Data Mining to Predict Climate and Groundwater Use Impacts on the Hydrology of Central Florida

Edwin A. Roehl, Jr.[1], Andrew M. O'Reilly[2], Paul A. Conrads[3], and Ruby C. Daamen[1]: [1]Advanced Data Mining Services, Greer, SC; [2]U.S. Geological Survey, Orlando FL; [3]U.S. Geological Survey, Columbia SC

South Carolina Water Resources Conference, 10–11 October 2012
Columbia, South Carolina

Introduction

In 2005, groundwater withdrawals averaging 762 million gallons per day (MGD) constituted 95 percent of the total amount of water withdrawn in the 5-county region surrounding the city of Orlando in central Florida (Figure 1). Groundwater recharge from rainfall into the well-drained karst terrain is the largest component of the water balance for the region's Floridan aquifer system (Sepúlveda and others, 2012). Consequently, variations in both rainfall and groundwater use can affect water levels and flows in aquifers, lakes, and springs. Several deterministic models have been developed to quantify cause-effect relationships and to help regulators and other stakeholders manage these regional resources. However, the models have been found to have difficulty simulating the complex interactions between the weather and the surface and subsurface environments in a karst terrain.

The goal of this project was to develop a decision support system (DSS) based on data mining results to complement the deterministic models. A DSS is a powerful, easy-to-use package that combines data, analytical results, predictive models, and supporting graphics that allows resource managers and stakeholders to evaluate alternative management strategies (Roehl and others, 2006).

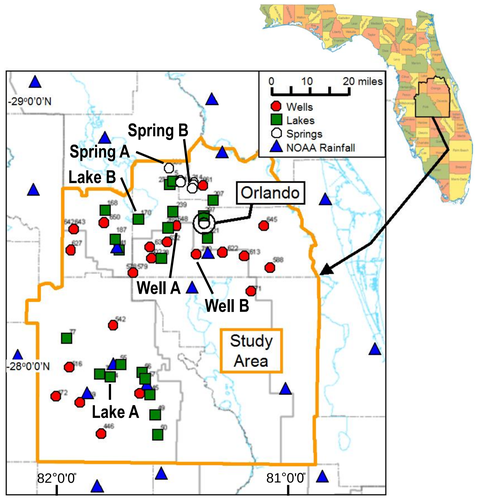

Figure 1. Study area showing locations of wells, springs, lakes, and NOAA climatic monitoring.

Description of the Data

Substantial historical hydrologic and climate data were available for data mining (Figure 1). Less complete groundwater use data was also available. They comprised:

- daily hydrographs for 23 wells (20 Floridan aquifer system, 3 surficial aquifer system), 22 lakes, and 6 springs;
- daily rainfall, air temperature, and estimated potential evapotranspiration from 18 National Oceanic and Atmospheric Administration (NOAA) sites; and
- monthly actual and estimated groundwater use representing utility pumping, phosphate mining, agriculture, citrus farming, golf course irrigation, and drainage well recharge.

The completeness (fewer missing data) and quality (more measured and less estimated data) of the data varied significantly. In general the NOAA meteorological data were the most complete and have the highest quality, followed by the well, lake, spring, and groundwater use data.

Technical Approach

Artificial neural networks (ANNs) are a multivariate, nonlinear curve fitting method from the field of Artificial Intelligence that is commonly used for industrial process modeling and control (Jensen, 1994). Because of delays in availability of groundwater use data, the data mining initially focused on determining the extent to which rainfall, air temperature, and potential evapotranspiration could explain daily variability in the hydrographs from 1942 through 2008. As a first step, an empirical, multi-layer perceptron ANN model was developed for each hydrograph. For inputs to the ANNs, the climate time series were decomposed into decorrelated spectral ranges that had window sizes from 30 days to six years to represent the dynamics of the spectral time periods.

Figure 2. ANN model schematic for each site. WL refers to the water levels of the wells and lakes. Q refers to the flow rates of the springs.

The ANN for each site was systematically trained by using sensitivity analyses to cull less predictive inputs. This 'training-sensitivity' process revealed that rainfall-derived inputs were the best climatic predictors. Temperature and potential evapotranspiration inputs were removed, resulting in 51 rainfall-only ANNs. For most sites, the data was bifurcated into training and testing data sets, the latter to provide independent statistics about model accuracy. This was not possible for some sites because their measurement population was too small.

The groundwater use impacts were subsequently modeled using inputs derived from aggregated data that summed all different types of groundwater use for each month. This approach was necessary because most of the groundwater use data were estimates whose temporal patterns varied little spatially, a problem for empirical modeling that relies on variability to be effective. The aggregation was also justified by the generally high hydraulic conductivity of the Floridan aquifer system that disperses localized impacts, and the one-month time step that dampens transient variability.

