Forecasting Water User Demand at Beaufort-Jasper Water and Sewer Authority

Edwin A. Roehl Jr., Advanced Data Mining, LLC[1]
Terry Murray, Beaufort-Jasper Water and Sewer Authority[2]

ABSTRACT

Demand influences virtually all aspects a water utility, and is controlled by the choices customers make. It varies by time-of-day, day of the week, the season, the weather, and changes in the customer base. Long-term variability in demand causes similar variability in production operations, billing and collection, and revenues, which in turn makes planning and investment to improve service, control costs, and meet anticipated needs much more difficult and risky.

Beaufort-Jasper Water and Sewer Authority (BJWSA) operates a coastal service area that has experienced rapid growth, and from 1998 to 2002 the region saw unprecedented drought. Consequently, the utility made significant investments in infrastructure and expanded production. 2003 brought a significant drop in demand and revenue when the drought ended and was followed by protracted cooler, wetter weather. The authors were charged with developing a better understanding of the factors that influence BJWSA's demand. They studied ten years of historical data using data mining methods including statistics, artificial neural network models, and multidimensional data visualization. The study successfully quantified the sensitivity of demand to changing growth, rates, and meteorological forcing. It also found that significant time delays between causes and effects indicate that reliable forecasting six months into the future is possible. The authors describe the study's methods and findings, so that other utilities can benefit from the approach.

INTRODUCTION

Demand influences virtually every aspect of BJWSA's operations. It is controlled by customer choice, e.g., "*Should I wash the car today?",* and varies with time-of-day, day of the week, the season, the weather, rates, and changes in the customer base. Variable demand causes variable production and revenue, which together make management planning and decision making to operate the utility and make capital investments more difficult and risky. BJWSA's service area has experienced sustained rapid growth, and from 1998 to 2002 the region experienced unprecedented drought. Consequently, the utility made significant investments in infrastructure and expanded production. 2003 brought a significant drop in demand and revenue when the drought ended and cooler, wetter weather followed. A data mining[3] study was initiated to quantify the factors that influence demand and evaluate the possibility of accurately forecasting demand over different time horizons. This report describes the study's technical approach and results, and the opportunities identified for forecasting and managing risk.

[1] 3620 Pelham Road, PMB 351, Greenville, SC 29615, 864 201-8679, ed.roehl@advdatamining.com
[2] P.O. Drawer 2149, 6 Snake Road, Okatie, SC 29910, 843 987-9212, terrym@bjwsa.org

[3] Data mining is defined as the search for valuable knowledge in massive volumes of data (Weiss and Indurkhya). Its toolkit combines technologies such as signal processing, statistics, multi-dimensional visualization, machine learning from the field of Artificial Intelligence (AI), and Chaos Theory. Data mining can solve complex problems that are unsolvable by other means.

BACKGROUND AND APPROACH

While it was known that BJWSA's demand varied because of seasonal irrigation, the large decline in 2003 that coincided with extended cool wet weather was a bit of a surprise. Therefore, meteorological forcing was known to be an important cause of variability. The utility had also adjusted wholesale and retail rates a number of times during the study period, and a quantification of the sensitivity of demand to both weather and rate variability was sought[4].

Sensitivity analysis quantifies the relationships between a dependant variable of interest and causal variables, e.g., we know demand is somehow dependant on ambient temperature and precipitation. Computing sensitivities requires defining the relationships between variables through modeling. Models generally fall into one of two categories, deterministic and empirical. Deterministic models are created from first-principles equations, while empirical modeling adapts generalized mathematical functions to fit a line or surface through data from two or more variables. Calibrating either type of model attempts to optimally synthesize a line or surface through the observed data. Calibrating models is made difficult when data has substantial measurement error or is incomplete, and the variables for which data is available may only be able to provide a partial explanation of the causes of variability. The principal advantages that empirical models have over deterministic models are they can be developed much faster and are more accurate when the modeled systems are well characterized by data. However, empirical models are prone to problems when poorly applied. Overfitting and multicollinearity caused by correlated input variables can lead to invalid mappings between input and output variables (Roehl et al, 2003).

