International Workshop on Advances in Hydroinformatics
4 – 7 June 2007, Niagara Falls Canada

CYBERINFRASTRUCTURE TECHNOLOGIES TO SUPPORT QA/QC AND EVENT-DRIVEN ANALYSIS OF DISTRIBUTED SENSING DATA

Yong Liu[1], Barbara Minsker[2], and David Hill[2]

[1]National Center for Supercomputing Applications, 1205 W. Clark St., Urbana, IL 61801, USA

[2]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Av., Urbana, IL 61801, USA

Abstracts

The deployment of distributed sensor networks in the environment provides unprecedented real-time data streams that are useful for building data-driven modeling/forecasting and decision support. However, poor quality of data, which cannot be used effectively in models, is not uncommon due to site/system failures or transmission failures. Except for some simple threshold cutoffs/filtering, most existing data quality and assurance approaches involve visual inspection by human experts and can not meet real-time decision support needs. Automated approaches for detecting sensor anomalies and alerting data managers in real time are needed to facilitate event-driven collaboration and management of sensors. This paper describes cyberinfrastructure technologies that enable event-triggered anomaly detection and workflow execution, which not only automates the quality assurance (QA) and quality control (QC) process, but also facilitates event-driven collaboration and decision analysis with event-triggered model and workflow execution. This approach can facilitate real-time adaptive monitoring and modeling, and ultimately real-time decision support. A case study is shown in which these cyberinfrastructure technologies were used to support QA/QC of sensors in Corpus Christi Bay, Texas. Implications of this approach to the national environmental observatory initiatives, such as the WATer and Environmental Research Systems (WATERS) network sponsored by the US National Science Foundation (NSF), are also discussed.

Introduction

The rapid development of sensor network technologies in the last few decades has seen many experimental applications in the environmental and hydrological monitoring arena, where physical, chemical and biological sensors have been placed in the field for "in-situ" sensing. Physical sensor networks that measure parameters such as temperature, pressure, humidity, light, sound etc. have been the most cost-effective and reliable sensors that have been used for monitoring the environment, while chemical and biological sensors are maturing rapidly for operational usage. For example, a recent review by Johnson et al. (2007) reported in-situ chemical sensor networks for the aquatic environment which can measure dissolved oxygen, methane and total gas tension, CO_2, pH, nitrate, and other nutrients. Some of these sensor technologies have been used in the field since 1997 in Monterey Bay, California (Johnson et al. 2007). Porter et al. (2005) has reviewed wireless sensor networks used for ecology, where some of the instrumentation methods are applied in the water environment. Wang et al. (2006) describes some geo-referenced environmental monitoring applications of wireless sensor networks. Predictions have been made in a recent Nature article (Butler 2006) that by 2020, in-situ sensing will measuring everything, everywhere. Although there are still many technical obstacles such as power consumption, cost, size, reliability and biofouling to be solved before some large-scale integrated environmental sensor networks can be deployed/augmented, nevertheless, the age of the physical world being intensively instrumented and digitized in real-time is on the horizon.

Compared to the rapid advancement of sensor technologies, information technologies that enable geographically-dispersed scientists and engineers to effectively and collaboratively use data from such distributed sensor networks for real-time decision support, modeling and forecasting are still in their infancy, let alone the

end-to-end support where users can easily find data history (also known as provenance) from sensors to workflows[1] to publications, and know what has been done to the data and by whom. In addition, event-driven analysis is often desirable as adaptive sampling/management is often needed when extreme environment events happen. Poor quality of data is not uncommon due to many factors such as system failures, unknown transmission errors or extreme environmental conditions. Except for some simple threshold cutoffs/filtering, most existing/traditional data quality assurance (QA) and quality control (QC) approaches involve manual visual inspection by human experts and can not meet real-time decision support needs. Technologies for automatic detection of sensor anomalies and alerting data managers in real time are needed to facilitate event-driven collaboration and management of sensors. This paper introduces research and cyberinfrastructure technology development underway at the National Center for Supercomputing Applications (NCSA) to address these needs, supported by the National Science Foundation towards their cyberinfrastructure vision (National Science Foundation 2007).

In the following sections, some related background information on the WATERS Network initiative and previous work on different aspects of using data from distributed sensor networks are discussed. Prototype cyberinfrastructure technologies are next presented, which integrate collaborative portals, knowledge networking tools, and workflow software with a back-end provenance and event management system to enable end-to-end data history from sensor to workflow to publication. These "cyberenvironments" are initially being developed to support emerging environmental observatory initiatives such as the WATERS Network, as well as for research projects on adaptive monitoring and hazard management. Sensor anomaly detection is used as an example QA/QC use case in one of the WATERS network testbed projects in Corpus Christi Bay, Texas, to show the benefits of using integrated cyberinfrastructure technologies. The data-driven algorithm that detects the anomaly values in the data streams is also described. Conclusions then consider future directions and the implications and relationships to other ongoing national observatory efforts.

