AMBROSia: An Overview and Recent Results

Leana Golubchik*, David Caron, Abhimanyu Das, Amit Dhariwal, Ramesh Govindan, David Kempe, Carl Oberg, Abhishek Sharma, Beth Stauffer, Gaurav Sukhatme and Bin Zhang
University of Southern California, Los Angeles, CA 90089

Submitted: January 2010; Accepted: January 2011

ABSTRACT

Observing systems facilitate scientific studies by instrumenting the real world and collecting corresponding measurements, with the aim of detecting and tracking phenomena of interest. A wide range of critical environmental monitoring objectives in resource management, environmental protection, and public health all require distributed observing systems. The goal of such systems is to help scientists verify or falsify hypotheses with useful samples taken by the stationary and mobile units, as well as to analyze data autonomously to discover interesting trends or alarming conditions. In our AMBROSia project, we focus on a class of observing systems which are *embedded* into the environment, consist of *stationary and mobile* sensors, and *react* to collected observations by reconfiguring the system and adapting which observations are collected next. In this paper, we give an overview of AMBROSia.

1. INTRODUCTION

Observing systems facilitate scientific studies by instrumenting the real world and collecting corresponding measurements, with the aim of detecting and tracking phenomena of interest. In our project, we focus on a class of observing systems which are (1) *embedded* into the environment, (2) consist of *stationary and mobile* sensors, and (3) *react* to collected observations by reconfiguring the system and adapting which observations are collected next. We refer to these as Reactive Observing Systems (ROS). The goal of ROS is to help scientists verify

*Corresponding author. leana@usc.edu

or falsify hypotheses with useful samples taken by the stationary and mobile units, as well as to analyze data autonomously to discover interesting trends or alarming conditions.

We explore ROS in the context of a marine biology application, where the system monitors, e.g., water temperature and light as well as concentrations of micro-organisms and algae in a body of water. Using a hybrid network of stationary and mobile sensors (refer to Section 3), communicating both via wired and wireless links, the system collects fine-grained measurements of interesting information in near real-time. An example use of such a system is the rapid identification of micro-organisms to predict the onset of algal blooms. Such blooms can have devastating economic consequences, as recently seen in [5].

However, current technology (and any realistic prediction of technologies in the near future) precludes sampling all possibly relevant data. For instance, bandwidth limitations between the stationary sensors make it impossible to collect all of the sensed data. Similarly, time and storage capacity constraints for the mobile entities severely curtail the number and locations of samples they can take.

To make good use of the limited resources, we need to develop a framework for ROS capable of optimizing and controlling the set of samples to be taken at any given time, taking into consideration the application's objectives and system resource constraints. To support such an optimization and control process, a significant part of the framework must be dedicated to the development of models of data, and their automatic validation or adaptation. By validation, we mean a process of verifying the accuracy of models, based on collected data, and subsequent discarding or updating/adaptation of those models. As part of the validation and adaptation process, the framework must also include a distributed support mechanism for locating data of interest. We refer to this framework as AMBROSia (Autonomous Model-Based Reactive Observing System).

We seek to develop AMBROSia as a multi-scale modeling framework for ROS. AMBROSia allows applications to construct inter-related models of varying spatio-temporal scope based on collected data. Guided by the models, the reactive elements of the system predict where interesting data and phenomena are likely to be found. In the process of constructing models, the system actively seeks most useful data to improve both the models and phenomenon detection and tracking. In a feedback cycle, this data acquisition is guided by previous, perhaps less precise, models. Thus, AMBROSia enables

optimal collection of measurements in a manner that respects system resource constraints, yet improves the overall fidelity of phenomenon detection and tracking.

An overview of our AMBROSia framework is given in Section 4 and our recent results and research directions are described in Section 5. Our vision for the future of AMBROSia is given in Section 6.

We note that the system we propose to develop is targeted at the marine application outlined above, and described in more detail in Section 2. However, we believe that many of the components we develop, as well as the general AMBROSia framework, may be quite useful in other settings.