The most common empirical approach is ordinary least squares (OLS), which relates variables using straight lines, planes, or hyper-planes whether the actual relationships are linear or not. (Ballard, 2003) suggests, *"Given the changing nature of technology and the globalization of business and financial markets, it is becoming increasingly important to be able to more quickly and accurately predict trends and patterns in data in order to maintain competitiveness. More specifically, it is becoming increasingly important for forecasting models today to be able to detect nonlinear relationships while allowing for high levels of noisy data and chaotic components.*" He reviewed the use of artificial neural networks (ANNs), a "machine learning" technique from the field of AI, in several financial prediction applications including securities management, fraud detection, risk modeling, stock price forecasting, and forecasting macroeconomic variables such as GDP[5]. (Charytoniuk et al, 2000) described how ANNs can be used to forecast electric power demand in markets transitioning to deregulation. Their approach created different customer classes, modeled the historical demand of each class with ANNs, and then aggregated the predicted demands to produce a total demand estimate. (Jensen, 1994) details of the "multi-layer perceptron" (MLP) ANN, the type used in the applications described by Ballard, Charytoniuk et al, and this study. MLP ANNs can synthesize functions to fit high-dimension, non-linear multivariate data. (Devine et al, 2003; Conrads and Roehl, 2004) describe their use in multiple applications to model and control combined man-made and natural systems including disinfection byproduct formation, industrial air emissions monitoring, and surface water systems impacted by point and non-point source pollution.

[4] An analysis of the impact of rate variability was performed, but space limitations preclude its inclusion here.

[5] Many types of ANNs are used to solve different kinds of problems, and are parts of the data mining toolkit.

The *"chaotic components"* mentioned by Ballard alludes to the dynamic nature of variable relationships that change in time. Chaos Theory provides a conceptual framework called *"state space reconstruction"* (SSR) for representing dynamic relationships. Data collected at a point in time can be organized as a vector of measurements, e.g., element one of the vector might be the demand, element two the rainfall, and so on. Engineers will say that a process evolves from one state to another in time, and that a vector of measurements, a.k.a. a "state vector", represents the process' state at the moment the measurements were taken. A sequence of state vectors represents a "state history." Mathematicians will say that the state vector is a point in a "state space" having a number of dimensions equal to the number of elements in the vector, e.g., eight vector elements equates to eight dimensions. Empirical modeling is the fitting of a multidimensional surface to the points arrayed in state space.

Chaos Theory proposes that a process can be optimally represented (reconstructed) by a collection of state vectors Y(t) using an optimal number of measurements, equal to "local dimension" d_L, that are spaced in time by integer multiples of an optimal time delay τ_d (Abarbanel, 1996)[6]. For a multivariate process of k independent variables:

$$Y(t) = \{[x_1(t), x_1(t - \tau_{d1}),\ldots, x_1(t - (d_{L1} - 1)\tau_{d1})],\ldots.,[x_k(t), x_k(t - \tau_{dk}),\ldots, x_k(t - (d_{Lk} - 1)\tau_{dk})]\} \qquad \text{eq. 1}$$

where each $x(t,\tau_{di})$ represents a different dimension in state space, and therefore a different element in a state vector. Values of d_L and τ_d are estimated analytically or experimentally from the data. The mathematical formulations for models are derived from those for state vectors. To predict a dependent variable of interest y(t) from prior measurements (a.k.a. forecasting) of k independent variables (Roehl et al, 2000):

$$\begin{aligned} y(t) = F\{&[x_1(t - \tau_{p1}), x_1(t - \tau_{p1} - \tau_{d1}),\ldots, x_1(t - \tau_{p1} - (d_{M1} - 1)\tau_{d1})], \\ &\ldots.,[x_k(t - \tau_{pk}), x_k(t - \tau_{pk} - \tau_{dk}),\ldots, x_k(t - \tau_{pk} - (d_{Mk} - 1)\tau_{dk})]\} \qquad \text{eq. 2} \end{aligned}$$

where F is an empirical function such as an ANN, each $x(t,\tau_{pi},\tau_{di})$ is a different input to F, and τ_{pi} is yet another time delay. For each variable, τ_{pi} is either: constrained to the time delay at which an input variable becomes uncorrelated to all other inputs, but can still provide useful information about y(t); constrained to the time delay of the most recent available measurement of x_i; or the time delay at which an input variable is most highly correlated to y(t). Here, the state space local dimension d_L of Equation 1 is replaced with a model input variable dimension d_M, which is determined experimentally. $d_M \leq d_L$, and tends to decrease with increasing k.

