Using "Big Data" to Optimally Develop Water Quality Temperature TMDLs for Expansive Areas

John B. Cook[1], Edwin Roehl, Jr.[2] and Paul A. Conrads[3]

AUTHORS: [1]Chief Executive Officer, Advanced Data Mining Intl, 3620 Pelham Rd., PMB 351, Greenville, SC 29615;843-513-2130; john.cook@advdmi.com; [2]Chief Technical Officer, Advanced Data Mining Intl, 3620Pelham Rd., PMB 351, Greenville, SC 29615; 864-201-8679; ed.roehl@advdmi.com; [3]Surface Water Specialist, U.S. Geological Survey, Gracern Rd, Suite 129, Columbia, SC 29210; 803-750-6141; pconrads@usgs.gov

REFERENCE: *Proceedings of the 2010 South Carolina Water Resources Conference,* held October 13-14, 2010, at the Columbia Metropolitan Convention Center.

Abstract. Natural systems exhibit complex behaviors that are driven by the earth's orbital motions, weather, and anthropogenic forcing. Modeling them on a large scale is challenging because behaviors vary discontinuously both spatially and in time. Modeling requires large amounts of old and new data, also known as "Big Data" that represent a diversity of causes and effects. Measured variables are either unchanging categorical or dynamic time series. Integrating multiple data types and reducing large numbers of variables to a select set of data often leads to subjective decision-making that has significant ramifications when applying state-of-the-art multi-step modeling approaches, for example, land-use models driving finite- difference/element flow models. This paper describes an alternative approach that employs numerically optimized data-mining algorithms to more accurately predict stream temperatures for a total maximum daily load to prevent thermal impairment of streams in Western Oregon. The methods include: 1) time series decomposition to discriminate chaotic and periodic time-series components attributable to different forcing functions; 2) time-series clustering to segment monitored sites by their dynamic behaviors; 3) non-linear, multivariate sensitivity analysis using multi-layer perceptron artificial neural networks (ANN) to determine the relative importance of categorical variables at predicting site-to-site behavioral variability; 4) spatially interpolating dynamic behaviors with ANNs; and 5) assembling an end-user application that integrates data, site attribute classifiers, and prediction models to model an expansive, behaviorally heterogeneous natural system.

INTRODUCTION

Natural resource managers commonly ask scientists to create predictive models of spatially expansive natural systems for planning their protection or management. This involves collecting large amounts of many types of old and new data, also known as "Big Data", for model development. The data should come from multiple locations that represent the diversity of behaviors across the natural system. Measured variables are either (practically) unchanging categorical, such as, geomorphology, or dynamic time series (signals), such as, stream temperature. Time-series variables usually have multiple periodic behavioral components driven by the earth's orbital motions. Periodicity is by definition highly predictable; however, time series also display dramatic spatial and temporal variability due to chaotic forcing by humans and weather. Chaotic behaviors are by definition only somewhat predictable, yet it is these behaviors that modelers strive to reproduce. Techniques such as band-pass and window-average filtering can *decompose* a time series to separate the periodic components, leaving behind chaotic components.

Conrads and Roehl (1999) found that multi-layer perceptron artificial neural network models (ANN) of the type described by Jensen (1994) offer a number of advantages over physics-based finite-difference 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 data. Coppola and others (2005) made some of the same observations after applying ANNs to forecast water levels at two monitoring wells in an aquifer affected by climatic variables and pumping.

An important benefit of physics-based finite-difference/element models is their ability to

provide spatially semi-continuous predictions from mesh nodes. Analogously, Dowla and Rogers (1996) used ANNs to predict three-dimensional land elevations from categorical coordinate data, and Conrads and others (2003) describe how dynamic ANN outputs for multiple locations can be interpolated as a post-processing step. Therefore, to model dynamic, spatially expansive natural systems, an approach for configuring ANNs to simultaneously predict spatial and temporal variability was developed. It involves:

- *ANN Modeling with a Stacked Database* – provides near optimal multivariate non-linear curve fitting of categorical and dynamic variables. A stacked database consists of categorical and time series variables to train ANNs for spatial interpolation. Each monitored site is represented by a block of rows denoting time stamps, and columns denoting candidate categorical and time series input variables, and time series output variables. The input and output variables and their column order are identical for all blocks. The blocks for each monitored site are stacked atop each other.
- *Sub- and Super- Models* – complex modeling problems are solved with relatively simple, near-numerically optimal sub-models of decomposed signals and behavioral classes of monitoring sites.
- *Signal Decomposition* – using filters, time series are decomposed into different frequency ranges, ascribable to different forcing functions that are more easily and accurately modeled with ANN sub-models.
- *Time Series Clustering* – produces numerically optimal segmentation of a large set of signals into classes, with each class comprising signals that behave similarly. Each behavioral class can then be modeled by an ANN sub-model. Typically, there are gradations of similarity among the different classes. A side benefit of time-series clustering is that it identifies redundant data, largely answering the question of, *"Which monitoring wells can be discontinued?"*
- *New Site Classification* – near-numerically optimal assignments of "new" sites, not used in model development, to behavioral classes so that the appropriate sub-models can be applied. Classification algorithm options include kriging, linear nearest neighbor, and non-linear ANN classifiers.

