7th International Conference on Hydroinformatics
HIC 2006, Nice, FRANCE

USING ARTIFICIAL NEURAL NETWORK MODELS TO INTEGRATE HYDROLOGIC AND ECOLOGICAL STUDIES OF THE SNAIL KITE IN THE EVERGLADES, USA

PAUL A. CONRADS
U.S. Geological Survey, 720 Gracern Road, Suite 129
Columbia, SC 29210, United States

EDWIN ROEHL
Advanced Data Mining, 3620 Pelham Road, PMB 351
Greenville, SC 29615, United States

RUBY DAAMEN
Advanced Data Mining, 3620 Pelham Road, PMB 351
Greenville, SC 29615, United States

WILEY M. KITCHENS
U.S. Geological Survey, Box 110485, Bldg. 810
Gainesville, FL 32611-0485, United States

Hydrologists and ecologists have been working in the Everglades on integrating a long-term hydrologic data network and a short-term ecological database to support ecological models of the habitat of the snail kite, a threatened and endangered bird. Data mining techniques, including artificial neural network (ANN) models, were applied to simulate the hydrology of snail kite habitat in the Water Conservation Area 3A of the Everglades. Hydroperiods of water depths have a significant affect on the nesting and foraging of the snail kite. Seventeen water-depth recorders are co-located at transects where extensive plant samples are collected. These continuous recorders were established in 2002. A long-term network of three water-level recorders has been maintained since 1991. Using inputs representing the three long-term gages, very accurate ANN models were developed as input to predict the water depths at the 17 short-term sites. The models were then used to hindcast water depths to 1991, resulting, much longer water-level record to help scientists better learn how the snail kite's habitat is affected by changing hydrology. A Decision Support System (DSS) was developed to disseminate the models in an easily used package. The DSS is a MS Excel™/VBA application that integrates the models and database with interactive controls and streaming graphics to run long-term simulations. As part of the Everglades restoration Interim Operating Plan (IOP), a regional hydrologic model is used to generate water levels for alternative flow regulation schedules. The alternative IOP water levels are input to the DSS to predict the hydrology of the snail kite habitat. The application demonstrates how very accurate empirical models can be built directly from data and readily deployed to end-users to support interdisciplinary studies.

INTRODUCTION

The restoration of the Everglades is one of the most ambitious ecosystem restorations undertaken [1]. The success of the restoration from a compartmentalized system of water impoundments to a flowing hydrologic system is dependent on the survivability of many threatened and endangered species. The snail kite (Rostrhamus sociabilis) is an endangered raptor with its distribution limited to South Florida [2]. The life cycle of the snail kite is highly dependent on the hydrology of the Everglades' wetlands in terms of habitat and diet. Water-depth fluctuations directly affect preferred vegetation (wet prairie) and principal food source, the aquatic apple snail (Pamacea paludosa). Scientists are studying the ecology of the snail kite in Water Conservation Area 3b (WCA3b) of the Everglades, the largest of designated critical habitats of the raptor (fig. 1). Much of the area is already serious degraded and various studies have documented the conversion of wet prairie to aquatic sloughs and losses of interspersed herbaceous and woody species essential for nesting habitat.

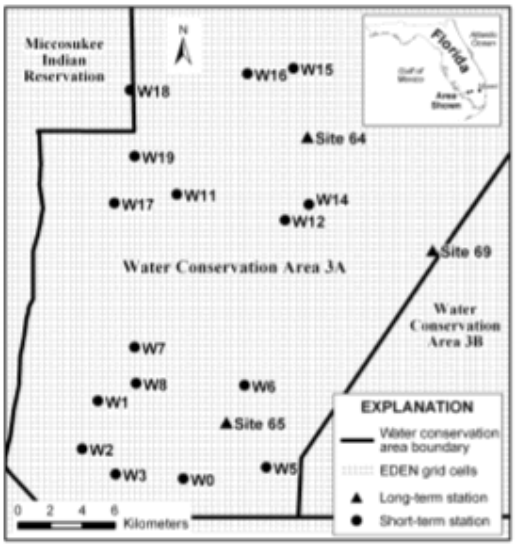

Figure 1. Mapping showing study area and location of continuous gaging stations.

