Development of a Neural Network Model for Dissolved Oxygen in the Tualatin River, Oregon

By Stewart A. Rounds, U.S. Geological Survey, Oregon Water Science Center, Portland, Oregon

Abstract

Many of the processes that affect dissolved oxygen concentrations in the Tualatin River — solubility, sediment oxygen demand, photosynthesis, respiration, biochemical oxygen demand, and reaeration — are controlled to some extent by physical and meteorological factors such as streamflow, air temperature, and solar radiation. To test the extent of that control, an artificial neural network model was constructed to predict dissolved oxygen concentrations in the Tualatin River at the Oswego Dam using only air temperature, solar radiation, and streamflow as inputs. Hourly dissolved oxygen concentrations have been collected at the Oswego Dam since 1991; the available dataset spans more than 10 years.

Feedforward neural network modeling techniques, the most widely used type, were applied to this dataset. Data were segregated into calibration, verification, and test subsets. Two neural network models were constructed in series: the first model simulated daily mean dissolved oxygen concentrations, while the second superimposed the daily periodic signals. The final calibrated neural network models predicted the dissolved oxygen concentration with acceptable accuracy, producing high correlations between measured and predicted values (r=0.83, mean absolute error < 0.9 mg/L).

By some measures, neural network model performance was better than that of a calibrated, mechanistic model of dissolved oxygen in the Tualatin River. As expected, however, dissolved oxygen concentrations affected by factors other than the physical and meteorological factors used as model inputs, such as large point-source ammonia releases, were not predicted well by the neural network model. Nevertheless, the neural network model demonstrated potential for use as a river management and forecasting tool to predict the effects of flow augmentation and near-term weather conditions on Tualatin River dissolved oxygen concentrations.

Factors Affecting Dissolved Oxygen

Dissolved oxygen concentrations in the Tualatin River (fig. 1) are affected by many physical factors and biological processes:

- Solubility
- Residence time
- Reaeration
- Algal respiration
- Photosynthesis
- Oxygen consuming reactions (BOD, SOD)

In addition, physical and meteorological factors such as temperature and residence time influence the effects of the biological processes.

Photosynthesis and respiration affect DO only when sufficient light energy is available and when streamflow is low enough (< 8.5 m^3/s) to allow sufficient time for the phytoplankton to grow while in the backwater reach (fig. 2).