[1] A workflow is a sequence of steps, such as data preparation and model execution, that are needed to accomplish a particular task.

Background and Related Work

The WATERS Network initiative is motivated by the national need to understand and restore lake, stream, and coastal water quality to achieve sustainable and secure water supply while improving and preserving aquatic habitats (Water Science and Technology Board, 2006). Supported by the US National Science Foundation (NSF) Engineering and Geoscience Directorates, the WATERS Network is being proposed through the Major Research Equipment and Facilities Construction (MREFC) program, which funds major community infrastructure. A design for the network is being created through a joint planning effort derived from initiatives of the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) and the Collaborative Large-scale Engineering Analysis Network for Environmental Research (CLEANER).

The WATERS Network will consist of tiered/multi-scale remote and embedded sensing networks at each of approximately 10 observatory sites in the U.S. that will enable researchers to answer critical questions regarding basin-scale transport and management of water, sediment, and contaminants, including integration of water studies in the natural environmental with water infrastructure, treatment technologies, and social processes. Real-time sensor data flow to models that enable adaptive monitoring and management are a key component of the WATERS Network as currently envisioned.

Processing the data flows from a large sensor network with hundreds of sensors, such as those proposed for the WATERS Network, has been likened to "drinking from a fire hose" (Porter et al. 2005). Finding anomalies and acting upon interesting pattern/events in the data flow is even more challenging. However, informed decisions cannot be made on the basis of unreliable data, and therefore certain levels of data quality must be assured. A monitoring system without adequate QA/QC runs the risk of not being able to control the quality of data, and not being able to assure accuracy and precision. QA/QC has thus become an essential part of all measurement systems in general and environmental observatories in particular, because such community initiatives require

especially high national or even international comparability of data.

Traditionally, QA/QC of environmental and water monitoring program follows certain protocols published by the Environmental Protection Agency (EPA) (e.g., EPA 1987, 1989a, 1989b, 1996, 1998, 2000), US Geological Survey (USGS) (e.g., USGS 2000a, 2000b) and/or the American Public Health Association (APHA) (e.g., APHA 1998). However, none of these methods work well for real-time sensor networks, where data streams are continuous and it is practically impossible to do in-lab experiments to manually detect outliers. One of the earlier efforts in performing QA/QC for real-time sensor data was the "Water on the Web" program for educational usage (http://waterontheweb.org/under/instrumentation/qaqc.html) during 1997-2004. They acknowledge that "*the QA/QC of near-real time remotely collected sensor data has provided challenges that were not present under traditional sampling regimes. ..... future efforts must be directed toward the unique problems posed by real-time data collection*".

Estrin et al. (2003) outlines some of the cyberinfrastructure needs for environmental sensor networks, which emphasizes developing error resilience of the sensor networks, new security and data management systems, metadata systems, and analysis and visualization software. In the past, most of the work related to environmental sensor network only addresses a subset of the problems. For example, a relational database was developed for the monitoring and analysis of watershed hydrologic data (Carleton et al. 2005). Ganesan et al. (2003; 2004) propose and design new data handling, compression and storage architectures for sensor networks. Vivoni and Camilli (2003) describe how to do real-time streaming of environmental field data using handheld computers and use GIS web services to display georeferenced field data. Some papers have addressed the problems of detecting outliers from streaming data. For example, Palpanas et al. (2003) describe some theoretical considerations on how to find outliers from streaming data without giving any concrete case studies. None of these works have addressed both the anomaly detection and the event-driven analysis, let alone the collaboration, provenance, and knowledge networking issues.

The integrated cyberenvironment developed in this project can support distributed yet collaborative QA/QC for observatories with large-scale distributed sensor networks. In the following sections, some of the technical details are discussed.

Event-Driven Cyberinfrastructure Technologies

The reactive nature of real-time environmental forecast and modeling requires an event-based system. Responding to real-time changes and events in a timely manner is one of the important requirements for adaptive sensing and management. Event-driven execution of workflow and event-driven collaboration are functionalities to facilitate such needs. Non-scheduled, event-driven collaboration will promote much faster turn-around time for research and projects. Event-Driven-Architecture (EDA) is adopted as a software architectural approach for meeting such requirements. EDA defines a methodology for designing and implementing applications and systems in which events transmit between loosely coupled software components and services. Building applications and systems around an event-driven architecture allows these applications and systems to be constructed in a manner that facilitates more responsiveness, since event-driven systems are, by design, more normalized to unpredictable and asynchronous environments (Michelson 2006).