The principal objective of the snail kite study in WCA3b is to separate plant community response due to typical seasonal and inter-annual variances in hydrologic regimes. The vegetative community structure of these sites is an expression of present and historic hydrologic conditions. A critical element of the study is to determine how the vegetative communities respond to temporal and spatial changes in hydrology. Water-level data from 1991 is available from three real-time gaging stations (referred to as "long-term" sites). In 2002, as part of the snail kite study, an array of 17 continuous water-depth monitors were co-located at transects where extensive plant samples are collected (referred to as "short-term" sites). To assist ecologists analyze the water depth and hydroperiods over a large range of hydrologic conditions and to integrate long-term ecological data, the hydrologic histories at the 17 transects were hindcast back to 1991 using artificial neural network (ANN) models. Figure 2 shows long-term (1991 to 2005) water-level data for Site 64 (fig. 1) and short-term (2002 to 2005) water-depth data for Site W8 (fig 1).

METHODS

The authors had previously developed ANN-based models of estuaries in Georgia and South Carolina. The type of ANN used was the multi-layered perceptron (MLP) described by Jensen [3], which is a multivariate, non-linear regression method based on machine learning. In a side-by-side comparison, Conrads and Roehl [4] found that ANN models had prediction errors 60-82 percent lower than those of a state-of-the-practice mechanistic model when predicting water temperature, specific conductance, and dissolved oxygen on Cooper River in South Carolina. Conrads and others [5] went on to use ANNs to estimate the impacts of nutrient loading from rainfall runoff and tidal marsh inundation on DO in the same waterway. In a regulatory application, Conrads and others [6] describe an ANN-based model for the permitting of three wastewater treatment plants that discharge into the Beaufort River estuary. In general, a high-quality predictive ANN models can be obtained when:

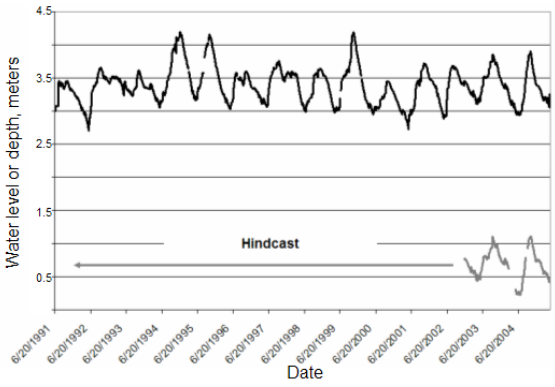

Figure 2. Graph showing long-term water-level and short-term water-depth data for the period 1991 to 2005.

- The data are well distributed throughout the state space (historical range of conditions) of interest,
- The input variables selected by the modeler share a lot of "mutual information" about the output variables,
- The form "prescribed" or "synthesized" for the model used to "map" (correlate) input variables to output variables is a good one. Machine learning techniques like ANNs synthesize a best fit to the data.

Data Set Preparation

Prior to analysis and modeling, the two data sets needed to be re-sampled to a common time interval, signals decorrelated, and additional variables computed. The long-term and short-term data sets were re-sampled to "time merge" the two sets. The short-term water-depth data are collected every 12 hours at 7:30 AM and 7:30 PM. The long-term water-level data are a daily mean water level. The short-term data was re-sampled and 7:30 AM data were used with the daily mean USGS data.

Often, explanatory variables share information about the behavior of a response variable. It is difficult, if not impossible, to understand the individual effects of these variables (sometime known as confounded or correlated variables), on a response variable. Empirical models have no notion of process physics, nor the nature of

interrelations between 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 data from the snail kite network measures water depths at the gaging stations. The USGS stations measure water levels to a known datum. To decorrelate the water-depth data and to set all the stations to a common datum for the analysis, Site 64 was used as the "standard" and the difference between Site 64 and the water depths sites was used as the time series for the analysis.

