Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data

JOHN L. CAMPBELL, LINDSEY E. RUSTAD, JOHN H. PORTER, JEFFREY R. TAYLOR, ETHAN W. DERESZYNSKI, JAMES B. SHANLEY, CORINNA GRIES, DONALD L. HENSHAW, MARY E. MARTIN, WADE M. SHELDON, AND EMERY R. BOOSE

Sensor networks are revolutionizing environmental monitoring by producing massive quantities of data that are being made publically available in near real time. These data streams pose a challenge for ecologists because traditional approaches to quality assurance and quality control are no longer practical when confronted with the size of these data sets and the demands of real-time processing. Automated methods for rapidly identifying and (ideally) correcting problematic data are essential. However, advances in sensor hardware have outpaced those in software, creating a need for tools to implement automated quality assurance and quality control procedures, produce graphical and statistical summaries for review, and track the provenance of the data. Use of automated tools would enhance data integrity and reliability and would reduce delays in releasing data products. Development of community-wide standards for quality assurance and quality control would instill confidence in sensor data and would improve interoperability across environmental sensor networks.

Keywords: computers in biology, informatics, instrumentation, environmental science

Streaming sensor networks have advanced ecological research by providing enormous quantities of data at fine temporal and spatial resolutions in near real time (Szewczyk et al. 2004, Porter et al. 2005, Collins et al. 2006). The advent of wireless technologies has enabled connections with sensors in remote locations, making it possible to transmit data instantaneously using communication devices such as cellular phones, radios, and local area networks. Advancements in cyberinfrastructure have improved data storage capacity, processing speed, and communication bandwidth, making it possible to deliver to end users the most current observations from sensors (e.g., within minutes after their collection). Recent technological developments have resulted in a new generation of *in situ* sensors that provide continuous data streams on the physical, chemical, optical, acoustical, and biological properties of ecosystems. These new types of sensors provide a window into natural patterns not obtainable with discrete measurements (Benson et al. 2010). Techniques for rapidly processing and interpreting digital data, such as webcam images in investigations of tree phenology (Richardson et al. 2009) and acoustic data in wildlife research (Szewczyk et al. 2004), have also enhanced our understanding of ecological processes. Access to near-real-time data has become important for human health and safety, serving as an early-warning system for hazardous environmental conditions, such as poor air and water quality (e.g., Glasgow et al. 2004, Normander et al. 2008), and natural disasters, such as fires (e.g., Hefeeda and Bagheri 2009), floods (e.g., Young 2002), and earthquakes (e.g., Hart and Martinez 2006). Collectively, these changes in the technological landscape are altering the way that environmental conditions are monitored, creating a platform for new scientific discoveries (Porter et al. 2009).

Although sensor networks can provide many benefits, they are susceptible to malfunctions that can result in lost or poor-quality data. Some level of sensor failure is inevitable; however, steps can be taken to minimize the risk of loss and to improve the overall quality of the data. In the ecological community, it has become common practice to post streaming sensor data online with limited or no quality control. That is, these data are often delivered to end users in a raw form, without any checks or evaluations having been performed. In such cases, the data are typically released provisionally with the understanding that they could change in the future. However, when provisional data are made publically available before they have been comprehensively checked, there is the potential for erroneous or misleading results.

BioScience 63: 574–585. ISSN 0006-3568, electronic ISSN 1525-3244. doi:10.1525/bio.2013.63.7.10

As streaming sensor data become more common, there is an increasing need for automated, algorithm-based methods to check and correct data to ensure that products posted online in near real time are of the highest quality. Sensor network technology is becoming mainstream at a time when scientific data quality is being increasingly scrutinized (COSEPUP 2009). These new data streams require automated quality assurance (QA) and quality control (QC; together, QA/QC), because the manual methods that ecologists have historically used are inadequate for the volumes of data produced by sensor networks and the time constraints imposed by near-real-time data processing. Automated QA/QC also expedites postprocessing (e.g., gap filling, drift correction) so that the final data are released sooner. Development of community-wide QA/QC standards will improve confidence in the data and will enhance the quality of cross-site syntheses.

Recent interest and investment in large-scale environmental observatories (e.g., the National Ecological Observatory Network [NEON], the Ocean Observatories Initiative, the Arctic Observatory Network, the Critical Zone Observatory) have highlighted a need for automated and standardized approaches to QA/QC as ecological research enters the era of "big data" (Hamilton et al. 2007, Lynch 2008). In ecology and related environmental sciences, most individual sensor network data sets are still relatively small (in the gigabyte range or less), although, in aggregate, their volume may approach the tera- to petabyte per year data sets found in other fields (e.g., astronomy, physics, medicine). However, a shift toward larger individual ecological data sets is occurring. For example, NEON sites are projected to collectively generate 200–400 terabytes per year (Schimel 2011). In addition to the sheer volume of these data, the large variety of sensor types can make them challenging to manage.

In the present article, we discuss QA/QC in the context of environmental sensor networks, drawing concepts from diverse disciplines. To understand complex ecosystem functioning, many different types of sensors are needed. Here, we focus on data collected with *in situ* sensors and do not discuss QA/QC for images, which is a topic that is being addressed by the photogrammetry and remote sensing community (e.g., Shuai et al. 2008, Honkavaara et al. 2009). We identify reasons that sensors fail and how these failures can be minimized or avoided. We then describe methods for detecting and flagging suspect data and procedures for incorporating corrective measures into data streams. Finally, we highlight the best practices and approaches for implementing automated QA/QC procedures to facilitate broader adoption by the ecological community.

QA versus QC

The concepts of QA and QC are often used together and are closely related, but each has a distinct meaning. The major difference is that QA is *process* oriented, whereas QC is *product* oriented. In the context of sensor networks, this leaves much room for interpretation, because it is difficult to determine when the data become a product. For our purposes, we define QA as a set of processes or steps taken to ensure that the sensor network and protocols are developed and adhered to in a way that minimizes inaccuracies in the data produced. The purpose of QA is to produce high-quality data while minimizing the need for corrective measures to improve data quality. QC, however, occurs after the data are generated and tests whether they meet the necessary requirements for quality outlined by the end users. QA is a proactive or preventive process to avoid problems; QC is a process to identify and flag suspect data after they have been generated.

