Dynamically Identifying and Tracking Contaminants in Water Bodies

Craig C. Douglas[1,2], Martin J. Cole[3], Paul Dostert[4], Yalchin Efendiev[4], Richard E. Ewing[4], Gundolf Haase[5], Jay Hatcher[1], Mohamed Iskandarani[6], Chris R. Johnson[3], and Robert A. Lodder[7]

[1] University of Kentucky, Department of Computer Science, 773 Anderson Hall, Lexington, KY 40506-0046, USA
hatch22@fastmail.us
[2] Yale University, Department of Computer Science, P.O. Box 208285 New Haven, CT 06520-8285, USA
douglas-craig@cs.yale.edu
[3] University of Utah, Scientific Computing and Imaging Institute, Salt Lake City, UT 84112, USA
{crj,mjc}@cs.utah.edu
[4] Texas A&M University, Institute for Scientific Computation, 612 Blocker, 3404 TAMU, College Station, TX 77843-3404, USA
richard_ewing@tamu.edu, {dostert,efendiev}@math.tamu.edu
[5] Karl-Franzens University of Graz, Mathematics and Computational Sciences, A-8010 Graz, Austria
gundolf.haase@uni-graz.at
[6] University of Miami, Rosenstiel School of Marine and Atmospheric Science, 4600 Rickenbacker Causeway, Miami, FL 33149-1098, USA
mohamed.iskandarani@rsmas.miami.edu
[7] University of Kentucky, Department of Chemistry, Lexington, KY, 40506-0055, USA
lodder@contactincontext.org

Abstract. We present an overview of an ongoing project to build a DDDAS for identifying and tracking chemicals in water. The project involves a new class of intelligent sensor, building a library to optically identify molecules, communication techniques for moving objects, and a problem solving environment. We are developing an innovative environment so that we can create a symbiotic relationship between computational models for contaminant identification and tracking in water bodies and a new instrument, the Solid-State Spectral Imager (SSSI), to gather hydrological and geological data and to perform chemical analyses. The SSSI is both small and light and can scan ranges of up to about 10 meters. It can easily be used with remote sensing applications.

1 Introduction

In this paper, we describe an intelligent sensor and how we are using it to create a dynamic data-driven application system (DDDAS) to identify and track contaminants in water bodies. This DDDAS has applications to tracking polluters,

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 1002–1009, 2007.

finding sunken vehicles, and ensuring that drinking water supplies are safe. This paper is a sequel to [1].

In Sec. 2, we discuss the SSSI. In Sec. 3, we discuss the problem solving environment that we have created to handle data to and from SSSI's in the field. In Sec. 4, we discuss In Sec. 5, we state some conclusions.

2 The SSSI

Using a laser-diode array, photodetectors, and on board processing, the SSSI combines innovative spectroscopic integrated sensing and processing with a hyperspace data analysis algorithm [2]. The array performs like a small network of individual sensors. Each laser-diode is individually controlled by a programmable on board computational device that is an integral part of the SSSI and the DDDAS.

Ultraviolet, visible, and near-infrared laser diodes illuminate target points using a precomputed sequence, and a photodetector records the amount of reflected light. For each point illuminated, the resulting reflectance data is processed to separate the contribution of each wavelength of light and classify the substances present. An optional radioactivity monitor can enhance the SSSI's identification abilities.

The full scale SSSI implementation will have 25 lasers in discrete wavelengths between 300 nm and 2400 nm with 5 rows of each wavelength, consume less than 4 Watts, and weigh less than 600 grams. For water monitoring in the open ocean, imaging capability is unnecessary. A single row of diodes with one diode at each frequency is adequate. Hence, power consumption of the optical system can be reduced to approximately one watt.

Several prototype implementations of SSSI have been developed and are being tested at the University of Kentucky. These use an array of LEDs instead of lasers.

The SSSI combines near-infrared, visible, and ultraviolet spectroscopy with a statistical classification algorithm to detect and identify contaminants in water. Nearly all organic compounds have a near-IR spectrum that can be measured. Near-infrared spectra consist of overtones and combinations of fundamental mid-infrared bands, which makes near-infrared spectra a powerful tool for identifying organic compounds while still permitting some penetration of light into samples [3].

The SSSI uses one of two techniques for encoding sequences of light pulses in order to increase the signal to noise ratio: Walsh-Hadamard or Complementary Randomized Integrated Sensing and Processing (CRISP).