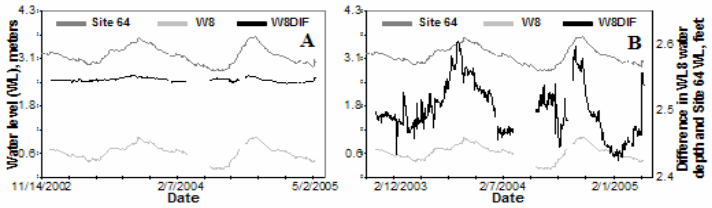

Figure 3. Plots showing water levels at Site 64, water depths at WL8, and the difference between the two time series (W8DIF). In figure 2a, the three time series are plotted on the same axis. In figure 2b, W8DIF is plotted on a separate axis to show the detail of the variability between the two signals.

Figure 3 shows the time series for the water level at Site 64, the water depth at W8, and the difference between the two time series (variable W8DIF). The variability of the difference between Site 64 and W8 is clearly seen in figure 3b where W8DIF is plotted on a separate axis. In addition to setting all the water- depth and water-level data to the same datum, using differences produces new signals that are less correlated than the original signals and reduces the multi-colinearity between the time series.

The 3 long-term sites are highly correlated and Sites 65 and 69 were decorrelated from Site 64. Decorrelation was accomplished in two steps (fig.4). The first step was to generate a Single Input Single Output (SISO) ANN model[1] using Site 69 as input and

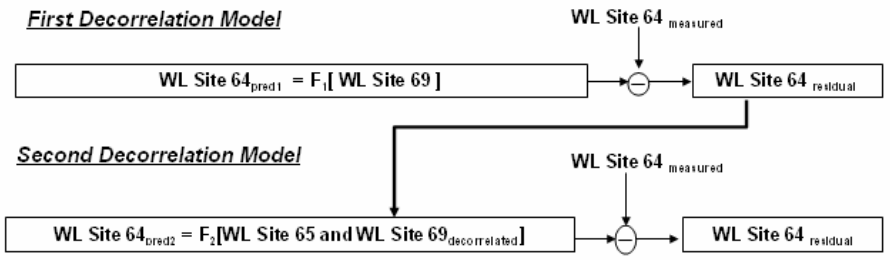

Figure 4. Schematic showing the decorrelation of water levels (WL) for Sites 69 and 64.

[1] The iQuest software was used in this study and is exclusively distributed by Advanced Data Mining, LLC, 3620 Pelham Road, PMB 351, Greenville, SC 29615-5044 Phone:864 201 8679

Site 64 as output. The residual error (the difference between predicted and measured values) is the "unshared" information between the two signals and the decorrelated signal for Site 69. The second step is to build an ANN model using Site 65 and the decorrelated signal for Site 69 to predict Site 64. The residual of this model is the decorrelated signal for Site 65.

For the long-term water level data, 2- and 3-day moving window averages (MWA) were used. To extract information from the time series on the movement or trajectory of the system, time derivatives of 1-, 2-, and 3-day time derivatives were computed on the daily and MWA values.

Modeling Approach

The predictions of water depths at the short-term sites are made in two steps. The first step is to develop ANN models to predict the water-depth difference (from Site 64) for each site. The second step is to subtract the predicted water-depth difference from Site 64 for the prediction at the short-term site. Each model uses combinations of two general types of input signals from the 3 long-term sites, a water level signal(s) (either the daily value or a MWA) and a time derivative signal(s). The input data sets are bifurcated into training and testing data sets using a zone averaging, or box, filter of the data. Using the zone average filter, all the data is used in the test dataset and a small selected sample of the data is used for the training dataset. The filter separates the dataset into user-specified number of zones or boxes and determines the input vectors with the highest information content and reserves these vectors for the training dataset. The percentage of training and testing data depended on the length of the dataset and the range of hydrologic conditions in the dataset. Typically, the zone averaging filter uses approximately 40 percent of the data for the training dataset.