Many QA/QC procedures can be automated (Shafer et al. 2000, Durre et al. 2010). For example, an automated QA procedure might monitor the cumulative depth of water in a rain gauge and alert a technician when it needs to be emptied. An automated QC procedure might identify anomalous spikes in the data and flag them. Even though it is almost always necessary to have some level of human intervention and inspection in QA/QC (Fiebrich and Crawford 2001, Fiebrich et al. 2006, Peppler et al. 2008), the inclusion of automated QA/QC is often an improvement, because it ensures consistency and reduces human bias. Automated QA/QC is also more efficient at handling the vast quantities of data that are being generated by streaming sensor networks and reduces the amount of human inspection required. Because automated QA/QC can be performed instantaneously (i.e., as the data are collected), inaccurate data are flagged and corrected more quickly than can be done manually. However, great care must be taken to ensure that valid data are not removed and that all processing steps are well documented so that they can be evaluated. Fully automated QC has limitations: There is a risk that real and potentially important phenomena will be ignored, such as when a real but extreme value is censored for falling outside an expected range. To ensure that this does not happen, data flagged as suspect should be reviewed carefully, and the raw (unmanipulated, preprocessed) data should always be saved.

QA

Environmental sensors can produce poor-quality data or fail completely for many reasons (Ganesan et al. 2004). They can be damaged or destroyed both by natural phenomena, such as floods, fire, lightning strikes, and animal activity (figure 1), and by malicious human activity (e.g., theft, vandalism). Sensors can also malfunction when they are not maintained properly or when they are operated in unsuitable environments. Loss or inadequate supply of electricity can cause sensor network faults, as can power surges (Suri et al. 2006). Even when sensors are working properly, the data can be corrupted during transmission because of factors such as adverse environmental conditions, an inadequate power supply, electromagnetic interference, and network congestion (Hill and Minsker 2006).

Figure 1. Examples of reasons that sensors malfunction: (a) a tower that broke at the Hubbard Brook Experimental Forest, New Hampshire, after a tree fell on a guy line (photograph: Ian Halm); (b) a meteorological station damaged by wildfire at the Sevilleta, New Mexico, Long Term Ecological Research Network site (photograph: Renee Brown); (c) a mud dauber wasp nest built on the underside of the antenna of a radar stream stage height recorder at the Baltimore, Maryland, Water and Environmental Research Systems test bed (photograph: Philip C. Larson); (d) mice nesting by a shaft encoder used to measure stream stage height at the Hubbard Brook Experimental Forest (photograph: Don Mower).

When sensors stop functioning altogether, the resulting loss of data may be obvious (with the notable exception of event detectors). What is more challenging is detecting subtle impairments during which data are still produced but with compromised quality. These types of problems may result from environmental conditions, such as excessive moisture or extreme temperatures that exceed the operating range of the sensor. Steps can be taken to avoid or at least minimize sensor failures when designing sensor networks and establishing protocols; however, the costs of these preventive measures must be balanced against the risk of data loss. Some QA procedures may seem excessive in terms of equipment expense or labor, but they may be warranted in circumstances in which there is a low tolerance for data loss.

In instances in which data are of crucial importance, there may be justification for installing replicate sensors

at the same location. An even higher level of precaution involves installing replicate sensors on separate dataloggers (or other data collection systems), each with an independent power supply. The National Oceanic and Atmospheric Administration's US Climate Reference Network is an example of a sensor network in which the primary meteorological measurements (air temperature, precipitation, soil moisture, soil temperature) are made with triple redundancy. The high-quality data being produced at the 114 sites in this network ensure that a national climate signal can be detected over the long term with a high degree of confidence (Diamond et al. 2013).

Aside from minimizing data loss due to sensor failure, sensor replication is useful for detecting subtle anomalies such as calibration drift, which are often difficult to discern with unreplicated sensors. Drift can occur as sensor components deteriorate over time because of age-related processes including corrosion, fatigue, and photodegredation. Sensor drift can also result from biofouling, which is especially common with submerged sensors and can be controlled with regular cleaning (Kotamäki et al. 2009). A minimum of three replicate sensors is typically needed to detect drift, because, with only two, it is difficult to determine which sensor is drifting. The use of roving reference sensors (i.e., sensors that are rotated around locations in the network) is a less expensive alternative to having multiple replicate sensors deployed simultaneously at each location.

Sensors require routine maintenance and scheduled calibration that, in some cases, can be done only by the manufacturer. Ideally, maintenance and repairs are scheduled to minimize data loss (e.g., snow-depth sensors repaired during the summer) or staggered in such a way that data from a nearby sensor can be used to fill gaps. In cases in which unscheduled maintenance is required, stocking replacement parts on site ensures that any part of the network can be replaced immediately.

Field technicians are often aware of sensor-related inaccuracies resulting from routine maintenance, repairs, or other interruptions of service. Tracking these events is crucial for identifying and understanding the origin of inaccurate data. The once-standard field notebooks are being routinely replaced by weatherproof handheld computers to document this information. At the Hubbard Brook Experimental Forest in New Hampshire, technicians track known events on handheld devices with electronic forms that have pull-down menus to ensure uniformity. When the technician returns from the field, the digital notes are downloaded and automatically synchronized with the sensor data using the date and time stamp. In some cases, even simpler automated methods suffice. For example, at the Virginia Coast Reserve, light sensors are installed in datalogger boxes to indicate when they were opened to perform maintenance.

Although site visits will always be necessary (Fiebrich et al. 2006), continuous monitoring of sensor network data and functionality can improve response times and provides insight into problems so that researchers can be better equipped to perform field repairs. This advancement has enhanced QA by making it possible to monitor systems and evaluate data from distant locations in near real time. When there are problems (e.g., low battery voltage, faulty data transmission), alerts can be automatically issued, sending e-mail, text, or telephone messages to respondents (Shafer et al. 2000). Remote sensor access also makes it possible for technicians to alter programs and to troubleshoot system failures as they occur.

QC

Robust automated QC procedures are essential for quickly identifying inaccurate data. Properly functioning QC procedures will accept valid data and reject invalid data. A false positive result occurs when good data are falsely marked as invalid, and a false negative occurs when erroneous data are accepted as valid. Analyses of the circumstances under which false positive and false negative errors occur provides information that can be used to further adjust the QC procedure to achieve optimal performance (Fiebrich and Crawford 2001, Durre et al. 2008). The efficacy of QC procedures can be tested with synthetic data sets or real data sets that contain seeded errors (Hubbard et al. 2007).

Automated QC methods are becoming increasingly necessary as the volume of data being collected by sensor networks grows. Manual methods may suffice for data sets at the megabyte scale; however, they are not practical at the giga- and terabyte scales that typify large sensor networks (Porter et al. 2012). QC procedures are specific to the type of data and the location at which they are collected. Consequently, there are no universal standards that are applicable in all circumstances. However, some common practices apply to most sensor data and can be customized by setting tolerances that are appropriate for the location and intended use of the data. The six QC tests listed in box 1 are typically applicable to most data generated by sensor networks and can be used to identify anomalies (figure 2).

