Non-Linear, Multivariate Forecasting of Hydrologic and Anthropogenic Responses to Meteorological Forcing

Edwin A. Roehl, Advanced Data Mining, LLC, Greer, SC, ed.roehl@advdatamining.com
Terry Murray, Beaufort-Jasper Water and Sewer Authority, Okatie, SC, USA, terrym@bjwsa.org

Abstract: Managers and users of natural resources often face two challenging problems. One is forecasting future natural system conditions for optimal resource allocation. Here, the natural system is comprised of the weather and a dependant hydrologic system that contains a water resource. The second problem is forecasting the behavior of a combined natural and man-made system, which also includes anthropogenic resource consumers. Even though detailed meteorological forecasting over weeks and months is impractical, hydrologic behaviors such as groundwater cycling can transpire over months and years. Alternatively, man-made systems exhibit behaviors that both lag and lead causal forcing, e.g., seasonal weather changes. This paper compares forecasting the behaviors of two systems. One is the upper Klamath Basin in Oregon and California where resource managers allocate water among competing interests, e.g., hydropower, farming, and fisheries. The second system is a water utility in coastal South Carolina, whose demand varies significantly with seasonal irrigation. A similar technical approach was used to model both systems, but the results are instructively dissimilar. The approach took signals of meteorological variables, basin inflows, and consumer demand and decomposed them into signal components that differentiated *standard* (seasonally periodic) behaviors from non-standard *chaotic* behaviors. Next, empirical process models were synthesized using *artificial neural networks* (ANN), a non-linear, multivariate curve fitting technique. The ANNs predicted chaotic output behaviors (basin inflow or consumer demand) from input chaotic meteorological signal components. Finally, prediction sensitivity to shifting the output forward in time relative to the inputs was determined. It was found that Klamath predictions decayed towards a predictability horizon of less than the minimum six-month forecast currently used by the water resource managers. Conversely, prediction accuracy of the part-anthropogenically driven water demand decayed far more slowly, easily straddling the critical six-month spring-to-fall irrigation season over which utility managers sought to forecast.

Keywords: forecast, hydrology, neural network

1. INTRODUCTION

When anthropogenic demand for a natural resource greatly exceeds the supply, environmental resource managers face the problem of allocating the resource among competing interests in a way that is both fair and environmentally protective. When the resource is water and the users are farmers, the farmers want to know how much water will be available before they plant to avoid monetary loss; however, the managers may not know the answer. Even though groundwater cycling can transpire over years, many hydrologic systems are affected predominantly by meteorological forcing on much shorter time scales.

An alternative problem occurs when there is ample supply and the demand is highly variable. It is usually desirable to know the demand ahead of time so that the resource can be provided in the most cost effective way. For example, a water utility might want to plan production and maintenance operations according to what is typically the most variable component of demand - water for irrigation. However, anthropogenic systems can exhibit behaviors that both lag and lead causal forcing, e.g., people can anticipate seasonal weather changes but not unusually dry or wet weather.

To solve either problem, the resource managers or the utility would need to forecast water availability or demand with meaningful accuracy. This paper compares forecast modeling of two systems. One is the upper Klamath Basin in Oregon and California where resource managers allocate water among hydropower, farming, fisheries, and other uses. The resource managers currently use statistical forecast models that predict six months or more into the

future. The second system is a water utility in coastal South Carolina, Beaufort-Jasper Water and Sewer Authority (BJWSA), whose demand varies with seasonal irrigation. The utility sought to determine the cause of a sudden drop in demand and revenue after several years of rapid growth. Similar technical approaches were used to model both systems, but the results are instructively dissimilar.

2. FORECAST MODELING

Sensitivity analysis quantifies the relationships between a dependant variable of interest and causal variables, e.g., we know water demand is somehow dependant on ambient temperature and precipitation. Computing sensitivities requires defining the relationships between variables with models, which are either deterministic, empirical, or both. 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 attempts to optimally synthesize a line or surface through the observed data. This is more difficult when the data is noisy or 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, Roehl et al [2003] describe how empirical models are prone to problems when poorly applied, such as overfitting and incorrect mappings caused by correlated input variables.