RESULTS

The final water-depth predictions at the 17 short-term sites were evaluated using four "goodness-of- fit" statistics; coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), and percent model error (PME) (Table 1). Model accuracy is often reported in terms of R^2 and is commonly interpreted as the "goodness of the fit" of a model. A second interpretation is one of answering the question, "How much information does one variable or a group of variables have about the behavior of another variable?" In the first context, an $R^2 = 0.6$ might be disappointing, while in the latter it is merely an accounting of how much information is shared by the variables being used. The R^2 for the models are very high (0.976 – 0.991) and indicates that the models explain almost all of the variability of measured data.

The RMSE is defined as the square root of the mean of the squared differences between the measured and predicted data. The RMSE for sites varied from 0.01 to 0.02 m. For the statistic to be relevant, RMSE should be evaluated with respect to the range of the output variable. A model may have a low RMSE but if the range of the output

Site	Statistic					
	n	R	R^2	Mean Error (m)	RMSE (m)	PME (%)
W0	694	0.995	0.990	0.009	0.02	2.6
W1	352	0.999	0.997	0.000	0.01	1.5
W2	594	0.992	0.985	0.008	0.02	3.0
W3	301	0.996	0.992	-0.003	0.02	2.7
W5	563	0.999	0.997	0.002	0.01	1.5
W6	682	0.994	0.988	0.011	0.02	3.0
W7	222	0.997	0.994	0.001	0.01	1.8
W8	690	0.997	0.995	0.004	0.02	1.8
W9	567	0.993	0.986	0.002	0.03	3.1
W11	658	0.996	0.992	0.009	0.02	2.3
W12	603	0.998	0.996	0.008	0.02	1.6
W14	674	0.998	0.996	0.017	0.02	2.2
W15	659	0.997	0.994	0.004	0.02	1.9
W16	377	0.997	0.994	0.001	0.02	1.8
W17	392	0.991	0.982	0.000	0.03	2.8
W18	613	0.988	0.976	0.011	0.04	3.8
W19	426	0.994	0.988	-0.006	0.03	2.7

Table 1. Summary statistics for final water-depths predictions at short-term sites
[n, number of samples; R, Pearson coefficient; R^2, coefficient of determination; RMSE, root mean square error; PME, percent model error]

variable is small, the model may be very accurate but for only a small range of conditions. The PME statistic divides the RMSE by the range of the measured data to determine the percent of error over the full range of modeled data. The PME varies from 1.5 to 2.7 percent. The average RMSE and PME for the model is 0.02 m and 2.4 percent, respectively. A plot of measured and predicted water depths for Site W11 is shown in Figure 5. Site W11 provides performance results in the middle of the range of model statistics. The plot shows that the model is able to simulate the full range of the measured data. The majority of the hindcasted water depths are within the range of the data used to train the model. There are periods in the hindcast when the model extrapolated past the range to the training data. Negative water-depths indicate ground-water levels during dry periods. Predictions of water depths are not continuous due to missing data for one or more of the index stations. For complete predictions, the missing data at the long-term stations would need to be filled with estimated data.

Development of Decision Support System

To maximize the usefulness of the ANN models and the hindcasted data to a broad range of users, a decision support system (DSS) was developed to integrate the historical data,

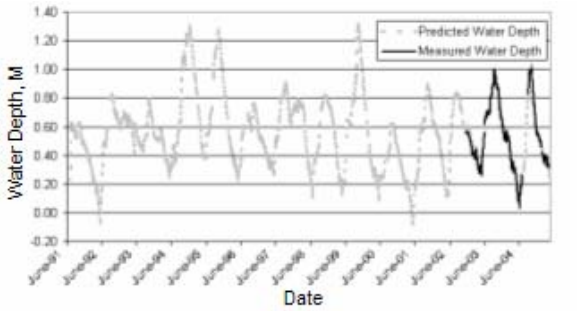

Figure 5. Measured (solid black trace) and predicted (dashed gray trace) water depth for Site W11. Period from June 1991 to June 2002 are the hindcasts from the model. Periods of missing predictions are due to missing data at one or more of the input stations.

