7th International Conference on Hydroinformatics
HIC 2006, Nice, FRANCE

HYDROLOGIC MODELING USING MULTIVARIATE STATE SPACE RECONSTRUCTION

EDWIN A. ROEHL JR.
Advanced Data Mining, LLC
3620 Pelham Road, PMB 351, Greenville, South Carolina 29650, USA

PAUL A. CONRADS
U.S. Geological Survey South Carolina Water Science Center
720 Gracern Road, Suite 127, Columbia, South Carolina 29210-7651, USA

State space reconstruction is an important tool used by dynamicists for characterizing and modeling dynamic processes. State space reconstruction indicates that the behavior of a dynamic process can be reconstructed from time series (signals) that describe a process's *state* at each point in time. Optimal reconstruction requires that states be represented by an optimal number of measurements spaced by an optimal time delay. Treatments of state space reconstruction typically deal with processes described by a single signal. State space reconstruction can be used to model dynamic, hydrologic systems that are usually described by multivariate data. This paper presents a multivariate form of state space reconstruction with an example hydrologic modeling problem of some complexity. Additional details are given on complementary techniques such as signal decomposition to differentiate spectral and chaotic behavioral components that evolve on different time scales, artificial neural networks to synthesize non-linear fits of multivariate data, and multidimensional visualization of data and models to reveal complex process physics.

INTRODUCTION

Natural resource managers face difficult challenges when managing the interactions between natural and man-made systems. Mechanistic models based on deterministic physical equations are often developed at considerable expense to evaluate options for using a resource while minimizing damage. The alternative modeling approach is empirical modeling, most often empirical least squares regression. Calibrating an empirical model is a process of fitting a function, such as a line or surface, through data from two or more variables. This can be difficult when the data are noisy or incomplete, and the variables for which data are available may only be able to provide a partial explanation of the causes of variability.

Functions are either prescribed or synthesized. The functions prescribed by mechanistic models are physical equations, which incorporate adjustable coefficients that are used to tune a model's predictions to match calibration data as closely as possible. Linear least squares regression prescribes straight lines, planes, or hyper-planes to fit calibration data. The potential problem with prescriptive modeling is that if the applied function is inherently unable to fit the variable relations manifest in the data, a representative model is unobtainable. There are many examples of mechanistic modeling projects that have consumed millions of dollars and many years of effort, yet the models were never accepted by the regulatory agencies and/or stakeholders.

The growing abundance of real-time data (signals) is creating new methods for understanding, monitoring, and controlling dynamic hydrologic processes. Data mining converts massive databases into valuable knowledge. When applied to real-time (or continuous) data, data mining uses special methods to represent complex behaviors that evolve in time, including signal processing, machine learning, multivariate visualization, and chaotic system analysis (Chaos) as described by Abarbanel [1]. This paper describes a multivariate form of the classical univariate state space reconstruction (SSR) from Chaos. Multivariate state space reconstruction (MSSR) provides a conceptual framework for synthesizing empirical models that optimally exploit multivariate data. MSSR is explained through the development of a predictive model at a gaging station on an upper reach of the lower Savannah River estuary.

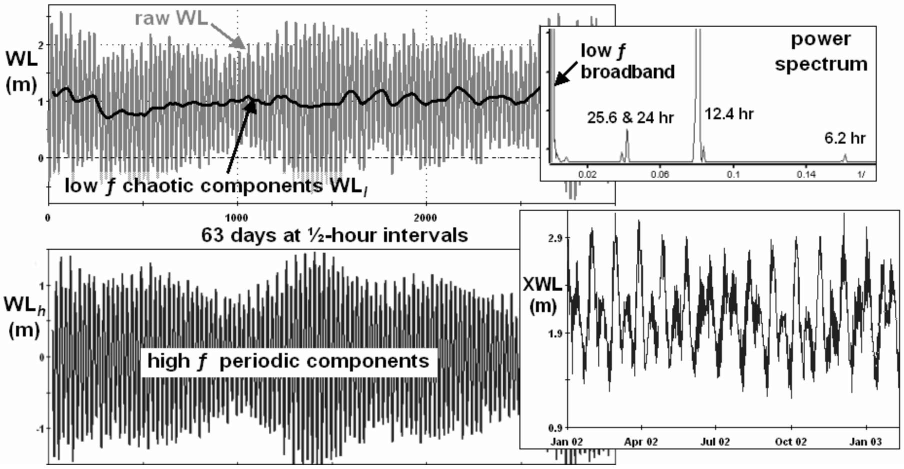

Figure 1: Decomposition of Savannah Harbor water level (WL) at U.S. Geological Survey (USGS) gaging station 02198980.