Beyond these simple QC procedures, methods from the machine-learning community are increasingly being adopted for use with ecological sensor data. These methods represent a data-driven approach to QC, wherein statistical models or classifiers are trained (i.e., they "learn") in an automated fashion using empirical data collected from sensors. This approach requires little knowledge about the sensor hardware or the phenomena being measured. However, it does require an archive of labeled data that contains examples of faulty data, clean data, or both for model training and validation. Discriminative algorithms, such as logistic regression, can encode a functional mapping from a set of inputs (sensor observations) to a set of output labels (*data-anomaly* or *normal* observations). Generative models, such as Bayesian networks, learn a joint probability distribution over the process, generating both the inputs and outputs. In ecology, artificial neural networks, support vector

Box 1. Six quality control tests applicable to most sensor network data.

Date and time. Each data point has a date and time associated with it. Because streaming sensor networks collect data in chronological order, the date–time pairs should be sequential. When data are collected at fixed intervals (e.g., hourly), it is possible to cross-check the recorded and expected date and time. When sensor data are automatically downloaded to a computer file system, comparing the last recorded date and time with the file system time stamp also provides a check for major datalogger clock errors and sensor failure. To use the time stamp effectively, it is important to know when it was applied (e.g., at the beginning, middle, or end of the sampling interval), and the datalogger clock must be calibrated regularly with a reference.

Range. A range check ensures that the data fall within established upper and lower bounds. These bounds can be absolute, based on the characteristics of the sensor or parameter measured (e.g., relative humidity must be between 0% and 100%) or based on historical or expected norms. Long-term data are helpful for setting appropriate bounds, providing information on extreme values (e.g., highest or lowest value ever recorded), statistical norms (e.g., an appropriate number of standard deviations from the mean), and similar metrics based on past observations. When no data exist, bounds may be established using data from nearby locations and refined as more data become available. Customized range tests can account for intra-annual variability, such as cyclical effects that occur over weeks, months, or seasons. For example, the long-term daily bounds used for air temperature measurements are narrower than the range for the entire year (figure 2a).

Persistence. When the same value is recorded repeatedly, it may be indicative of a bad sensor or other system failure. For example, wind speed typically changes continually; a constant value over a period of time therefore suggests that a problem has occurred (figure 2b).

Change in slope. A check for a change in slope tests whether the rate of change is realistic for the type of data collected. A sharp increase or decrease over a very short time interval (i.e., a spike or step function) may indicate that the sensor was disturbed or has malfunctioned (figure 2c).

Internal consistency. Checks for consistency evaluate differences between related parameters, such as ensuring that the minimum air temperature is less than maximum air temperature or that snow water equivalent is less than snow depth (figure 2d). Consistency checks can also determine whether data were collected under unsuitable conditions for a specific sensor. Examples include water temperatures recorded when the sensor was not submerged (i.e., based on corresponding water depth measurements) or when incoming solar radiation was recorded at night (i.e., based on the time of day).

Spatial consistency. If no replicate sensors exist, intersite comparisons are useful, whereby data from one location are compared with data from nearby identical sensors (figure 2e). Several different tests for spatial consistency have been employed in streaming sensor network applications, including spatial regression (e.g., Hubbard and You 2005), differences in the statistical distributions of neighboring stations (e.g., Collins et al. 2006), and the Barnes objective analysis (e.g., Fiebrich and Crawford 2001).

machines, decision trees, and probabilistic models have all become popular machine-learning approaches (Hill and Minsker 2006, Olden et al. 2008, Dereszynski and Dietterich 2011). Software packages such as MATLAB (MathWorks, *www.mathworks.com/products/matlab*) and WEKA (Waikato Environment for Knowledge Analysis; University of Waikato; *www.cs.waikato.ac.nz/ml/weka*; Hall et al. 2009) contain libraries for applying machine-learning algorithms to data. However, even with this advanced software, it can be challenging to apply these tools correctly. Care must be taken to

Figure 2. Examples of sensor quality control failures (see box 1): (a) A range test: An erratic air temperature value (in degrees Celsius) below the lower limit established using long-term (57-year) data. For much of this record, air temperature was measured with a chart-recording hygrothermograph that has been replaced with a digital thermistor (photograph: John L. Campbell). (b) Persistence: Snow and ice coated the anemometer propeller causing a constant zero wind speed reading (in meters per second), despite measurable wind at the time (photograph: Al Levno). (c) Change in slope: An inoperable wiper caused biofilm to develop on a submerged optical sensor that measures turbidity (with the left cylinder in the photograph) and fluorescent dissolved organic matter (FDOM, measured in millivolts; with the right cylinder) in the stream. A sharp increase in FDOM occurred after the sensor was cleaned (photograph: Manual Rosario Torres). (d) Internal consistency: An anomalous snow depth value measured (in centimeters) with an ultrasonic sensor dropped below the snow water equivalent (SWE) measured with a snow pillow. Snow depth should always exceed SWE, which is a measure of the amount of water in the snowpack (photograph: John L. Campbell). (e) Spatial consistency: Air temperature (in degrees Celsius) sensors mounted vertically on a tower at 1.5 and 4.5 meters (m) from the bottom produced comparable values until the snowpack covered the lower sensor, insulating it from fluctuating air temperatures (photograph: Al Levno).