2. MONITORING MARINE ECOSYSTEMS

A wide range of critical environmental monitoring objectives in resource management, environmental protection, and public health all require distributed observing systems. Here we focus primarily on a marine biology application. Our application's primary long term scientific goal is to understand, and ultimately predict, the conditions under which specific populations of marine microorganisms develop in nature. A fundamental requirement for attaining this objective is the correlation of environmental conditions with microorganismal abundances at the small spatial and temporal scales that are relevant to the organisms. This is not possible with current technology and methodological approaches. Sampling the environment with high resolution and identifying microorganisms in situ in near-real time will constitute a revolutionary advance in the study of the ecology of marine microbial species. In addition, the rapid identification of aquatic microorganisms will be extremely valuable for the early detection of harmful organisms and the mitigation of their effects on the environment and the human population.

Marine microorganisms such as viruses, bacteria, microalgae, and protozoa have a major impact on the ecology of the coastal ocean. For example, blooms of harmful and/or toxic algae (e.g., red, brown and green tides) in aquatic ecosystems have increased dramatically on a global scale in recent years [2, 3]. These events result in the loss of human life each year, and economic losses in the billions of dollars due to effects on fisheries and tourism. Likewise, the increasing encroachment of humans along coasts has resulted in the recognition of potential public health issues as a consequence of the introduction of pathogenic microorganisms into these waters from land runoff, storm drains and sewage outflow. Similar concerns exist regarding the potential for contamination of drinking water supplies with harmful, pathogenic or nuisance

microbial species. Today, the environmental factors that stimulate the growth of such microorganisms are still poorly understood, and tests for their abundances are not sufficiently rapid to detect the onset of major outbreaks.

We now give a brief illustration of the types of hypotheses which are of interest, i.e., this serves as an example of a potential AMBROSia application. Of course, our goal is not to evaluate these specific hypotheses, but rather to design and build a system capable of aiding scientists in the evaluation of hypotheses.

A popular hypothesis to explain accumulations of harmful microalgae near the shore (resulting in 'red' tides) includes the release of cysts from the sediments, growth in the water column, and then accumulation near shore as a result of favorable weather conditions. Thus, [13] suggests that a combination of winds — specifically wind speed, direction and duration that result in transport of the population towards the coast — may lead to red tides along the coast. Alternate theories posit the importance of upwelling events (the movement of deep water and nutrients contained in deep water into surface waters) or the breaking of internal waves that propagate along subsurface density discontinuities, as contributors to these coastal phenomena.

The Need for a Hybrid Sensing Approach. In tracking the phenomena described above, we encounter the challenge that the relevant locations can not be predicted at deployment time, and indeed, the phenomena themselves migrate over time. At the same time, the application requires prediction and sampling in near real-time. Neither of the traditional approaches for studying spatiotemporal phenomena can adequately deal with both these challenges. Statically deployed sensor networks may not have sensors in the most relevant locations at a given time (and an excessively high sensor density is likely to disturb the phenomenon). On the other hand, a system consisting purely of mobile sensor nodes tends to be too slow in tracking the phenomenon, in particular over a large area.

We therefore propose a hybrid approach, combining a larger number of static sensors with a few mobile sensing robots. The static sensors will be able to monitor basic attributes in near-real time, and infer potentially interesting undersampled locations. These can then be sampled more densely by the mobile sensors. In addition, the mobile entities will be able to collect samples for offline evaluation by human experts in a laboratory. This allows the system to track attributes for which no autonomous sensing devices have been devised yet.

3. A REACTIVE OBSERVING SYSTEM

We have constructed a suite of ten sensing/sampling nodes for deployment in natural aquatic ecosystems. The system consists of ten stationary 'nodes' that sense pertinent environmental characteristics, collate those measurements into a 2/3-D picture of the ecosystem, and guide a small autonomous surface vehicle to desired sampling locations to retrieve samples. Figure 1 shows (from left to right) the sampling system, the robotic boat, one buoy, and the electronics chassis that mounts inside the buoy. The sampling system is a six-port device custom built for this project. A 36-port version has been designed and is being tested. The prototype boat is a modified RC airboat, equipped with a Garmin GPS and compass for navigation. All modules and sensors have been integrated and connected to the main boat processor board. The main board on both the buoys and the boat is an Intel Stargate. A peripheral basic stamp module is used as the sensor interface. The current sensor suite on each buoy consists of an array of thermistors for sampling temperature at different depths and a fluorometer that can measure the concentration of chlorophyll-a (an indicator of phytoplankton abundance). Communication is based on AODV over an 802.1 1b wireless connection. Moreover, a land-based weather station will be constructed and integrated into the network to provide pertinent meteorological information for data interpretation.