RESULTS

The study period was January 1, 1994 to July 31, 2004 (3,865 days), which started with the onset of consistent data collection and archival at BJWSA. It straddles five years of record drought from 1998 to 2002, and El Ninos in 1998 and 2003 that brought sustained rains. Daily weather observations were obtained from the NOAA National Data Center for five stations in the region around BJWSA's service area. The observed variables were daily maximum and minimum ambient temperatures (T_{max}, T_{min}) and daily cumulative precipitation (Precip). The original intent was to evaluate how spatial as well as temporal weather variability affect demand; however, problems such as gaps and baseline shifts were found in all of the data. Data from Yemassee was

[6] In Chaos Theory, d_L and t_d are called "dynamical invariants", and are analogous to the amplitude, frequency, and phase angle of periodic time series.

least afflicted and deemed the only data set immediately useful[7]. The study evaluated forecasting 7, 30, and 90 to 180 days into the future. Because of space limitations, only the 90 to 180-day forecast modeling is described.

Correlation analysis quickly revealed that demand is highly correlated to T_{max}, somewhat less to Precip, and that T_{min} contributed very little information not accounted for in other variables. Figure 1 shows actual and *"standard"* 90-day MWA T_{max} (T_{max}S) and Precip (PrecipS). T_{max}S and PrecipS are calculated by computing an average value for each variable for each day of the year. Some summers and winters are warmer than others, and T_{max} is much less variable than Precip. Most years exhibit spring rains, followed by a dryer period, and then rainfall peaks in the latter half of the summer. Some years receive more rain than others. High winter rains of the El Ninos of 1998 and 2003 are apparent. In most years Precip peaks one to two months after T_{max} peaks. This is not true for 2003 when Precip rose unusually early. Also note that 2003 was cooler than the previous five years. The early Precip and lower T_{max} after five years of drought probably led customers to irrigate less in 2003.

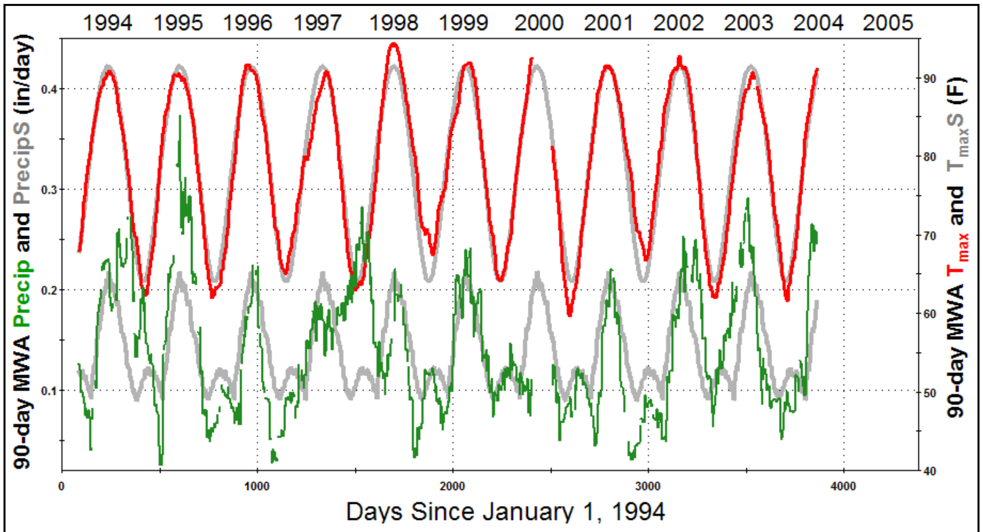

Figure 1: 90-day MWA Standards T_{max}S and PrecipS, and Actual of T_{max} and Precip. R^2 for T_{max}S vs. PrecipS = 0.71.