ANN models, simulation controls, statistical analysis, and output. The DSS was developed as a Microsoft Excel™/Visual Basic for Applications (VBA) program. Figure 6 shows the basic architectural elements of the DSS. The DSS is operated through a graphical user interface (GUI) composed of menus and controls that requires no typing. This makes the DSS easy to use and eliminates the need to trap user errors. The GUI also provides graphical outputs of measured and predicted hydrologic behaviors. The main application (Simulator) within the DSS performs all WL predictions and statistical analysis. The Simulator uses a total of 19 ANN models (2 decorrelation models and 17 water depth models). The Simulator reads data and writes output files for the various run-time options that can be selected by the user through the system's GUI. Users can also select a variety of statistics to calculate for a given simulation.

User-defined hydrographs for Site 64, 65, and 69 can be inputted to the DSS to evaluate alternative water-management scenarios. As part of the Everglades restoration Interim Operating Plan (IOP), a regional hydrologic model is used to generate water levels for alternative flow regulation schedules. The alternative IOP water levels are input to the DSS to predict the hydrology of the snail kite habitat. Ecologist and water-resource managers can statistically analyze the predicted water depths to determine impacts on vegetative communities and, ultimately, on the snail kite.

SUMMARY

Hydrology has a significant affect on the nesting and foraging of the threatened and endangered snail kite. ANN models were developed using the three long-term gages as input to simulate the water levels at the 17 short-term sites and predict 11 years of hindcasted water depths. To facilitate the technical transfer of historical data and predictive models, a DSS MS Excel application was developed that would allow a broad range of uses to have equal access to the analytical tools. For ecologists, the DSS will

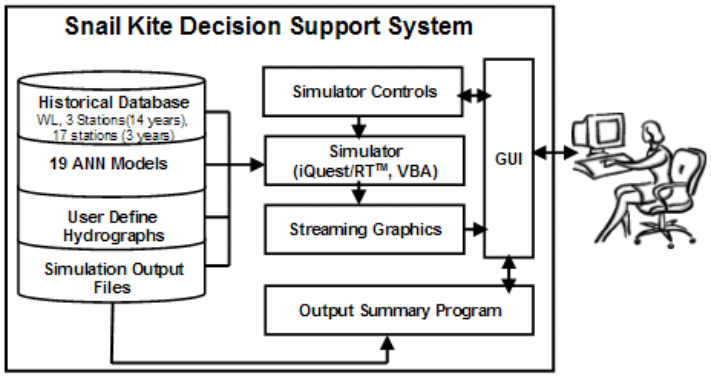

Figure 6. Schematic showing Snail kite Decision Support System (DSS) architecture

allow them to generate extended hydrologic records to increase the predictive capabilities for evaluating the snail kite habitat to changing hydrology. For water-resource managers, the DSS will allow them to evaluate alternative hydrologic regimes. The application demonstrates how very accurate empirical models can be built directly from data and readily deployed to end-users to support interdisciplinary studies.

REFERENCES

[1] U.S. Army Corps of Engineers, "*Central and South Florida Project; Comprehensive Review Study, Final Integrated Feasibility Report and Programmatic Environmental Impact Statement*", Jacksonville, FL: U.S. Army Corps of Engineers, November 1994

[2] Kitchens, W.M., 2001, "Estimation of Critical Parameters in Conjunction with Monitoring the Florida Snail Kite Population", ***http://sofia.usgs.gov/proposals/2001/kitep01.html***

[3] Jensen, B.A., "*Expert systems - neural networks, Instrument Engineers' Handbook*" Third Edition, Chilton, Radnor PA, 1994.

[43] Conrads, P.A., and Roehl, E.A., "*Comparing physics-based and neural network models for predicting salinity, water temperature, and dissolved-oxygen concentration in a complex tidally affected river basin*", paper presented at the South Carolina Environmental Conference, Myrtle Beach, March 15-16, 1999.

[5] Conrads, P.A., Roehl, E.A, and Cook, J.B., "*Estimation of Tidal Marsh Loading Effects in a Complex Estuary*," American Water Resources Association Annual Conference, New Orleans, May 2002.

[6] Conrads, P.A., E.A. Roehl, and Martello, W.P.,"*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.