Periodicity, Noise, Chaos, and Signal Decomposition

Processes exhibit three types of behavior — periodic, chaotic, and noise. Process signals are a combination of behaviors that are superimposed upon each other. For example, the coastal water levels exhibit multiply periodic behaviors caused by the gravitational forces applied by the earth, moon, and sun. Coastal water levels are also affected by chaotic and random influences, such as wind and storms. Theoretically, periodic behaviors are perfectly predictable. Chaos describes physical processes that are highly sensitive to small changes in boundary conditions. They can change between different behaviors for little apparent reason. The weather is both multiply periodic and chaotic. The weather affects nearly everything, and apart from seasonal and diurnal behaviors is impossible to predict beyond a week or so. Chaotic processes are somewhat predictable and SSR has been developed to describe them. Noise is random behavior and is by definition unpredictable. Random behavior may be a consequence of simply not having the information to make it predictable. If so, *unmeasured disturbance variables* are at play, and identifying and measuring them would make the process more predictable.

Figure 1 illustrates signal decomposition, which uses filtering and other methods to decompose signals into components that manifest different behaviors. Periodic components are separated by spectral filters, leaving chaotic and random components behind. Chaotic components are removed using more complex filters based on empirical forecast models. Removing all of the predictable components leaves behind the noise. Signal decomposition provides a quantitative accounting of the predictable and unpredictable components. Referring to Figure 1, the power spectrum (upper right) of the raw water level (WL) signal (upper left) shows peaks at ½, 1, and 2 times the 12.4-hour tidal cycle, which are caused by the earth's rotation and the moon's orbit. A spectral filter decomposed the raw WL into high (lower left) and low (upper left) frequency (*f*) component signals, WL_h and WL_l respectively. The tidal range (XWL, lower right) is calculated from the raw WL and is multiply periodic, exhibiting 14-day, 28-day, semi-annual, and annual cycles caused by the motions of the earth, moon, and sun. WL_l contains periodicity on semi-annual and longer time scales and manifests chaotic meteorological forcing on shorter time scales.

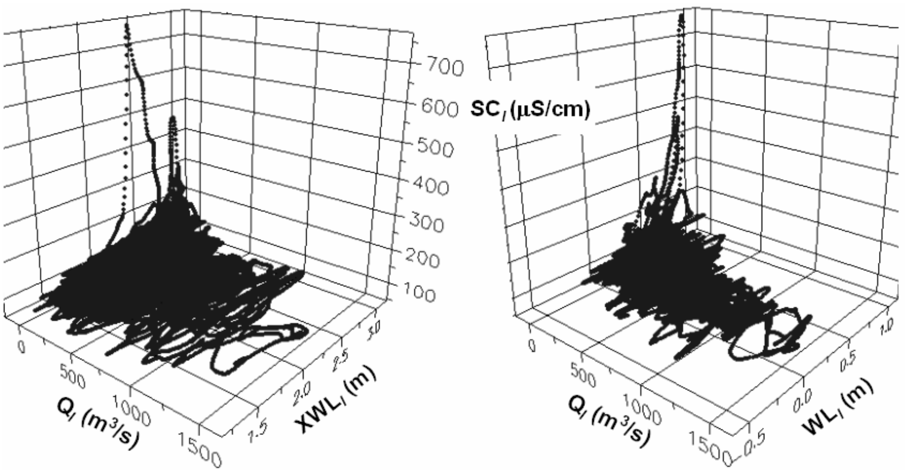

Figure 2. Tracks of estuary specific conductance (SC_l) with freshwater flow (Q_l) and tide range (XWL_l, left), and Q_l and water level (WL_l, right). Subscript "*l*" denotes low *f* signal components of measured times series after filtering to remove high *f* components > 1/day.