Figure 2 shows the 30-day MWA demand (Q). It shows dramatic growth between 1998 and 2000, the shape of the annual demand curve changes from year to year, and the increased peak-to-trough difference from 1999 onwards indicates increased demand for

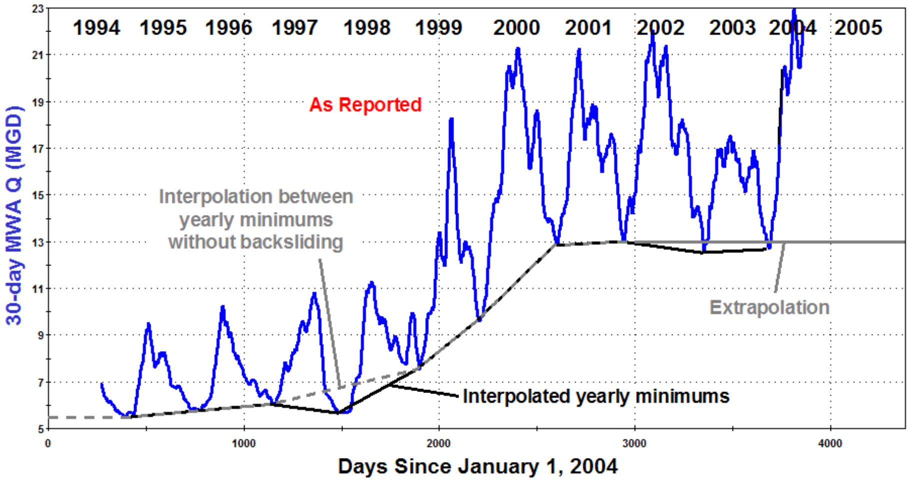

Figure 2: 30-day MWA demand and Baseline. Shown are interpolated yearly minimum demand, and interpolated yearly minimum demand without backsliding.

[7] Station-to-station correlations for each variable type were found to be high, indicating that errors and gaps at one station could be reconstructed by correlating to time series at other stations.

seasonal irrigation. The 2003 Q is also seen to be significantly lower than in neighboring years. Overlain onto Q are options for a "baseline demand" (Baseline), which is the minimum Q for the calendar year that occurs in late winter. The black line connects each year's minimum while the gray line (selected for this study) only changes when a new "high low" is observed (no backsliding).

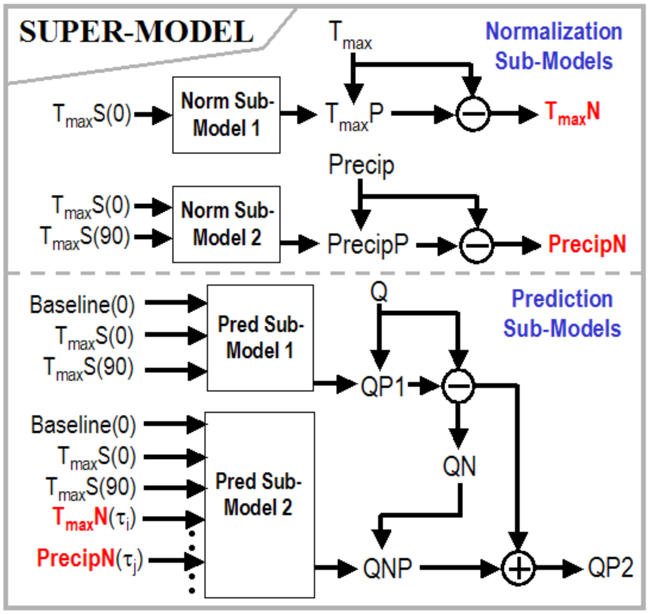

Figure 3: Super-model architecture used to predict demand QP2.

Figure 3 shows that the forecasting model would be a "super-model" comprised of four "sub-models", each having a specific purpose. Figure 4 shows demand predictions QP1 made by an ANN[8] (Prediction Sub-Model-1) using only the Baseline, $T_{max}S$, and PrecipS as inputs[9]. While the model is statistically accurate, QP1 does not track measured Q well through the decline in 2003. The black line is the model's prediction error, which is also the "normalized" demand (QN) after having Baseline, $T_{max}S$ and PrecipS "components" largely removed. Normalizing variables tends to amplify untypical behaviors for study. Figure 5 shows Q, and T_{max} ($T_{max}N$) and Precip (PrecipN) after being normalized using Normalization Sub-Models 1 and 2 respectively. Normalization Sub-Model 1[10] uses only $T_{max}S$ as an input. Normalization Sub-Model 2[11] also uses $T_{max}S$ as an input, but at two time delays, effectively decorrelating $T_{max}N$ and PrecipN. Annual peak Qs are marked with dotted lines for comparison by inspection to low $T_{max}N$ and PrecipN. The difficulty in ascertaining clear relationships between these variables underscores the need for creating a process model.

