Development of Inferential Sensors for Real-time Quality Control of Water-level Data for the Everglades Depth Estimation Network

Ruby C. Daamen[1], Edwin Roehl, Jr.[2] and Paul A. Conrads[3]

AUTHORS: [1]Managing Partner, Advanced Data Mining Intl, 3620 Pelham Rd., PMB 351, Greenville, SC 29615; ruby.daamen@advdmi.com; [2]Chief Technical Officer, Advanced Data Mining Intl, 3620 Pelham Rd., PMB 351, Greenville, SC 29615; ed.roehl@advdmi.com; [3]Hydrologist, U.S. Geological Survey, Gracern Rd, Suite 129, Columbia, SC 29210; 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. The Everglades Depth Estimation Network (EDEN) is an integrated network of real-time water-level gaging stations, ground-elevation models, and water-surface models designed to provide scientists, engineers, and water-resource managers with current (2000-present) water-depth information for the entire freshwater portion of the greater Everglades. The generation of EDEN water-level surfaces is derived from real-time data. Real-time data are automatically checked for outliers using minimum, maximum, and rate-of-change thresholds for each station. Smaller errors in the real-time data, such as gradual drift of malfunctioning pressure transducers, are more difficult to immediately identify with visual inspection of time-series plots and may only be identified during on-site inspections of the gages. Correcting smaller errors in the data often is time consuming and water-level data may not be finalized for several months. To provide water-level surfaces on a daily basis, EDEN needed an automated process to identify errors in water-level data and to provide estimates for missing or erroneous water-level data.

A technology often used for industrial applications is "inferential sensor." Rather than installing a redundant sensor to measure a process, such as an additional water-level gage, an inferential sensor, or virtual sensor, is developed that estimates the processes measured by the physical sensor. The advantage of an inferential sensor is that it provides a redundant signal to the sensor in the field but without exposure to environmental threats. In the event that a gage does malfunction, the inferential sensor provides an estimate for the period of missing data. The inferential sensor also can be used in the quality assurance and quality control of the data. Inferential sensors for gages in the EDEN network are currently (2010) under development. The inferential sensors will be automated so that the real-time EDEN data will continuously be compared to the inferential sensor signal and digital reports of the status of the real-time data will be sent periodically to the appropriate support personnel. The development and application of inferential sensors is easily transferable to other real-time hydrologic monitoring networks.

INTRODUCTION

The Everglades Depth Estimation Network (EDEN) is an integrated network of approximately 250 real-time water-level gaging stations, ground-elevation models, and water-surface models designed to provide scientists, engineers, and water-resource managers with current (2000-present) water-depth information for the entire freshwater portion of the greater Everglades (Telis, 2006). The U.S. Geological Survey Greater Everglades Priority Ecosystems Science program provides support for EDEN with the goal of providing quality-assured hydrologic data for the Comprehensive Everglades Restoration Plan (CERP) (U.S. Army Corps of Engineers, 1999). Presented on a 400-square-meter grid spacing, the EDEN offers a consistent and documented data set that can be used by scientists and managers to: (1) guide large-scale field operations, (2) integrate hydrologic and ecological responses, and (3) support biological and ecological assessments that measure ecosystem responses to the CERP. These data establish a large data set of baseline conditions prior to the implementation of the CERP that offers investigators a single repository for historic hourly water-level data.

While EDEN data are of great importance to many scientific and resource management activities, some of the massive amounts of data being collected by EDEN are inaccurate for reasons such as sensor malfunction, data communication errors, and other types of hardware issues. Detecting these issues can be time consuming and problematic, especially when they are not obvious by inspection, such as detecting drift. It can be time consuming to correct these types of problems because of the remoteness of the monitoring sites and the expense of having qualified technical personnel travel to the gages. In order for these data to be used for important assessments they need to be validated and sometimes corrected, further adding to the expense and time required to disseminate the data. A technology often used for industrial applications is the inferential sensor. Rather than installing a redundant sensor to measure a process, such as an additional water-level gaging station, an

