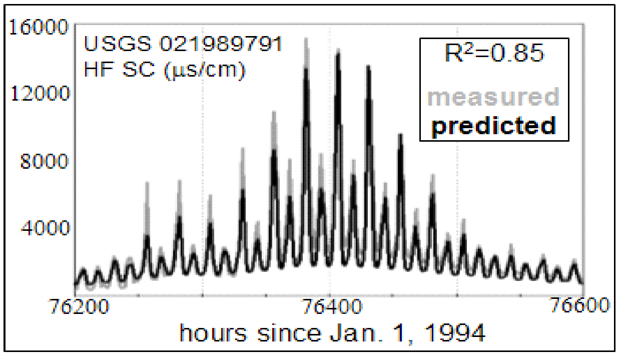

Figure 12. Measured and predicted HF riverine SC. 16.6 days are shown during the drought.

riverine and marsh stations. This provided one set of ANNs that linked the river's main-channel behaviors to tidal forcing and freshwater flows, and a second set that linked main-channel behaviors to those in backwaters and the marsh. Figures 11, 12, and 13 show SC predictions at a riverine gage and a nearby marsh gage. The R^2 values for the SC predictions at most of the gages were between 0.8 and 0.9. The R^2 values for the WL predictions were generally above 0.9.

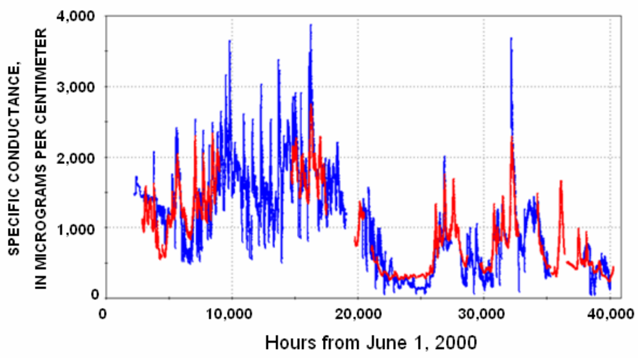

Figure 13. Measured (blue) and predicted (red) marsh SC. Gaps mark missing input data. Marsh parameters are very difficult to monitor for extended periods because of the physical instability of gaging sites.

Roehl and others (2003) describe the use of 3D response surfaces to visualize the functional forms of multivariate interactions as learned by ANNs. A surface is generated by selecting and stepping two inputs across their historical ranges, while "unshown" inputs are set to values of interest; for example, minimums, maximums, or means. Figures 14 and 15 show response surfaces representing the behaviors at a riverine gage and a nearby marsh gage. While the behavior at the riverine gage is highly non-linear with respect to freshwater flows and tides, the marsh response to the riverine SC is relatively linear. This suggests the reasonableness of using ANNs trained with backwater and marsh data collected only during the drought, but driven by riverine predictions from ANNs trained over widely ranging conditions, to extrapolate to non-drought conditions.

Development of the Decision Support System

Daamen and Roehl (2005) describe how the execution of the large number of Savannah area sub-models was orchestrated by a custom DSS. The DSS integrates the super-model with an 11½ year database, comprising more that 200,000 records of half-hourly measurements, for running long-term simulations. Similar to the Beaufort River DSS, the M2M provides a graphical user interface, streaming graphics, several freshwater flow input options, and output file generation to allow stakeholders of varying technical backgrounds to evaluate alternative scenarios under the widely ranging conditions manifest in such a long historical record.

The M2M provides the interface to integrate the output from the 3DM as input for the MSM. Figure 10 shows that the 3DM is linked to the M2M super-model through an output file. The file contains WL and SC biases for the main gaging sites. The biases are calculated by subtracting 3DM predictions representing proposed channel geometries from predictions generated using the actual historical conditions. Figure 10 shows that riverine and marsh predictions at gaged sites are interpolated to generate a 2D contour map of SC on a grid of the study area. The interpolation is performed using rules written for each grid cell. The rules accommodate the area's topological features and the different transport mechanisms of channels and marshes. The interpolation and visualization are performed in a custom post-processor that imports output from the DSS and writes interpolated values to an output file. The post-processor converts SCto salinity, and provides different options for time-averaging the predictions. Output from the post-processor can be imported into the MSM so that plant ecologists can evaluate the impacts of predicted salinity changes.