Within the context of this EDA technology discussion, an event is defined as a message generated by an object, describing an aspect of the system's state or history (e.g., temperature at a specific geographic location at a particular time). A sensor anomaly event, thus, can be considered as a sensor reading that appears to deviate markedly from other members of the data sample in which it occurs. There are usually three components from the event-processing point-of-view: event generators, event broker, and downstream event-driven consumers. An event generator can be a component in the system (such as a workflow, a portlet, or a sensor) that produces events when appropriate (e.g., when an anomalous data point is identified by a workflow). The event broker usually does filtering, routing and deciding where to send the event. A downstream event consumer can be any other component in the cyberinfrastructure system (such as a desktop

dashboard that summarizes events for users, a portlet, a workflow etc.).

In the current implementation, a publish/subscribe mechanism is used, based on the Java Message Service (JMS) standard (http://www.onjava.com/pub/a/onjava/2001/05/03/jms_primer.html) to allow the propagation of the event information from the event generator to the consumer via an event broker (ActiveMQ is used for this purpose. See http://activemq.apache.org/ for more information about the JMS broker). For example, if a user wants to be notified by email whenever a sensor anomaly occurs, he would subscribe to a specific "sensor anomaly" topic whose messages will only be sent out to the subscribers when an anomaly reading happens. A workflow component that can detect anomalies in a data stream can also subscribe to a specific topic, such as a wind speed data stream from a wind sensor, so that it can continuously process the data whenever it arrives.

The EDA forms the backbone of the cyberinfrastructure and allows other components, described next, to interact and communicate. This essentially creates an event-driven cyberinfrastructure system.

Integrated Cyberenvironment

An integrated cyberenvironment promotes individual innovation and group collaboration. To meet this goal, the following components are integrated in the system:

1) ***A collaborative portal***: a web-based portal called CyberCollaboratory (Liu et al. 2006) has been created to provide a centralized access point for users to find relevant groups, data, tools, documents, etc. and collaborate with each other based on common interests. Since it is web-based, users only need to have a web browser (such as Internet Explorer or FireFox) to access it. The CyberCollaboratory supports customizable community spaces for different groups, where each group can have its own collection of tools and users. Users in these online virtual organizations can write individual blogs or post announcements to group blogs. A mediaWiki-based (http://www.mediawiki.org) wiki system for supporting collaborative writing was also provided. Discussion boards (or forums) are also being used and integrated with e-mail listservs so that bi-directional posting (email to forum and forum to email) is supported. Other documents such as Word or PDF file or PowerPoint slides are kept in document libraries where users can also post threaded comments on individual files. In addition, other tools such as text chat, data mining tools, or visualization tools can be integrated as portlets (a portlet is a software component that resides inside the portal). Figure 1 shows the community space for the WATERS Network Project Office in the CyberCollaboratory, where over 100 researchers across the U.S. have been using this site to support planning for the WATERS network.

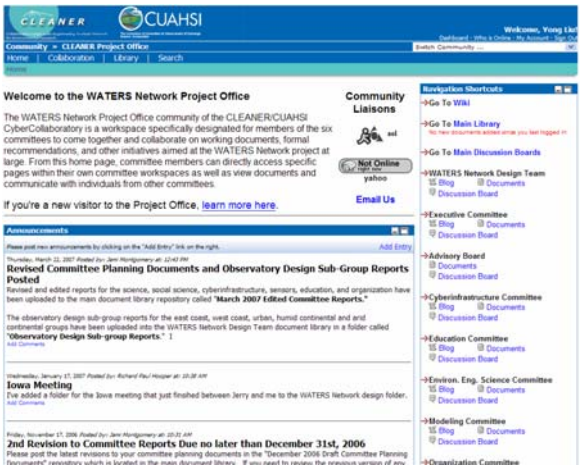

Figure 1. Screenshot of WATERS Network Project Office community space in the CyberCollaboratory.

Currently, several other projects, including the Corpus Christi Bay WATERS Testbed Observatory project have their own spaces in the CyberCollaboratory.

An accompanying component to the CyberCollaboratory called CyberDashboard (a Java standalone desktop application, which users can download and run on their desktop computers) tracks users' activities (file uploads, discussion board postings, users recently logged in, etc) and allows individuals to monitor events within communities of interest so they can remain connected to their groups without always logging in.

2) ***A scientific workflow tool***: Scientific workflows are usually defined as "*networks of analytical steps that may involve, e.g., database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high-performance cluster computers*" (Ludäscher et al. 2006). A scientific workflow integration tool called CyberIntegrator (Marini et al. 2006) was

created to allow users to link different external tools such as Excel spreadsheets, ArcGIS modules or C++ codes (such as the sensor anomaly detection code) together with different input data sets. This tool can be either launched from within the CyberCollaboratory as an Applet running inside the web browser or as a standalone Java desktop application.