In a Walsh-Hadamard sequence multiple laser diodes illuminate the target at the same time, increasing the number of photons received at the photo detector. The Walsh-Hadamard sequence can be demultiplexed to individual wavelength responses with a matrix-vector multiply [4]. Two benefits of generating encoding sequences by this method include equivalent numbers of on and off states for each sequence and a constant number of diodes in the on state at each resolution point of a data acquisition period.

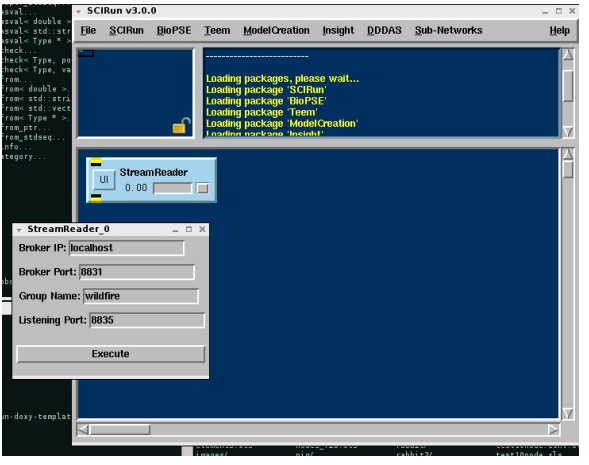

Fig. 1. SCIRun screen with telemetry module in forefront

CRISP encoding uses orthogonal pseudorandom codes with unequal numbers of on and off states. The duty cycle of each code is different, and the codes are selected to deliver the highest duty cycles at the wavelengths where the most light is needed and lowest duty cycle where the least light is needed to make the sum of all of the transmitted (or reflected) light from the samples proportional to the analyte concentration of interest.

3 Problem Solving Environment SCIRun

SCIRun version 3.0 [5,6], a scientific problem solving enironment, was released in late 2006. It includes a telemetry module based on [7], which provides a robust and secure set of Java tools for data transmission that assumes that a known broker exists to coordinate sensor data collection and use by applications. Each tool has a command line driven form plus a graphical front end that makes it so easy that even the authors can use the tools.

In addition there is a Grid based tool that can be used to play back already collected data. We used Apple's XGrid environment [8] (any Grid environment will work, however) since if someone sits down and uses one of the computers in the Grid, the sensors handled by that computer disappear from the network until the computer is idle again for a small period of time. This gives us the opportunity to develop fault tolerant methods for unreliable sensor networks.

The clients (sensors or applications) can come and go on the Internet, change IP addresses, collect historical data, or just new data (letting the missed data fall on the floor). The tools were designed with disaster management [9] in mind and stresses ease of use when the user is under duress and must get things right immediately.

A new Socket class was added to SCIRun, which encapsulates the socket traffic and is used to connect and transfer data from the server. The client handshakes with the server, which is informed of an ip:port where the client can be reached, and then listens on that port. Periodically, or as the server has new data available, the server sends data to the listening client.

The configuration for SCIRun was augmented to include libgeotiff [10]. SCIRun then links against this client and has its API available within the modules. This API can be used to extract the extra information embedded in the tiff tags in various supported formats. For example, position and scale information can be extracted so that the images can be placed correctly.

To allow controller interfaces to be built for the SSSI, a simulation of the device has been written in Matlab. This simulation follows the structure of the firmware code and provides the same basic interface as the firmware device. Data files are used in place of the SSSI's serial communication channel to simulate data exchange in software. Matlab programs are also provided to generate sample data files to aid in the development of Hadamard-Walsh and CRISP encodings for various SSSI configurations. The simulation also provides insight into the SSSI's firmware by emulating the use of oversampling and averaging to increase data precision and demonstrating how the data is internally collected and processed. The simulation can be used for the development of interfaces to the SSSI while optimization and refinement of the SSSI firmware continues.

SCIRun has a Matlab module so that we can pipe data to and from the SSSI emulator. As a result, we can tie together the data transfer and SSSI components easily into a system for training new users and to develop virtual sensor networks before deployment of a real sensor network in the field.