The groundwater use data were processed into spectral ranges similarly to the rainfalls. The 51 groundwater use ANNs simulate the monthly-averaged prediction errors (residuals) of the rainfall ANNs. The residuals represent the portion of the variability in the hydrographs that is not explained by rainfall ANNs. For all but a few sites, testing data were not used because of low measurement populations resulting from changing the time step from daily to monthly. Figure 2 shows that the outputs of each site's ANN pair are summed to compute a final prediction.

Results and Discussion

Daily Rainfall ANNs - Accuracy statistics (based on testing data when available) for the rainfall ANNs indicate the average coefficient of determination (R^2) is highest for the wells, followed by the lakes and then the springs (Table 1). The average percent error is lowest for the wells and higher for lakes and springs. Lines fitted to the rainfall ANN residuals by least-squares with respect to time (green lines in Figure 3) denote their long-term trends and suggest long-term changes in water use, land use, and other factors. The long-term decreasing trend of Well A is accurately predicted using rainfall ANNs because rainfall in the western portion of the study area was observed to decline. The ANN poorly replicates the long-term trend and more extreme high frequency variability of Well B, which may be caused by pumping and shallow water-table dynamics, respectively.

All six springs are clustered at the northern center of the study area (Figure 1). The springs were sporadically measured for most of the study period, but more frequent measurements were made in the last decade. Spring discharge "flat-lining," or consecutive days of identical flows, is possibly due to procedures used to estimate daily data from direct measurements and were removed. The elevated flows around day 11,000 at Spring A are seen at other sites having data for this period and are not predicted by the ANNs. Spring B's high frequency variability during the last decade is not accurately predicted possibly due to more localized rainfall events not observed in any of the 18 NOAA rainfall gages.

The longer-term up and down trending at Lake A is accurately predicted using rainfall ANNs. At Lake B, the minimum water levels around day 22,000 are not predicted, and possibly indicate a period of high pumping during a sustained drought.

Monthly Groundwater Use ANNs - Limited improvement in prediction accuracy was gained by incorporation of groundwater use. The R^2 values for summed rainfall and usage ANNs (Table 2) are similar to those for rainfall ANNs (Table 1), but for monthly time steps. The average R^2 values for the groundwater use ANNs generally are low (Table 2), possibly because: the usage impacts are low most of the time at most sites, actual usages are not accurately represented in the largely estimated data, and(or) the variability in the rainfall ANN residuals manifest forcing that is not represented in the usage data or the ANNs. However, R^2 values tended to be higher for the springs, suggesting larger usage impacts. Limited measurement population precluded using testing data for usage ANNs.

Table 1. Statistics for rainfall ANNs. %Error = 100 * root mean square error/historical range.

	#Sites	#With Test Data	R^2 Max	R^2 Min	R^2 Avg	%Error Max	%Error Min	%Error Avg
Wells	23	23	0.91	0.56	0.82	13.6	4.8	7.1
Springs	6	2	0.78	0.49	0.63	10.7	6.4	9.1
Lakes	22	14	0.89	0.46	0.72	13.1	7.5	9.5

Table 2. Statistics for usage ANNs (Training R^2), and summed rainfall and usage ANNs (Sum R^2).

	#Sites	Training R^2 Max	Training R^2 Min	Training R^2 Avg	ANN Sum R^2 Max	ANN Sum R^2 Min	ANN Sum R^2 Avg
Wells	23	0.31	0.00	0.08	0.91	0.67	0.85
Springs	6	0.56	0.04	0.27	0.74	0.37	0.56
Lakes	22	0.35	0.00	0.12	0.90	0.32	0.72