3) ***A knowledge network***: The social networking capability provided by web sites such as MySpace (http://www.myspace.com) and Facebook (http://www.facebook.com) has received much attention recently. Such capability has been leveraged in this system with more extended and contextualized scientific information (Green et al. 2006). A tool called CI-KNOW (CyberInfrastructure Knowledge Network on the Web) has been integrated into the system that mines users' activities and interactions when using the CyberCollaboratory or the CyberIntegrator and presents the results either in a graphical network or as user recommendations. Users can browse a contextualized scientific knowledge network in a clickable graphical network format, where relevant documents, publications, people, workflow, data, sensor, etc. are readily accessible and recommended that are related to users' current tasks. Such network creation is supported by the provenance and metadata service described next.

4) ***Provenance and metadata service***: Provenance and metadata service provides the semantic information about who uses/associates with what data/resources at what time. The current approach taken in this system is to enforce global unique identifiers and the initial implementation has focused on agreement on unique global user identifiers across different components (Futrelle et al. 2006). The underlying Tupelo 2 system provides a log-style API (Application Programming Interface) for different components (e.g., wiki, discussion board, document library, or workflow tool) to record RDF (Resource Definition Framework) triples. An RDF triple is in the form of <subject><predicate><object>, where the subject identifies a particular resource or entity such as a user or a tool, and the predicate expresses a relationship between the subject and the object. For example, one way to represent the notion "George has run the sensor anomaly detection model" is as a triple: <George> <has run> <the sensor anomaly detection model>. Such a triple is encoded as an XML (eXtensible Markup Languages) string so that computers can process it. The RDF information can be used by CI-KNOW to construct the scientific knowledge network that graphically represents relationships as discussed previously.

Sensor Anomaly Detection Case Study

To show the benefits of these integrated components and services, consider a sensor anomaly detection case study in Corpus Christi Bay, Texas. This case study focuses on the Corpus Christi Bay (CCBay) WATERS Testbed Observatory, where an interdisciplinary team of hydrologists, environmental engineers, information technologists, and biologists are currently collaborating to improve understanding of hypoxia in the bay (Minsker et al. 2006a). Hypoxia occurs when dissolved oxygen (DO) in aquatic environments is reduced to less than 30% saturation or (~ 2 mg/L) where most fish cannot live. Hypoxia seems to be correlated with salinity-induced stratification, but the causes of stratification and spatial and temporal patterns of hypoxia are currently uncertain. The objectives of the testbed project are to: (1) explore how sensor data can be used to guide adaptive sampling, (2) create improved models of hypoxia, coupling numerical hydrodynamic and oxygen models with data mining methods, (3) demonstrate how these information sources can be integrated into emerging cyberinfrastructure tools to create an environmental information system (EIS) for collaborative research and decision support.

Figure 2 shows the geospatial locations of the Corpus Christi Bay study area and the available sensor platforms in that region. There are multiple sensors administered by different organizations, including the Shoreline Environmental Research Facility (SERF, a research facility affiliated with Texas A&M University College Station), the Texas Coastal Ocean Observation Network (TCOON), and the Harte Research Institute (HRI) at Texas A&M University-Corpus Christi, (marked as UTMSI in Figure 2.). Ojo et al. (2007) gives details on the design concepts and field implementation of the SERF sensor platforms in CCBay.

The following QA/QC scenario for CCBay could be envisioned for using the cyberinfrastructure technologies discussed previously:

A data manager, George Smith, subscribes to an alert service that provides a message whenever the sensor data streams show anomaly values. The alerts are triggered by an anomaly detection workflow running on the server, which continuously processes data streams from Corpus Christi Bay in near-real-time. One day, George gets an email saying that there are many anomalous readings in the windspeed measurements from one of the sensors. George logs into the CyberCollaboratory, where he joins an ongoing discussion with his collaborators, who also received the alert, and are looking at the real-time data plots and anomaly values on the sensor monitoring page in the CyberCollaboratory. The group agrees that the anomaly detection algorithm appears to be malfunctioning after looking at nearby sensors that also measure windspeed. Using the CI-KNOW knowledge network, George discovers another sensor anomaly algorithm used for windspeed data at a different observatory. He then clicks on that sensor anomaly algorithm in the knowledge network (see Figure 3), which launches the CyberIntegrator workflow engine in the web browser and loads the new sensor anomaly detection algorithm into the CyberIntegrator. He then changes the input data stream to be the windspeed data stream from Corpus Christi Bay. He then launches this new workflow to the Observatory server by clicking the "publish" button, which begins running the workflow on the CCBay windpseed data stream immediately. Now the new sensor anomaly detection workflow produces a new anomaly data stream that is accessible for subscription or use by all other users who have permission.

In the following paragraphs, some technical details are discussed on how such a scenario can be supported using the technologies described previously.