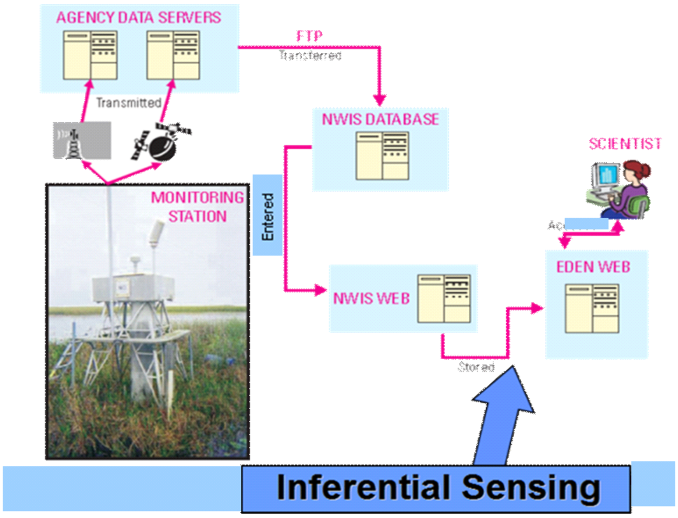

Figure 1. Location of the inferential sensor within the data stream of EDEN.

inferential sensor, or virtual sensor, is developed that estimates the processes measured by the physical sensor. The inferential sensor typically is an empirical or mechanistic model using inputs from one or more proximal gages. The advantage of using an inferential sensor is that it provides a redundant signal to the sensor in the field but without exposure to the environmental threats (floods or hurricanes, for example). In the event that a physical sensor does malfunction, the inferential sensor provides an estimate for the period of missing or erroneous data. The inferential sensor also can be used in the quality assurance and quality control of the data. The virtual signal can be compared to the real-time data and if the difference between the two signals exceeds a certain tolerance, corrective action can be taken. Inferential sensors for gages in the EDEN network are currently under development. The inferential sensors (fig. 1) will be automated so that the real-time EDEN data will continuously be compared to the inferential sensor signal and digital reports of the status of the real-time data will be sent periodically to the appropriate personnel.

METHODS

The inferential sensor will sequence two algorithms to automatically analyze the real-time data. The first layer implements a Statistical Process Control (SPC) series of 14 univariate filters (table 1) (Cook et. al.,, 2008). Univariate filters provide information about the quality and behavior of the data for each parameter and combined with post-processing of the filter outputs with logic integrates information for multiple sensors to validate measurements and generate intelligent notifications for system managers. For example, if only one EDEN sensor were to exhibit odd behavior, but neighboring sensors do not, then a physical sensor issue is likely the problem. If multiple virtual sensors exhibit odd behavior, then systematic network issues or network maintenance may be occurring.

The second algorithm addresses synthesizing measurements to augment actual measurements determined to be erroneous or unreliable. As it is not known at any given date and time what sites will have reliable data available, it is necessary that empirical models be created “on the fly”. A matrix of Pearson coefficients is calculated using the most recent 90 days of filtered data and candidates gaging stations to be used as inputs to an empirical model for a given site are selected based on degree of correlation. Filtered data are used to remove the influence of any outliers on the calculated Pearson coefficients. Ninety days was selected to capture any seasonal changes in relations between sites. The selected signals must then be automatically decorrelated from each other. The first approach made the input site that was most highly correlated a “standard” signal and then decorrelated the other input sites by computing their differences from the standard signal. This approach has been used successfully by the authors in many hydrology projects (Conrads et. al., 2006), but proved unsuccessful in this case. A new approach was needed to ensure that model inputs are decorrelated. A statistical technique known as Principal Component Analysis (PCA, Joliffe, 2002), which has been widely used in data analysis and compression, was selected.