The most common empirical technique is ordinary least squares (OLS), which relates variables using straight lines, planes, or hyper-planes whether the actual relationships are linear or not. In a review of using of *artificial neural networks* (ANNs) in several financial applications, Ballard [2003] states, *"Given the changing nature of technology… 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."* Charytoniuk et al [2000] described how ANNs can be used to forecast electric power demand. Jensen [1994] details the *multi-layer perceptron* (MLP) ANN, the type used in the applications described by Ballard and Charytoniuk et al. MLP ANNs can synthesize non-linear functions to fit multivariate data.

Ballard's *nonlinear relationships* (among variables) and *chaotic components* (of variables) suggests a need for forecasting approaches that can handle the complex, dynamic variable relationships. 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 a flow, element two the rainfall, and so on. As a process evolves from one state to another in time, *state vectors* represent the process' behaviors. A sequence of vectors represents a *state history*. A state vector can be a coordinate in a *state space* having dimensions for each vector element. Empirical modeling is the fitting of a multidimensional surface to the points arrayed in state space.

Abarbanel [1996] describes how process behavior can be optimally 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. The expression below describes a multivariate process of k independent variables, where each $x_k(t,\tau_{di})$ represents a different dimension in state space.

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

For univariate systems characterized by a single signal, Abarbanel suggests estimating τ_d using the *first minima of the average mutual information function*. Other techniques include using the *first zero crossing of the autocorrelation function* and *peak-to-peak intervals*. When a complex signal is decomposed into simpler components using spectral filtering, these techniques will give similar estimates of τ_d for each component. Abarbanel suggests the *local false nearest neighbors test* to estimate d_L. It uses an empirical function, i.e., linear or quadratic, to map prior measurements to the next measurement. Thus, d_L is determined experimentally and equals the number of prior measurements that parsimoniously gives the best prediction.

The following expression predicts $y_p(t)$ of a measured dependent variable of interest y(t) from prior measurements (a.k.a. forecasting) of k independent variables, where F is an empirical function, hereafter an ANN.

$y_p(t)=F\{[x_1(t-\tau_{p1}),x_1(t-\tau_{p1}-\tau_{d1}),\ldots,x_1(t-\tau_{p1}-(d_{M1}-1)\tau_{d1})],\ldots,[x_k(t-\tau_{pk}),x_k(t-\tau_{pk}-\tau_{dk}),\ldots,x_k(t-\tau_{pk}-(d_{Mk}-1)\tau_{dk})]\}$

Each $x_k(t,\tau_{pi},\tau_{di})$ 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. y(t) can be a superposition of disparate behaviors $y_j(t)$ originating from different forcing functions, such that $y_p(t) = \Sigma y_{pj}(t) = \Sigma F_j$.

3. FORECASTING KLAMATH INFLOWS

Risley et al [2005] describe how water resource managers in the upper Klamath Basin, located in south-central Oregon and northeastern California, must annually allocate limited water supplies among competing demands such as farm irrigation, endangered fish habitat, and hydropower. They rely on 6-month or longer forecasts of the total flow into Klamath Lake for the irrigation season ending October 1. The forecasts are produced monthly from January through May from linear principle component regression models that use measured snow-water equivalent and precipitation signals from several sites as inputs. The models predict the total volume of water that will pass five key flow gages upstream of the lake. In recent years, inaccurate forecasts have undermined confidence in the water allocation process, leading to a study to improve the forecasts using ANNs.

The data available to forecast streamflow Q at one of the gages (Sprague) included: daily precipitation (P) from 15 locations; daily snow-water equivalent (S) from 17 locations; and daily maximum and minimum air temperatures (T) from 15 locations. The daily measurements were transformed into 13-week moving window averages (MWA) to remove high frequency variability for forecasting flows aggregated over several months. Figure 1 shows Q from 1978-2003. Its outstanding feature is an apparent 14-year sinusoidal component represented here as f_{sine}. Also shown is a periodic *standard* component Q_{stand} calculated as the average flow for each day of the year. The Pacific Decadal Oscillation index (PDO) is a sometimes-useful long-term indicator of meteorological trends. It was found that the PDO and f_{sine} were poorly correlated and that PDO was not useful in this application.