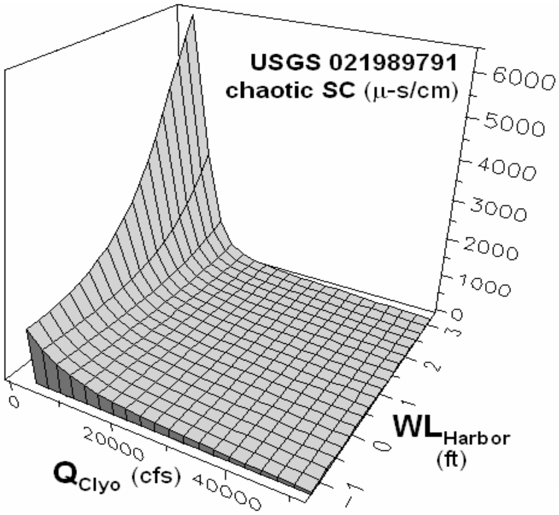

Figure 14. 3D response surface generated with a chaotic model of SC. The spikes in Figure 11 occur at low Q_{Clyo} and high WL_{Harbor}

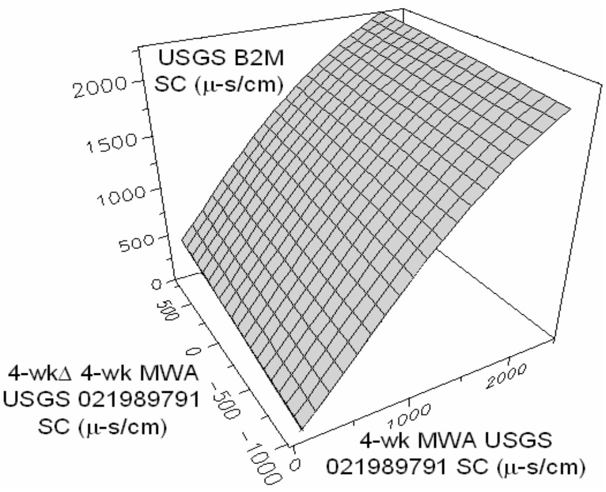

Figure 15. 3D response surface generated with a model of marsh SC at USGS B2M. The response at B2M to long-term (4-week MWA) SC at nearby riverine gage 021989791 is nearly linear. Not surprisingly, marsh SC increases if riverine SC has been high for some time, as indicated by the 4-week change(Δ) in the 4-week MWA of the riverine SC.

Discussion of Case Study II

The M2M leverages and integrates millions of dollars of field data collection and modeling performed over more than a decade by several scientific organizations and disciplines. A divide-and-conquer super-model solution, enabled by signal decomposition and accurate ANN sub-models, allowed a large amount of disparate data and intermediate works to be optimally used in their entirety. By integrating the databases and simulation models from different disciplines, the M2M integrated the knowledge of the different research groups. The packaging of the super-model and data in a DSS makes the scientific products immediately accessible and useful to all stakeholders.

CASE STUDY 3: PREDICTING SALINITY INTRUSION ALONG THE GRAND STRAND OF SOUTH CAROLINA

The Pee Dee River Basin, with approximately 12,700 square miles of drainage area in eastern North and South Carolina supplies freshwater to the Grand Strand along the South Carolina coast from Little River Inlet to the north and Winyah Bay to the south (U.S. Geological Survey, 1986) (fig. 16). 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. During the drought between 1998 and 2002, salinity intrusion forced a municipal intake to close until increased streamflow moved the freshwater-saltwater interface downstream from the intake.