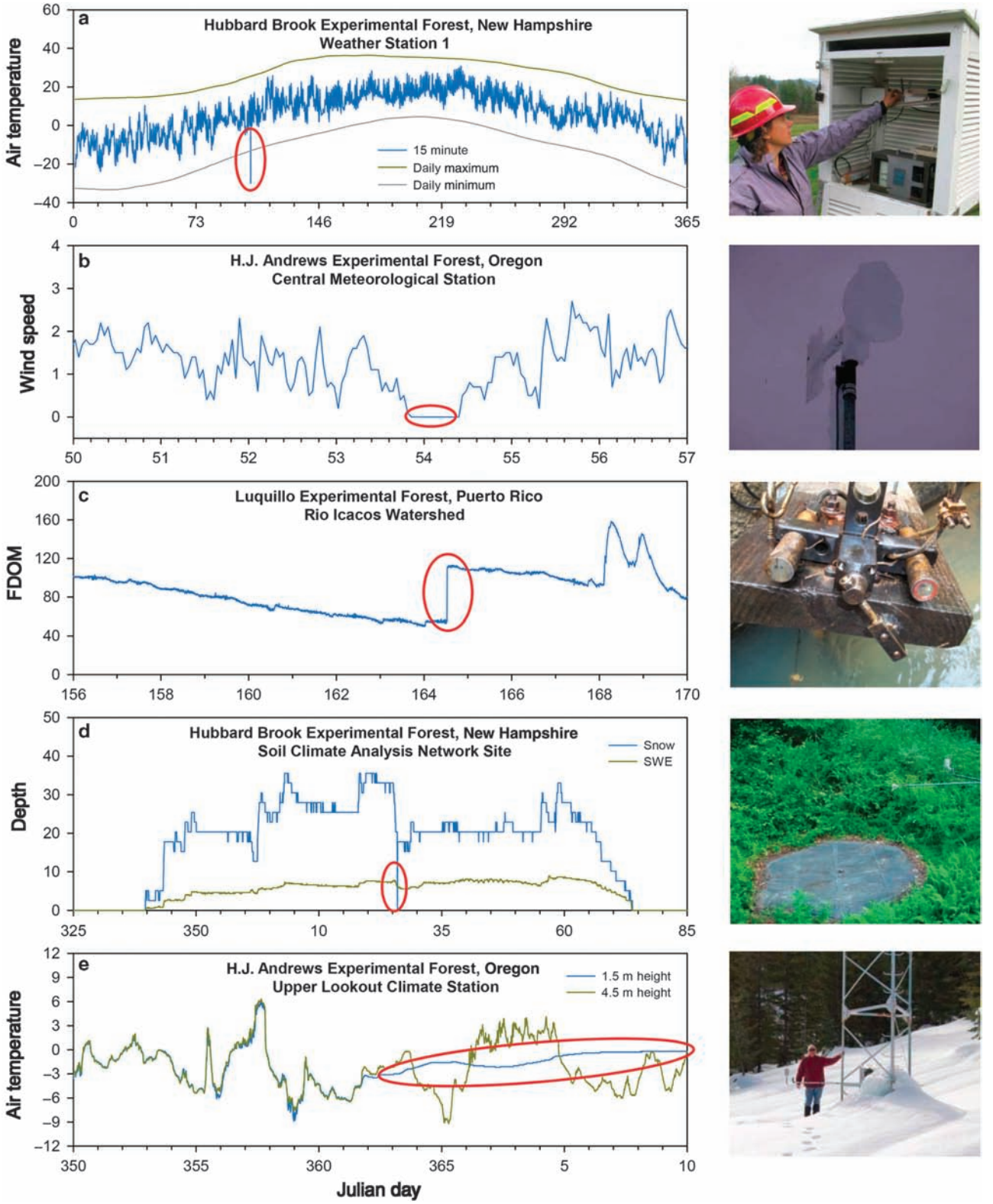

Figure 2. Please see the facing page for the caption.

properly divide the data into training, validation, and test sets. Proper tuning of learned models is also essential to ensure that they generalize well to unseen data and future observations (Solomatine and Ostfeld 2008).

Correcting inaccurate data

Defective or missing data are unavoidable and require decisions on whether to remove, adjust, or replace them with an estimated value. In some cases, the data contain inaccuracies

but are still usable following modification. An example is instrument drift: When the time course of drift is known and is derived from replicate sensor readings or calibrations, simple corrective algorithms can be applied (Horsburgh et al. 2010). However, intermediate or nonlinear drift is more common in environmental applications and is difficult to remedy. For example, abrupt changes in readings following sensor maintenance or recalibration indicate the amount of drift that has occurred but not the starting point and rate of drift.

When it is not feasible to correct the data or when sensors fail completely, gaps arise in the record. Filling these gaps may enhance the data's fitness for use, meaning that it can help meet specific objectives identified by data users (e.g., for calculating annual net fluxes). However, gap filling can be a complex endeavor and can lead to misinterpretation and inappropriate data use. The decision about whether to fill gaps and the selection of the method with which to do so are subjective and depend on factors such as the length of the gap (e.g., days, weeks, months), the level of confidence in the estimated value, and how the data are being used. Many different imputation techniques exist, including linear interpolation, estimates based on historical data, relationships with other stations, and results from process-based models (e.g., Horsburgh et al. 2010). Some corrections can be applied in near real time, whereas others (e.g., sensor drift) can be done only after an adequate period of time has elapsed for the trend or pattern to be sufficiently characterized. Nevertheless, automated anomaly detection minimizes the response time required to detect drift (Moatar et al. 2001), thereby expediting the corrective process.

Flagging suspect data

Flags or qualifiers convey information about individual data values, typically using codes that are stored in a separate field to correspond with each value. Flags can be highly specific to individual studies and data sets or standardized across all data for a program or agency. No community-wide flagging standards currently exist for environmental data; however, there are often commonalities in the information given (table 1). Flags can be used to identify suspect data by indicating failure of one or more QC checks. Moreover, they can provide information about how the data were processed, such as the method used for filling gaps. Flags can indicate data that exceed the acceptable measurement range of the instrument. They can also serve as reminders for scheduling maintenance. For example, when sensors are calibrated, the date can be stored, which enables a flag to be triggered when the next calibration is due. Similarly, flags can indicate sensor data that were collected with an expired calibration.

Flags may be applied at different stages of QC. Some flags may be only for internal purposes, so that the data quality analyst can review suspect data before releasing them to the public. Flags can also become a permanent part of the record, providing crucial knowledge about the data to ensure their appropriate use. In some cases, flags are used to reflect a decision about whether data should be used or rejected. However, these judgments may not be appropriate in all cases, because standards may vary depending on the scientific goals (Daly et al. 2005). A more informative, albeit challenging, approach is to use flags to indicate uncertainty in the value so that the user can decide what is suitable for the intended purpose.

Sensor data contain multiple sources of uncertainty. These uncertainties can be reported individually or combined into a single estimate using standard statistical methods (e.g., the root-mean-square error of the individual component errors; Lehrter and Cebrian 2010). Uncertainties may arise from measurement error associated with the sensor device itself or with the periodicity of sampling. Manufacturers of sensors typically report the measurement error of the instrument. When sufficient resources permit, instrument

Table 1. Examples of flags used to provide information about the data collected.

Type of flag	Example
Internal	
Missing value	No measured value available because of equipment failure or another reason
Low battery	Sensor battery dropped below a threshold
Calibration due	Sensor needs to be sent back to the manufacturer for calibration
Calibration expired	Value was collected with a sensor that is past due for calibration
Invalid chronology	One or more nonsequential date or time values
Persistent value	Repeated value for an extended period
Above range	Value above a specified upper limit
Below range	Value below a specified lower limit
Slope exceedance	Value much greater or lower than the previous value, resulting in an unrealistic slope
Spatial inconsistency	Value greatly differed from values collected from nearby sensors
Internal inconsistency	Value was inconsistent with another related measurement
Detection limit	Value was below the established detection limit of the sensor
External	
Pass	Value passed all quality control tests and is considered valid
Estimated	Estimated value from a model or other sources
Missing	Missing value
Uncertainty	Estimate of uncertainty of the value expressed as a percent

Note: Internal flags are for field technicians and data quality analysts; *external* flags are what the public sees.

uncertainty can be determined by analyzing the measurements obtained from replicate sensors. Selection of the sample interval results in temporal uncertainty, which increases with interval length (e.g., Harmel and King 2005). Even though it may be desirable to have data with a fine temporal resolution, both the volume of data collected and the available supply of power may impose limitations (Suri et al. 2006).

