Dynamic Contaminant Identification in Water

Craig C. Douglas[1,2], J. Clay Harris[3], Mohamed Iskandarani[4], Chris R. Johnson[5], Robert Lodder[3], Steve Parker[5], Martin J. Cole[5], Richard Ewing[6], Yalchin Efendiev[6], Raytcho Lazarov[6], and Guan Qin[6]

[1] University of Kentucky, Department of Computer Science, 773 Anderson Hall, Lexington, KY 40506-0046, USA.
[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 Kentucky, Department of Chemistry, Lexington, KY, 40506 USA.
claymay27@gmail.com, lodder@contactincontext.org.
[4] University of Miami, Rosenstiel School of Marine and Atmospheric Science, 4600 Rickenbacker Causeway, Miami, FL 33149-1098, USA.
mohamed.iskandarani@rsmas.miami.edu.
[5] University of Utah, Scientific Computing and Imaging Institute, Salt Lake City, UT 84112, USA. {crj,sparker,mjc}@cs.utah.edu.
[6] Texas A&M University, Institute for Scientific Computation, 612 Blocker, 3404 TAMU, College Station, TX 77843-3404, USA. {richard_ewing,guan.qin}@tamu.edu and{efendiev,lazarov}@math.tamu.edu

Abstract. We describe how we plan to convert a traditional data collection sensor and ocean model into a DDDAS enabled system for identifying contaminants and then reacting with different models, simulations, and sensing strategies in a symbiotic manner. The sensor is just as useful in water as it would be on Mars for material identification. A successful terrestrial application of the sensor will lead to many new applications of the device and possible technology transfer to the private sector.

1 Introduction

The Solid-State Spectral Imager (SSSI) is a new instrument to gather hydrological and geological data and to perform chemical analyses. It is suitably small and light to mount in remote sensing applications, and can scan ranges of up to 10 meters. 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 [1].

Ultraviolet (UV), 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.

Several prototype implementations of SSSI have been developed and are being tested at the University of Kentucky. A full-scale implementation of SSSI is being designed with 25 lasers in discrete wavelengths between 300 nm and 2400

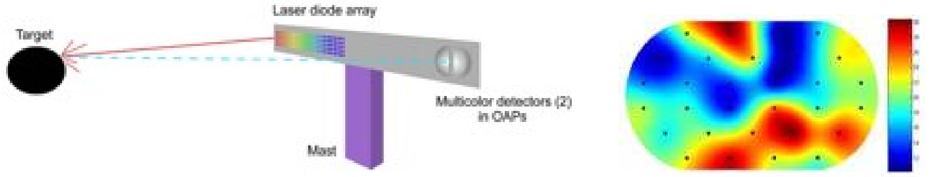

Fig. 1. SSSI emiter and collected spectrum.

nm with 5 rows of each wavelength. This full-scale version is designed to consume less than 4 Watts and weigh less than 600 grams. The rugged laser diodes and detectors allow SSSI to be packaged in a small, rugged, space qualified package. For water monitoring in the open ocean, imaging capability is not needed, and a single row of diodes (with one diode at each frequency) is sufficient, and power consumption of the optical system can be reduce to approximately one watt.

The SSSI combines near-infrared, visible, and ultraviolet spectroscopy with a powerful statistical classification algorithm to detect and identify contaminants in water. Virtually every organic compound (e.g., polycyclic aromatic hydrocarbons, paraffins, carboxylic acids, and sulfonic acids) has a near-IR spectrum that can be measured, including two classes of terrestrial biomarkers, lipids, and amino acids. Near-infrared spectra consist of overtones and combinations of fundamental mid-infrared bands, giving near-infrared spectra a powerful ability to identify organic compounds while still permitting some penetration of light into samples [2].

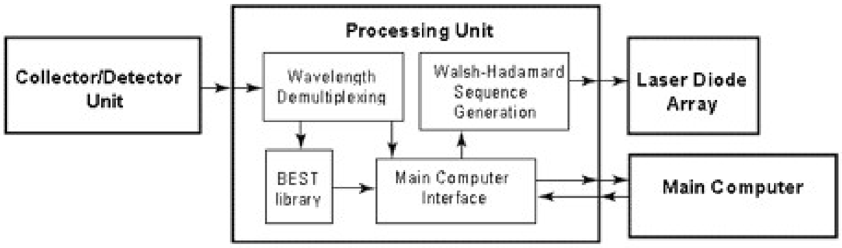

Fig. 2. SSSI processing

To further increase the signal-to-noise ratio, the SSSI uses Walsh-Hadamard or CRISP encoding sequences of light pulses. 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 [3]. 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.