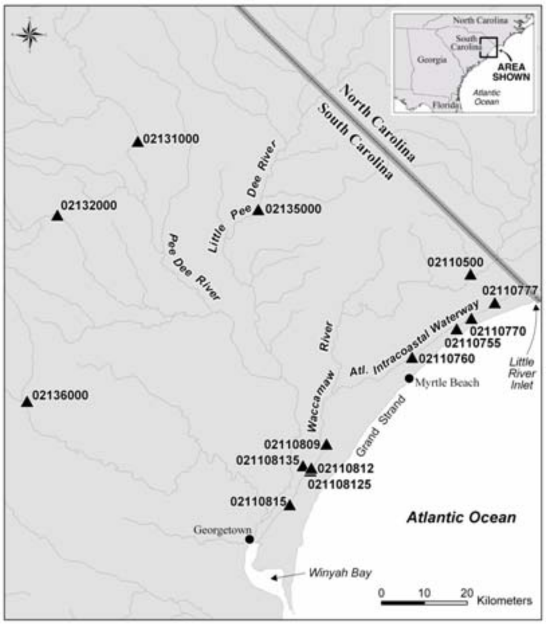

Figure 16. Location of study area and continuous gaging stations (triangles).

The North Carolina reservoirs are currently being re-licensed by the Federal Energy Regulatory Commission (FERC) for a 50-year operating permit. The water has important commercial value for generation of electric power, waterfront property development, water supply, assimilative capacity, navigation, and recreation. A coalition of stakeholders including Alcoa Power, Progress Energy, the Pee Dee River Coalition, and the South Carolina Department of Natural Resources sought to model the system's hydrodynamics and determine the minimum flow needed to protect coastal intakes.

Approach

The problem of estimating the salinity intrusion of the Pee Dee River is very similar to the salinity issues addressed in the Savannah River estuary study. A similar approach to the Savannah River estuary study was used for the Pee Dee River study. A large difference in the

Pee Dee River study is the releases from the reservoirs are 160 miles upstream with several intervening large tributaries rather than streamflow 30 miles upstream of the area of interest, as with the Savannah River estuary study.

Data Sets and Data Preparation. The USGS maintains a real-time stream-gaging network of WL and SC recorders in the Pee Dee and Waccamaw River Basins. For the streamflow stations, there is greater than 50 years of record at the majority of the stations. For the coastal water-quality stations, there is greater than 15 years of WL and SC data. Data from the Grand Strand network are a valuable resource for addressing the critical conditions for salinity encroachment on the Pee Dee and Waccamaw Rivers. During the past 15 years of data collection, the estuarine system has experienced various extreme conditions including large 24-hour rainfalls, the passing of major offshore hurricanes, and drought conditions.

For this study, a subset of the USGS data were used including nine coastal gaging stations that provided WL and SC data and five upland gaging stations that provided streamflow data. The data spanned 17½ years, but not all of the gaging stations were operating concurrently. The database for the study was augmented with rainfall data from six regional meteorological stations, and coastal wind speed and direction data from one additional meteorological station. The resulting database comprises 17½ years of hourly data (150,000+ time stamps) for 27