Sensor Map Portlet

First, a sensor map portlet was created in the CCBay testbed community space in the CyberCollaboratory by integrating Google Map (http://maps.google.com/) and the locations of the sensor platforms (latitude and longitude) in the study area, shown in Figure 4. This type of integration technique is called mashup of web applications. The resulting portlet has a list of sensor platform names on the right and a live google map ("live" means that all functionalities of Google Map are available.) on the left with clickable icons for different sensor platforms.

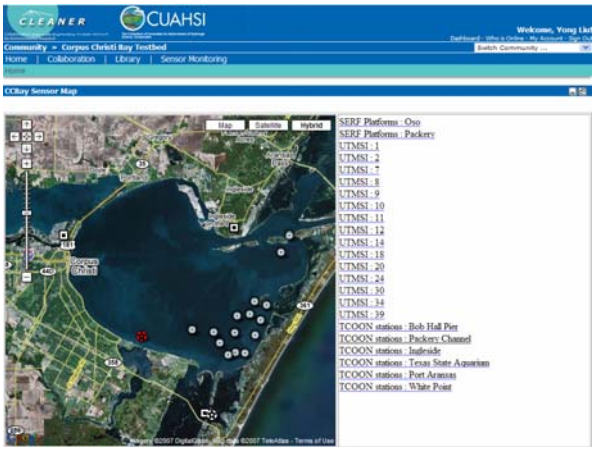

Figure 2. Study area and sensor locations in Google Map for the CCBay testbed observatory in the CyberCollaboratory.

When a user clicks on a sensor platform icon on the map or a name on the left, a popup window will show up on the map with two tabs. The first tab shows basic information about the sensor platform such as the list of sensors available at this platform and a picture of the platform. The second tab is for user subscription to a specific data stream. For example, Figure 4 shows the information window (the top screenshot) of the SERF Oso sensor platform that has a list of sensors such as ADCP, Microcat, MetStation, FL3, Optode, and LISST. The bottom screenshot in Figure 3 shows the subscription window for a particular sensor "MetStation" that measures windspeed. Users can then subscribe to either the raw data stream or the anomaly data stream. The raw data stream will come directly from the sensor, while the anomaly data stream is the result of running a sensor anomaly detection workflow. The notification method for subscribers can be either through email or the CyberDashboard that the user runs on his/her desktop. For demonstration purposes, historical data were used to simulate the windspeed data stream (Minsker et al. 2006b), but the technology can be hooked up with real-time data when it is available. In the future, CUAHSI WaterOneFlow web services (CUAHSI, 2006) will be integrated to allow direct access to real-time data stored on other servers.

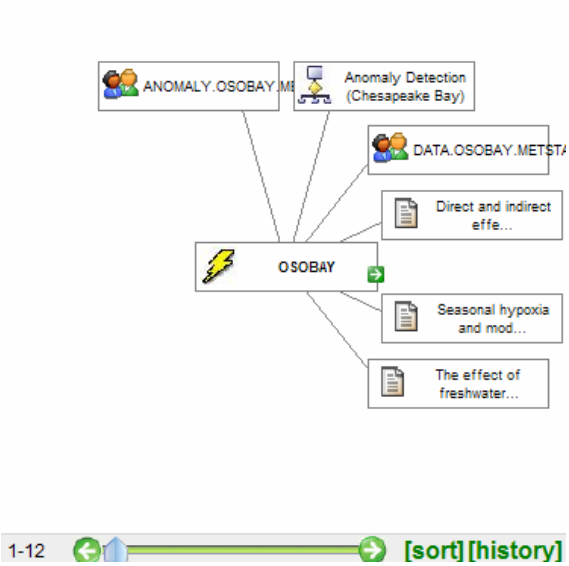

Figure 3. Knowledge network for the sensor anomaly detection case study

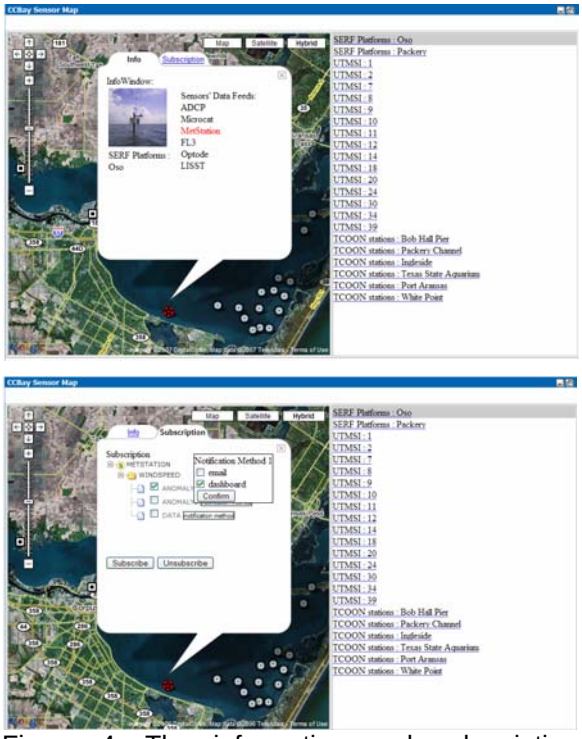

Figure 4. The information and subscription windows for the SERF Oso sensor platform.