In addition to measurement error, uncertainties may arise from missing data and the methods used to fill gaps. The length of the gap will influence the error term (e.g., Richardson and Hollinger 2007), which is important when streaming sensor data are temporally aggregated (e.g., calculating a daily value from data collected at 5-minute intervals). When a model is used to estimate a value, there is uncertainty in the input data used to run the model, in the model parameters, and in the model's representation of observed processes. The interpretation of estimates of uncertainty is confounded by the many possible options for quantifying and reporting uncertainty. It is essential that the sources of uncertainty be evaluated and that the methods used to quantify uncertainty be thoroughly described in the metadata, which is a nontrivial task because of the many potential sources of uncertainty associated with each data set and the complexity of error propagation.

Implementation of QA/QC

QA and QC procedures can be applied at various stages as data flows from sensors to the end user. Examples include simple programs that run on dataloggers in the field, standalone computer programs that run after data are transmitted to a server, and queries that are applied in a relational database management system. Currently, off-the-shelf software solutions for implementing QC are fairly limited and have not kept pace with advances in hardware. However, tools are increasingly being developed for this purpose (see box 2).

In recent years, scientific workflow systems (e.g., Kepler, Pegasus, Taverna, VisTrails) have been used to implement QA/QC in near real time (Liu et al. 2007, Barseghian et al. 2010, Porter et al. 2012). Automated scientific workflows can compile raw data from various field sensors and perform a series of computations that are executed sequentially (figure 4). Scientific workflow systems are generally flexible, in that they can include a combination of programs and scripts written in different computer languages. They also typically include visualization tools for organizing procedures and evaluating the output generated. An important strength of scientific workflow systems is that they are useful for tracking the provenance (lineage) of the data and processing steps—information that can be collected as part of the workflow (e.g., Altintas et al. 2006, Belhajjame et al. 2008).

As with all scientific data, it is important to document sensor network data provenance in sufficient detail to allow replication (Lerner et al. 2011). The reliability of the data is based, in part, on this capacity to reproduce data products. Preservation of the original data set in its raw, unmanipulated form is crucial for reproducing any subsequent procedures performed on the data. Documentation should describe the QC level (e.g., raw data, qualifiers added, problematic data removed or corrected and the gaps filled) and contain all the necessary information used to generate the data, such as the source file used, data-rejection criteria, gap-filling method, and model parameters. This information enables the data user to carefully scrutinize the data and determine whether the data-processing methods used are appropriate for the particular application. Reviewing uncorrected data may help identify real phenomena that could not be observed with the corrected data. Assigning unique identifiers to various versions of the input data, workflows, QC programs, and models is necessary for retracing steps, so that the procedures can be replicated and reevaluated. As software tools improve in the future, some of the burden of creating and maintaining provenance information is expected to shift from the scientist to the tools themselves.

In recent years, various metadata standards have been developed for environmental data and can be applied to sensors that produce streaming data. SensorML (Sensor Model Language), EML (Ecological Metadata Language), and WaterML (Water Markup Language) are all common metadata standards that use Extensible Markup Language (XML). XML is a flexible and widely used standard for encoding information in a format that is both human and machine readable, which facilitates its use in Internet applications.

Conclusions

Sensor networks are increasingly being used to monitor ecosystems and will soon become the standard approach to recording ecological phenomena around the world. These sensor networks will require rapid and comprehensive QA/QC to ensure the data's quality and usefulness. Automated QA/QC procedures will be essential for dealing with this data deluge, because they can quickly process data and identify and correct problems in near real time without introducing human error. Some procedures, such as corrections for drift, can be done only *post facto*, so it is unlikely that provisional data releases will ever be entirely eliminated. Moreover, it is improbable that QA/QC will become completely automated and replace human decisionmaking and intervention in the foreseeable future. Although the need for automated QA/QC is compelling, it can be challenging to implement. For instance, it is difficult to set QA/QC tolerances in such a way as to minimize false positive and false negative errors, especially under changing environmental conditions. Expert knowledge is often required to make appropriate decisions about how to treat data flagged as problematic. Despite the limitations of automated QA/QC, it can minimize the amount of human intervention required, can improve the quality of the data, and can allow for final data products to be released more rapidly.

Box 2. The Georgia Coastal Ecosystems Data Toolbox for MATLAB.

Innovative software for the quality control (QC) of ecological data has been developed at the Georgia Coastal Ecosystems (GCE) Long Term Ecological Research Network site (Sheldon 2008). The MATLAB-based GCE Data Toolbox (*https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox*) automates the processing of data collected by a wide variety of datalogger systems, from initial acquisition through QC and the distribution of data sets and plots. QC flags can be automatically assigned using rule-based expressions defined for each data column (i.e., sets of algebraic or statistical comparisons that return character flags for values meeting specified criteria). Flagging expressions can contain references to multiple data columns, supporting complex, multicolumn dependency checks (e.g., flagging of all measured values when a hydrographic instrument is out of the water, determined by depth or pressure values). The example in figure 3a shows the QC Flag Criteria Editor menu, where the expressions are defined. In addition to automated rule-based flagging, flags can be assigned manually, in a spreadsheet-like data editor; graphically, by selecting data points with the mouse; and algorithmically, on the basis of parameterized models referencing external data. The graph and enlarged inset (figure 3b) show the interactive visualization tool that indicates flags (in red) that are automatically updated when values are changed.