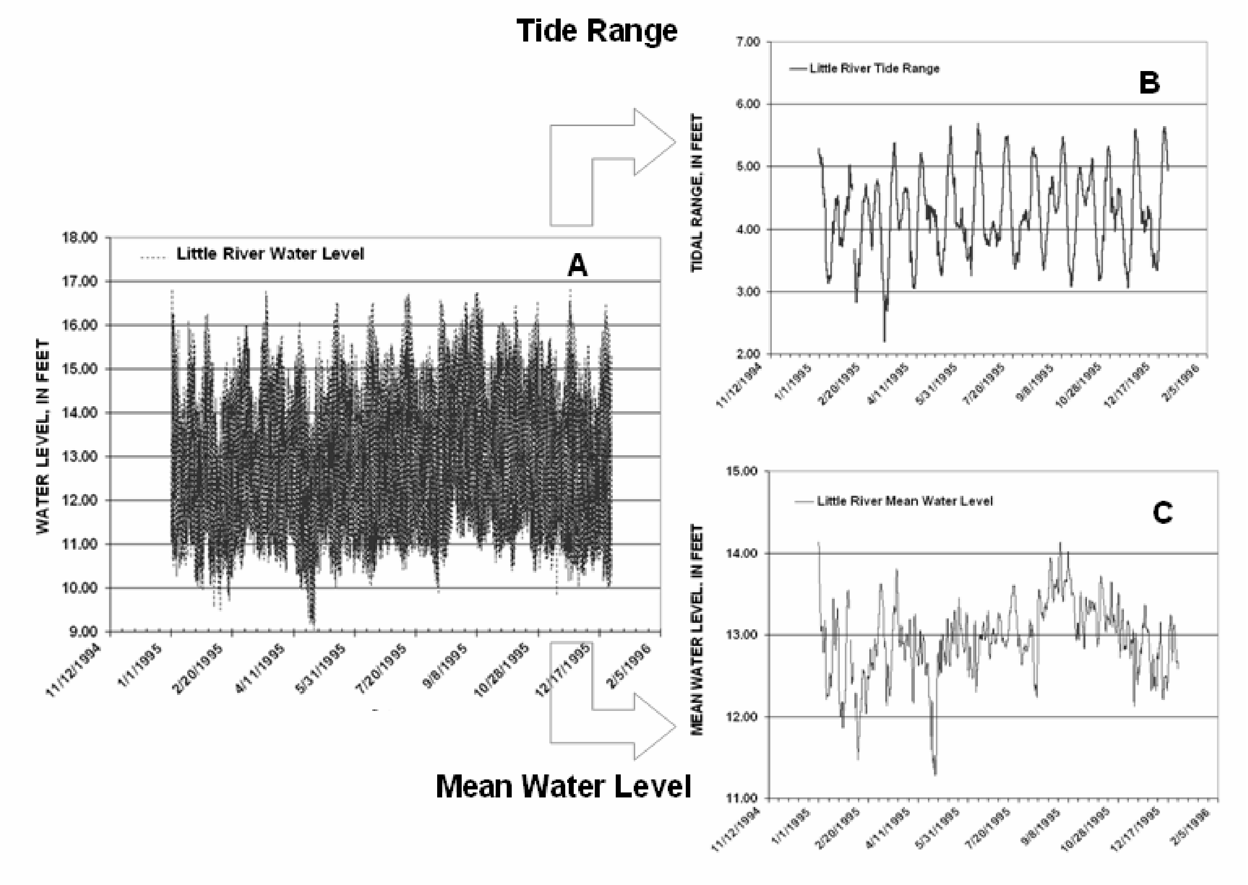

Figure 17. Tidal water-level signal (A), decomposition into a periodic signal of tide range (B), and a chaotic signal of mean water level (C).

measured variables. A similar approach to data preparation that was used in the Beaufort River and Lower Savannah River estuary studies were used for the Pee Dee River Study. The periodic and chaotic components of the WL and SC signals were decomposed and the XWL signals computed. As shown in Figure 17, the measured water level at Little River Inlet (station 02110777) (fig. 17A) was decomposed into its periodic signal of XWL time series (fig. 17B) and its chaotic signal of mean WL time series (fig. 17C).

Historically, streamflow in the Pee Dee River varies between 700 and 215,000 cubic feet per second (ft^3/s) (Cooney and others, 2003). Salinity in the lower Pee Dee River is constantly responding to changing streamflow and tidal conditions. Figure 18A shows the daily mean SC values for the Hagley Landing gaging station (Station 02110815, fig. 16) and daily mean streamflow for Pee Dee River at Pee Dee (Station 02131000, fig. 16) for the 1983 to 2003 water years[3]. The period includes a full range of flows for the system from high flows of the El Niño in 1998 and 2003 (approximately, 43,000 and 98,000 ft^3/s, respectively) to the low flows of the extended drought in the Southeast from 1998 to 2002 (fig. 18B). During periods of medium and high flows (streamflow greater than 7,000 ft^3/s), the SC values are low. During periods of low flow (streamflow less than 3,000 ft^3/s), values of SC values increase with increased salinity intrusion. During the low-flow periods prior to the high-flow El Niño of 1998, salinity intrusion with SC values ranging from 10,000 to 15,000 microsiemens per centimeter (µs/cm) were not uncommon. After the high flow of 1998 and during the extended drought, flows were even lower and remained lower for extended periods, which resulted in greater salinity at Hagley Landing with daily mean SC values greater than 15,000 µs/cm.