Figure 2 shows the predicted $Q_{p1}=F_1(f_{sine},Q_{stand})$, where F_1 is an ANN. R^2=0.64. Also shown is the *residual* error Q_r= measured – predicted flow. Consider that: f_{sine} and Q_{stand} represent average long-term and seasonal behavioral components; Q_r is a chaotic component that represents non-periodic meteorological forcing; seasonal periodicity aside, weather is unpredictable for the forecast horizons of interest here; there must be a significant delayed response of Q to P and S for these variables to be useful in forecasting.

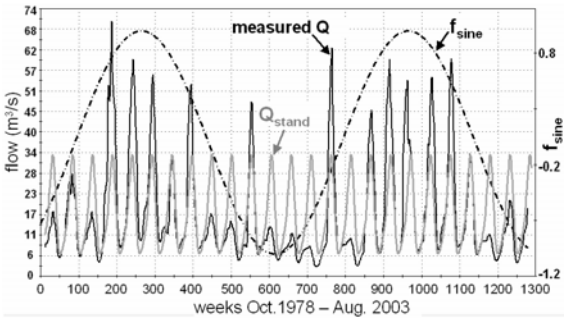

Figure 1. Measured Q, Q_{stand}, and f_{sine}

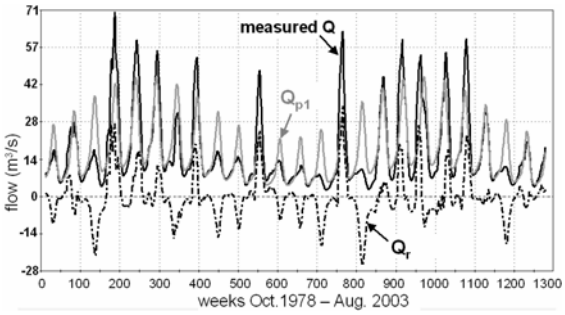

Figure 2. Measured Q, Q_{p1}, and Q_r.

The meteorological variables were *normalized* by subtracting standard components that were calculated for each of them, to produce P_{ni}, S_{nj}, and T_{nk}. A cross-correlation matrix was generated in which coefficients of determination R^2 were calculated for all variable pairings. Q_r was included in the matrix at τ_p spaced at 2 to 4 week intervals. The matrix indicated that: all of the normalized meteorological variables were highly correlated, suggesting a dearth of unique information among them; the residual was more highly correlated to P_{ni}

and S_{nj} than to T_{nk}; and correlations peaked at τ_p up to 10 weeks. Figure 3 shows how the R^2 varied with τ_p when Q_r was linearly correlated to the most influential meteorological variables. Note that the peak R^2 is only 0.36, that R^2 decays more slowly with P than S and that R^2 falls below 0.1 for all five variables before 24 weeks.

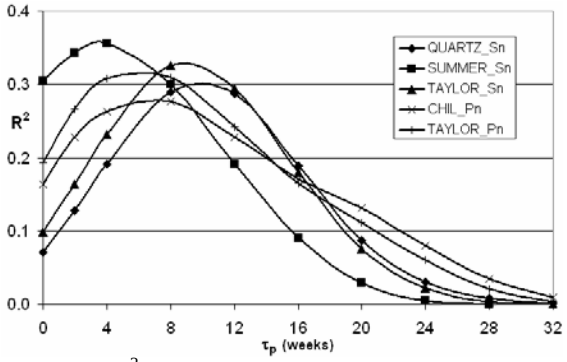

Figure 3. R^2 of the Q_r versus normalized S and P at increasing τ_p at sites QUARTZ, SUMMER, TAYLOR, and CHIL.

$Q_{p2} = F_2(P_{ni},S_{nj})$ was iteratively developed to predict Q_r. In its final form it used inputs CHIL_P_n at τ_p=8 weeks with τ_d=13, and QUARTZ_S_n at τ_p=12 weeks with τ_d=13. With two inputs per meteorological variable, d_M=2 for both CHIL_P_n and QUARTZ_S_n. τ_d is usually ¼-year for systems dominated by seasonal forcing, which can be verified using autocorrelation. CHIL_P_n was selected first because its correlation to Q_r decayed the most slowly. Placing the residual error from a prototype ANN, which used only CHIL_P_n inputs, in a cross correlation matrix indicated the use of QUARTZ_S_n and its delay. Subsequent attempts at using other meteorological inputs were unfruitful. The combined prediction was $Q_p=Q_{p1}+Q_{p2}$.