States, Vectors, Points, and Space

Typically, a variable x(t) is measured at constant time intervals, allowing its evolution to be observed over time. A future value of x can be forecast using a function to fit recent measurements and extrapolating forward in time. A problem with this approach is that it accommodates only one variable at a time, making it inadequate for reproducing the behaviors of multivariate processes. Like trending, Chaos generally uses multiple measurements to characterize process behavior. Some useful concepts from Chaos described:

- Chaotic processes transit from one unique *state* to another in time. This is unlike periodic processes, which revisit the same states at constant time intervals.
- Each state is characterized by a collection of measurements called a *state vector*. The vector's *elements* can represent one or more variables, such that the vector is either univariate or multivariate. Multiple measurements from the same signal can be assigned to elements to represent inertial effects.
- Each vector element represents a different dimension in a Euclidian *state space*. For example, a vector having five elements is said to be 5-dimensional and lies in a 5-dimensional state space. The coordinates given by the element values of a vector represents a *point* in state space.
- Figure 2 shows that as a process changes in time, it leaves a *track* of points in state space, representing a *state history*. A process's recent state history can be used to forecast near-term future states by curve fitting.
- A new process state is derived largely, but not entirely, from previous states. Unknown disturbance variables influence state transitions, so that process behavior can never be completely characterized or predictable.

MULTIVARIATE STATE SPACE RECONSTRUCTION

SSR is the means by which complex dynamic processes can be represented in straightforward geometric terms for analysis, visualization, and modeling. Abarbanel [1] describes how a process's behavior can be reconstructed "*in a space of vectors*" Y(t). Y(t) is described by Eq. (1), where d is the number of vector elements equal to the state space dimension, and τ_d is a time delay that equally spaces measurements x(t) in time. The variables d and τ_d are called *dynamical invariants*, and are analogous to the amplitude, frequency, and phase of a periodic signal.

$$Y(t)=[x(t),x(t-\tau_d),\ldots,x(t-(d-1)\tau_d)] \qquad (1)$$

Note that the choice of τ_d affects the interdependence of vector elements $x(t,\tau_d)$, and that a poor choice of d would over- or under-specify the reconstruction. Abarbanel suggests estimating τ_d using the "*first minima of the average mutual information function.*" The authors advocate using the "*first zero crossing of the autocorrelation function*," hereafter autocorrelation, for each component derived from decomposing a more complex signal. The decomposition is necessary for multivariate modeling anyway, and both techniques will give similar estimates of τ_d for suitably simple components.

Abarbanel suggests the "*local false nearest neighbors test*" to estimate d, which uses an empirical function, such as a linear or quadratic, to map prior measurements to the next measurement. Thus, d is determined experimentally and equals the number of prior measurements that parsimoniously gives the best prediction. This applies equally well on a variable-by-variable basis to multivariate processes.

Eq. (2) describes the state vectors for a multivariate process of k independent variables x_k having d_k and τ_{dk}, and each element $x_k(t,\tau_{di})$ represents a state space dimension.

$$Y(t)=\{[x_1(t),x_1(t-\tau_{d1}),\ldots,x_1(t-(d_1-1)\tau_{d1})], ...,[x_k(t),x_k(t-\tau_{dk}),\ldots,x_k(t-(d_k-1)\tau_{dk})]\} \qquad (2)$$

Eq. (3) adapts Eq. (2) to predict $y_p(t)$ of a measured dependent variable of interest $y(t)$ from prior measurements (also known as forecasting) of k independent variables, where F is an empirical function. The authors suggest that F be a *multi-layer perceptron artificial neural network* (ANN) model of the type described by Jensen [2].

