

# A Generic Multi-scale Modeling Framework for Reactive Observing Systems: An Overview\*

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  
`leana@cs.usc.edu`

**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 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 our project in the context of a marine biology application.

## 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 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

---

\* 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 CCR-0120778, the National Oceanic and Atmospheric Administration Grant NA05NOS47812228, and the NSF EIA-0121141 grant.

stationary and mobile sensors, 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 microorganisms to predict the onset of algal blooms. Such blooms can have devastating economic consequences, as recently seen in [4].

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<sup>1</sup> or adaptation. 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.

The system we propose to develop is targeted at the marine application outlined above, and described in more detail below. 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

---

<sup>1</sup> 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.

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 [1, 2]. 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, [6] 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 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 802.11b 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.

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 more

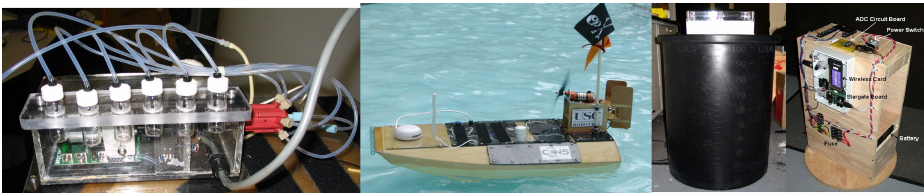


Fig. 1. Experimental Setup

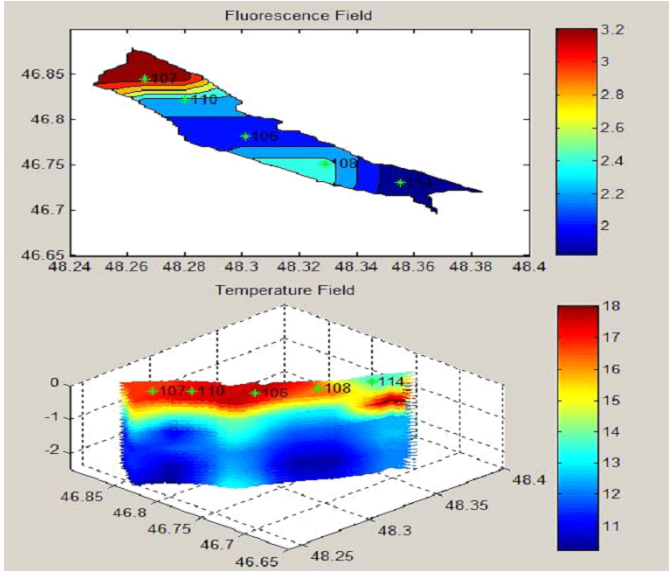


Fig. 2. Empirical Data from Lake Fulmor, James Reserve

commonplace and affordable, the scale of the system will grow significantly, and may eventually monitor large coastal expanses with 100s or 1000s of nodes and 10s 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

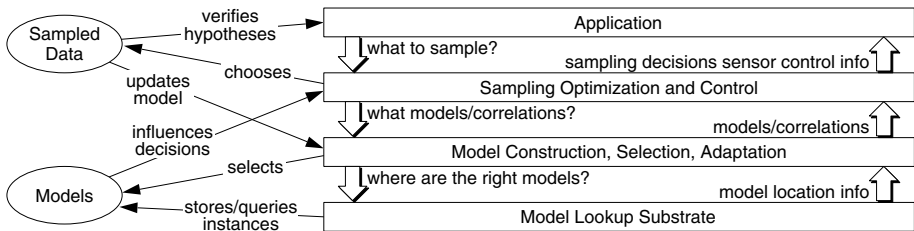


Fig. 3. System Overview

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 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”. A simple 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<sup>2</sup>. Thus, 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”  $k$  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 [7]: Given a set of  $n$  random variables  $X_1, X_2, \dots, X_n$  (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 [3, 7]. Other variants of this problem have been independently proposed in the mathematics community, such as the sparse approximation problem [8]; theoretical results [5, 8] here have been limited to the special case of nearly orthogonal dictionaries where greedy and convex relaxation methods have been shown to provide  $(1 + \epsilon)$  approximation bounds to the optimal solution.

---

<sup>2</sup> 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.

**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 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.

## References

- [1] D.M. Anderson. Turning back the harmful red tide. *Nature*, 388:513–514, 1997.
- [2] D.M. Anderson and D.J. Garrison. The ecology and oceanography of harmful algal blooms. *Limnol. Oceanogr.*, 42(5:2):1009–1305, 1997.
- [3] Kurt M. Anstreicher, Marcia Fampa, Jon Lee, and Joy Williams. Maximum-entropy remote sampling. *Discrete Applied Mathematics*, 108:211–226, 2001.
- [4] 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.
- [5] A. Gilbert, S. Muthukrishnan, and M. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms*, 2003.
- [6] 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.
- [7] Alan Miller. *Subset Selection in Regression*. Second edition, 2002.
- [8] Joel Tropp. *Topics in Sparse Approximation*. PhD thesis, The University of Texas at Austin, 2004.