Simulation of Salinity Intrusion

Similar to the approach taken in the Beaufort River and Lower Savannah River estuary studies, subdividing a complex modeling problem into sub-problems and then addressing each one is a means to achieving the best possible result. For the Pee Dee study, individual ANN models for SC were developed for nine continuous coastal streamgages. The models were developed in two stages. The first stage modeled the chaotic, lower-frequency portion of the signal, as represented by the filtered SC signals. The second stage modeled the periodic, higher-frequency, hourly SC, using the predicted SC as a carrier signal. Each model uses three general types of signals, or time series: streamflow signal(s), WL signal(s), and XWL signal(s). The signals may be of the measured series values, filtered values, and/or a time derivative of the signals. The available datasets for developing the models were randomly bifurcated into training and testing datasets. Some small datasets were not bifurcated to maximize the information content in a signal. All ANN models were carefully evaluated to ensure the models did not “overfit” the data.

Eighteen models were developed: nine daily models and nine hourly models. Generally, the daily models had R^2 values ranging from 0.62 to 0.96. The hourly models had R^2 values ranging from 0.69 to 0.92. An example of the measured and predicted daily and hourly SC response models are shown in Figure 19. The daily model is able to simulate the sharp SC spikes (fig. 19A) and the hourly model is able to simulate the high-frequency SC response (fig 19B).