A sample of the data collected by the experimental setup shown in Figure 1 is depicted in Figure 2. This figure plots the spatial patterns of chlorophyll and temperature in Lake Fulmor at James Reserve, California. It also shows the fluctuations in phytoplankton and temperature in the vicinity of one of the buoys.

While our current experimental system is on the scale of 10s of stationary nodes, and only one boat, we envisage that as the technology becomes commonplace and affordable, the scale of the system will grow significantly, and may eventually monitor large coastal expanses with 100 s or 1000 s of

Figure 1. Experimental setup.

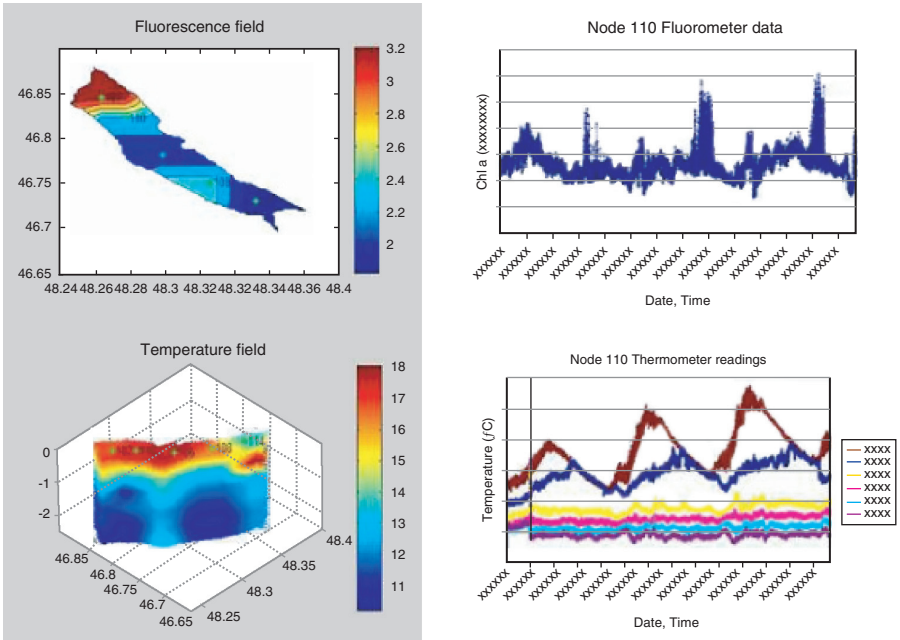

Figure 2. Empirical data from Lake Fulmor, James Reserve.

nodes and 10 s of boats. Hence, AMBROSia will be designed to scale gracefully to this size.

Similarly, at the moment, many attributes of interest require analysis of samples in a laboratory, conducted by an expert biologist. However, we anticipate that as analysis technology improves, the system will increasingly be able to determine quantities of interest autonomously. Hence, our system design will allow for the easy inclusion of additional attributes.

4. HIGH-LEVEL VIEW OF OUR APPROACH

Figure 3 gives a schematic view of AMBROSia. At the core of AMBROSia is a component for the construction, selection, and adaptation of models. Based on the chosen models, a separate unit optimizes future samples and controls their acquisition. The results of these samples in turn affect the decisions of the modeling unit, thus constituting a strong feedback loop in the system. Sampling and control are of course also influenced directly by the requirements of the application, and samples are frequently collected for application-specific purposes, e.g., for scientists to act on. In order to make good decisions about model updates, the model construction unit also requires access to data and

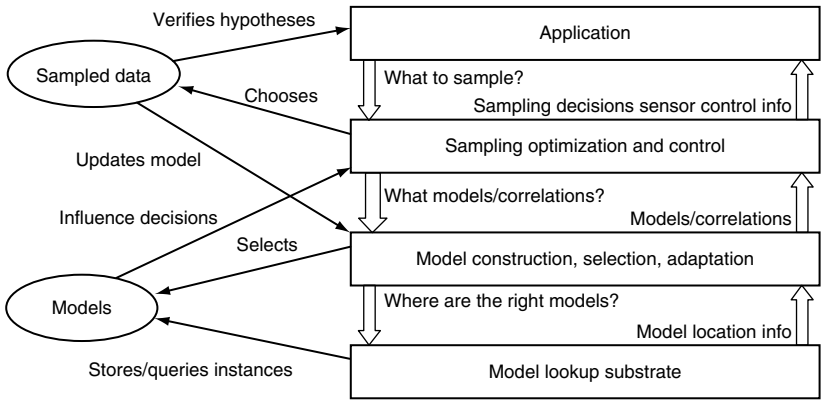

Figure 3.

fine-grained models stored at individual nodes. In order to find the relevant data, it relies on a model lookup substrate.