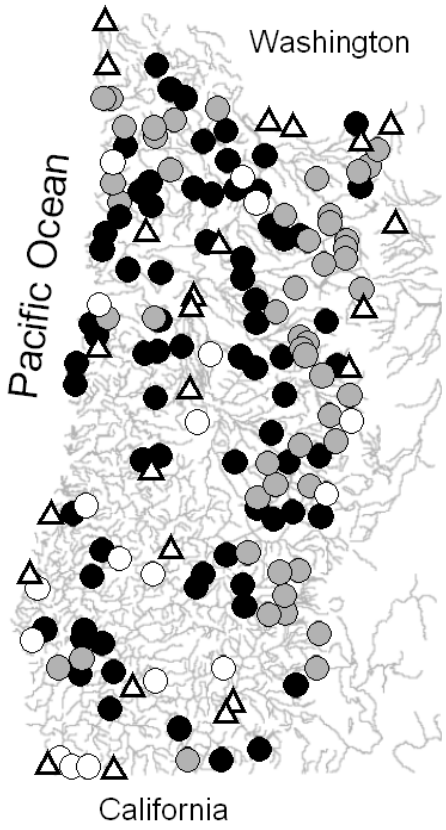

Figure 1. Western Oregon study area. Class 1, 2, and 3 sites are circles in white, gray, and black respectively. Triangles mark climatic and snowpack monitoring sites.

APPLICATION OF APPROACH

Risley and others (2003) describe using this approach to model "unimpaired" temperatures in small streams in the western third of Oregon to support federal and state efforts to estimate total maximum daily loads (TMDL) to prevent thermal impairment of streams in Western Oregon. The "Big Data" used in this project were comprised of:

- *Stream Temperature (ST)* - hourly time series from 148 unimpaired sites recorded from June to September 1999 (fig. 1). The sites were located on streams that drained basins ranging from 0.3 to over 300 square kilometers (km^2). Site elevations ranged from 7 to 1,445 meters (m) above mean sea level. Six of the 148 sites

were randomly withheld from model development for validating results.

- *Climate* – 65 hourly time series of air temperature, dew-point, solar radiation, barometric pressure, snowpack, and precipitation from 25 locations.
- *Stream Habitat and Basin Attributes* – 34 categorical variables that included stream bearing, gradient, canopy cover, wetted widths, depth, and bed substrate; and basin topographic and vegetation characteristics such as size and forest cover.

Additional technical details for this project included:

- The original objective was to predict maximum daily ST; however, initial attempts to model daily maximums directly were less successful than modeling the hourly ST and picking the daily maximum. This indicated a need for three cascaded sub-models (where outputs from one model are used to for inputs to a subsequent model) for each behavioral class to predict categorical, chaotic, and hourly STs.
- A large list of candidate categorical and dynamic inputs whose interrelationships and predictive performance were unknown. Many of the variables were highly correlated.
- New site classification could not be based solely on spatial coordinates because of the influences of categorical habitat and basin attributes.

Signal decomposition of the hourly water temperature time series $ST_{Hi}(t)$ involved the following. The categorical (static) components at the sites ST_{Si} = the historical mean of $ST_{Hi}(t)$. The chaotic components $ST_{Ci}(t)$ = the 24-hour moving window averages of $ST_{Hi}(t)$. $ST_{Ci}(t)$ was then normalized as $ST_{CNi}(t) = ST_{Ci}(t) - ST_{Si}$. $ST_{Hi}(t)$ was normalized as $ST_{HNi}(t) = ST_{Hi}(t) - ST_{CNi}(t) - ST_{Si}$.

$ST_{Ci}(t)$ were clustered into three classes (1,2, and 3) using time-series clustering. Here, the STs of all the sites were cross-correlated to produce a matrix of Pearson correlation coefficients. Each row and column represented a different site and its behavioral similarity to each of the other sites. The rows were then clustered using the k-means algorithm. The number of classes, k, is determined by the sensitivity of the root mean square error to k.

Class 1 sites were generally located in warmer climate regions at lower elevations and in the southern portion of the study area (fig.1). This includes the Klamath Mountains ecoregion and the Willamette River valley lowlands. Class 2 sites were more predominant at higher elevations, particularly in the Cascade Mountains (fig.1). Class 3 sites were widely distributed at middle elevations (fig.1).

The climatic hourly time series, $C_{Hi}(t)$, were decomposed into chaotic components $C_{Ci}(t)$ = 24-hour moving window averages of $C_{Hi}(t)$, and then normalized hourly $C_{HNi}(t) = C_{Hi}(t) - C_{Ci}(t)$. Each type of climatic variable was measured at multiple stations. These tended to be highly correlated station-to-station, so they were decorrelated by setting one station as a "standard" and calculating difference between the standard and the other stations.