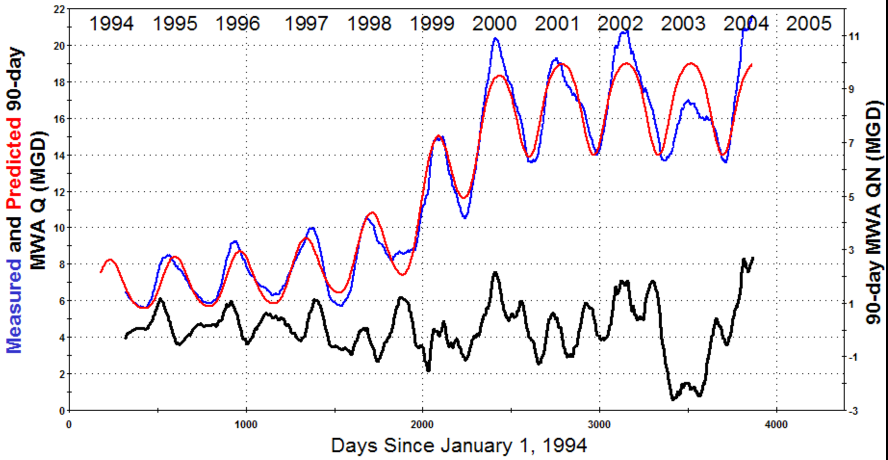

Figure 4: Measured, Predicted, and Normalized 90-day MWA Demand from 1994. Measured versus Predicted R^2 = 0.96 and root man square error (RMSE) = 0.95 MGD for the entire period, diminishing to 0.71 and 1.1 MGD for 2000 onward.

[8] Referring to Equation 2, k=3 input variables: Baseline, $T_{max}S$, and PrecipS. All τ_p=0, all d_M=1.

[9] The models and plots in this paper were generated using the iQuest^{TM} data mining software.

[10] k=1, τ_p=0, d_M=1

[11] k=1, τ_p=0, τ_d=90, d_M=2

Figure 6 shows cross correlation plots of QN versus $T_{max}N$ and PrecipN, which are created by time-stepping one time series relative to the other and calculating the Pearson coefficient R at each time step. The (+/-) signs of R confirm that QN increases with ambient $T_{max}N$ and decreases with PrecipN (sprinklers are on when it's hot and dry). With respect to $T_{max}N$, R^2 grows from a significant 0.17 at τ_p=0 to a large peak of 0.32 at τ_p=140 days. With respect to PrecipN, R^2 grows from an insignificant 0.01 at τ_p=0 to a large peak of 0.28 at τ_p=187 days.

These results show that large changes in prevailing weather patterns affect demand for a long time and that forecasting to a horizon or up to six months is possible.