Models. At a basic level, a "measurement model" is a representation of a sequence of measurements taken by a sensor. These measurements are sampling some phenomenon, such as temperature or chlorophyll, over a period of time. (Notice that the notion of a model thus extends and subsumes the notion of individual data items and measurements.) Due to resource constraints, the representation will frequently be approximate, and could take the form of a time series, histogram, distribution functions, Hidden Markov Models, or decision tree, for instance. In principle, measurement models can be drawn from any desired class of models, as the overall system architecture does not prescribe specific models to use. We do note that good models would not only accurately represent the observed data but would also have predictive value and/or be able to extract interesting features from the observed data.

The measurement models are distributed over the physical nodes, which perform and store the measurements. The nodes maintain detailed models. Some forms of these models are reported to other nodes in the system (for analysis). Nodes can make requests for more or less detailed model forms (including full data), based on need and resource availability.

The measurement models are complemented by "external models". External models are provided by human application experts, and capture prior knowledge about the physics underlying the phenomena. For example, an external model could predict how wind affects water movement between two stationary observation nodes.

Naturally, more powerful inferences and predictions can be made based on combinations of models from multiple sensors, as well as external models. We call such combinations "composite models". An elementary composite model in our example application would be a description of the evolution of the average temperature across a cluster of nodes over time.

Model Lookup. In order to form useful and meaningful composite model, nodes must be able to efficiently locate relevant models at other nodes. Thus, an important component of our research will be the design of a flexible framework for model location, for the purpose of querying or updating them. This lookup system will be akin to distributed naming systems and sensor databases, but due to the higher complexity of our notion of models, will significantly extend the approaches used in those settings. The design of this component will be integrated into the modeling framework; dynamically instantiated composite models will help guide the lookup operations.

Sampling Optimization and Control. Sophisticated algorithms are needed to control the mobile nodes with changing task assignments in an uncertain and dynamic environment. Moreover, measurement, external, and composite models together allow us to make predictions about future states of the system. These predictions are needed to dynamically control the observation system, and determine the best set of measurements to perform next, subject to resource and time constraints. Newly obtained measurements, along with historic data, are also used to dynamically adjust the models. This direct feedback loop requires that the model validation process be automated and made part of the running system. Thus, an important goal here is to develop a framework for autonomous adaptation of models based on collected measurements. Another important aspect of the proposed research is the design of algorithms for selecting samples that will likely be of high use to the driving scientific application, or for improving the quality of models. As an illustration, we now give one example formulation of such a problem.

The sampled data at each of the sensor nodes might have varying correlations with each other and with the result of a user query, and hence would make varying degrees of contributions toward resolving the query. In order to minimize processing and communication costs for sensor-network query resolutions, it is crucial to decide how to trade off accuracy in the query-result against sampling a smaller subset of sensor nodes instead. In particular, if at most k sensors can be queried for a given query, what would be the "best" $\boldsymbol{k}$ sensors to choose that would still predict the given query with sufficiently low error, given a priori statistical knowledge about the correlation between the query result and measurements at the sensor nodes?

In its simplest form, this problem can be mathematically formulated as a subset-selection problem in linear regression [14]: Given a set of n random variables $X_1, X_2, \ldots, X_n$ (corresponding to measurements at individual sensor nodes) and a predictor random variable Z (corresponding to the query), we are required to select the best k out of the n random variables, such that Z can be predicted by a linear combination of these k random variables with minimum least square prediction error. The only information we are given are the statistical variances and covariances between the X_i and Z variables, obtained from previous data. While this problem has been well known in the statistics community, theoretical progress so far has been limited to greedy and local-search heuristics, without a rigorous analysis of error bounds and time complexities involved [4, 14]. Other variants of this problem have been independently proposed in the mathematics community, such as the sparse approximation problem [21]; theoretical results [11, 21] here have been limited to the special case of nearly orthogonal dictionaries where greedy and convex relaxation methods have been shown to provide $(1 + \varepsilon)$ approximation bounds to the optimal solution.