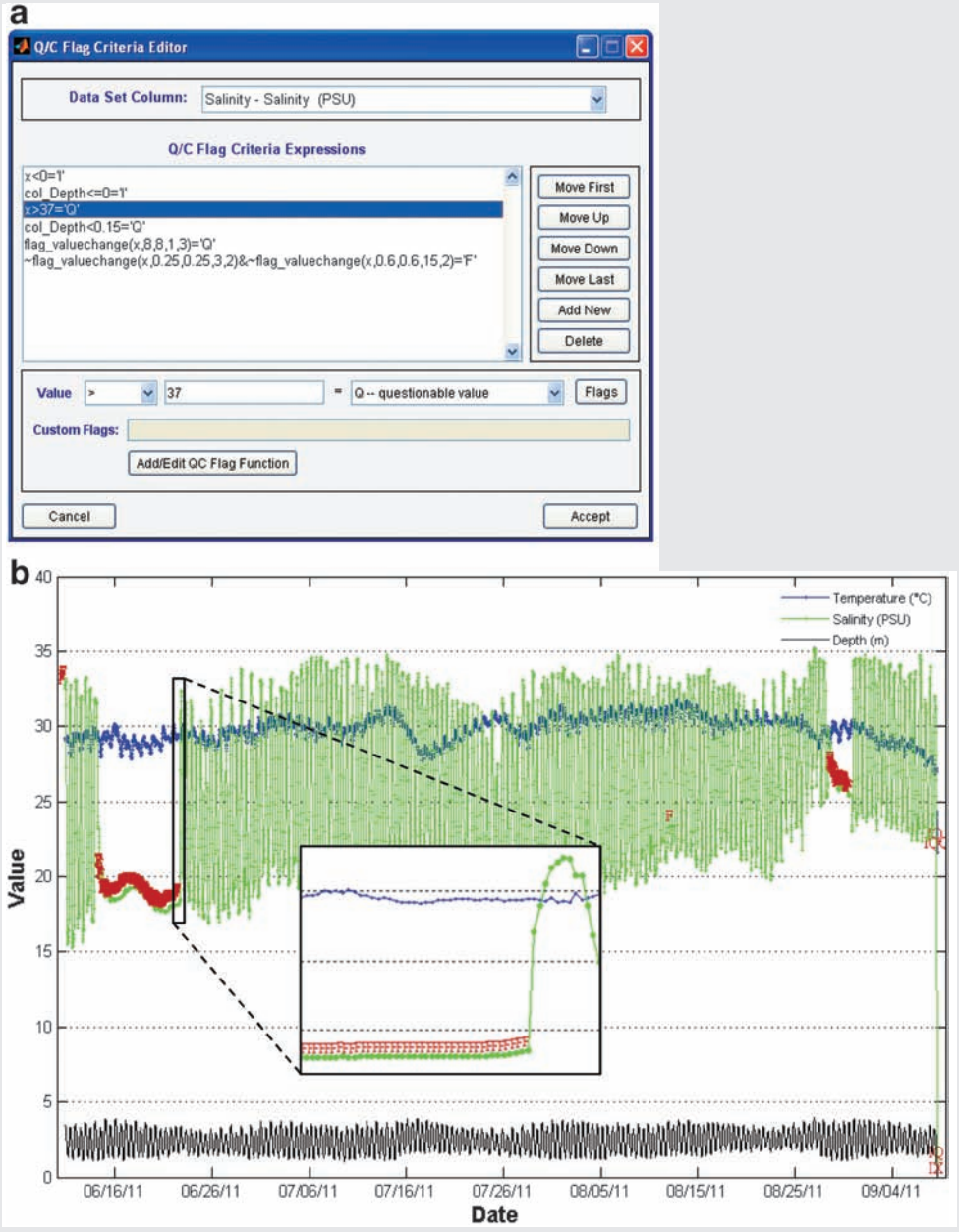

Figure 3. (a) The QC (quality control) Flag Criteria Editor menu. (b) Example sensor readings. The sections in red show data flagged as problematic. Abbreviations: °C, degrees Celsius; m, meters; PSU, practical salinity units.

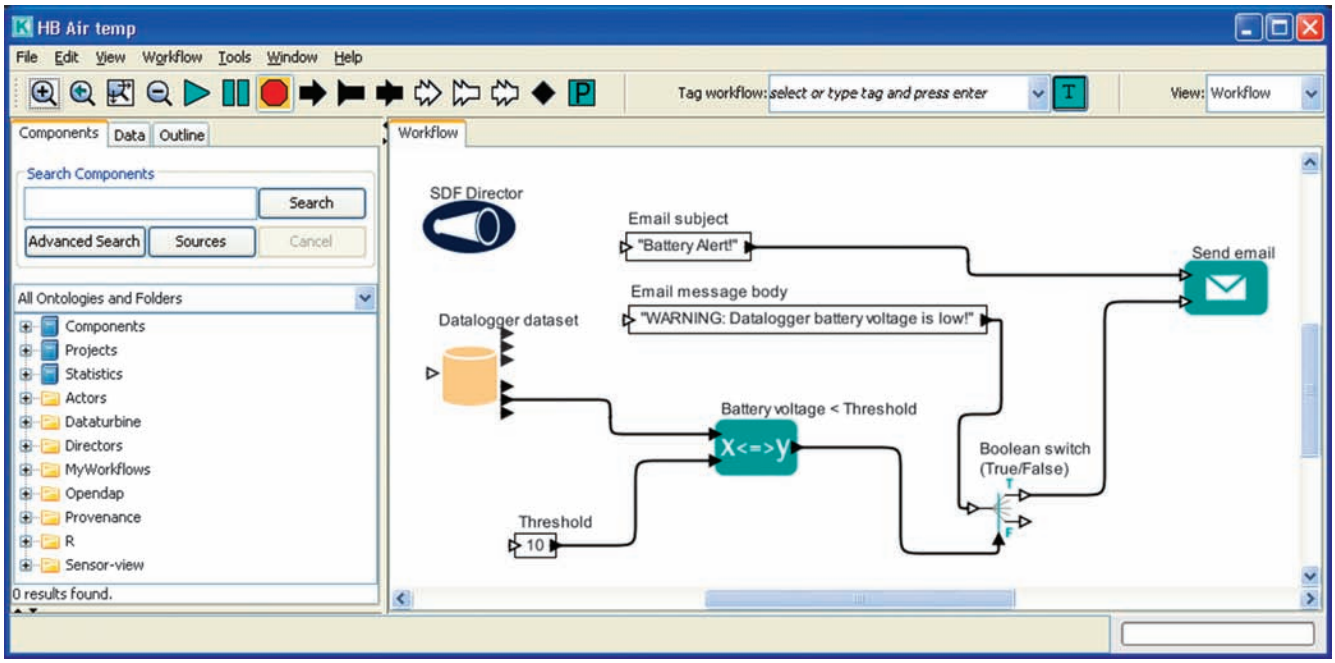

Figure 4. A simple example of a Kepler scientific workflow that checks the battery voltage of a datalogger. The workflow reads the last record of a datalogger file and compares the battery voltage to a critical threshold value. When the battery voltage drops below the threshold, an e-mail alert message is automatically sent to a technician.

Scientific workflow systems are making it easier to implement QA/QC, but they still require custom programming. The development of generic and facile QA/QC software and visualization tools in the future will facilitate the adoption of QA/QC procedures in environmental sensor network applications.

Improvements in sensor technology are simplifying the implementation of automated QA/QC. Sensors with separate logging systems that collect data at fixed intervals are being replaced by intelligent sensors, which have advanced learning and adaptation capabilities. These sensors contain microprocessors that act on environmental cues (e.g., light, sound, motion), thus eliminating superfluous data collection and processing. Intelligent sensors consume less energy, because they operate for shorter periods of time, which minimizes the chance that the sensors will fail as a result of power shortages. Intelligent sensors may also have embedded diagnostic capabilities to monitor their performance and function. Direct two-way communication with sensors will enable technicians to identify problems and take corrective actions as those problems arise.