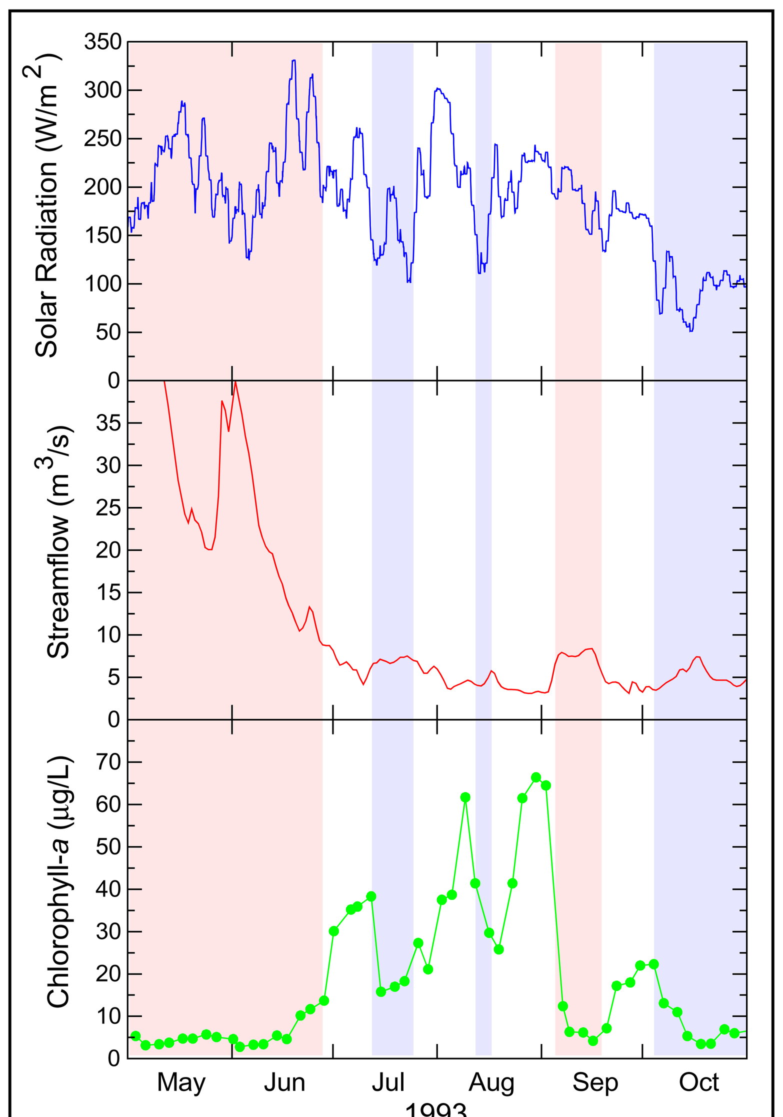
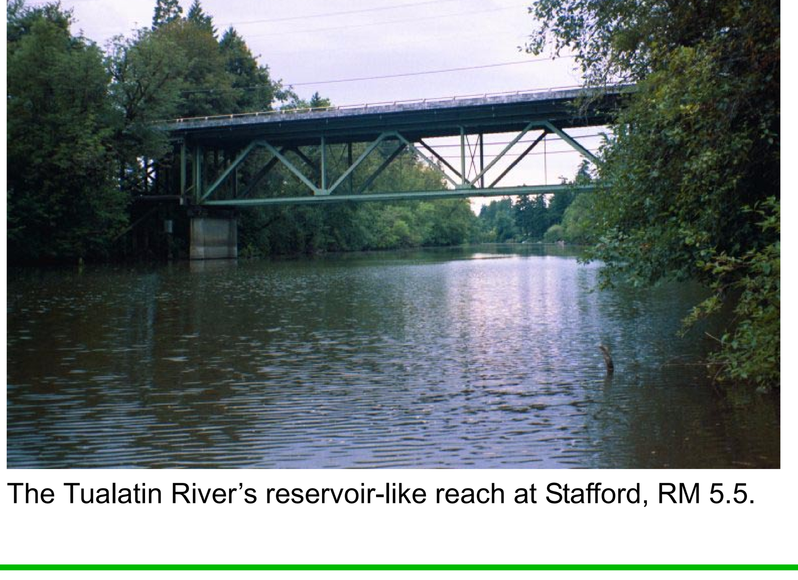


Figure 2. Favorable streamflow and light conditions are necessary before sizable algal blooms can occur in the Tualatin River. Shaded periods are unfavorable for algal growth due to high flow (red) or low light (blue) conditions.



The Tualatin River's reservoir-like reach at Stafford, RM 5.5.