Sensor Anomaly Detection Algorithms

To detect sensor anomalies in environmental data streams, a suite of new data-driven approaches are being developed and tested that will be deployed in the event-driven CI in the future. The methods can be divided roughly into two categories: (1) autoregressive model-based anomaly detection and (2) Bayesian anomaly detection. An overview of these methods and initial findings are given below.

The autoregressive model-based anomaly detection methods use a data-driven autoregressive model of the sensor data stream to act as a simulated redundant sensor whose measurements can be compared with those of the actual sensor. The classification of a measurement as anomalous is based on the difference between the model prediction and the sensor measurement. The performance of several data-driven modeling approaches were compared, including nearest neighbor, clustering, perceptron, and artificial neural networks for providing the autoregressive model. The variance of the "now-cast" predictions was calculated using 10-fold cross validation, and deviation of the sensor data from its corresponding now-cast prediction that is greater than the variance of the model will indicate that the data point was caused by either sensor mal-operation or a process anomaly. It was shown that this method, along with a neural-network model of the sensor data stream, outperformed the other modeling methods for detecting errors in Corpus Christi windspeed sensor data. In fact, these anomaly handling strategies identified a significant number of erroneous measurements in the windspeed data that manual QA/QC had failed to detect. The errors had durations ranging from 1 second to several minutes and affected approximately 6% of the data. After cleaning the errors in the training data, an assessment of several autoregressive anomaly detection strategies identified the best performing strategy to be a neural network detector using a 95% prediction interval, which had false positive and false negative rates of only 1% and 2%, respectively. The autoregressive model-based methods, however, are limited because they cannot consider several data streams at once and because missing values in the data stream render them incapable of classifying measurements that immediately follow the missing values.

The Bayesian anomaly detection methods address these shortcomings by employing dynamic Bayesian networks (DBNs). DBNs are Bayesian networks with network topology that evolves over time by adding new state variables

to represent the system state at the current time. Filtering (e.g. Kalman filtering or Rao-Blackwellized particle filtering) can then be used to infer the expected value of unknown system states, as well as the likelihood that a particular sensor measurement is anomalous. Measurements with a high likelihood of being anomalous are classified as such. Like the autoregressive model based methods, these methods perform fast, incremental evaluation of data as it becomes available, scale to large quantities of data, and require no a priori information regarding process variables or types of anomalies that may be encountered, and they can be easily deployed on large networks of heterogeneous sensors. However, unlike the autoregressive model based methods, these methods can operate on a single sensor data stream, or they can consider several data streams at once using all of the streams concurrently to perform coupled anomaly detection, and they are robust in the presence of missing values. Ongoing work is currently investigating the Bayesian methods' abilities, using both coupled and uncoupled detection, to perform QA/QC on the Corpus Christi sensor data.

Automated Event-Driven QA/QC Workflow

Sensor anomaly detection algorithms such as those described above, as well as other event-driven workflows such as real-time models, can be integrated into the CyberIntegrator and run on the server continuously to process the data stream. The workflow is configured so that it subscribes to the incoming data stream from the sensor and continuously processes the data to find any possible anomalies and produce its findings as a new data stream: the anomaly data stream. The anomaly data stream typically consists of the abnormal data value and other metadata description about the value (such as the time and location of the measurement).

Furthermore, users can change the workflow parameters on the fly and publish the new workflow to a server that can then produce new data streams for community use, as described in the above scenario where George loads a different sensor anomaly workflow and changes its parameters so that it accepts the data stream of his choice. This goes beyond just publishing and sharing workflow templates for others to use locally, but allows users to start adding new resources to an observatory system. This is central to end-user customization and can facilitate individual's innovation and community collaboration. Users of such an integrated cyberenvironment for environmental observatories are no longer just passive consumers (e.g., getting data from the observatory), but also active participants and contributors. This resembles what the users' roles are in a typical Web 2.0 environment (e.g., Wikipedia is a user-generated encyclopedia on the web) (see also Atkins 2007, NCSA interview scripts). The cyberinfrastructure technologies developed in this project create a distributed yet collaborative QA/QC system to support large scale distributed sensor network.