4 Accurate Predictions

The initial deployment of the sensor network and model will focus on estuarine regions where water quality monitoring is critical for human health and environmental monitoring. The authors will capitalize on an existing configuration of the model to the Hudson-Raritan Estuary to illustrate the model's capabilities (see [1] for details). We will consider passive tracer driven by external sources:

$$\frac{\partial C(x,t)}{\partial t} - L(C(x,t)) = S(x,t),\ C(x,0) = C^0(x)\ x \in \Omega,$$

where C is the concentration of contaminant, S is a source term and L is linear operator for passive scalar (advection-diffusion-reaction). L involves the velocity field which is obtained via the forward model based on the two-dimensional Spectral Element Ocean Model (SEOM-2D). This model solves the shallow water equations and the details can be found in our previous paper [1]. We have developed the spectral element discretization which relies on relatively high degree (5-8th) polynomials to approximate the solution within flow equations. The main features of the spectral element method are: geometric flexibility due to its unstructured grids, its dual paths to convergence: exponential by increasing polynomial degree or algebraic via increasing the number of elements, dense

computational kernels with sparse inter-element synchronization, and excellent scalability on parallel machines.

We now present our methodology for obtaining improved predictions based on sensor data. For simplicity, our example is restricted to synthetic velocity fields. Sensor data is used to improve the predictions by updating the solution at previous time steps which is used for forecasting. This procedure consists of updating the solution and source term history conditioned to observations and reduces the computational errors associated with incorrect initial/boundary data, source terms, etc., and improves the predictions [11,12,13]. We assume that the source term can be decomposed into pulses at different time steps (recording times) and various locations. We represent time pulses by $\delta_k(x,t)$ which corresponds to contaminant source at the location $x = x_k$.

We seek the initial condition as a linear combination of some basis functions $C^0(x) \approx \tilde{C}^0(x) = \sum_{i=1}^{N_D} \lambda_i \varphi_i^0(x)$. We solve for each i,

$$\frac{\partial \varphi_i}{\partial t} - L(\varphi_i) = 0,\ \varphi_i(x,0) = \varphi_i^0(x).$$

Thus, an approximation to the solution of $\frac{\partial C}{\partial t} - L(C) = 0,\ C(x,0) = C^0(x)$ is given by $\tilde{C}(x,t) = \sum_{i=1}^{N_D} \lambda_i \varphi_i(x,t)$. To seek the source terms, we consider the following basis problems

$$\frac{\partial \psi_k}{\partial t} - L(\psi_k) = \delta_k(x,t),\ \psi_k(x,0) = 0$$

for ψ and each k. Here, $\delta_k(x,t)$ represents unit source terms that can be used to approximate the actual source term. In general, $\delta_k(x,t)$ have larger support both in space and time in order to achieve accurate predictions. We denote the solution to this equation as $\{\psi_k(x,t)\}_{k=1}^{N_c}$ for each k. Then the solution to our original problem with both the source term and initial condition is given by

$$\tilde{C}(x,t) = \sum_{i=1}^{N_D} \lambda_i \varphi_i(x,t) + \sum_{k=1}^{N_c} \alpha_k \psi_k(x,t).$$

Thus, our goal is to minimize

$$F(\alpha,\lambda) = \sum_{j=1}^{N_s} \left[\left(\sum_{k=1}^{N_c} \alpha_k \psi_k(x_j,t) + \sum_{k=1}^{N_D} \lambda_k \varphi_k(x_j,t) - \gamma_j(t) \right)^2 \right] + \sum_{k=1}^{N_c} \tilde{\kappa}_k \left(\alpha_k - \tilde{\beta}_k \right)^2 + \sum_{k=1}^{N_D} \hat{\kappa}_k \left(\lambda_k - \hat{\beta}_k \right)^2, \tag{1}$$

where N_s denotes the number of sensors. If we denote $N = N_c + N_d$, $\mu = [\alpha_1, \cdots, \alpha_{N_c}, \lambda_1, \cdots, \lambda_{N_D}]$, $\eta(x,t) = [\psi_1, \cdots, \psi_{N_c}, \varphi_1, \cdots, \varphi_{N_D}]$,

$\beta = \left[\tilde{\beta}_1, \cdots, \tilde{\beta}_{N_c}, \hat{\beta}_1, \cdots, \hat{\beta}_{N_D}\right]$, and $\kappa = [\tilde{\kappa}_1, \cdots, \tilde{\kappa}_{N_c}, \hat{\kappa}_1, \cdots, \hat{\kappa}_{N_D}]$ then we want to minimize

$$F(\mu) = \sum_{j=1}^{N_s} \left[\left(\sum_{k=1}^{N} \mu_k \eta_k (x_j, t) - \gamma_j (t) \right)^2 \right] + \sum_{k=1}^{N} \kappa_k (\mu_k - \beta_k)^2 .$$

This leads to solving the least squares problem $A\mu = R$ where

$$A_{mn} = \sum_{j=1}^{N} \eta_m (x_j, t) \eta_n (x_j, t) + \delta_{mn} \kappa_m,$$

and

$$R_m = \sum_{j=1}^{N} \eta_m (x_j, t) \gamma_j (t) + \kappa_m \beta_m.$$

We can only record sensor values at some discrete time steps $t = \{t_j\}_{j=1}^{N_t}$. We want to use the sensor values at $t = t_1$ to establish an estimate for μ, then use each successive set of sensor values to refine this estimate. After each step, we update and then solve using the next sensor value.

