REAL-TIME QUALITY CONTROL (QC) PROCESSING, NOTIFICATION, AND VISUALIZATION SERVICES, SUPPORTING DATA MANAGEMENT OF THE INTELLIGENT RIVER©

David L. White, Julia L. Sharp, Gene Eidson, Shashank Parab, Farha Ali, Sam Esswein

AUTHORS:

Dr. David L White, Director of Environmental Informatics, Cyberinfrastructure Technology Integration, Clemson Computing and Information Technology; Dr. Julia L. Sharp, Assistant Professor, Applied Economics & Statistics; Dr. Gene Eidson, Center for Watershed Excellence; Shashank Parab, School of Computing; Farha Ali, School of Computing; Sam Esswein, Department of Forestry and Natural Resources, Clemson University, Clemson, SC 29634

REFERENCE:

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

Abstract. Quality Assurance (QA) of real-time environmental data is realized through the planning and implementation of standard operating procedures. A critical component of any QA program supporting real-time data is Quality Control (QC). These processes need to be performed using automated and manual methods. The need for automatic QC has become even more critical as web-based technology has enabled access to data by end-users following data collection and limited resources can hinder the timeliness and scope of manual QC methods. Ultimately, the end goal of the Intelligent River© QA program is to produce research quality data that is available to end users following data collection. To support this goal, automated QC methods were implemented that provide signals, realized with a flagging mechanism in the event of a failure or anomaly and in some cases correcting data for systematic errors. These QC processes are organized into levels, with each level performing more robust checks and corrections as the data are entering the Intelligent River© network. Level-1 algorithms apply simple heuristics to identify invalid or suspect observations. Specifically, observations with erroneous timestamps originating from the same device, data that exceed expected high and low thresholds for any given site location, missing observations, and observations exhibiting excessive variability originating from the same device are flagged. Flagged observations are republished to a level-2 process responsible for correcting the errors when appropriate. These QC processes are augmented by tools that support automated notification such as email and RSS feeds that can be tuned to varying levels of notification dependent on the end user needs. QC error notifications are ingested by web-based mapping technologies and visualized in a similar manner as the environmental data observations. Further, a level-3 product (in development) focuses on the identification of sensor drift and overall sensor performance, resulting in the automated publication of a PDF report that details a series of statistical analyses on any given sensor.

INTRODUCTION

The Intelligent River© project is developing and operating environmental and hydrological observation systems to support research and provide real-time monitoring, analysis, and management of natural resources in the Southeast (http://www.intelligentriver.org). The Intelligent River© was conceived as a massive-scale *macroscope* capable of providing high-precision visibility into the physical world for purposes of enabling advanced environmental and hydrological modeling, high-fidelity visualization, and scientifically-based decision-making processes. This vision calls for unparalleled data density, both temporally and spatially, an end-to-end hardware/software infrastructure engineered to support real-time monitoring and management of environmental resources across multiple landscapes from fresh and marine resources to farms and urban infrastructures.

Advanced monitoring networks are increasingly necessary to support the management of water resources from local to global scales. Urban and agricultural runoff

and even pharmaceutical/personal care products are contributing to biological and chemical contamination of surface and groundwater (Smyth et. al., 2008). Climate Change is affecting global and regional water resources thus putting additional pressures on human populations and societies (Thompson et. al., 2006). Ultimately, these issues will require an integrated management approach to provide resource estimation, provisioning, and pricing (Eidson et. al., 2009).

Monitoring networks will need to support variable spatial and temporal scales to provide data across a spectrum of observation types from biological to physical. The supporting sensor infrastructures will need to monitor contaminant loads, release and hold water resources for human consumption, monitor ecological function, and access industrial and agricultural needs. This necessitates advanced software architecture(s) to provide for access to quality assured data in real-time that are available to multiple data users, configurable with other data providers, documented with industry standard metadata, and usable across a suite of mobile, web, desktop and modeling tools.