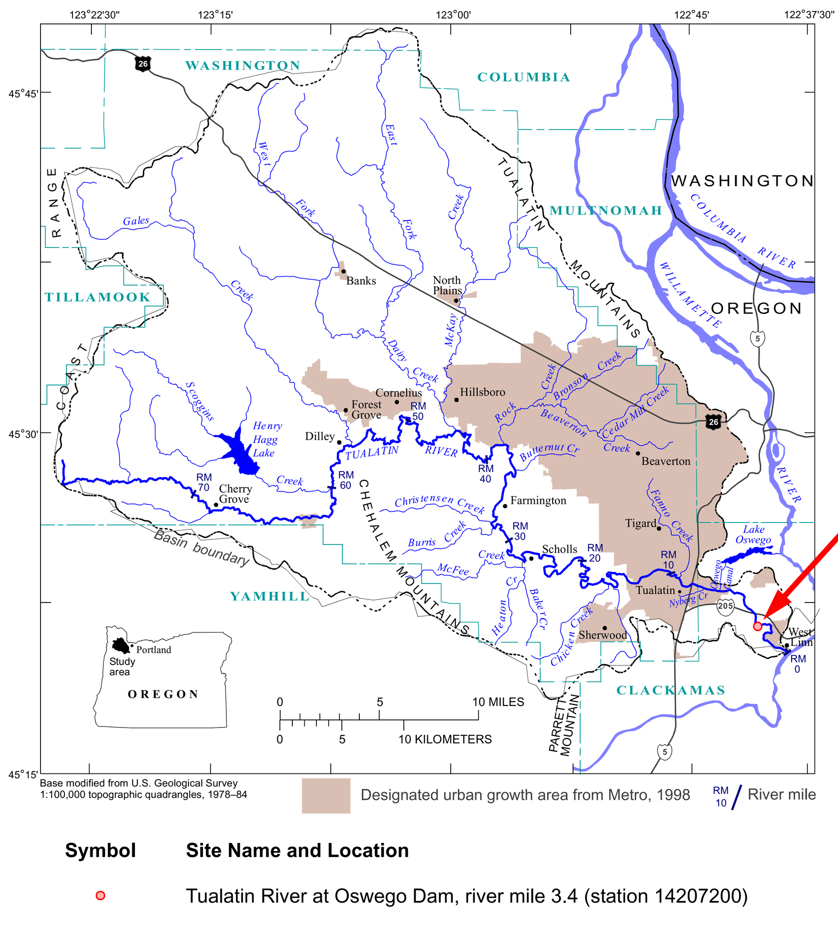
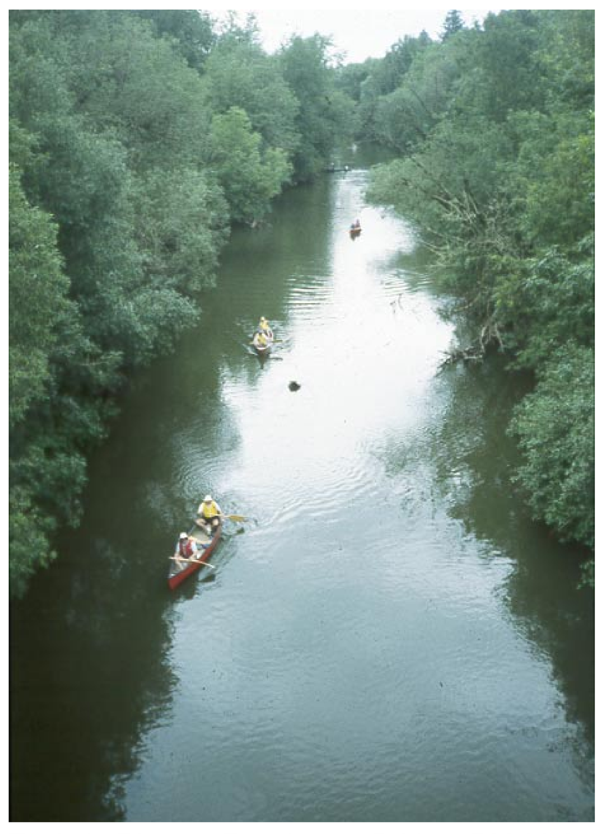


Figure 1. Map of Tualatin River Basin.

The Tualatin River at Oswego Dam, river mile (RM) 3.4.



Tualatin River at Farmington Bridge, RM 33.3.

Objectives and Approach

The purpose of this study was to determine the extent to which the DO concentration in the Tualatin River at the Oswego Dam could be predicted solely from physical and meteorological measurements such as streamflow, air temperature, solar radiation, and rainfall, using multiple linear regression and artificial neural network modeling techniques. If successful, these models would be used to create a real-time DO forecasting tool.

Artificial Neural Networks

An artificial neural network (ANN) is a mathematical structure designed to mimic the information processing functions of a network of neurons in the brain. ANNs are highly parallel systems that process information through many interconnected units that respond to inputs through modifiable weights, thresholds, and mathematical transfer functions. Each unit processes the pattern of activity it receives from other units, then broadcasts its response to still other units.

ANNs are particularly well suited for:

- large datasets
- complex, nonlinear relations
- pattern recognition

ANNs are able to find and identify complex patterns in datasets that may not be well described by a set of known processes or simple mathematical formulae. They are not constrained by any preconceived algorithms or relations among inputs.