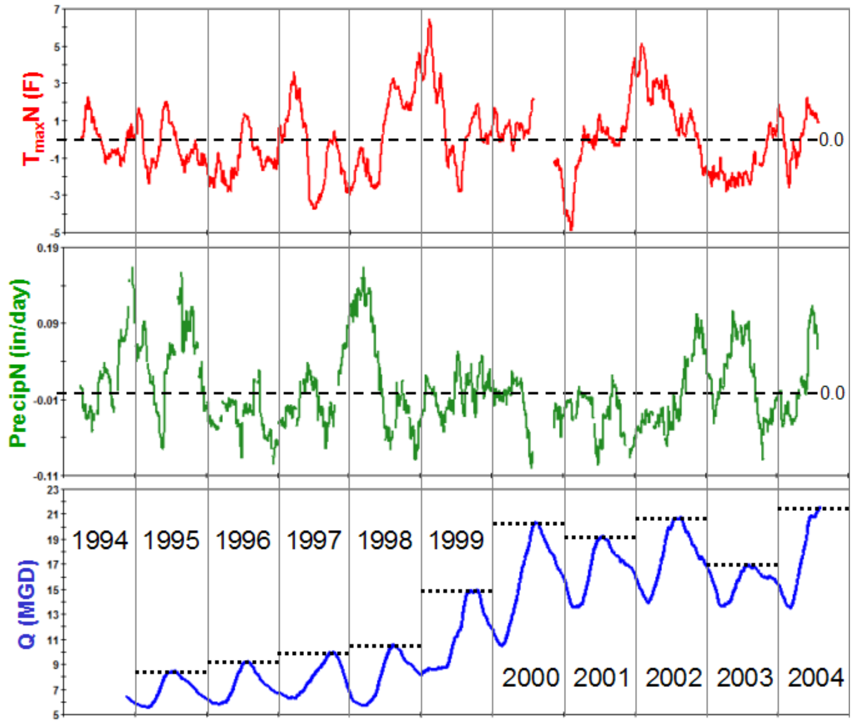

Figure 5: $T_{max}N$, PrecipN, and Q. R^2 for $T_{max}N$ vs. and PrecipN is 0.098 (previously 0.20 for T_{max} vs. Precip).

"Prediction Sub-Model 2" predicts the normalized demand QN (QNP) from inputs for Baseline, $T_{max}S$, $T_{max}N$, and PrecipN[12]. The $T_{max}S$ inputs provide time-of-year information that interacts with the normalized weather variables to boost prediction accuracy. The prediction sub-model outputs QP1 and QNP are summed to produce the final prediction QP2.

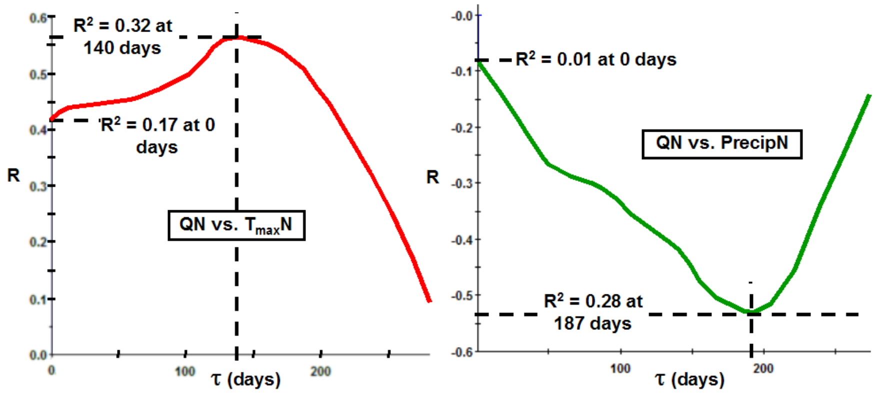

Figure 6: Cross correlation plots of QN versus $T_{max}N$ and PrecipN.

A model can be configured to forecast a parameter at a future time by shifting τ_{pi} and "retraining" the affected ANN sub-models. Here, only Prediction Sub-Model 2 uses real-time inputs and has to be retrained. Figures 7 and 8 show the predictions QNP and QP2 at τ_p= 0, 90, 120, 150, 180 days respectively from 2000 onward. The ANNs

[12] k=5 input variables: Baseline using τ_p=0, d_M=1; $T_{max}S$ using τ_p=0, τ_d=90, d_M=2; 30-day MWA $T_{max}N$ (all other inputs use 90-day MWA) using τ_p=0, d_M=1; $T_{max}N$ using τ_p=30, τ_d=90, d_M=2; and PrecipN using τ_p=190, d_M=1.

readily represent the conditions that led to the significant demand decline in 2003. The R^2 and RMSE (root mean squared error) for QNP were 0.71 and 1.1 MGD (million gallons/ day) for 2000 onward, which improved to 0.95 and 0.36 MGD for QP2. The errors of the forecast models ($\tau_p > 0$) at the beginning of the time series and adjacent to the large data gap indicate unlearned behavior because of missing data. The expected per-day forecast errors for τ_p=0, 90, 180 days are 0.36, 0.56, 0.60 MGD respectively. Over 90 and 180 days, the expected cumulative errors are ±50 and ±108 MG. For comparison, given an average flow of 17 MGD, the demand over 90 and 180 days would be 1,530 and 3,060 MG respectively, corresponding to error percents of 3.3% and 3.5% respectively. These high accuracies result from the seasonally periodic nature of demand, the very high correlation between demand and weather variables, and the use of ANNs in an SSR framework that can accommodate highly complex, nonlinear variable interactions.