Quality Assurance (QA) of real-time data is a critical component to ensure the accuracy of data collection that benefits both data producers and data consumers. A focus of the Intelligent River© has been the development of Quality Control (QC) procedures that support the identification and documentation of potential erroneous data. These processes run in real-time and are checking and flagging potentially erroneous data prior to data archival. Additional processes are designed to provide analyses at longer-term scales that allow data managers to examine sensor performance using statistical methods. A key feature of these efforts is to support the visualization of QC products. We have developed desktop software that allows data managers to access data and QC flags and also produce a configurable PDF report for each sensor.

METHODS

A novel QA/QC application listens to real-time data and provides checks to ensure data falls within acceptable limits, providing a flagging mechanism in the event of a failure or anomaly and in some cases corrects data for systematic errors. This real-time QA/QC application is performed as a series of steps termed "levels" that identify and then correct problematic data. Each step is an independent software application written in JAVA. Level 0 are the "raw" data that first enter the sequence of data integrity checks. Level 0 data are parsed and examined for identical timestamps or for all zeroes, and for meteorological data a variability check is performed. Data are collected at high sample rates leading to difficulty in buffering data in the cellular modems and resulting in several records with identical timestamps. In addition, some data records report all zeroes for each environmental measurement. In the level 0 variability check, the difference between the current and immediate past observation is calculated. This result is compared with the expected maximum change for meteorological observations (World Meteorological Organization, 2004) and values over a maximum change are flagged. The flagging mechanism for all of these checks consists of a two-character code for each measurement value indicating the status such as a duplicate timestamp or a step check failure. This process creates the Level 1 data set that is then processed by the Level 2a application and archived.

Following initial data flagging, the Level 2a process corrects the duplicate and all zero data. For the duplicate errors, a mean value is calculated for all of the duplicate timestamps and a single value reported, and all zero records are removed. Corrected data are documented by a second character-based QA/QC string that uses a similar two-character code for each measurement value, indicating if actions were performed on any given measurement. This Level 2a product is written as a separate data set and to the production database (PostgreSQL) for web dissemination and the Oracle QA/QC database. Level 2b is a non-real time QA/QC process that will be deployed in the future using *R*, an open source statistical software package (R Core Development Team, 2010). The Level 2b software application will use the Level 2a data product from the Oracle database using Oracle JDBC connections to *R*. In Level 2b, we perform statistical analyses to examine for outliers and drift from the previous 24 hours. These tests include locally weighted regression to fit smooth curves to the data. Locally weighted regression smoothing splines (LOWESS) weight observations closest to a particular independent variable value more highly and use these observations to predict the response, thus forming a smooth curve. In addition, further tests include quartile examination and Individual Moving Range control charts for outlier detection and Cumulative Sum (cusum) Control Charts for drift detection (Montgomery, 1997). The output from this level is exported to a PDF file using Sweave a tool that embeds the R code and output for complete data analyses in LaTeX documents. The PDF document is generated from the previous day and be ready for analysis by project staff. The product from Level 2b is primarily targeted for internal management of the sensor network to indentify problematic sensors that are failing over-time due to sensor drift or power failures, but also to document data integrity.

A Level 3 QA/QC product is the current final step of the QA/QC processes. The product from Level 3 will be

similar to that of Level 2b, except that it will run at longer time periods and will include the use of local correlation (LoCo) score to identify longer-term drift (Papdimitriou et. al., 2006). The local correlation score is computed using an exponential window to estimate the local autocovariance matrix so that observations closest to the time considered are weighted more than observations further away. The score considers the correlation between observations at multiple sensors and the autocovariance between observations within a sensor to discover sensor drift and describe overall sensor performance. The LoCo score is implemented in *R*. Output from this level is exported using Sweave, a tool that embeds R code and output in LaTeX where it can be compiled to a PDF file. The PDF report consists of details on statistical analyses for any given sensor.