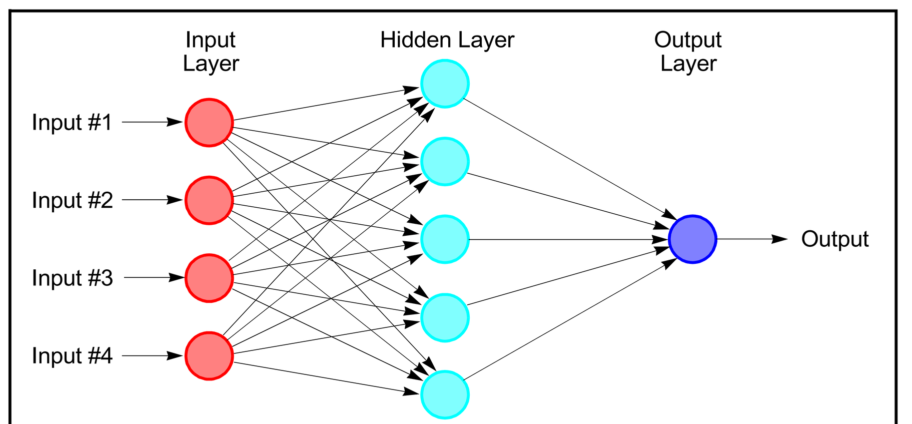


Figure 3. A representation of a 3-layer feedforward artificial neural network with four inputs, 5 hidden nodes, and one output.

Training an ANN is a mathematical exercise that optimizes all of the ANN's weights and threshold values, using some fraction of the available data. Optimization routines can be used to determine the ideal number of units in the hidden layer and the nature of their transfer functions. ANNs "learn" by example; as long as the input dataset contains a wide range of the types of patterns that the ANN will be asked to predict, the model is likely to find those patterns and successfully use them in its predictions.

Table 1. Goodness-of-fit statistics for models predicting DO at the Oswego Dam.

Model Type	Time Scale	Number of Data Points	Mean Absolute Error (mg/L)	Root Mean Square Error (mg/L)	Correlation Coefficient (r)
Multiple Linear Regression	low-frequency	40,388	1.29	1.69	0.589
Artificial Neural Network (ANN)	low-frequency	40,388	0.83	1.14	0.837
	final hourly	40,372	0.86	1.21	0.831

Data Preparation and Decorrelation

To maximize the signals in the input data that will help to predict the output, it is critical to examine the data for periodicity, cross-correlations, and important time lags.

Periodicity

Each parameter's data were analyzed by Fourier transform to determine the presence of periodic signals. Solar radiation, air temperature, and DO all had strong periodic signals at daily time scales; periods of 24 and 12 hours characterized the most important signals. Streamflow appeared to have useful signals at time scales longer than a day or two, but only weak patterns at shorter time scales. Figure 4 illustrates typical power spectrums from these data.

Strong signals at daily time scales can obscure important correlations and time lags in the data; therefore, the short and long time scale signals in the data were separated. A low-pass filter was used to remove the 24-hour and shorter periodic signals from each time series, preserving any periodic signals at time scales longer than one day.

Long-term patterns and short-term periodicity in the data were simulated with separate models.

Cross-Correlations and Time Lags

Multiple linear regression and ANN techniques work best with independent inputs. To test for interdependence, the data were correlated against one another using linear regression techniques with an imposed time lag (fig. 5).

DO has its highest correlation with the solar insolation rate that occurred about 2 days previous. That time lag has a physical basis because the available solar energy affects the amount of DO produced by photosynthesis, and the effects of very sunny or very cloudy days on algal growth are not immediate.

Many of the DO cross-correlations are minimized at a time lag on the order of 12 days, which is the typical summer residence time in the backwater reach of the Tualatin River.