$$\begin{aligned} y_p(t)=F\{&[x_1(t-\tau_{p1}),x_1(t-\tau_{p1}-\tau_{d1}),\ldots,x_1(t-\tau_{p1}-(d_{M1}-1)\tau_{d1})],\\ &....,[x_k(t-\tau_{pk}),x_k(t-\tau_{pk}-\tau_{dk}),\ldots,x_k(t-\tau_{pk}-(d_{Mk}-1)\tau_{dk})]\}\end{aligned} \tag{3}$$

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 x_k, 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 dimension d of Eq. (2) is replaced with a model input variable dimension d_M. Generally, $d_M \leq d$, and tends to decrease with increasing k. The variable $y(t)$ can be a superposition of disparate behaviors $y_j(t)$ originating from different forcing functions, such that

$$y_p(t) = \Sigma y_{pj}(t) = \Sigma F_j \tag{4}$$

A study by Conrads and Roehl [3] found that ANN models offered a number of advantages over mechanistic models in reproducing the dynamic flow and water-quality behaviors in an estuary. Most importantly, the ANNs gave much better prediction accuracy when using the same input and output variables and data. ANNs are a curve-fitting technique that synthesizes continuously differentiable, multivariate non-linear functions to near-optimally fit measurements that represent complex process behavior. Roehl *and others* [4] state that the perceived shortcomings of ANNs generally result from misapplication, for example, failure to decorrelate input variables. Conrads *and others* [5] describe a method for non-linearly decorrelating variables for estuary ANN models.

EXAMPLE APPLICATION

Figure 3 shows historical salinity and flow to illustrate maximum salinity intrusions at an inland gaging station on the lower Savannah River estuary. Specific conductance (SC) is a field measurement that is often used to compute salinity concentration. The intrusions were manifest as SC spikes that coincided with record low freshwater flows during a 4½-year drought. The spikes are also apparent in the three-dimensional (3D) tracks shown in Figure 2, and are an example of chaotic behavior — a gradual reduction in freshwater flow (Q) eventually elicits a sudden and dramatically different behavior. Note that even though the spikes occurred at 28-day intervals as XWL was peaking, Figure 2 shows that they occurred near the midrange of XWL rather than at the historical maximum.

Two separate 3D mechanistic flow models of the lower Savannah River estuary were developed to predict this and other hydrodynamic behaviors [6] [7]. Both had difficulty in capturing the on/off nature of these SC spikes in the upper reaches of the estuary. Therefore, these data provide an excellent test of the MSSR and ANN approach advocated here.

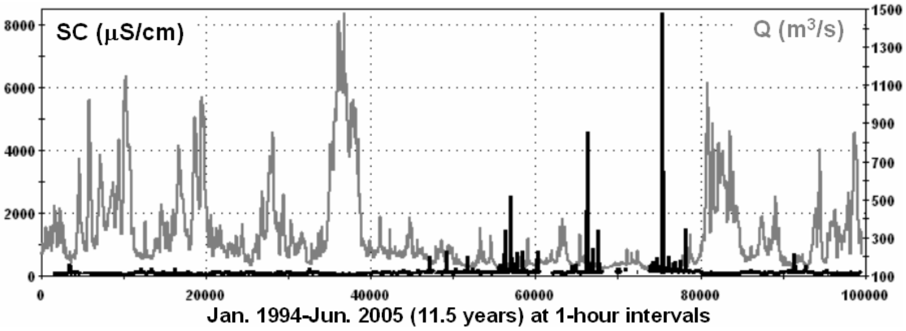

Figure 3. Hourly SC at USGS gaging station 02198840 and Q.

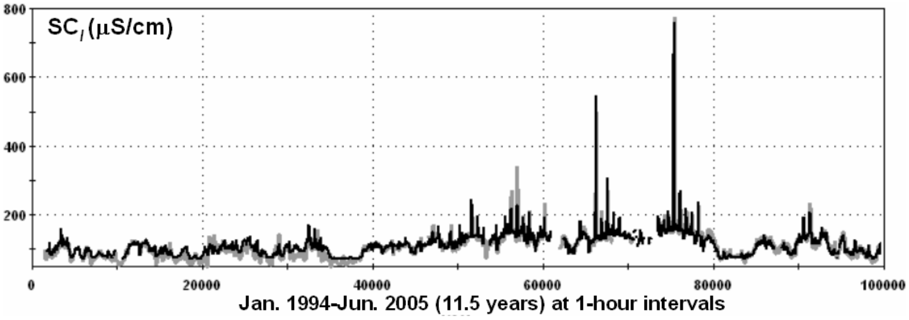

Figure 4. Measured (gray) and predicted (black) SC$_l$.