Summary. The novel aspect of our framework is the strong feedback between the data acquisition process and the modeling process. Traditional sensing systems simply model all attributes of the environment, or a user-specified relevant subset. In contrast, our system will make autonomous data acquisition decisions, which in turn inform the models that guide future decisions. We believe that such a tightly coupled approach will lead to significantly more useful data being collected with the same limited resources.

5. RECENT RESULTS AND RESEARCH DIRECTIONS

In this section we give a more detailed description of recent research results and directions we have been pursuing, in the context of AMBROSia.

As already noted, one of the core functionalities of AMBROSia is the selection of samples which (1) can be retrieved at reasonably low energy cost, and (2) yield as much information as possible about the system. The second property in particular will change dynamically: in reaction to past measurements, different observations may be more or less useful in the future. At any given time, the system must select the most informative samples to retrieve based on the model at that point. We briefly outline a mathematical formulation of this problem and results to date in Section 5.1.

In ROS accurate measurements are useful to scientists seeking a better understanding of the environment. However, it may not be feasible to move the static sensor nodes after deployment. In such cases, mobile robots could be used

to augment the static sensor network, hence forming a robotic sensor network. In such networks, an important question to ask is how to coordinate the mobile robots and the static nodes such that estimation errors are minimized. Our recent efforts on addressing this question are briefly outlined in Section 5.2.

The successful use of ROS, as envisioned in AMBROSia, partly depends on the system's ability to ensure the collected data's quality. However, various sensor network measurement studies have reported transient faults in sensor readings. Thus, another important goal in AMBROSia is automated high-confidence fault detection, classification, and data rectification. As a first step towards that goal, we explore and characterize several qualitatively different classes of fault detection methods, which are briefly outlined in Section 5.3.

5.1. A Mathematical Formulation of Sample Selection

Mathematically, our sample selection problem can be modeled naturally as a subset selection problem for regression: Based on the small number of measurements X_i taken, a random variable Z (such as average temperature, chlorophyll concentration, growth of algae, etc.) is to be estimated as accurately as possible. Different measurements X_i, X_j may be partially correlated, and thus partially redundant, a fact that should be deduced from past models. In a pristine and abstract form, the problem can thus be modeled as follows:

We are given a covariance matrix $\mathbf{C}$ between the random variables X_i and a vector $\mathbf{b}$ describing covariances between measurements X_i and the quantity Z to be predicted (C and $\mathbf{b}$ are estimated based on the model). In order to keep the energy sampling cost small, the goal is to find a small set S (of size at most k) so as to minimize the *mean squared prediction error* [8, 11] $\mathrm{Err}(Z,S) := \mathrm{E}[(Z - \Sigma_{i \in S} \alpha_i X_i)^2]$ where the α_i are the optimal regression coefficients specifically for the set S selected.

The selection problem thus gives rise to the well-known *subset selection problem for regression* [14], which has traditionally had many applications in medical and social studies, where the set S is interpreted as a good predictor of Z. Finding the best set S of size k is NP-hard, and certain approximation hardness results are known [6, 15]. However, despite its tremendous importance to statistical sciences, very little was known in terms of approximation algorithms until recent results by Gilbert et al. [10] and Tropp [21] established approximation guarantees for the very special case of nearly independent X_i variables.

In ongoing work, we are investigating several more general cases of the subset selection problem for regression, in particular with applications to

selecting samples to draw in sensor network environments. Over the past year, we have obtained the following key results [6]:

Theorem 1. *If the pairwise covariances between the* $\boldsymbol{X_i}$ *are small (at most* $1/6\boldsymbol{k}$*, if* $\boldsymbol{k}$ *variables can be selected), then the frequently used Forward Regression heuristic is a provably good approximation.*

The quality of approximation is characterized precisely in [6], but omitted here due to space constraints. This result improves on the ones of [10, 21], in that it analyzes a more commonly used algorithm, and obtains somewhat improved bounds. The next theorem extends the result to a significantly wider class of covariance matrices, where several pairs can have higher covariances.