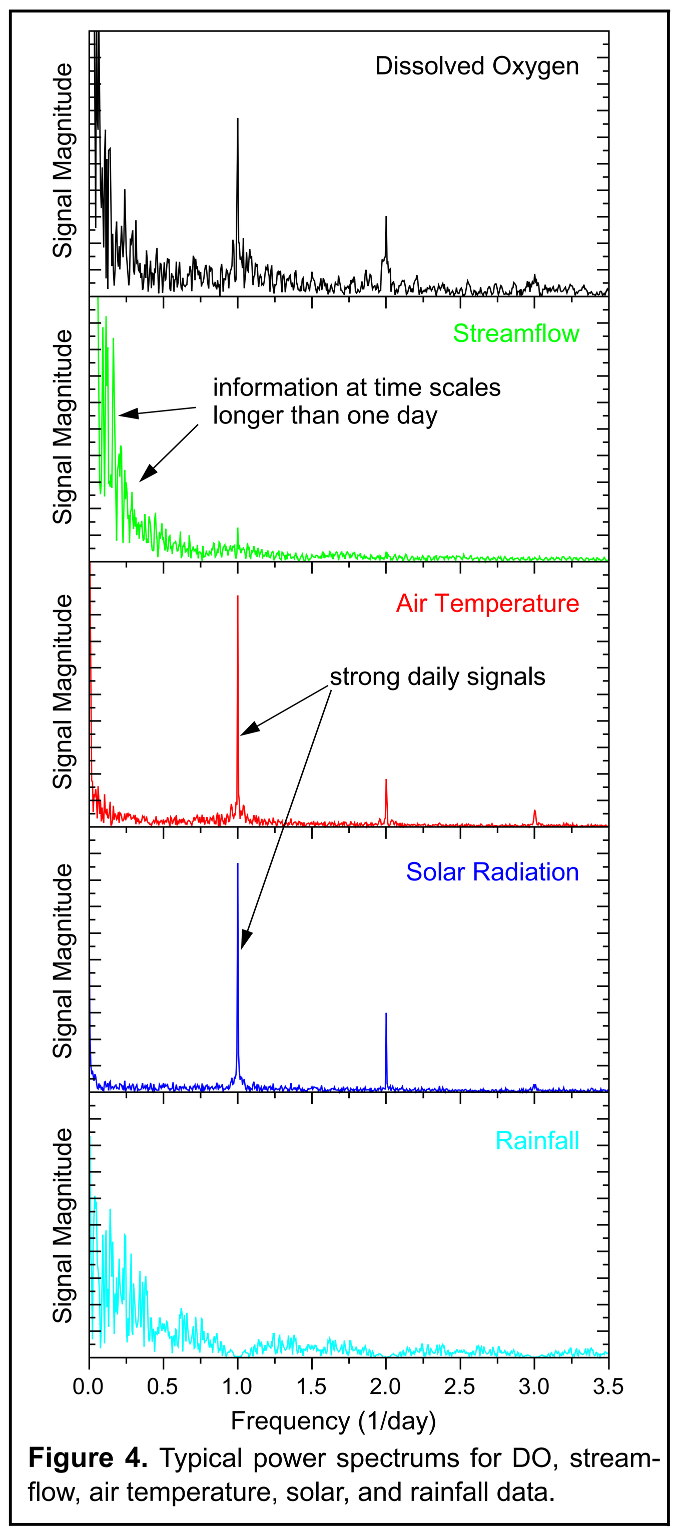


Figure 4. Typical power spectrums for DO, streamflow, air temperature, solar, and rainfall data.

Low-Frequency (Daily Mean) Models

Long- and short-term patterns were simulated with separate models (fig. 6). After optimization, the low-frequency ANN model required only 8 inputs:

- low-pass (lp) filtered data: lp-Q, lp-S, lp-AT
- low-pass filtered data from 12 days ago: lp-Q-12, lp-AT-12
- low-pass filtered & lagged data: lp-S-lag (1.75 days)
- day-of-year, year

where Q, S, and AT stand for streamflow, solar radiation, and air temperature. Time-lagged inputs were calculated as differences.

Multiple Linear Regression

Multiple linear regression is a special case ANN model that uses linear transfer functions and no hidden layers. Patterns in the data, however, were highly nonlinear and the regression did not perform well (table 1).