“The central idea of principal component analysis is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables” (Joliffe, 2002). In simple terms, PCA is performed by calculating the eigenvalues and eigenvectors of the covariance matrix of the assembled data set. Each eigenvector has an associated eigenvalue. When sorted by eigenvalue (highest to lowest), the first eigenvector (PC) explains most of the variance in the original variables. As all eigenvectors of a symmetric matrix (in this case the covariance matrix) are orthogonal to each other, each PC is decorrelated from any other PC.

UNIVARIATE FILTER	CHECK DESCRIPTION	PRECEDENCE	WATER LEVEL LIMIT (ft.)
LOST_SIGNAL	no signal	1	NA
GT_RNG_UL	x(t) > signal range Upper Range Limit	2	15.19
LT_RNG_LL	x(t) < signal range Upper Range Limit	3	6.99
GT_UCL	x(t) > signal Upper Control Limit	4	14.73
LT_LCL	x(t) < signal Upper Control Limit	5	8.56
Sn_LT_L	flatlined: x'(t) = x(t)=x(t-1); SUM[(|x'(t)|,…,|x'(t-n+1)|] < Limit	6	0.00
D1_GT_L_1	vfast vlarge increase: x(t)-x(t-1) > Limit	7	1.92
D1_LT_L_1	vfast vlarge decrease: x(t)-x(t-1) < Limit	8	-2.34
D1Sn_GT_L_1	fast vlarge increase: x'(t)=x(t)-x(t-1); Sum[x'(t),…x'(t-n+1)] > Limit	9	1.98
D1Sn_LT_L_1	fast vlarge decrease: x'(t)=x(t)-x(t-1); Sum[x'(t),…x'(t-n+1)] < Limit	10	-2.52
D1_GT_L_2	vfast large increase: x(t) - x(t-1) > Limit	11	1.69
D1_LT_L_2	vfast large decrease: x(t) - x(t-1)< Limit	12	-0.25
D1Sn_GT_L_2	fast large increase: x'(t)=x(t)-x(t-1); Sum[x'(t),…x'(t-n+1)] > Limit	13	1.98
D1Sn_LT_L_2	fast large decrease: x'(t)=x(t)-x(t-1); Sum[x'(t),…x'(t-n+1)] < Limit	14	-0.27

Table 1. Univariate filter descriptions. Filters are applied in order of precedence. Limit values shown are for illustration only. The limits are uniquely set for each gaging station and each parameter in the gaging network.

RESULTS

Twelve sites were selected to test the use of PCA coupled with multivariate linear regression to predict water levels (WL). The sites were selected to represent the different types of locations (marsh, canal, marsh structure and canal structure) as well as those with highly correlated candidates and those with few or no highly correlated candidates. The data set included hourly data from 4/2009 – 12/2009. The real-time water-level data were first run through the univariate filters. Water levels at each of the 12 sites were predicted over the data set time period using PCA and multivariate linear regression. Data from up to five candidates were included in the models. Correlations and regressions used the most recent 90 days of data. Figures 2, 3, and 4 show results from three of the sites to highlight various aspects of the study. Displayed in each graph are measured WL, filtered WL, predicted WL using linear regression, and predicted WL using PCA and multivariate linear regression. In figure 2, little improvement is seen by adding additional sites. L31N1 has a number of highly correlated sites (coefficient of determination [R^2] 0.99 or greater). In figure 3, a better estimate is seen using five similar sites and PCA over one site and regression. Figure 4 highlights the ability of the predictions to pick up erroneous data that were missed by the univariate filtering.