Next, we present a representative numerical result. We consider contaminant transport on a flat surface, a unit dimensionless square, with convective velocity in the direction $(1, 1)$. The source term is taken to be 0.25 in $[0.1, 0.3] \times [0.1, 0.3]$ for the time interval from $t = 0$ to $t = 0.05$. Initial condition is assumed to have the support over the entire domain. We derive the initial condition (solution at previous time step) by solving the original contaminant transport problem with some source terms assuming some prior contaminant history.

To get our observation data for simulations, we run the forward problem and sample sensor data at every 0.05 seconds for 1.0 seconds. We sample at the following five locations: $(0.5, 0.5)$, $(0.25, 0.25)$, $(0.25, 0.75)$, $(0.75, 0.25)$, and $(0.75, 0.75)$.

When reconstructing, we assume that there is a subdomain $\Omega_c \subset \Omega$ where our initial condition and source terms are contained. We assume that the source term and initial condition can be represented as a linear combinations of basis functions defined on Ω_c. For this particular model, we assume the subdomain is $[0, 0.4] \times [0, 0.4]$ and we have piecewise constant basis functions. Furthermore, we assume that the source term in our reconstruction is nonzero for the same time interval as $S(x, t)$. Thus we assume the source basis functions are nonzero for only $t \in [0, 0.05]$.

To reconstruct, we run the forward simulation for a 4×4 grid of piecewise constant basis functions on $[0, 0.4] \times [0, 0.4]$ for both the initial condition and the source term. We then reconstruct the coefficients for the initial condition and source term using the approach proposed earlier. The following plot shows a comparison between the original surface (in green) and the reconstructed surface (in red). The plots are for $t = 0.1$, 0.2, 0.4 and 0.6. We observe that the recovery at initial times is not very accurate. This is due to the fact that we have not

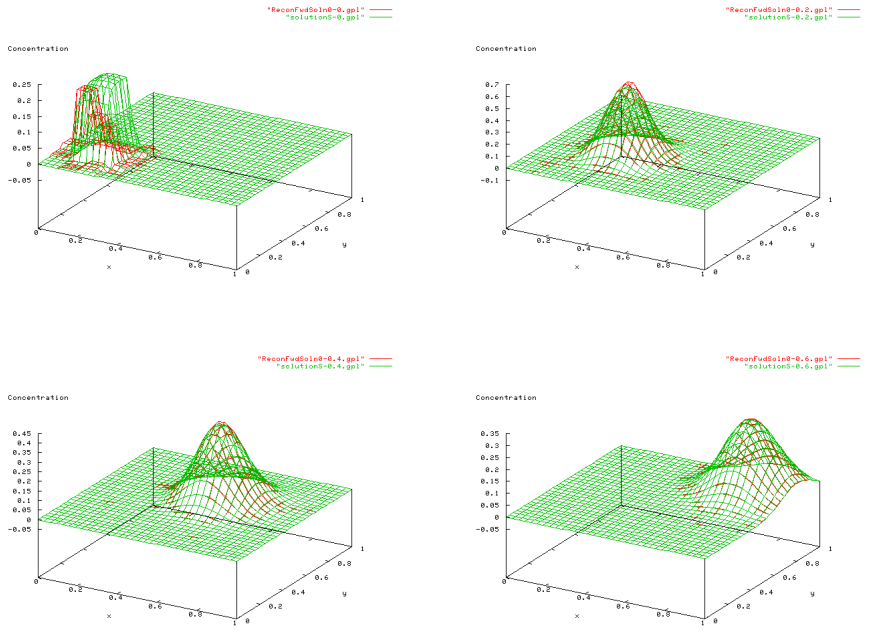

Fig. 2. Comparison between reconstructed (red) solution and exact solution at $t = 0.1$ (upper left), $t = 0.2$ (upper right), $t = 0.4$ (lower left), and $t = 0.6$ (lower right)

collected sufficient sensor data. As the time progresses, the prediction results improve. We observe that at $t = 0.6$, we have nearly exact prediction of the contaminant transport.