Sensors are becoming smaller and less expensive, which, when combined with advancements in communications and data storage and transfer, will make it possible to increase the number of sensors deployed for environmental monitoring. Sensor redundancy will become more feasible as a result of these developments, which will make it easier to confirm values and to fill gaps. In the future, sensors will also have the capability of storing crucial metadata. For example, sensors currently being developed for NEON contain embedded metadata, with information such as the sensor's type, identification number, and calibration data. These metadata are automatically uploaded to a database when sensors are installed, to reduce tracking and calibration errors.

Intelligent sensor nodes can actively transmit data to a sensor network server, enabling its integration with ecological data from other networks. The expansion of sensor networks and improvements to streaming data middleware and applications are making it possible to compare data sets across sensor networks. Metadata standardization has also improved our ability to compare data sets. However, no standards currently exist for QA/QC, which raises questions about the reliability of the data sets being compared and the interpretation of the results derived from them. Current QA/QC protocols and procedures are being developed independently, which is inefficient and costly. Studying complex ecosystem behavior requires many different types of sensors that are administered by organizations representing diverse disciplines. To make the best use of available resources and to promote interoperability among the growing number of sites and networks producing streaming sensor data (e.g., the US Long Term Ecological Research Network, NEON, the US Forest Service Experimental Forests and Ranges network, the US Geological Survey Real-Time Water Data program, the US Natural Resources Conservation Service Soil Climate Analysis network, the US Environmental Protection Agency Clean Air Status and Trends Network, the US Department of Energy

Box 3. Some best practices for quality assurance and quality control (QA/QC) of streaming environmental sensor data.

Automate QA/QC procedures.

Maintain an appropriate level of human inspection.

Replicate sensors.

Schedule maintenance and repairs to minimize data loss.

Have ready access to replacement parts.

Record the date and time of known events that may affect measurements.

Implement an automated alert system to warn about potential sensor network issues.

Retain the original unmanipulated data.

Ensure that the data are collected sequentially.

Perform range checks on numerical data.

Perform domain checks on categorical data.

Perform slope and persistence checks on continuous data.

Compare the data with data from related sensors.

Correct the data or fill gaps, if that is prudent.

Use flags to convey information about the data.

Estimate uncertainty in the value, if that is feasible.

Provide complete metadata.

Document all QA/QC procedures that were applied.

Document all data processing (e.g., correction for sensor drift).

Retain all versions of the input data, workflows, QC programs, and models used (data provenance).

Atmospheric Radiation Measurement network), there is a pressing need for the development and adoption of QA/QC standards and best practices. For example, it would be useful to establish the basic QC tests that should be applied, the provenance information to collect, and the naming conventions for flags. Some of the more general best practices identified in this article are summarized in box 3 and may serve as a starting point for the development of QA/QC standards. Further advances toward QA/QC standardization will ensure the reliability of the data used in future synthetic activities.

Acknowledgments

The authors would like to thank David Hollinger and Nicholas Grant for providing helpful comments on an earlier draft of this manuscript. Funding was provided by the Northeastern States Research Cooperative and the National Science Foundation (NSF) through a cooperative agreement to the US Long Term Ecological Research Network Office (grant no. DEB-0832652) and NSF grants no. DEB-0620443, no. DEB-1237733, no. DEB-0822700, no. DEB-0823380, no. DEB-1114804, and no. OCE-0620959.

References cited

Altintas I, Barney O, Jaeger-Frank E. 2006. Provenance collection support in the Kepler scientific workflow system. Pages 118–132 in Moreau L, Foster I, eds. Proceedings of the International Provenance and Annotation Workshop. Lecture Notes in Computer Science, vol. 4145. Springer.

Barseghian D, et al. 2010. Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. Ecological Informatics 5: 42–50.

Belhajjame K, Wolstencroft K, Corcho O, Oinn T, Tanoh F, William A, Goble C. 2008. Metadata management in the Taverna workflow system. Pages 651–656 in Priol T, Lefevre L, Buyya R, eds. CCGRID 2008: Eighth IEEE International Symposium on Cluster Computing and the Grid. Institute of Electrical and Electronics Engineers.

Benson BJ, Bond BJ, Hamilton MP, Monson RK, Han R. 2010. Perspectives on next-generation technology for environmental sensor networks. Frontiers in Ecology and the Environment 8: 193–200.

Collins SL, et al. 2006. New opportunities in ecological sensing using wireless sensor networks. Frontiers in Ecology and the Environment 4: 402–407.

[COSEPUP] Committee on Science, Engineering, and Public Policy. 2009. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. National Academies Press.

Daly C, Redmond K, Gibson W, Doggett M, Smith J, Taylor G, Pasteris P, Johnson G. 2005. Opportunities for improvements in the quality control of climate observations. Paper presented at the 15th American Meteorological Society Conference on Applied Climatology; 20–23 June 2005, Savannah, Georgia. (7 May 2013; *https://ams.confex.com/ams/15AppClimate/techprogram/paper_94199.htm*)

Dereszynski EW, Dietterich TG. 2011. Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. ACM Transactions on Sensor Networks 8 (art. 3). doi:10.1145/1993042.1993045

Diamond HJ, et al. 2013. U.S. Climate Reference Network after one decade of operations: Status and assessment. Bulletin of the American Meteorological Society. (19 April 2013; *http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00170*) doi:10.1175/BAMS-D-12-00170

Durre I, Menne MJ, Vose RS. 2008. Strategies for evaluating quality assurance procedures. Journal of Applied Meteorology and Climatology 47: 1785–1791.

Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS. 2010. Comprehensive automated quality assurance of daily surface observations. Journal of Applied Meteorology and Climatology 49: 1615–1633.

Fiebrich CA, Crawford KC. 2001. The impact of unique meteorological phenomena detected by the Oklahoma Mesonet and ARS Micronet on automated quality control. Bulletin of the American Meteorological Society 82: 2173–2187.

Fiebrich CA, Grimsley DL, McPherson RA, Kesler KA, Essenberg GR. 2006. The value of routine site visits in managing and maintaining quality data from the Oklahoma Mesonet. Journal of Atmospheric and Oceanic Technology 23: 406–416.

Ganesan D, Cerpa A, Ye W, Yu Y, Zhao J, Estrin D. 2004. Networking issues in wireless sensor networks. Journal of Parallel and Distributed Computing 64: 799–814.

Glasgow HB, Burkholder JM, Reed RE, Lewitus AJ, Kleinman JE. 2004. Real-time remote monitoring of water quality: A review of current applications, and advancements in sensor, telemetry, and computing technologies. Journal of Experimental Marine Biology and Ecology 300: 409–448.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: An update. SIGKDD Explorations 11: 10–18.