Figure 4 shows how Q_{p2} and Q_p R^2s decay with τ_p. Figure 5 compares Q_{p2} at τ_p=0 and 24 weeks to Q_r. For up to 8 weeks R^2 remains constant as the ANN input delays are shortened to compensate for increasing τ_p. Beyond 8 weeks the R^2 of Q_{p2} declines quickly, such that beyond 24 weeks R^2<0.1. Similarly, the R^2 of Q_p asymptotically approaches the delay-insensitive R^2=0.64 of Q_{P1}. The correlations shown in Figure 3 were typical of those at the other four inflows.

The study also evaluated auto-regression ANNs that used prior measurements of flow to forecast future flows. While more accurate at shorter τ_p, they were overtaken at longer delays by the models that used meteorological inputs because of the delayed responses of Q to P and S. It was also known from field studies that the area's groundwater cycling transpired over multiple years. Concerted attempts to use inputs representing much longer behaviors were fruitless, suggesting that the long-term groundwater contribution to inflow is small.

These results indicate that at best there is only a small benefit in using meteorological variables to forecast flows into the upper Klamanth Basin at horizons of six months and beyond. Revisiting Figure 1, while the statistical accuracy of forecasts over time might seem reasonable, in any given year a forecast could be quite poor.

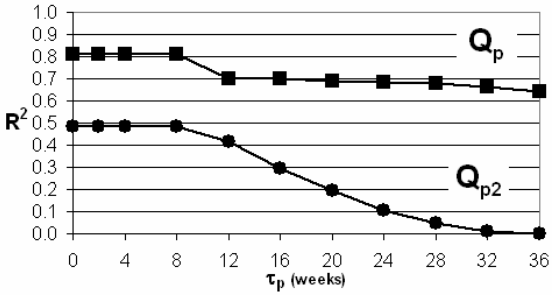

Figure 4. R^2 of Q_{p2} and Q_P versus τ_p.

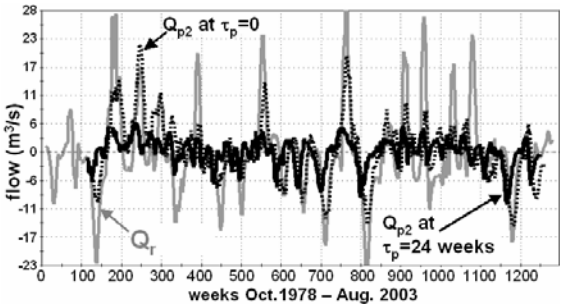

Figure 5. Q_r and Q_{p2} at τ_p=0 and 24 weeks.

4. FORECASTING BJWSA DEMAND

Demand (Q) influences virtually every aspect of BJWSA's operations. It is controlled by customer choice and varies with time-of-day, day of the week, the season, the weather, and changes in the customer base. Variability in Q causes variability in production and revenue, which together make management planning and decision making more difficult. As shown in Figure 6, BJWSA saw a rapid increase in Q as its service area grew. At the same time its region experienced a record drought from 1998 to 2002. Consequently, the utility made significant investments and expanded production.

2003 brought a significant drop in Q and revenue when the drought ended and cooler, wetter weather followed. A study was undertaken to quantify the factors that influence Q and evaluate the possibility of accurately forecasting demand.