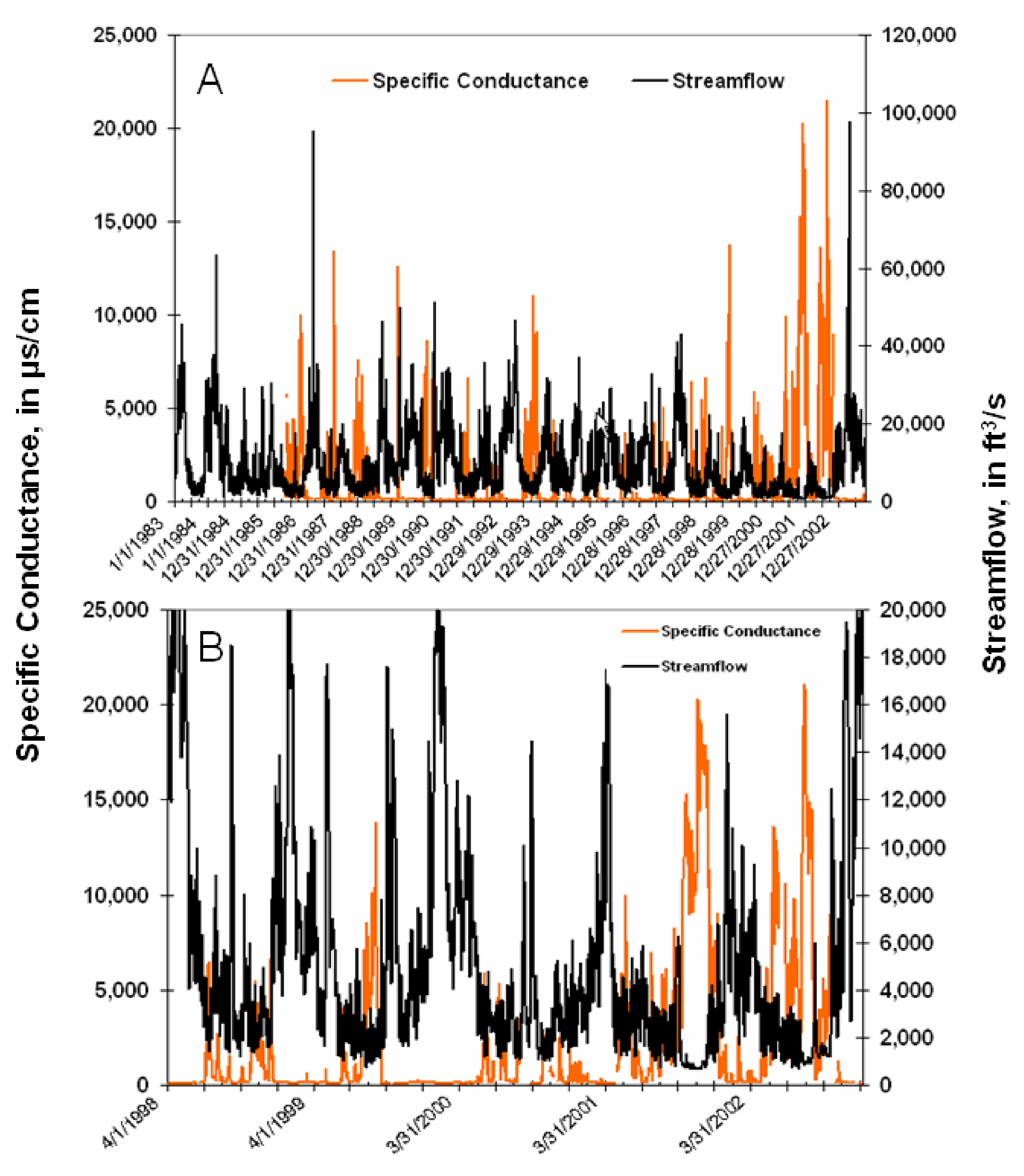

Figure 18. Graphs showing Pee Dee River flow at Station 02131000 and specific conductance response at Hagley Landing (Station 02110815) for the period 1983 to 2003 (A) and April 1998 to December 2002 (B).

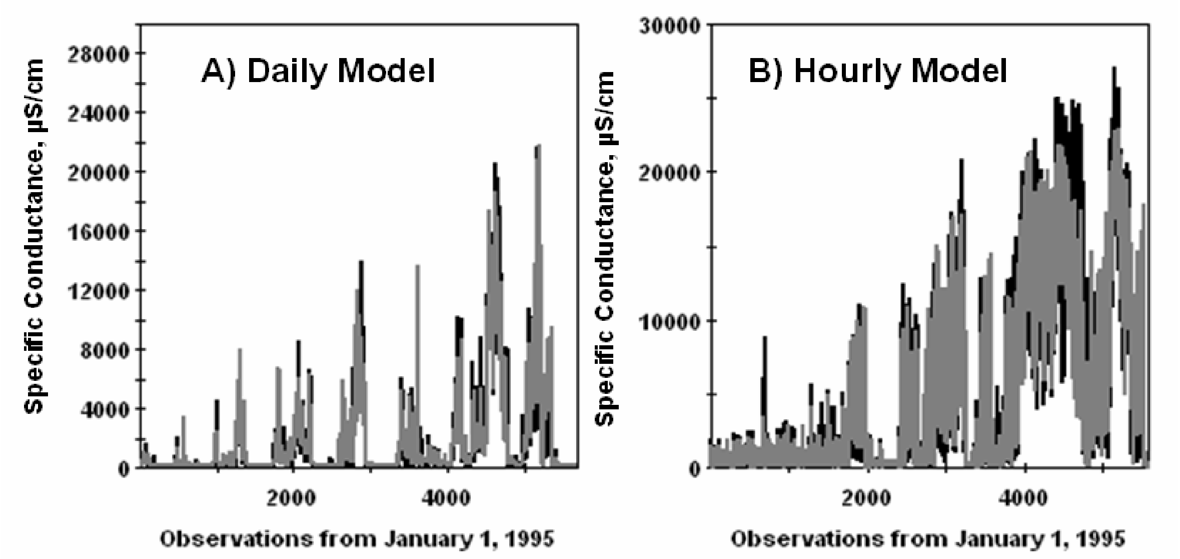

Figure 19. Graphs showing measured (black trace) and predicted (gray trace) specific conductance for Hagley Landing (Station 02110815). Results for the daily model are shown on the left (A) and the hourly model on the right (B).

Development of Decision Support System

The model in the DSS is a "super-model" that represents the whole system. The super model is composed of the 18 "sub-models" of the daily and hourly models for each gaging station. These sub-models are then incorporated into a "super-model" application that integrates the model controls, model database, and model outputs. This produces predictive models that are customized to the unique circumstances and data of a particular system. The DSS has at least two executions of the super-model. One generates predictions using actual historical input conditions, which are used to compute prediction errors and graphically depict accuracy. The second execution generates "*What if?"* predictions using user-established controllable inputs.

