Maximizing Water Quality Data Value (Newsletter)*

March 31, 2011

Everyone with an NPDES or other water discharge permit periodically measures the chemical composition of waters leaving the permit area and reports results to their regulator. For some, knowing no chemical exceeded a concentration threshold is sufficient. Other permit holders, along with regulators and environmental NGOs, want more information. Are concentrations increasing or decreasing over time? Are there associations between two or more chemicals or other factors? If so, by what time does one lead or lag the other? Much time, effort, and money is consumed by sampling, analyzing, recording, and reporting. It is good business to extract all useful information from the data. Spreadsheet line plots of concentration versus time yield no value other than whether thresholds were exceeded and concentrations change from one sample to the next. To really understand water quality data requires statistical time series modeling.

While this newsletter focuses on water quality data all environmental, operational, and business data that are measured sequentially in time can be examined for meaning. Technically sound and legally defensible time series models allow managers and executives to make better informed decisions and have higher confidence in their forecasts.

Time series have three components to be quantified and understood. First, the overall trend: are values increasing or decreasing? If so, is the trend significantly different from a flat line? Second, are seasonal variations of importance to understanding system dynamics? Third, how much of the values are from random events?

In most historical time series the observations are serially dependent and correlated with time. That is, a value observed or measured at time t' is dependent upon the observation or measurement at time $t-1$. Much of a time series analysis is aimed at explaining this correlation and the main features in the data, using appropriate statistical models and descriptive methods. These insights can be used to forecast future values of the measured variable. There are also correlations of variables that need to be understood; for example, precipitation and stream discharge, chemical concentration and current velocity, or concentrations of two chemicals. These associated variables may lead or lag each other by varying times. Quantifying the lead/lag in a statistical model increases understanding of system dynamics.

Trends in time series can be classified as stochastic or deterministic. Stochastic trends show inexplicable changes in direction and high serial correlation with random error. This type, common in financial and climatic data, are simulated using random walk or auto-regressive process models. Deterministic trends have plausible physical explanation and are modeled more robustly. For example, an increasing trend in water chemistry concentration may be related to an decreasing treatment efficacy. Deterministic trends and seasonal variation can be modeled by regression.

Environmental monitoring data often violate the regular periodic measurement requirement. Samples may be missed because the stream is frozen or covered in several meters of snow, access is blocked, the river is flooding and moving too fast, the channel is dry, the sample bottle fell off the vehicle roof, the expensive laboratory equipment failed. There are many examples of high variability in sampling frequency, and almost all can be fit with an appropriate robust statistical model. These analytical models are not included in commercial statistical software but they allow real world data to provide maximum insights and value to decision-makers.