The modeling approach uses two ANN *sub-models* (F_1 and F_2) that together compose a *super-model*. F_1 and F_2 will generate SC_{lp} and SC_{hp}, which are the predicted values of the low and high frequency SC (SC_l and SC_h, respectively). Total prediction $SC_p = SC_{lp} + SC_{hp}$. The sub-models are described below. The WL and XWL inputs are shown in Figure 1. The streamflow Q was measured at an upstream gaging station near Clyo, Georgia. The τ_p and τ_d values below are in units of hours.

The F_1 input configuration is: WL_l at τ_p=0, τ_d=190, d_M=3; XWL_l at τ_p=0, τ_d=120, d_M=2; and Q_l at τ_p=48, d_M=1. The τ_d of WL_l was determined by autocorrelation after removal of components having periods > 3 months. The τ_d of XWL_l was determined by autocorrelation. The τ_p of Q_l was determined by cross correlation with SC_l. For all inputs to F_1 and F_2, values of d_M were determined experimentally. Figure 4 shows measured and predicted SC_l, and an R^2=0.88.

Note that the prediction accuracy is generally better after hour 40,000, indicating that the general quality of the field measurements improved over time with better equipment and maintenance practices. Consider that statistical measures of accuracy are commonly cited in ways that assume that measurements are more accurate than model predictions; however, also consider the value of averaging measurements to improve accuracy of noisy data. The fitting of empirical models tends to ignore the noise in data; therefore, model predictions can in some cases be more accurate at representing behavior than noisy measurements.

Figure 5 shows a *3D response surface* generated by F_1. When compared to the 3D track at right in Figure 2, it is apparent that the functional form of F_1 does match the data. Response surfaces are generated by ranging two of a model's inputs and calculating the output response. *Unshown* inputs are set to a constant value. Here, the delayed WL_l and the XWL_l inputs were set to their historical means.

F_2 required the calculated variables: the high *f* SC component $SC_h = SC - SC_{lp}$; and the normalized tidal range $XWL_n = XWL - XWL_l$. The F_2 input configuration is: WL at τ_p=1, τ_d=3, d_M=3; XWL_n at τ_p=0, d_M=1; and Q_l at τ_p=48, d_M=1. The τ_p of WL was determined by cross correlation with SC_h, and τ_d was determined by autocorrelation. The R^2 for F_2 is 0.72. The R^2 of the combined predictions SC_p shown in Figure 6 is 0.77. The prediction does not quite match the single SC measurement greater than 8,000 microsiemens/cm. The next highest SC value is 4,800 microsiemens/cm. The F_2 response surface is shown at right in Figure 5. The SC spikes are predicted at peaking SC_p and WL.

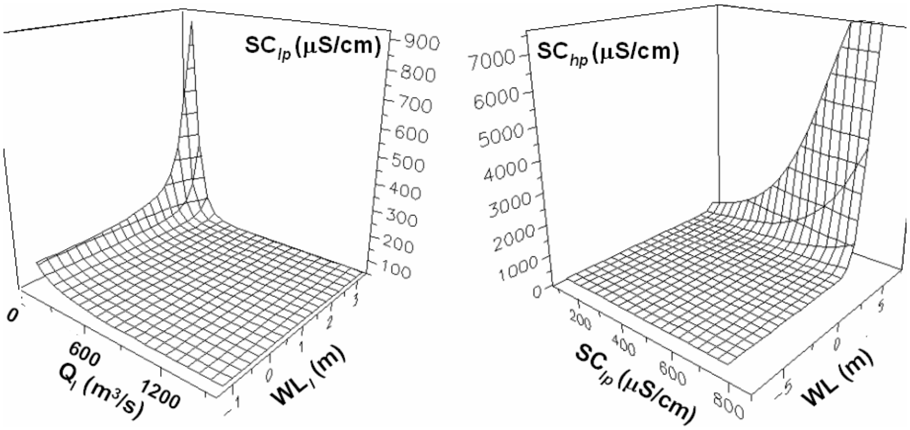

Figure 5. 3D response surfaces of F_1 and F_2. WL (right) τ_p=4.