Automated QC processes are vital functions that support the QA program for the Intelligent River program. However, automated QC procedures are of little use without the ability to visualize and examine data once flagged and to inspect data for trends and other potential errors. Thus, a key component that we developed was the QC dashboard. This is an Adobe® Air® product that runs outside of a web browser but functions in a similar manner as a web-based application. The overarching goal was to develop a tool that allows the user to query, plot and export data. In addition to reporting data, QC flags are available for review and QC Level 1 flags are plotted as overlay points on the line plots. Level 1 and Level 2 flags are exported along with time stamp and data value. As use of this tool increased, we found that we needed a comment and reporting tool to document system, platform and sensor errors and/or failures. This functionality was included into the QC dashboard and reporting is integrated into the Intelligent River© web site.

DISCUSSION

Monitoring technology is increasingly becoming ubiquitous across industry, agriculture and the environment. Our efforts are focused on the development of robust quality assurance procedures that ultimately can support diverse and dense sensor networks. To meet this goal there are several processes that are critical to support a well rounded quality assurance program including: 1) development and maintenance of metadata for data and instrument hardware; 2) performing quality control using automated and manual procedures; 3) documentation of system and instrument hardware performance that is accessible and acted upon by program personnel; and 4) final verification / certification of data quality [5]. Our focus has been on the development of automated QC processes. These efforts have resulted in processes that increase in complexity with the ultimate goal of providing a rigorous infrastructure that allow data management personnel to monitor network performance and provide high quality data meeting the needs of multiple user groups.

CONCLUSION

The use of automated QC procedures has enabled our ability to identify primary data quality concerns in a real-time environment, which in turn, is supported by a layer of notification services including email, mapping and RSS. Ideally, these types of processes are enabling a greater level of efficiency in the Intelligent River© by providing feedback to program personnel concerning sensor performance over varying time frames. Our future efforts will continue to focus on the development of automated QC procedures at the middleware level to identify potential sensor drift, through the use of statistical procedures. Ultimately, these processes serve as tools and facilitate publication of verified data and support program management.

ACKNOWLEDGEMENTS

This work was supported through the Clemson PSA Remote Sensing Initiative, Clemson Computing and Information Technology, and the NSF (CNS-0745846). The authors gratefully acknowledge these agencies for their support.

LITERATURE CITED

Eidson, G.W., S.T. Esswein, J.B. Gemmill, J.O. Hallstrom, T.R. Howard, C.J. Post, C.T. Sawyer, K.C. Wang, and D.L. White, 2009. The South Carolina Digital Watershed: End-to-end support for realtime management of water resources, The 4th International Symposium on Innovations and Real-time Applications of Distributed Sensor Networks, May 18-21, 2009, Hangzhou China.

Friebrich, C.A., R.A., McPherson, C.C., Fain, J.R., Henslee, and P.D., Hurlbut, 2005. An End-to-End Quality Assurance System for the Modernized COOP Network. Preprints, 15th Conference on Applied Climatology, Savannah, GA, American Meteorological Society, CD-ROM, 3.3

Montgomery, D. C., 1997. Introduction to Statistical Quality Control. John Wiley & Sons, Inc. 3rd edition.

Open Geospatial Consortium Inc. OpenGIS©: Sensor Model Language (SensorML) Implementation Specification (2007). OGC 06-028r5. Open Geospatial Consortium Inc..

Papadimitriou, S., J. Sun, and P.S., Yu. P, 2006. Local Correlation Tracking in Time Series. Sixth IEEE International Conference on Data Mining, pp. 456-465.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Smyth, S.A., L. Lishman, S. Kleywegt, L.M. Svoboda, H-B. Lee, and P. Seto, 2008. Pharmaceuticals and Personal Care Products in Canadian Municipal Wastewater, Proceedings of the Water Environment Federation, WEFTEC 2008: Session 41 through Session 50, pages 3505-3518(14).

Thompson, L.G., E. Mosley-Thompson, H. Brecher, M. Davis, B. León, D. Les, P. Lin, T. Mashiotta, and K. Mountain, 2006. Abrupt tropical climate change: Past and present, Proceedings of the National Academy of Sciences of the United States of America, 103(28), 10536-10543.

World Meteological Organization (WMO), 2004. Guidelines on Quality Control Procedures for Data from Automatic Weather Stations. CBS/OPAG-IOS (ET AWS-4)/Doc. 4(1), Annex 1, p.4