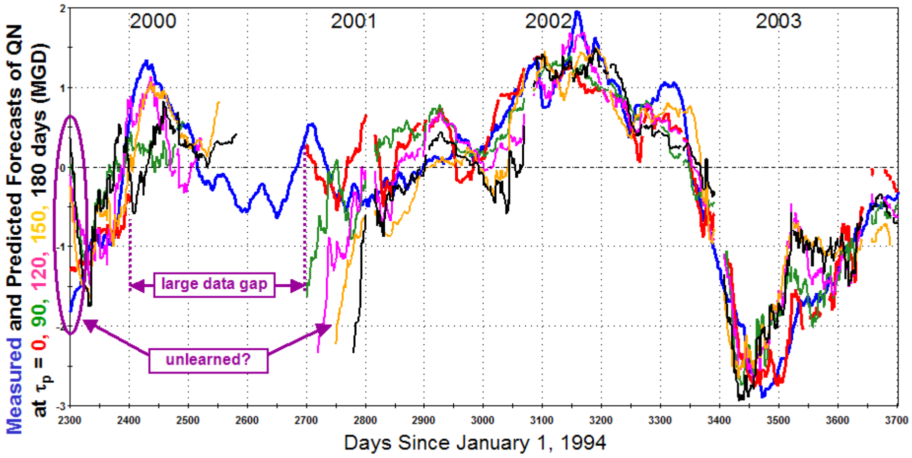

Figure 7: Measured, predicted, and forecast QN. At τ_d = 0, 90, 120, 150, 180 days, R^2 = 0.90, 0.77, 0.79, 0.80, 0.72 and RMSE = 0.36, 0.56, 0.52, 0.51, 0.60 MGD respectively.

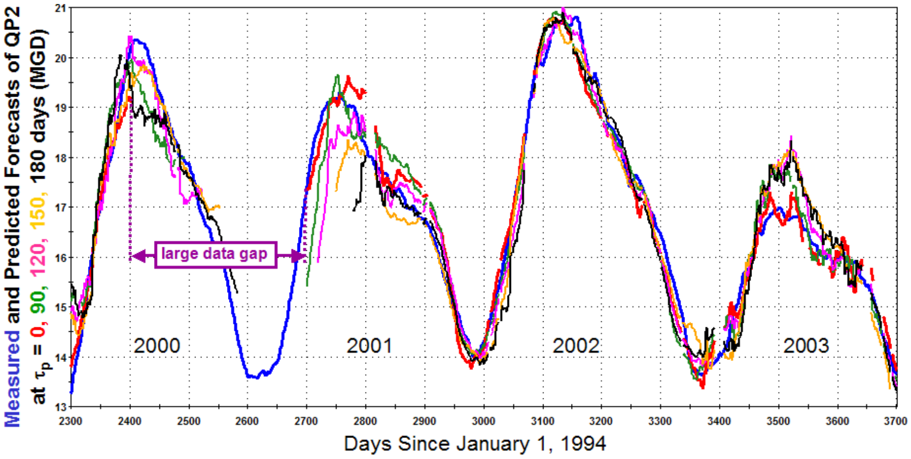

Figure 8: Measured, predicted, and forecast QP2. At t_d = 0, 90, 120, 150, 180 days, R^2 = 0.95, 0.87, 0.83, 0.84, and 0.83 respectively.