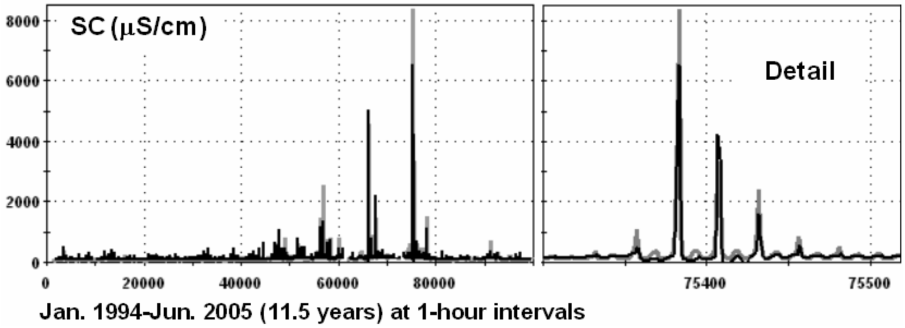

Figure 6. Measured (gray) and predicted (black) SC. Detail of maximum SC intrusion shown on the right.

CONCLUSIONS

The example model was developed from only two measured input signals, Q(t) and WL(t), yet the model actually fits 11½ years of data with a certain statistical measure of accuracy. The example model is the product of:

- *Signal Decomposition* – that transforms a complex signal into simpler components that represent different forcing functions evolving on different time scales.
- *Multivariate State Space Reconstruction* – represents dynamic multivariate behaviors using an optimal number of measurements, d_M, for each input variable, which are optimally spaced in time by a time day τ_d, and are input to a model at another optimal time delay τ_p.
- *Artificial Neural Network Models* – that synthesize multivariate non-linear functions to fit multivariate non-linear data.
- *Super-Model Architecture* – allows output behaviors evolving on different time scales to optimally modeled independently by sub-models, and have their predictions superposed.
- *Visualization* – of data and model responses to reveal a system's process physics.

The example model outperforms recent mechanistic models that predict the same output signal by a substantial margin. While the mechanistic models represent an entire spatially expansive system, individual ANN models for multiple gaging sites can provide expanded coverage, and give insights into process data and behavior that are useful in any modeling endeavor.

REFERENCES

[1] Abarbanel, H.D.I., "*Analysis of Observed Chaotic Data",* Springer-Verlag, Inc., New York, (1996).

[2] Jensen, B.A., "*Expert systems - neural networks, Instrument Engineers' Handbook Third Edition",* Chilton, Radnor PA, (1994).

[3] Conrads, P.A. and Roehl, E.A., "Comparing physics-based and neural network models for predicting salinity, water temperature, and dissolved oxygen concentration in a complex tidally affected river basin", South Carolina Environmental Conference, Myrtle Beach, (1999), pp 1-7.

[4] Roehl, E.A., Conrads, P.A. and Cook, J.B., "Discussion of using complex permittivity and artificial neural networks for contaminant prediction", *Journal of Environmental Engineering*, November, (2003), pp 1069-1071.

[5] Conrads, P.A., Roehl, E.A. and Martello, W.P., "Development of an empirical model of a complex, tidally affected river using artificial neural networks", Water Environment Federation TMDL Specialty Conference, Chicago, Illinois, November, (2003), pp 1-33.

[6] Applied Science Associates and Applied Technology and Management, "Hydrodynamic and Water Quality Modeling of the Lower Savanna River Estuary," report to the Georgia Ports Authority, Savannah, Georgia, (1998).

[7] Tetra Tech, Inc., "Development of the EFDC hydrodynamic model for the Savannah Harbor", report to the U.S. Army Corps of Engineers, Savannah District, (2005).