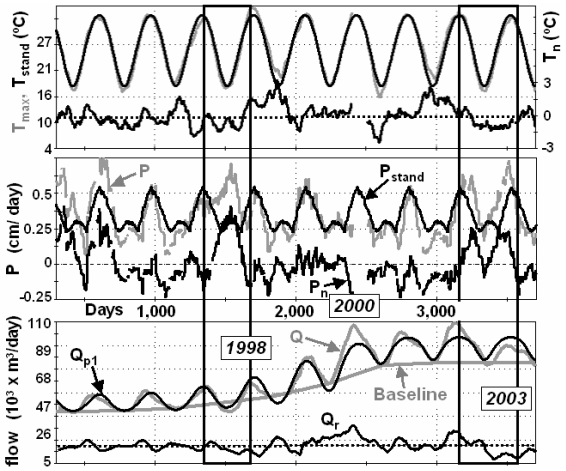

Figure 6. T_{max}, T_{stand}, T_n, P, P_{stand}, P_n, Q, Baseline, Q_{p1}, and Q_n. Boxes mark El Ninos (note P_n).

The study period was January 1994 through February 2004, which started at the onset of consistent data collection at BJWSA. It straddles the drought and El Ninos in 1998 and 2003, which brought sustained rains. Daily weather observations were obtained from the NOAA National Data Center. The observed variables were the daily cumulative precipitation (P) and the daily maximum and minimum air temperatures (T_{max}, T_{min}). An initial correlation analysis revealed that that T_{min} contributed very little information not manifest in P and T_{max}.

Figure 6 shows the results of processing the signals in a manner similar to that used in the Klamath study. 90-day MWAs were applied and *standard* and *normalized* components calculated to produce T_{max}, T_{stand}, T_n and P, P_{stand}, P_n. A 90-day MWA was also applied to Q. A demand Baseline was computed by interpolating each year's lowest Q, occuring in late February or early March, but not allowing backsliding. Note that some summers and winters are warmer than others, and T_{max} is much less variable than P. 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, especially during the El Ninos. In most years P peaks one to two months after T_{max} peaks. This was not true in 2003 when P rose unusually early. Also note that 2003 was cooler than the previous five years. The early P and lower T_{max} after five years of drought probably led customers to irrigate less in 2003.

Figure 6 shows the predicted $Q_{p1}=F_1(T_{stand},\text{Baseline})$, with two T_{stand} inputs at $\tau_p=0$ and $\tau_d=13$ weeks. $R^2=0.95$. Also shown is the Q_{p1} residual error Q_r. Q_r represents the normalized Q, which is seen to exhibit a time-lagged inverse reponse to the El Ninos. A sensitivity analysis conducted with Q_{p1} found meteorological forcing had a proportionately greater affect on Q at the recent higher Baseline, which was consistent with a dramatic increase in the number of residential customers, who typically water lawns. Therefore, it was decided to focus on the behaviors of greater commercial interest from 2000-on at the sacrifice of having only fours years of data to work with.

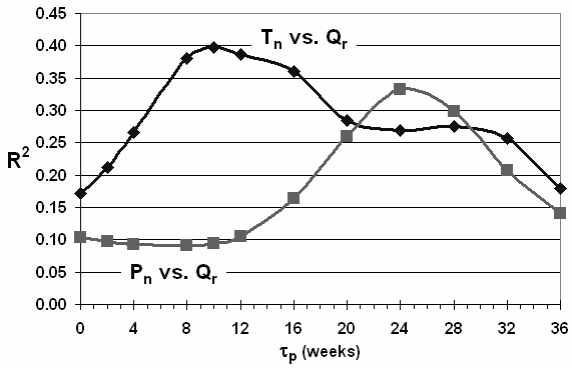

Figure 7. Cross correlation plots of T_n and P_n versus Q_r.

Q_{p1} was retrained with the shorter dataset and omitting the Baseline input. The new model's $R^2=0.68$ was lower because its statistic did not reflect predicting the 10-year high-range/low-noise Q baseline. Figure 7 shows cross correlation plots of T_n and P_n versus Q_r, where Q_r was the residual error of Q_{p1}. The peak T_n $R^2=0.39$ at $\tau_p=10$ weeks, and remains above 0.25 beyond six months. The peak P_n $R^2=0.33$ at $\tau_p=24$ weeks. Unlike the Klamath results shown in Figure 3, it is apparent that meteorology has a strong, long-delayed, and lasting impact on BJWSA's demand.