Theorem 2. *If the pairs of variables* $\boldsymbol{X_i}$ *with high covariance (exceeding* $\Omega(1/4\boldsymbol{k})$*) form a tree, then a provably good approximation can be obtained in polynomial time using rounding and dynamic programming.*

While this result significantly extends the cases that can be approximated, it is not directly relevant to measuring physical phenomena. Hence, we also study the case of sensors embedded in a metric space, where the covariance between sensors' readings is a monotone decreasing function of their distance. The general version of this problem is the subject of ongoing work, but [5] contains a promising initial finding:

Theorem 3. *If the sensors are embedded on a line (in one dimension), and the covariance decreases roughly exponentially in the distance, then a provably good approximation can be obtained in polynomial time.*

The algorithm is again based on rounding and a different dynamic program, and makes use of some remarkable properties of matrix inverses for this type of covariance matrix. At the moment, we are working on extending these results to more general metrics (in particular, two-dimensional Euclidean metrics), and different dependencies of covariances on the distance.

5.2. Scalar Field Estimation

Sensor networks provide new tools for observing and monitoring the environment. In aquatic environments, accurately measuring quantities such as temperature, chlorophyll, salinity, and concentration of various nutrients is useful to scientists seeking a better understanding of aquatic ecosystems, as well as government officials charged with ensuring public safety via appropriate hazard warning and remediation measures. Broadly speaking, these quantities of interest are scalar fields. Each is characterized by a single scalar quantity which varies spatiotemporally. Intuitively, the more the readings near the location where a field estimate is desired, the less the reconstruction error.

In other words, the spatial distribution of the measurements (the *samples)* affects the estimation error. In many cases, it may not be feasible to move the static sensor nodes after deployment. In such cases, one or more mobile robots could be used to augment the static sensor network, hence forming a sensor-actuator network or a robotic sensor network.

The problem of adaptive sampling: An immediate question to ask is how to coordinate the mobile robots and the static nodes such that the error associated with the estimation on the scalar field is minimized subject to the constraint that the energy available to the mobile robot(s) is bounded. Specifically, if each static node makes a measurement in its vicinity, and the total energy available to the mobile robot is known, what path should the mobile robot take to minimize the mean square integrated error associated with the reconstruction of the entire field? Here we assume that the energy consumed by communications and sensing is negligible compared to the energy consumed in moving the mobile robot. We also assume that the mobile robot can communicate with all the static nodes and acquire sensor readings from them. Finally, we focus on reconstructing phenomena which do not change temporally (or change very slowly compared to the time it takes the mobile robot to complete a tour of the environment).

The domain: We develop a general solution to the above problem and test it on a particular set up designed to monitor an aquatic environment. The experimental set up is a system of anchored buoys (the static nodes), and a robotic boat (the mobile robot) capable of measuring temperature and chlorophyll concentrations. This test bed is part of the NAMOS (Networked Aquatic Microbial Observing System) project (http://robotics.usc.edu/ ~namos), which is used in studies of microbial communities in freshwater and marine environments [8, 19].

Contributions: We propose an adaptive sampling algorithm for a mobile sensor network consisting of a set of static nodes and a mobile robot tasked to reconstruct a scalar field. Our algorithm is based on local linear regression [17, 10]. Sensor readings from static nodes (a set of buoys) are sent to the mobile robot (a boat) and used to estimate the Hessian Matrix of the scalar field (the surface temperature of a lake), which is directly related to the estimation error. Based on this information, a path planner generates a path for the boat such that the resulting integrated mean square error (IMSE) of the field reconstruction is minimized subject to the constraint that the boat has a finite amount of energy which it can expend on the traverse. Data from extensive (several km) traverses in the field as well as simulations validate the performance of our algorithm.

We are currently working on how to determine the appropriate resolution to discretize the sensed field. One interesting observation from the simulations and experiments is that when the initial available energy is increased, the estimation errors decrease rapidly and level off instead of decreasing to zero. Theoretically, when the energy available to the mobile node increases, more sensor readings can be taken and hence, the estimation errors should keep decreasing. By examining the path generated by the adaptive sampling algorithm, we found that when the initial energy is enough for the mobile node to go through all the 'important' locations, increasing the initial energy does not have much effect on the estimation error. We plan to investigate advanced path planning strategies and alternative sampling design strategies in future work.

5.3. Faults in Sensor Data

With the maturation of sensor network software, we are increasingly seeing longer-term deployments of wireless sensor networks in real world settings. As a result, research attention is now turning towards drawing meaningful scientific inferences from the collected data [20]. Before sensor networks can become effective replacements for existing scientific instruments, it is important to ensure the quality of the collected data. Already, several deployments have observed faulty sensor readings caused by incorrect hardware design or improper calibration, or by low battery levels [16, 20].