A single categorical sub-model that used only categorical variable inputs to interpolate ST_S for all three classes was used. For each class, chaotic sub-models were trained to interpolate $ST_{CNj}(t)$ from categorical and chaotic climatic inputs. Similarly, hourly sub-models were trained to interpolate $ST_{HNj}(t)$ from categorical and hourly climatic inputs. Input variables were selected according to their predictive performance. $ST_{Hi}(t)$ and $ST_{Ci}(t)$ predictions were summations of the categorical and normalized chaotic and hourly predictions. The critical input variables included air temperature, riparian shade, site elevation, and percent of forested area in the basin.

Figure 2 shows measured and predicted $ST_{Hi}(t)$ at the "best" and "worst" of the six validation sites. Both predictions track the climatically-forced dynamic behaviors; however, the Fisher Creek predictions are offset from the measurements by an average of 2.4 degrees Celsius (°C). The offset is due largely to the error in the predicted categorical ST, suggesting that overall model error is a consequence of the process by which habitat and basin attributes are determined. A second potential cause of the offset is related to the procedure used to select validation sites, such as random selection as was used here. A validation site whose attributes are unique and unlearned will be poorly represented by an empirical model.

A non-linear classifier comprised of three ANNs, one for each class, was created to select the appropriate categorical+chaotic+hourly sub-

model triplet for a new site. Each class' ANN was trained to predict a binary digit (0 or 1) depending on whether or not a new site's habitat and basin attributes matched those of its member sites. Programmed logic was used to resolve ambiguous cases.

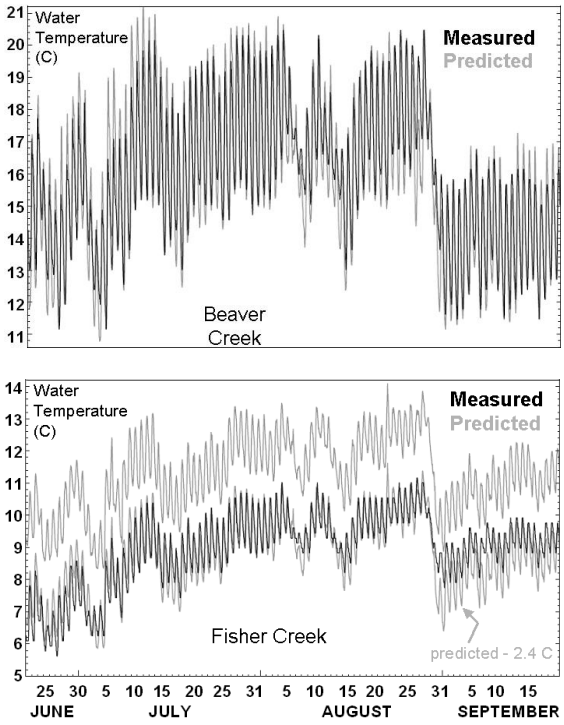

Figure 2. Measured and predicted stream temperature (ST) at two validation sites.

CONCLUSIONS

This paper provides an overview of leveraging "Big Data" with a divide-and-conquer approach to empirically model spatially heterogeneous, dynamic behaviors across expansive regions. The "Big Data" comprises many types of categorical and time series data that should be used to the fullest possible extent with minimal subjectivity about which data are important. Time-series clustering provided a numerically optimal solution to segmenting the many ST time series into classes. An ANN-based nonlinear classifier provided a numerically optimal means to classify new sites for model runs. ANNs use an inherently non-linear, multivariate architecture and error minimizing training algorithm to fit data representing complex behaviors. Their performance is improved by decomposing time series into static and dynamic components and modeling them separately. Modeling behavioral classes separately avoids prediction errors caused by fitting discontinuous behaviors with continuous functions. ANNs can be trained to spatially interpolate with a stacked training database that combines static and time-series variables. The best predictor variables can be found by systematically adding and removing candidates and tracking statistical measures of prediction accuracy. ANN sub-models are easily assembled into super-models that can be integrated with a database and control program to form run-time applications.

LITERATURE CITED

Conrads, P.A., and E.A. Roehl,1999, Comparing physics-based and neural network models for predicting salinity, water temperature, and dissolved-oxygen concentration in a complex tidally affected river basin, paper presented at the South Carolina Environmental Conference, Myrtle Beach, March 15-16.

Conrads, P.A., E.A. Roehl, and W.P. Martello, 2003, Development of an empirical model of a complex, tidally affected river using artificial neural networks," Water Environment Federation TMDL Specialty Conference, Chicago, Illinois, November.

Dowla, F.U. and L.L. Rogers, 1996, *Solving Problems in Environmental Engineering and Geosciences with Artificial Neural Networks*, MIT Press, 159-172, Cambridge, MA.

Coppola, E.A., A.J. Rana, M.M. Poulton, F. Szidarovszky, and V.W. Uhl, 2005, A neural network model for predicting aquifer water level elevations, Ground Water 43(2), 231-241.

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

Risley, J.C., E.A. Roehl and P.A. Conrads, 2003, Estimating water temperatures in small streams in western Oregon using neural network models, U.S. Geological Survey Water-Resources Investigations Report 02-4218.