Artificial Neural Network

Optimization yielded an ANN with one hidden layer containing seven nodes. ANN predictions were markedly better than the linear model and in many cases better than a USGS process-based model, with a mean absolute error of only 0.83 mg/L and a correlation coefficient of 0.837 (fig. 7, table 1).

The ANN model captured the most important patterns in the data, producing remarkable fits to the measured DO considering that the predictions were based only on streamflow, air temperature, solar radiation, year, and day-of-year. The most important predictor variables were lp-Q, day-of-year, lp-S, and lp-S-lag, respectively.

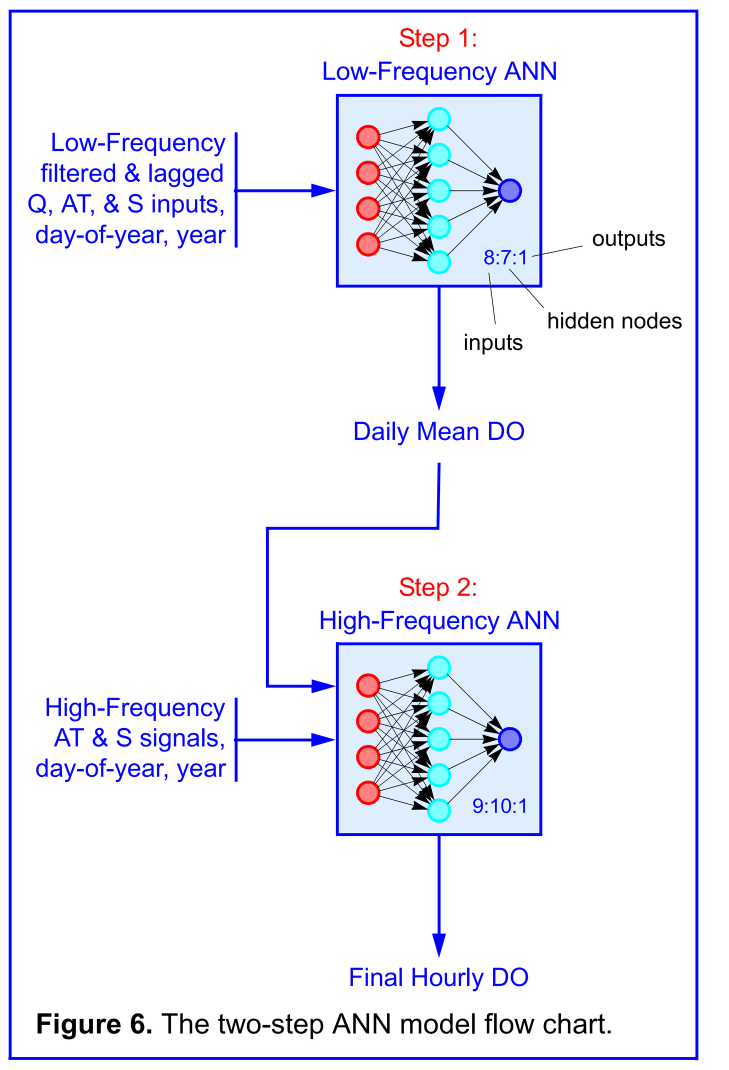


Figure 6. The two-step ANN model flow chart.

Final Hourly Model

High-frequency signals in the data were separated from low-frequency signals by subtracting the low-pass filtered data from the original data. High-pass AT and S inputs were included at several time lags to capture their 12- and 24-hour signals; the streamflow data had no useful high-frequency signals (fig. 4).

Final training and optimization yielded a high-frequency ANN model with nine inputs:

- output from the low-frequency ANN,
- high-pass (hp) filtered AT & S (3 time lags each),
- day-of-year, and year.

The high-frequency ANN used one hidden layer with 10 nodes.

The final model captured both the long-term and daily patterns in the measured DO data, producing a mean absolute error of 0.86 mg/L and a correlation coefficient of 0.831 (table 1).

Figure 8 illustrates the typical daily variations that the model produced in the final hourly DO. These predictions are accurate enough to be useful and can form the basis for a real-time DO forecasting tool.