Hamilton MP, Graham EA, Rundel PW, Allen MF, Kaiser W, Hansen MH, Estrin DL. 2007. New approaches in embedded networked sensing for terrestrial ecological observatories. Environmental Engineering Science 24: 192–204.

Harmel RD, King KW. 2005. Uncertainty in measured sediment and nutrient flux in runoff from small agricultural watersheds. Transactions of the ASAE 48: 1713–1721.

Hart JK, Martinez K. 2006. Environmental sensor networks: A revolution in the Earth system science? Earth-Science Reviews 78: 177–191.

Hefeeda M, Bagheri M. 2009. Forest fire modeling and early detection using wireless sensor networks. Ad Hoc and Sensor Wireless Networks 7: 169–224.

Hill DJ, Minsker BS. 2006. Automated fault detection for in-situ environmental sensors. In Gourbesville P, Cunge J, Guinot V, Liong SY, eds. Hydroinformatics 2006: Proceedings of the Seventh International Conference on Hydroinformatics. Research Publications. (7 May 2013; *http://emsa.ncsa.illinois.edu/Documents/Conference/Hill_Minsker_HIC2006.pdf*)

Honkavaara E, et al. 2009. Digital airborne photogrammetry: A new tool for quantitative remote sensing? A state-of-the-art review on radiometric aspects of digital photogrammetric images. Remote Sensing 1: 577–605.

Horsburgh JS, Jones AS, Stevens DK, Tarboton DG, Mesner NO. 2010. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. Environmental Modelling and Software 25: 1031–1044.

Hubbard KG, Guttman NB, You J, Chen Z. 2007. An improved QC process for temperature in the daily cooperative weather observations. Journal of Atmospheric and Oceanic Technology 24: 206–213.

Hubbard KG, You J. 2005. Sensitivity analysis of quality assurance using the spatial regression approach—A case study of the maximum/minimum air temperature. Journal of Atmospheric and Oceanic Technology 22: 1520–1530.

Kotamäki N, Thessler S, Koskiaho J, Hannukkala AO, Huitu H, Huttula T, Havento J, Järvenpää M. 2009. Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in southern Finland: Evaluation from a data user's perspective. Sensors 9: 2862–2883.

Lehrter JC, Cebrian J. 2010. Uncertainty propagation in an ecosystem nutrient budget. Ecological Applications 20: 508–524.

Lerner B, Boose E, Osterweil L, Ellison A, Clarke L. 2011. Provenance and quality control in sensor networks. Pages 98–103 in Jones MB, Gries C, eds. Proceedings of the Environmental Information Management Conference. University of California, Santa Barbara.

Liu Y, Minsker B, Hill D. 2007. Cyberinfrastructure technologies to support QA/QC and event-driven analysis of distributed sensing data. Paper presented at the International Workshop on Advances in Hydroinformatics; 4–7 June 2006, Niagara Falls, Canada.

Lynch C. 2008. Big data: How do your data grow? Nature 455: 28–29.

Moatar F, Miquel J, Poirel A. 2001. A quality-control method for physical and chemical monitoring data: Application to dissolved oxygen levels in the River Loire (France). Journal of Hydrology 252: 25–36.

Normander B, Haigh T, Christiansen JS, Jensen TS. 2008. Development and implementation of a near-real-time web reporting system on ground-level ozone in Europe. Integrated Environmental Assessment and Management 4: 505–512.

Olden JD, Lawler JJ, Poff NL. 2008. Machine learning methods without tears: A primer for ecologists. The Quarterly Review of Biology 83: 171–193.

Peppler RA, et al. 2008. An overview of ARM Program Climate Research Facility data quality assurance. Open Atmospheric Science Journal 2: 192–216.

Porter J[H], et al. 2005. Wireless sensor networks for ecology. BioScience 55: 561–572.

Porter JH, Nagy E, Kratz TK, Hanson P[C], Collins SL, Arzberger P. 2009. New eyes on the world: Advanced sensors for ecology. BioScience 59: 385–397.

Porter JH, Hanson PC, Lin C-C. 2012. Staying afloat in the sensor data deluge. Trends in Ecology and Evolution 27: 121–129.

Richardson AD, Hollinger DY. 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record. Agricultural and Forest Meteorology 147: 199–208.

Richardson AD, Braswell BH, Hollinger DY, Jenkins JP, Ollinger SV. 2009. Near-surface remote sensing of spatial and temporal variation in canopy phenology. Ecological Applications 19: 1417–1428.

Schimel D. 2011. The era of continental-scale ecology. Frontiers in Ecology and the Environment 9: 311.

Shafer MA, Fiebrich CA, Arndt DS, Fredrickson SE, Hughes TW. 2000. Quality assurance procedures in the Oklahoma Mesonetwork. Journal of Atmospheric and Oceanic Technology 17: 474–494.

Shuai Y, Schaaf CB, Strahler AH, Liu J, Jiao Z. 2008. Quality assessment of BRDF/albedo retrievals in MODIS operational system. Geophysical Research Letters 35 (art. L05407). doi:10.1029/2007GL032568

Solomatine DP, Ostfeld A. 2008. Data-driven modelling: Some past experiences and new approaches. Journal of Hydroinformatics 10: 3–22.

Suri A, Iyengar SS, Cho E. 2006. Ecoinformatics using wireless sensor networks: An overview. Ecological Informatics 1: 287–293.

Szewczyk R, Osterweil E, Polastre J, Hamilton M, Mainwaring A, Estrin D. 2004. Habitat monitoring with sensor networks. Communications of the ACM 47: 34–40.

Young PC. 2002. Advances in real time flood forecasting. Philosophical Transactions of the Royal Society A 360: 1433–1450.

John L. Campbell (jlcampbell@fs.fed.us) and Lindsey E. Rustad are affiliated with the Northern Research Station of the US Department of Agriculture (USDA) Forest Service, in Durham, New Hampshire. John H. Porter is affiliated with the Department of Environmental Sciences at the University of Virginia, in Charlottesville. Jeffrey R. Taylor is affiliated with the National Ecological Observatory Network, in Boulder, Colorado. Ethan W. Dereszynski is affiliated with the Computer Science Department at Oregon State University, in Corvallis. James B. Shanley is affiliated with the New England Water Science Center, US Geological Survey, in Montpelier, Vermont. Corinna Gries is affiliated with the Center for Limnology at the University of Wisconsin–Madison. Donald L. Henshaw is affiliated with the Pacific Northwest Research Station of the USDA Forest Service, in Corvallis, Oregon. Mary E. Martin is affiliated with the Complex Systems Research Center at the University of New Hampshire, in Durham. Wade M. Sheldon is affiliated with the Department of Marine Sciences at the University of Georgia, in Athens. Emery R. Boose is affiliated with Harvard University's Harvard Forest, in Petersham, Massachusetts.