Given these observations, and the realization that it will be impossible to always deploy a perfectly calibrated network of sensors, an important research direction for the future will be automated detection, classification, and root-cause analysis of sensor faults, as well as techniques that can automatically scrub collected sensor data to ensure high quality. A first step in this direction is an understanding of the prevalence of faulty sensor readings in existing real-world deployments.

We focus on a small set of sensor faults that have been observed in real deployments: single-sample spikes in sensor readings (SHORT faults), longer duration noisy readings (NOISE faults), and anomalous constant offset readings (CONSTANT faults). Given these fault models, our work makes the following two contributions.

Detection Methods. We have explored three qualitatively different techniques for automatically detecting such faults from a trace of sensor readings. Rule-based methods leverage domain knowledge to develop heuristic rules for detecting and identifying faults. Linear Least-Squares Estimation (LLSE) based methods predict "normal" sensor behavior by leveraging sensor

correlation, flagging deviations from the normal as sensor faults. Finally, learning-based methods (based on Hidden Markov Models) are trained to statistically detect and identify classes of faults.

Our findings indicate that these methods sit at different points on the accuracy/robustness spectrum. While rule-based methods can detect and classify faults, they can be sensitive to the choice of parameters. By contrast, the LLSE method is a bit more robust to parameter choices but relies on spatial correlations and cannot classify faults. Finally, our learning method (based on Hidden Markov Models) is cumbersome, partly because it requires training, but it can fairly accurately detect and classify faults. We also explored hybrid detection techniques, which combine these three methods in ways that can be used to reduce false positives or false negatives, whichever is more important for the application. These results are omitted for brevity and the interested reader is referred to [18].

Evaluation on Real-World Datasets. We applied our detection methods to real-world data sets. Here, we present results from the NAMOS data set [1], where we examine the fraction of faulty samples in a sensor trace.

Nine buoys with temperature and chlorophyll concentration sensors (fluorimeters) were deployed in Lake Fulmor, James Reserve for over 24 hours in August 2006. Each sensor was sampled every 10 seconds. We analyzed the measurements from chlorophyll sensors for the prevalence of faults.

The predominant fault was a combination of NOISE and CONSTANT caused by hardware faults in the ADC (Analog-to-Digital Converter) board. Figure 4(a) shows the measurements reported by buoy 103. We applied the

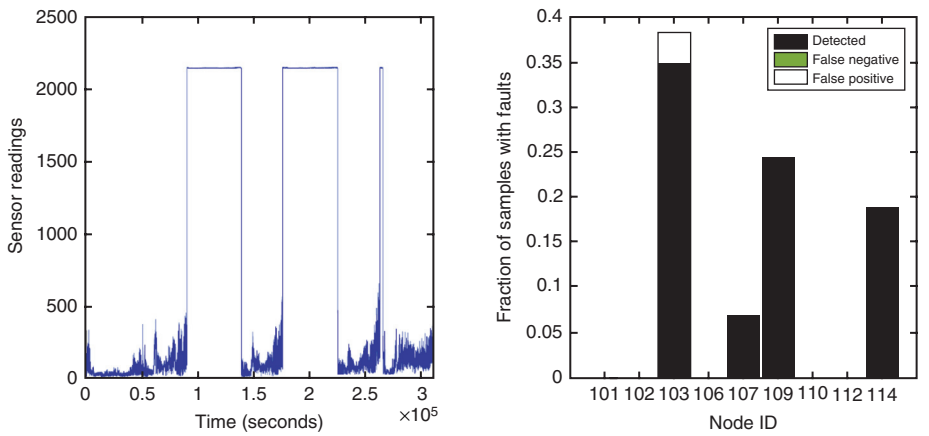

Figure 4. NAMOS.

NOISE Rule to detect samples with errors. Figure 4(b) shows the fraction of samples corrupted by faults. The sensors at 4 buoys was affected by the ADC board fault and in the worst case, at buoy 103, 35% of the reported values were erroneous. We could not apply LLSE and HMM method because there was not enough data to train the models (data was collected for 24 hours only). For results on other data sets, please refer to [18].

Our study informs the research on ensuring data quality. It is evident from Figure 4(b) that sensor faults can affect a significant fraction of samples, and careful attention needs to be paid to engineering the deployment and to analyzing the data. Furthermore, our detection methods could be used as part of an online fault diagnosis system, i.e., where corrective steps could be taken during the data collection process based on the diagnostic system's results.