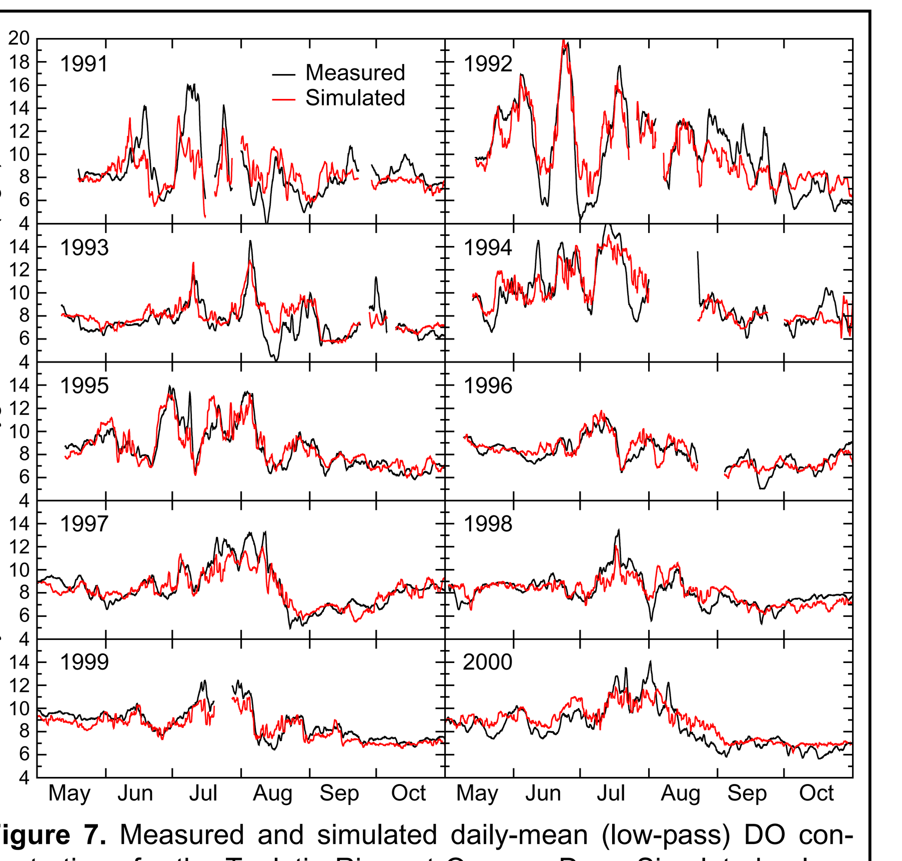


Figure 7. Measured and simulated daily-mean (low-pass) DO concentrations for the Tualatin River at Oswego Dam. Simulated values were calculated by the low-frequency ANN model (8 inputs, 1 hidden layer with 7 processing units, 1 output).

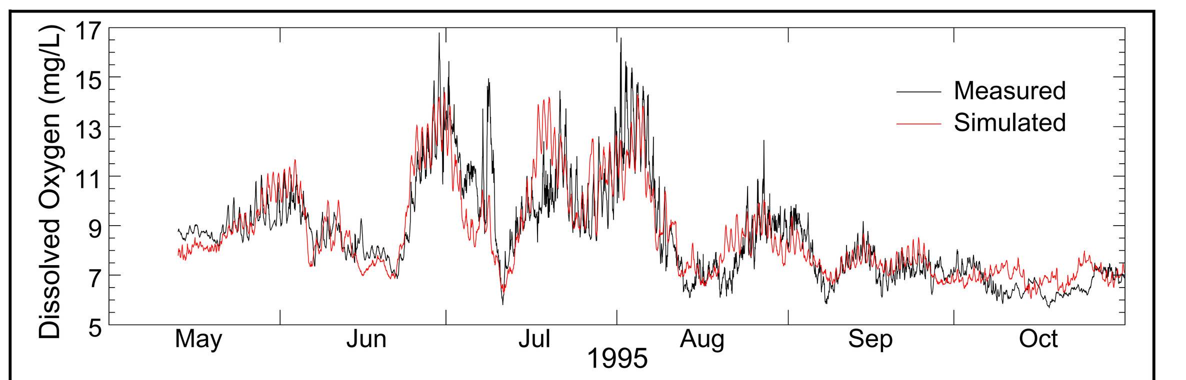


Figure 8. Measured and simulated hourly DO concentrations for the summer of 1995 in the Tualatin River at Oswego Dam. Simulated values were calculated by the final hourly ANN model (9 inputs, 1 hidden layer with 10 processing units, 1 output).