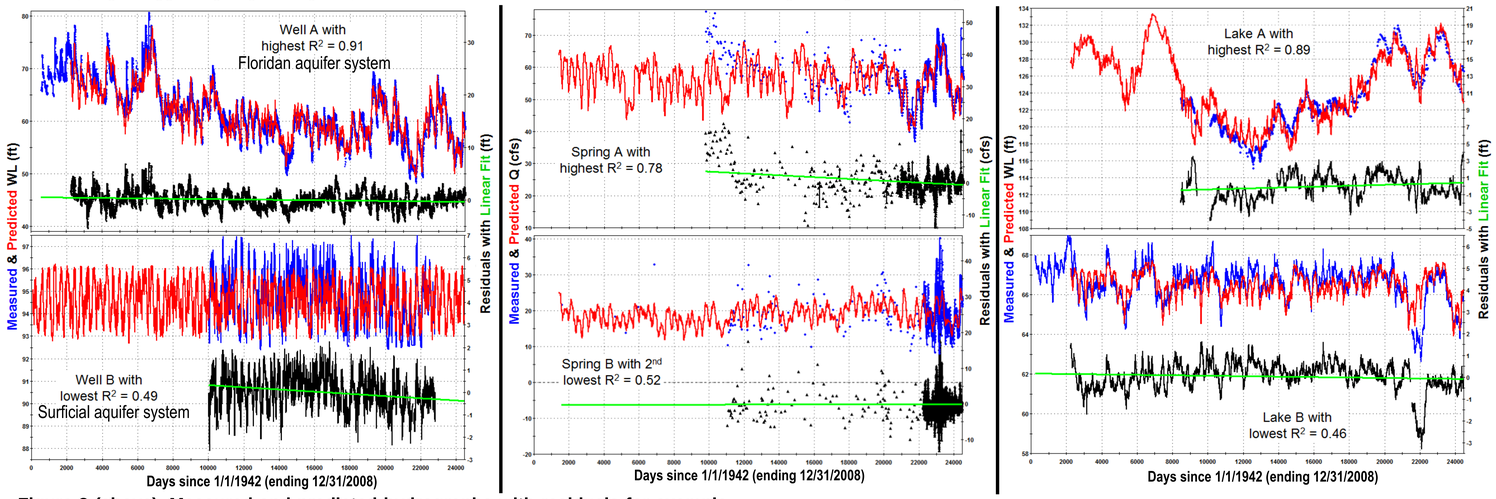

Figure 3 (above). Measured and predicted hydrographs with residuals for example wells, springs, and lakes based on daily rainfall ANNs. The well and lake examples are those having the highest and lowest R^2. The spring examples are those having the highest and second lowest R^2. Site locations shown in Figure 1.

Decision Support System (DSS)

A DSS was developed in Microsoft Excel™. It integrates the 102 rainfall and usage ANNs with the historical database, and provides user controls (Figure 4) and streaming graphics to allow users to run simulations having alternative rainfall and groundwater use scenarios (Figure 5). The DSS executes at a monthly time step from 1965 through 2008.

Conclusions

For nearly all sites, groundwater use was found to explain much less of the observed variability in hydrographs than climatic forcing, although relative groundwater use impacts are greater during droughts. These results may be affected by the relatively poor completeness and quality of the groundwater use data. Nevertheless, results indicate that consideration of both climate variability and groundwater use in predictions of future hydrologic system behavior would benefit the sustainable management of the resource. The ANN models were embedded in a DSS that will be distributed to resource managers and other stakeholders.

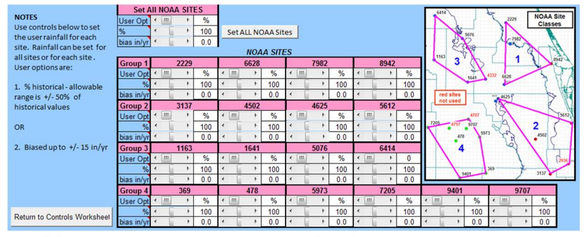

Figure 4. DSS rainfall set point controls. Rainfall data are modulated as either a percentage of historical values or using a constant bias. As shown in the map at right, the sites were grouped based on k-means clustering of 1,440-day moving window averages.

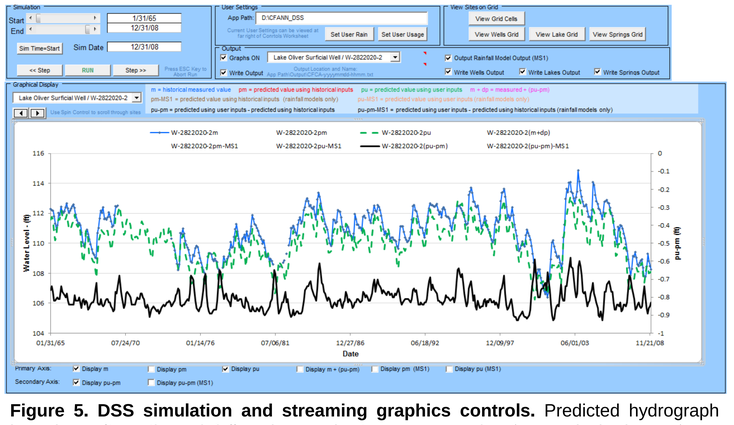

Figure 5. DSS simulation and streaming graphics controls. Predicted hydrograph based on alternative rainfall and groundwater use scenarios (green dashed curve) are visualized with the historical hydrograph (blue curve). Black curve indicates the difference between scenario and historical hydrographs.

Acknowledgments

The authors thank the St. Johns River Water Management District, South Florida Water Management District, and the Southwest Florida Water Management District for supporting this project. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Jensen, B.A., 1994, Expert systems - neural networks, Instrument Engineers' Handbook Third Edition: Chilton, Radnor PA, p. 48-54.

Roehl, E.A., Daamen, R.C., and Conrads, P.A., 2006, Features of advanced decision support systems for environmental studies, management, and regulation: *in* Voinov, A., Jakeman, A.J., Rizzoli, A.E., eds., Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modelling and Software, International Environmental Modelling and Software Society, Burlington, VT, July 2006.

Sepúlveda, Nicasio, Tiedeman, C.R., O'Reilly, A.M., Davis, J.B., and Burger, Patrick, 2012. Groundwater flow and water budget in the surficial and Floridan aquifer systems in east-central Florida: U.S. Geological Survey Scientific Investigations report 2012–5161, 214 p.

ADM's ADVANCED DATA MINING SERVICES

USGS science for a changing world