$Q_{p2}=F_2(T_n,P_n,T_{stand})$ was developed to predict Q_r. Per Figure 7, the ANN used inputs T_n at $\tau_p=10$ weeks and P_n at $\tau_p=24$ weeks. It also used T_{stand} inputs at $\tau_p=0$ and $\tau_d=13$ weeks to indicate the time of year. The combined prediction was $Q_p=Q_{p1}+Q_{p2}$. Figure 8 shows how Q_{p2} and Q_p R^2s decay with τ_p. Figure 9 compares Q_p at $\tau_p=12$ and 24 weeks to the measured Q. Like the Klamath, for up to 8 weeks R^2 remains constant as the ANN input delays are shortened to

compensate for increasing τ_p. Unlike the Klamath, where beyond 8 weeks the R^2 of Q_{p2} declined quickly, BJWSA's Q decays slowly out to 28 weeks. Q_p will approach the delay-insensitive R^2=0.68 of Q_{P1} as the R^2 of Q_{p2} approaches 0.

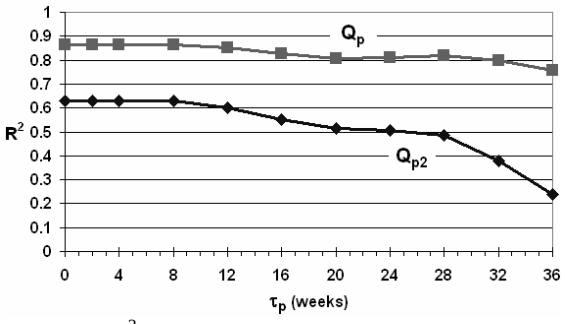

Figure 8. R^2 of Q_{p2} and Q_P versus τ_p.

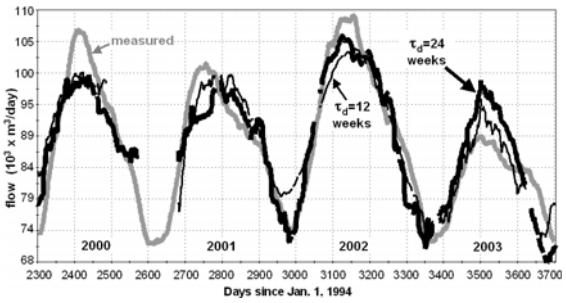

Figure 9. Measured Q and Q_p at τ_p=12 and 24 weeks. Q_p at τ_p=0 was essentially the same as that at 12 weeks.

CONCLUSIONS

Forecasting the behaviors of two systems six months or more into the future was evaluated using multi-layer perceptron artificial neural network models, whose input-output architectures were defined by dynamical behaviors revealed through multivariate state space reconstruction. It was found in the purely natural system of Klamath inflows that the available meteorological data cannot support accurate six-month forecasts. It was also found that in any given year a forecast could be off by a wide margin. The primary reason for this is that the system's process physics causes the vast majority of the water that comes into the system to flow past the gages inside the forecast horizon.

Unlike the Klamath system, the second system combined a natural system with anthropogenic resource consumers. It was found that meteorological forcing had a strong, long-delayed, and lasting impact on consumer demand that made it feasible to accurately forecast 90-day average demand six months or more into the future. The reason for this is that people generally reduce irrigation when nature is doing for them, and tend to turn the water back on only when they perceive a need for it.

5. ACKNOWLEDGEMENT

The authors wish to sincerely thank John Risley and the U.S. Geological Survey in Portland, Oregon for providing the Klamath data and background information.

6. REFERENCES

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

Ballard, R., Forecasting with neural networks – a review, *National Social Science J.*, Feb. 24, 2003

Charytoniuk, W., E.D. Box, W.J. Lee, M.S. Chen, P. Kotas., and P. Van Olinda, Neural-network-based demand forecasting in a deregulated environment, *IEEE Transactions on Industry Applications*, 36(3), 2003

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

Risley, J.C., M.W. Gannett, J.K. Lea, and E.A. Roehl, An analysis of statistical methods for seasonal flow forecasting in the upper Klamath River basin of Oregon and California, U.S. Geological Survey Scientific Investigations Report 2005-5177, 2005.

Roehl, E.A., P.A. Conrads and J.B. Cook, Discussion of "Using complex permittivity and artificial neural networks for contaminant prediction," *Journal of Environmental Engineering*, 1069-107, November 2003.