To account for the uncertainties associated with sensor measurements, we consider an update of initial condition and source terms, within a Bayesian framework. The posterior distribution is set up based on measurement errors and prior information. This posterior distribution is complicated and involves the solutions of partial differential equations. We developed an approach that combines least squares with a Bayesian approach, such as Metropolis-Hasting Markov chain Monte Carlo (MCMC) [14], that gives a high acceptance rate. In particular, we can prove that rigorous sampling can be achieved by sampling the sensor data from the known distribution, thus obtaining various realizations of the initial data. Our approach has similarities with the Ensemble Kalman Filter approach, which can also be adapted in our problem. We have performed numerical studies and these results will be reported elsewhere.

5 Conclusions

In the last year, we have made strides in creating our DDDAS. We have developed software that makes sending data from locations that go on and off the Internet and possibly change IP addresses rather easy to work with. This is a stand alone package that runs on any devices that support Java. It has also

been integrated into newly released version 3.0 of SCIRun and is in use by other groups, including surgeons while operating on patients. We have also developed software that simulates the behavior of the SSSI and are porting the relevant parts so that it can be loaded into the SSSI to get real sensor data. We have developed algorithms that allow us to achieve accurate predictions in the presence of errors/uncertainties in dynamic source terms as well as other external conditions. We have tested our methodology in both deterministic and stochastic environments and have presented some simplistic examples in this paper.

References

1. Douglas, C.C., Harris, J.C., Iskandarani, M., Johnson, C.R., Lodder, R.A., Parker, S.G., Cole, M.J., Ewing, R.E., Efendiev, Y., Lazarov, R., Qin, G.: Dynamic contaminant identification in water. In: Computational Science - ICCS 2006: 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings, Part III, Heidelberg, Springer-Verlag (2006) 393–400
2. Lowell, A., Ho, K.S., Lodder, R.A.: Hyperspectral imaging of endolithic biofilms using a robotic probe. Contact in Context **1** (2002) 1–10
3. Dempsey, R.J., Davis, D.G., R. G. Buice, J., Lodder, R.A.: Biological and medical applications of near-infrared spectrometry. Appl. Spectrosc. **50** (1996) 18A–34A
4. Silva, H.E.B.D., Pasquini, C.: Dual-beam near-infrared Hadamard. Spectrophotometer Appl. Spectrosc. **55** (2001) 715–721
5. Johnson, C.R., Parker, S., Weinstein, D., Heffernan, S.: Component-based problem solving environments for large-scale scientific computing. Concur. Comput.: Practice and Experience **14** (2002) 1337–1349
6. SCIRun: A Scientific Computing Problem Solving Environment, Scientific Computing and Imaging Institute (SCI). http://software.sci.utah.edu/scirun.html (2007)
7. Li, W.: A dynamic data-driven application system (dddas) tool for dynamic reconfigurable point-to-point data communication. Master's thesis, University of Kentucky Computer Science Department, Lexington, KY
8. Apple OS X 10.4 XGrid Features, Apple, inc. http://www.apple.com/acg/xgrid (2007)
9. Douglas, C.C., Beezley, J.D., Coen, J., Li, D., Li, W., Mandel, A.K., Mandel, J., Qin, G., Vodacek, A.: Demonstrating the validity of a wildfire DDDAS. In: Computational Science - ICCS 2006: 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings, Part III, Heidelberg, Springer-Verlag (2006) 522–529
10. GeoTiff. http://www.remotesensing.org/geotiff/geotiff.html (2007)
11. Douglas, C.C., Efendiev, Y., Ewing, R.E., Ginting, V., Lazarov, R.: Dynamic data driven simulations in stochastic environments. Computing **77** (2006) 321–333
12. Douglas, C.C., Efendiev, Y., Ewing, R.E., Ginting, V., Lazarov, R., Cole, M.J., Jones, G.: Least squares approach for initial data recovery in dynamic data-driven applications simulations. Comp. Vis. in Science (2007) in press.
13. Douglas, C.C., Efendiev, Y., Ewing, R.E., Ginting, V., Lazarov, R., Cole, M.J., Jones, G.: Dynamic data-driven application simulations. interpolation and update. In: Environmental Security Air, Water and Soil Quality Modelling for Risk and Impact Assessment. NATO Securtity through Science Series C, New York, Springer-Verlag (2006)
14. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer-Verlag, New York (1999)