Automated Sensor Monitoring Portlet

A sensor monitoring portlet was also developed to visualize the real-time sensor data stream in the CyberCollaboratory. The sensor monitoring plot can automatically plot/update the real-time sensor data streams (both raw and derived) inside the web browser in real-time without the user pressing any refresh button. This helps users to identify any questionable event in the data stream visually, as well as monitoring the measurement graphically (see Figure 5).

Conclusions

This paper presented a novel cyberinfrastructure to support distributed and collaborative QA/QC and event-driven analysis for distributed sensor networks. An event-driven-architecture was adopted to build the integrated cyberenvironment. The sensor anomaly detection case study at the Corpus Christi Bay of Texas demonstrated the functionalities and benefits of such a system.

A successful Observatory system such as the WATERS network needs a highly effective QA/QC system. The prototype cyber-infrastructure technologies and algorithms presented here can help to meet such demands. Furthermore, the same event-driven analysis capability can support real-time data mining, modeling, and decision support that goes beyond just QA/QC. Event-triggered modeling can provide automated forecasting that will enable researchers to adaptively monitor infrequent events and significantly improve the lead time needed for emergency management such as stormwater management, pollution events, and flood control.

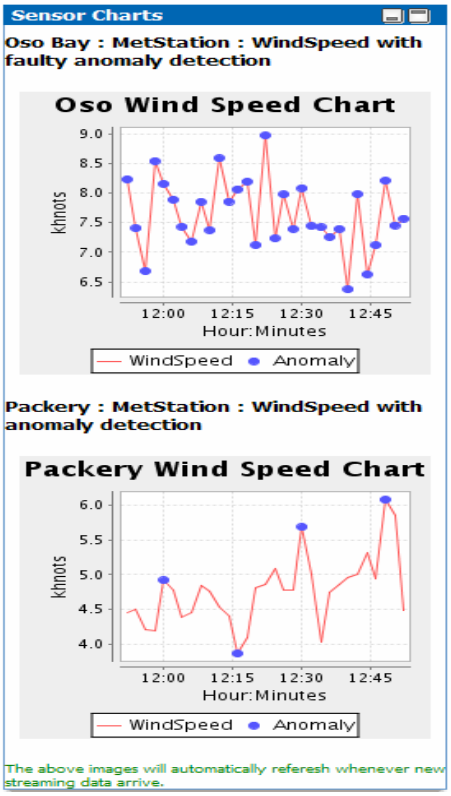

Figure 5. Sensor monitoring portlet showing both the raw data stream (Windspeed in knots) and simulated anomaly values in two different locations (Oso and Packery).

Acknowledgments

The authors thank the entire NCSA ECID (Environmental Cyberinfrastructure Demo) project team and the CCBay testbed observatory team, who contributed significantly to the approaches and findings discussed in this paper. Funding for this work came from the National Science Foundation (BES-0414259, BES-0533513, and SCI-0525308) and the Office of Naval Research (N00014-04-1-0437).

References

APHA. 1998. Standard methods for the examination of water and wastewater. 20th edition. American Public Health Association, Washington, D.C.

Atkins, D. 2007. Where the rubber hits the road: NCSA Interview. NCSA Access Magazine. Available: http://access.ncsa.uiuc.edu/Stories/Atkins/

Butler,D. 2006. 2020 computing: Everything, everywhere. Nature, 440:402-405. Available: http://www.nature.com/news/2006/060320/pdf/440402a.pdf

Carleton,C. J., Dahlgren, R. A., and Tate,K. W. 2005. A relational database for the monitoring and analysis of watershed hydrologic functions: I. Database design and pertinent queries. Computers & Geosciences, 31(4):393-402.

Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI). 2006. WaterOneFlow Webservices. [Online]. http://www.cuahsi.org/his/webservices.html

EPA. 1987. Handbook of methods for acid deposition studies-Laboratory analysis for water chemistry. EPA/600/4-87-026

EPA. 1989a. Preparing perfect project plans. US EPA Risk Reduction Engineering Laboratory, Cincinnati, OH, EPA/600/9-89/087.

EPA.1989b. Handbook of methods for acid deposition studies-Field operations for surface water chemistry. EPA/600/4-89-020.

EPA. 1996. The Volunteer Monitor's Guide to: Quality Assurance Project Plans. EPA 841-B-96-003, Sep 1996, U.S. EPA, Office of Wetlands, Washington, D.C. 20460, USA (http://www.epa.gov/owowwtr1/monitoring/volunteer/qappexec.htm)

EPA. 1998. Guidance for Quality Assurance Project Plans EPA QA/G-5. EPA/600/R-98/018, Feb 1998 (http://www.epa.gov/quality1/qs-docs/g5-final.pdf). U.S. EPA, Washington, D.C. 20460

EPA. 2000. Delivering timely water quality information to your community: The Lake Access-Minneapolis project. EPA/625/R-00/012, September 2000, U. S. Environmental Protection Agency, Office of Research and Development, Cincinnati, OH, 45268, USA.