The Pee Dee River DSS provides for simulations corresponding to the most recent and 6½ years of higher-quality data, at daily or hourly time steps. Streamflow inputs can be set by the user to be a constant or a percentage of the historical measurements. User-defined hydrographs can also be run. The Pee Dee River DSS also provides a constrained optimizer that automatically modulates streamflow to match user-established maximum SC setpoints. The setpoints can be applied on a daily or hourly basis. The Pee Dee River DSS also provides built-in documentation that describes the variables and user controls. Documentation appears in pop-up windows as the mouse is moved in the GUI.

Discussion of Case Study III

Effective environmental management of water resources relies on the information available from various sources including monitoring data, data analysis, and predictive models. Ultimately, Pee Dee River stakeholders wanted to understand the causes of the large salinity intrusions and to determine the minimum flows that should be required in the FERC license. To facilitate the technical transfer of historical data and predictive models for the Pee Dee Basin, a DSS was developed that would allow stakeholders to have equal access to the analytical tools. Stakeholders of various technical backgrounds were able to access results from the study to transform data into usable information to enhance understanding and decision making. Stakeholders were able to determine a minimum flow to protect the intakes for a large range of hydrologic conditions. Stakeholders also realized that during extreme hydrologic conditions, the municipalities should have contingency plans rather than required unrealistic flows from the reservoirs to protect their intakes.

CONCLUSIONS

The three 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 WRF will meet the wastewater needs of the community for the next few decades. 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 four 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. The M2M will help guide decision makers in writing the Environmental Impact Statement for the proposed deepening of Savannah Harbor. 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.

END NOTES

[1] An ANN model is a flexible mathematical structure capable of describing complex nonlinear relations between input and output datasets. The architecture of ANN models is loosely based on the biological nervous system (Hinton, 1992). Although there are numerous types of ANNs, the most commonly used type of ANN is the multi-layer perceptron (MLP) (Rosenblatt, 1958).

[2] Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

[3] Water year is the 12-month period October 1 through September 30. The water year is designated by the calendar year in which it ends.

REFERENCES

Abarbanel, H.D.I., 1996, Analysis of Observed Chaotic Data, Springer-Verlag New York, Inc., New York, 4-12.

Burden, R.L., Faires, J.D. and Reynolds, A.C., 1981, Numerical Analysis 2nd Edition, Prindle, Weber & Schmidt, Boston, pp. 112-113.

Conrads, P.A., E.A. Roehl, and Martello, W.P., 2003, Development of an empirical model of a complex, tidally affected river using artificial neural networks, Water Environment Federation TMDL Specialty Conference, Chicago, Illinois, November 2003.

Cooney, T.W., Drewes, P.A., Ellisor, S.W., Lanier, T.H., and Melendez, Frank, 2003, Water Resources Data South Carolina Water Year 2002, U.S. Geological Survey Water-Data Report SC-02-1

Daamen, R.C., and E.A. Roehl, 2005, Integrating multiple databases and estuary models into a comprehensive software tool for regulatory support, South Carolina Environmental Conference, Myrtle Beach, March 2005.

Hinton, G.E., 1992, How neural networks learn from experience," Scientific American, September 1992p.145-151.

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.

Roehl, E.A., P.A. Conrads, and T.A. Roehl, 2000, Real-time control of the salt front in a complex, tidally affected river basin, Proceedings of the Artificial Neural Networks in Engineering Conference, St. Louis, p. 947-954

Roehl, E.A., P.A. Conrads, and J.B. Cook., 2003, Discussion of using complex permittivity and artificial neural networks for contaminant prediction, J. Environmental Engineering, p. 1069-1071, November 2003.

Rosenblatt, F., 1958, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological Review, 65, p. 386-408.

South Carolina Department of Health and Environmental Control, 1998 State of South Carolina 303d list for 1998, Bureau of Water, EPA Approved, June 1998

U.S. Geological Survey, 1981, Technical memorandum 81.11, Reston, VA, 1981.

U.S. Geological Survey, 1986, National water summary 1985; hydrologic events and surface-water resources, U.S. Geological Survey Water-Supply Paper 2300, 506 p.