Figure 9 shows the "response surface" (Roehl et al, 2003) of Prediction Sub-Model 2, which reveals the functional form of its mapping of T_{max}N at τ_p=30 days and PrecipN at τ_p=190 days to QNP. Note that the response surface is non-linear (non-planar), and remember that this sub-model has seven inputs[10]. All but the variables selected for the horizontal axes, here T_{max}N and PrecipN, are "unshown". The values to which unshown variables are set affects the surface's shape. Here they correspond to mid summer: T_{max}S90 at τ_p=0 and 90 days = 78 and 90 F respectively; Baseline = 13 MGD; 30-day MWA T_{max}N = 0; and T_{max}N at τ_p=120 days = 0. The vertical range of QNP ≈ 4 MGD. The horizontal plane at QNP = 0.0 marks a boundary between above and below average demand. The response surface shows that demand QNP is greatest when conditions are warmest and hottest (a); demand is lowest when conditions are coolest and wettest (b); demand is less sensitive to $T_{max}N$ when PrecipN is low (c); and demand is less sensitive to PrecipN when $T_{max}N$ is high (d). No surprises, but reassuring to see that the functional relationships follow expectations and are now quantified.

CONCLUSIONS

This study determined that more than 90% of demand variability is attributable to the Baseline and weather, which are already being monitored. Prediction accuracy was significantly improved by augmenting the Baseline and standard weather variables with real-time weather inputs. The possibility of accurately forecasting demand three to six months into the future is supported by initially high model performance statistics that decline slowly as the prediction date is push forward into the future. This would allow a utility to forecast through late winter and spring to predict demand for the warm weather months having the greatest year-to-year variability. Further, the model's "What ifs?" capability, e.g., *"What if it stops raining...or starts raining...a lot?"* could make it a useful addition to a utility's risk management strategy.

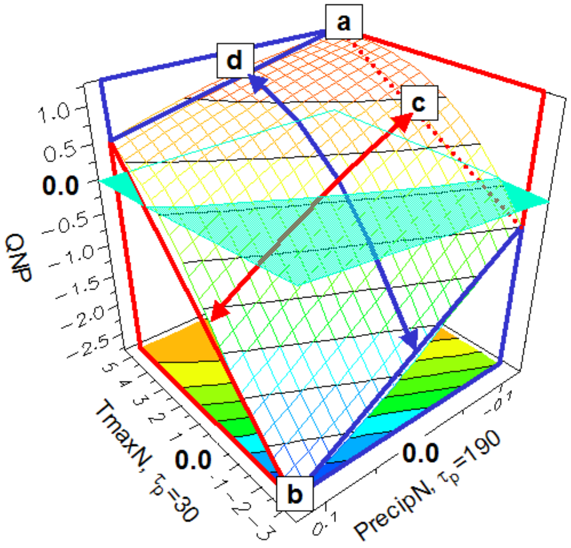

Figure 9: ANN model response surface.

REFERENCES

Abarbanel, H.D.I., 1996, Analysis of Observed Chaotic Data, Springer-Verlag New York, Inc., New York, 4-12, 39.

Ballard, R., 2003, "Forecasting with Neural Networks – A Review," National Social Science J., Feb. 24, 2003

Conrads, P.A. and Roehl, E.A., 2004, "Integration of Data Mining Techniques with Mechanistic Models to Determine the Impacts of Non-Point Source Loading on Dissolved Oxygen in Tidal Waters," South Carolina Environmental Conference, Myrtle Beach, March 2004.

Charytoniuk, W., Box, E.D., Lee, W.J., Chen, M.S., Kotas, P., and Van Olinda, P., 2000, "Neural-Network-Based Demand Forecasting in a Deregulated Environment," IEEE Transactions on Industry Applications, 36(3)

Devine, T.W., Roehl, E.A., and Busby, J.B., 2003, "Virtual Sensors - Cost Effective Monitoring," Air and Waste Management Association Annual Conference, June 2003

Jensen, B.A., 1994, Expert Systems - Neural Networks, Instrument Engineers' Handbook Third Edition, Chilton, Radnor PA.

Roehl, E.A., Conrads, P.A., and Roehl, T.A., 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, 947-954

Roehl, E.A., Conrads, P.A., and Cook, J.B., 2003, "Discussion of Using Complex Permittivity and Artificial Neural Networks for Contaminant Prediction," J. Env. Eng., Nov. 2003, pp. 1069-1071

Weiss, S.M. and Indurkhya, N., 1997, Predictive Data Mining: A Practical Guide, Morgan Kaufmann.