Estrin, D.,Michener, W., Bonito, G. 2003. Environmental Cyberinfrastructure Needs for Distributed Sensor Networks: A Report from a National Science Foundation

Sponsored Workshop, 12–14 August 2003, Scripps Institute of Oceanography. Available: http://www.lternet.edu/sensor_report/

Futrelle, J., Myers, J., Minsker, B., and Bajcsy, P. 2006. Community-based Metadata Integration for Environmental Research. Presented at the Seventh International Conference on Hydroscience and Engineering (ICHE-2006), Philadelphia, PA, September 10-13, 2006.

Ganesan, D., Estrin, D., and Heidemann, J. 2003. DIMENSIONS:Why do we need a new data handling architecture for sensor networks? ACM SIGCOMM Computer Communication Review 33:143–148.

Ganesan, D., Ratnasamy, S., Wang, H., and Estrin, D. 2004. Coping with irregular spatio-temporal sampling in sensor networks. ACM SIGCOMM Computer Communications Review, 34(1): 125-130.

Green,H. D., Contractor, N. S., and Yao, Y. 2006. CI-KNOW: Cyberinfrastructure Knowledge Networks on the Web. A Social Network Enabled Recommender System for Locating Resources in Cyberinfrastructures. Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

Johnson, K. S., Needoba, J. A., Riser, S. C., and Showers, W. J. 2007. Chemical Sensor Networks for the Aquatic Environment. Chemical Reviews, 107(2):623 – 640.

Liu, Y., Downey, S., Minsker, B., Myers, J., Wentling, T., and Marini, L. 2006. Event-Driven Collaboration through Publish/Subscribe Messaging Services for Near-Real- Time Environmental Sensor Anomaly Detection and Management. Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J., Zhao, Y. 2006. Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience, Special Issue on Workflow in Grid Systems, 18(10):1039-1065.

Marini,L., Minsker,B., Kooper,R., Myers, J., and Bajcsy, P. 2006. CyberIntegrator: A Highly Interactive Problem Solving Environment to Support Environmental Observatories. Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

Michelson, Brenda M. 2006. Event-Driven Architecture Overview: Event-Driven SOA is Just Part of the EDA Story. Patricia Seybold Group, February 2006. Available: http://dx.doi.org/10.1571/bda2-2-06cc.

Minsker, B., Coopersmith, E., Hodges, B., Maidment, D., Bonner, J., and Montagna, P. 2006a. An Environmental Information System for Hypoxia in Corpus Christi Bay: A WATERS Network Testbed. Presented at AGU 2006 Fall Meeting, San Francisco, CA, December 11-15, 2006

Minsker, B., Myers, J., Marikos, M., Wentling, T., Downey,S., Liu, Y., Bajcsy,P., Kooper, R., Marini, L., Contractor, N., Green, H., Yao, Y., Futrelle, J.. 2006b. Environmental CyberInfrastructure Demonstrator Project: Creating Cyberenvironments for Environmental Engineering and Hydrological Science Communities. Presented at Supercomputing Conference 2006 (SC06), Tampa, FL, November 13-17, 2006.

National Science Foundation. 2007. Cyberinfrastructure Vision for 21st Century Discovery. Available: http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

Ojo,T. O., Bonner,J. S., Page, C. 2007. A Rapid Deployment Integrated Environmental and Oceanographic Assessment System (IEOAS) for Coastal Waters: Design Concepts and Field Implementation. Environmental Engineering Science, 24(2): 160-171.

Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunonpulos, D. 2003. Distributed Deviation Detection in Sensor Networks. SIGMOD Record, 32(4):77-82.

Vivoni, E.R., and Camilli, R. 2003. Real-time streaming of environmental field data. Computers & Geosciences, 29:457-468.

Wang, N., Zhang, N., Wang, M. 2006. Wireless sensors in agriculture and food industry - Recent development and future perspective. Computers and Electronics in Agriculture, 50 (1):1-14

Water Science and Technology Board (WSTB), Division on Earth and Life Studies, National Research Council of the National Academies. (2006). CLEANER and NSF's Environmental Observatories. National Academy Press. 2006. Available:

http://orsted.nap.edu/openbook.php?record_id=11657&page=R1

USGS. 2000a. Guidelines and standard procedures for continuous water-quality monitors: Site selection, field operation, calibration, record computation, and reporting. R.J. Wagner, H.C. Mattraw, G.F. Fritz and B.A. Smith. U.S. Geological Survey Techniques of Water-Resources Investigations Report 00-4252 (http://water.usgs.gov/pubs/wri/wri004252/). U.S. Geological Survey, Reston, Virginia, USA.

USGS. 2000b. National field manual for the collection of water-quality data. U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chaps. A1-A9, 2 v., variously paged.