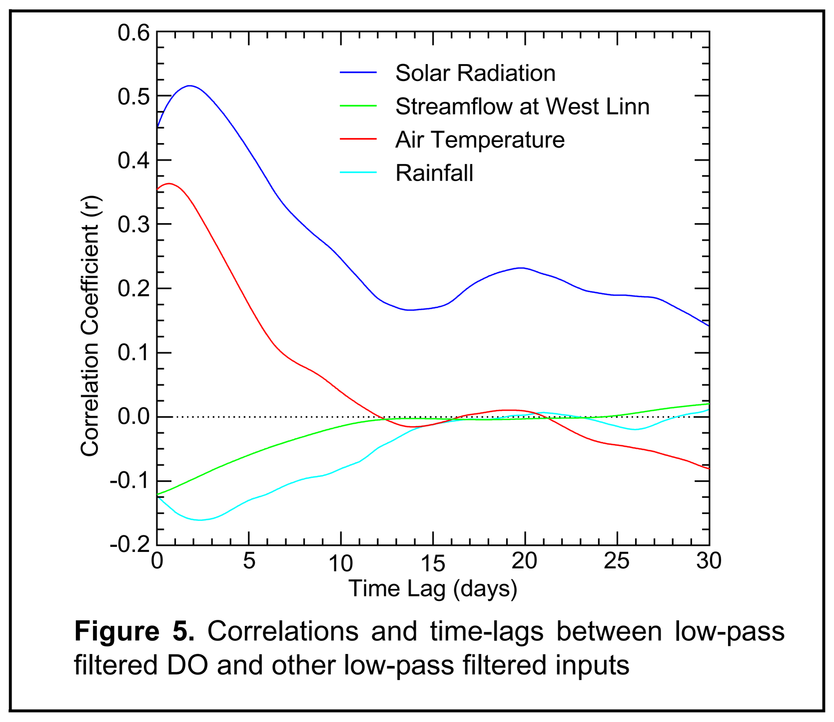


Figure 5. Correlations and time-lags between low-pass filtered DO and other low-pass filtered inputs

Lee Falls on the Tualatin River in the Coast Range Mountains.

Conclusions

Artificial neural network models were developed to simulate daily mean and hourly DO concentrations in the Tualatin River at the Oswego Dam. The DO at that site is affected by its solubility as well as biological processes such as algal photosynthesis and respiration, sediment oxygen demand, biochemical oxygen demand, and ammonia nitrification. The effects of these biological processes on DO, however, are constrained by physical and meteorological factors such as streamflow, air temperature, and solar radiation. Neural network and regression models were built to predict DO based on these factors, using data from May-October of 1991-2000.

- Multiple linear regression models failed to capture the long-term patterns in the DO data, producing poor results.
- Neural network models were successful in predicting patterns in the DO data on daily, weekly, and seasonal time scales. Separate models were used to simulate the low- and high-frequency patterns in the data.
- ANN model performance was good, with mean absolute errors less than 0.9 mg/L. Approximately 70% of the variation in the DO data was captured by the final ANN model.
- ANN predictions often were better than those from a USGS process-based model of the Tualatin River (not shown). As applied to the Tualatin River, however, ANN and process-based models have different purposes. The process-based model is most useful for providing insight into how the river works, identifying important processes, and testing the effects of point-sources and management strategies. The ANN model has tremendous potential as a forecasting tool, but yields less insight into the specifics of riverine processes.

Future work will focus on incorporating these and other ANN models into real-time water-quality forecasting tools. Such tools will provide important information to river managers, particularly as they make decisions regarding the proper level of flow augmentation.