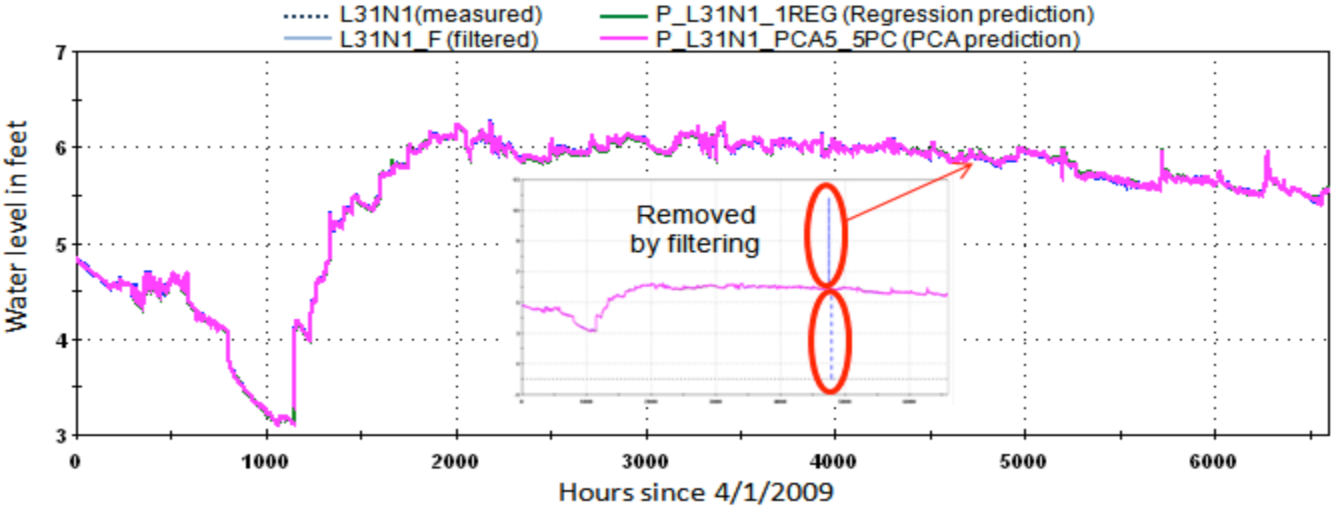

Figure 2. Gage L31N1 in Everglades National Park. Univariate filtering removed the large spikes shown in the inset.

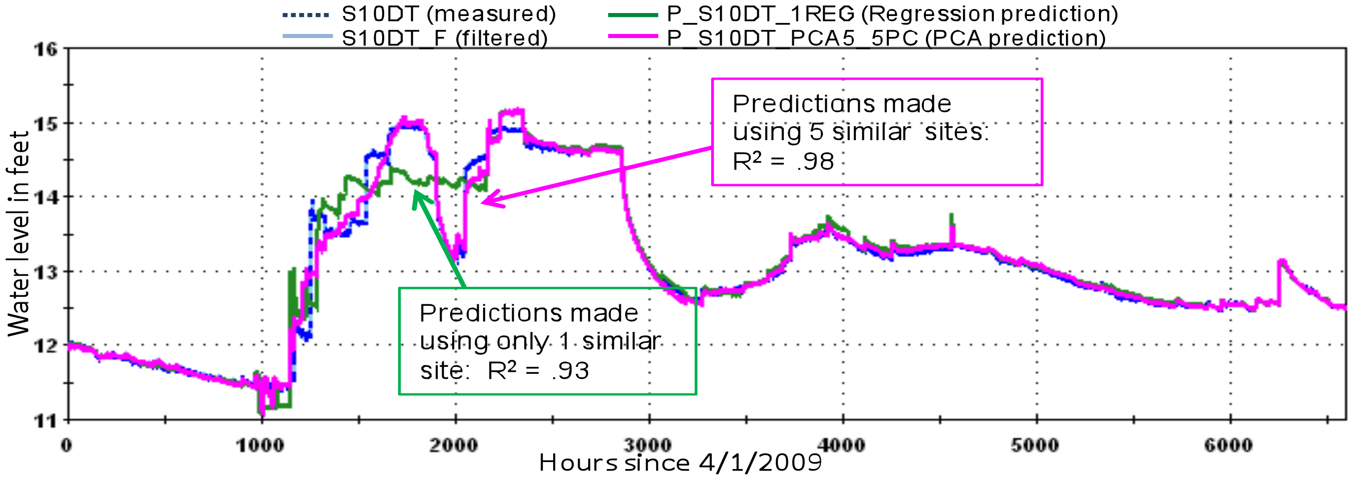

Figure 3. Gage S10DT in Water Conservation Area 2.