6. VISION

AMBROSia will aid scientific research by facilitating the testing of scientific hypothesis. It will provide timely predictions of sampling needs. For instance, it might predict that there is a need to increase (in time and/or space) chlorophyll measurements in a particular region in preparation for a possible algae bloom. This prediction might be made based on newly received temperature measurements and wind model predictions. It will also provide tracking information for dynamic phenomena. For instance, it might detect red tide movement and predict better sampling regions for mobile nodes. Overall, our vision for AMBROSia is that it will facilitate observation, detection, and tracking of scientific phenomena that were previous only partially (or not at all) observable and/or understood.

7. ACKNOWLEDGMENTS

This research has been funded by the NSF DDDAS 0540420 grant. It has also been funded in part by the NSF Center for Embedded Networked Sensing Cooperative Agreement CCR0120778, the National Oceanic and Atmospheric Administration Grant NA05NOS478, and the NSF EIA–0121 141 grant.

REFERENCES

1. NAMOS: Networked Aquatic Microbial Observing System. *http://robotics.usc.edu/~namos.*
2. D.M. Anderson. Turning back the harmful red tide. *Nature,* 388: 513–514, 1997.
3. D.M. Anderson and D.J. Garrison. The ecology and oceanography of harmful algal blooms. *Limnol. Oceanogr.*, 42(5:2): 1009–1305, 1997.

4. K. Anstreicher, M. Fampa, J. Lee, and J. Williams. Maximum-entropy remote sampling. *Discrete Applied Mathematics*, 108(3): 21 1–226, 2001.
5. Pam Belluck. Red tide shuts shellfish areas in new england. *New York Times, http://www.nytimes.com/2005/06/04/national/04tide.html*, June 4, 2005.
6. A. Das and D. Kempe. Algorithms for Subset Selection in Linear Regression, In *the Proceedings of the Symposium on Theory of Computing (STOC)*, 2008.
7. G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *Journal of Constructive Approximation*, 13: 57–98, 1997.
8. Amit Dhariwal, Bin Zhang, Carl Oberg, Beth Stauffer, Aristides Requicha, David Caron, and Gaurav S. Sukhatme. Networked aquatic microbial observing system. In *the Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, May 2006.
9. G. Diekhoff. *Statistics for the Social and Behavioral Sciences.* Wm. C. Brown Publishers, 2002.
10. Jianqing Fan. Local linear regression smoothers and their minimax efficiences. *The Annals of Statistics*, 21(1): 196–216, 1993.
11. A. Gilbert, S. Muthukrishnan, and M. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. ACM-SIAM Symposiun on Discrete Algorithms*, 2003.
12. R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, 2002.
13. D. J. McGilliguddy, R. P. Signell, C. A. Stock, B. A. Keafer, M.D. Keller, R.D. Hetland, and D.M. Anderson. A mechanism for offshore initiation of harmful algal blooms in the coast Gulf of Maine. *Journal of Plankton Research*, 25(9): 1131–1138, 2003.
14. A. Miller. *Subset Selection in Regression.* Chapman and Hall, second edition, 2002.
15. B. Natarajan. Sparse approximation solutions to linear systems. *SIAM Journal on Computing*, 24: 227–234, 1995.
16. N. Ramanathan, L. Balzano, M. Burt, D. Estrin, E. Kohler, T. Harmon, C. Harvey, J. Jay, S. Rothenberg, and M. Srivastava. Rapid Deployment with Confidence: Calibration and Fault Detection in Environmental Sensor Networks. Technical Report 62, CENS, April 2006.
17. D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3): 1346–1370, 1994.
18. A. Sharma, L. Golubchik, and R. Govindan. On the Prevalence of Sensor Faults in Real World Deployments. Technical Report 07–888, Computer Science, University of Southern California, 2007.

19. Gaurav S. Sukhatme, Amit Dahriwal, Bin Zhang, Carl Oberg, Beth Stauffer, and David Caron. The design and development of a wireless robotic networked aquatic microbial observing system. *Environmental Engineering Science*, 2007.
20. Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, and Wei Hong. A Macro-scope in the Redwoods. In *SenSys '05: Proceedings of the 2nd international conference on Embedded networked sensor systems,* pages 51–63, New York, NY, USA, 2005. ACM Press.
21. J. Tropp. *Topics in Sparse Approximation.* PhD thesis, University of Texas, Austin, 2004.