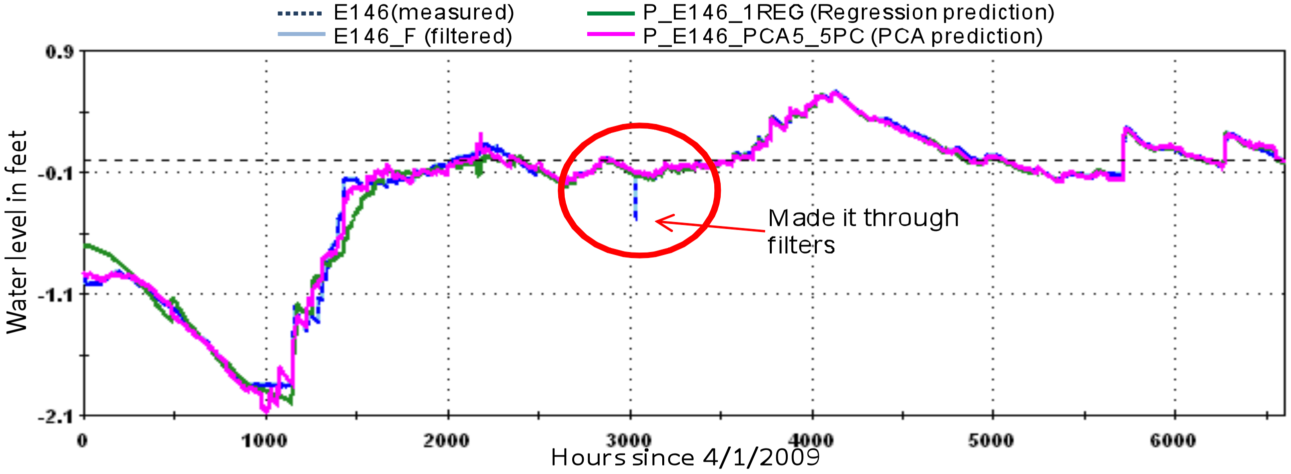

Figure 4. Gage E146 in Everglades National Park.

CONCLUSIONS

Initial results confirm that the use of filters and empirical models using PCA and multivariate linear regression, combined with post-processing logic can comprise an inferential sensor that will provide both accurate estimates of data when missing and quality assurance of the data. Developing empirical models "on the fly" ensures that the greatest number of available gaging stations is used for developing the inferential sensor. The development and application of inferential sensors is easily transferable to other real-time hydrologic monitoring networks.

LITERATURE CITED

Conrads, P.A. and Roehl, E.A., Daamen, R.C., and Kitchens, W.M., 2006, Using artificial neural network models to integrate hydrologic and ecological studies of the snail kite in the Everglades, USA, Hydroinformatics 2006, edited by Philippe Gourbesville, Jean Cunge, Vincent Guinot, Shie-Yui Liong, Vol. 3, p.1651-1658

Cook, J., Daamen, R., Roehl, E., Cho, S., Carlson, K., Byer, D., Byrne, J., and Cline, M., 2008,"Distribution System Security and Water Quality Improvement Through Data Mining", Report for American Water Works Association Research Foundation Project 3086.

Joliffe, I.T, (2002), *Principal Component Analysis*, *Springer Series in Statistics*, Second Addition (2002).

Telis, P.A. 2006, The Everglades Depth Estimation Network (EDEN) for Support of Ecological and Biological Assessments: U.S. Geological Survey Fact Sheet 2006-3087, 4 p.

U.S. Army Corps of Engineers, 1999, Central and Southern Florida Project Comprehensive Review Study: Jacksonville, Florida, Final Integrated Feasibility Report and Programmatic Environmental Impact Statement.