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.

The initial deployment of the sensor 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. As shown in Fig. 3 the model domain includes the Lower and Middle Hudson River, the Hudson-Raritan Estuary, Newark Bay and Long Island sound. In the initial experimentation stage only a portion of the grid will be for fast prototyping of the different elements in the DDDAS system.

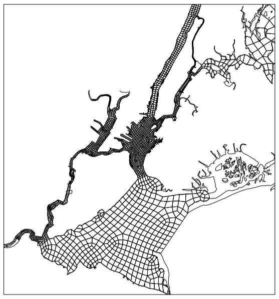

Fig. 3. Spectral element grid showing elemental partition of the New York/Newark Bay estuarine system.

The forward model is based on the two-dimensional Spectral Element Ocean Model (SEOM-2D) which solves the shallow water equations:

$$\boldsymbol{u}_t + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \boldsymbol{f} \times \boldsymbol{u} + g \nabla \eta = \frac{\boldsymbol{\tau}_w - \boldsymbol{\tau}_d}{\rho h} + \frac{\nabla \cdot (h \nu \nabla \boldsymbol{u})}{h} \tag{1}$$

$$h_t + \nabla \cdot (h \boldsymbol{u}) = Q \tag{2}$$

where $h = H + \eta$ is the layer thickness, and H and η are the resting depth and interface displacement, respectively, $\boldsymbol{u}$ is the depth-average velocity; g is the gravity coefficient, $\boldsymbol{f}$ is the Coriolis parameter, ν the viscosity, $\boldsymbol{\tau}_w$ the surface wind stress, $\boldsymbol{\tau}_b$ the bottom drag, and Q is an area mass source. An advection-diffusion equation tracks the evolution of passive tracers:

$$T_t + \boldsymbol{u} \cdot \nabla T = \frac{\nabla \cdot (\alpha h \nabla T)}{h} \tag{3}$$

where α is the diffusion coefficient, and T stands for a generic tracer. The model can be forced through winds, tides, and lateral injection of mass at inflow boundaries (e.g. river input).

The spectral element discretization is an *h-p* type finite element method which relies on relatively high degree (5-8th) polynomials to approximate the solution within each element. 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.

A sample tidal calculation is performed using a grid that encompasses the Newark/New York bays regions, the Long-Island sound, and a substantial portion of the Hudson River. The model is forced with tidal elevation obtained from tide gauges located on the eastern edge of the Long-Island sound and in Sandy Hook.

The SSSI is reprogramable in the field. When an interesting chemical trace is discovered, the reaction from the application overseeing the SSSI is two-fold: (a) invoke an appropriate application, and (b) request that the SSSI look for specific other chemical traces. There is a symbiotic relationship between the sensor network and the application simulation that is typical in a DDDAS.

Consider finding gasoline or diesel fuel in a body of water. This can be a sign of innocuos pollution from a boat. Depending on what other traces are found, it could be an indication that a boat sank recently nearby. The SSSI needs to be reprogrammed in the latter case and a search and locate application must be invoked to find the sunken boat. Emergency services, the coast guard, and the news media may also need to be automatically informed of progress.

The SSSI has a modest amount of memory and computing capacity on board. Some of the computing and decision making will be put onto the SSSI over time, thus reducing the amount of time needed to reprogram the device.

2 Data assimilation and accurate predictions

Data collection is initiated by a signal sent to the serial interface of the SSSI. All data is collected by the SSSI using a phototransistor connected to an operational amplifier circuit. The analog signal is converted with an on-board 0.5-5 V analog to digital converter at 12 bits. Each scan consists of 256 data points collected in both the on and off states of 25 Hadamard encoded light sequences. The result is 50 total states with 12800 data points collected for each scan. The corresponding values of the on and off states for each Hadamard coded light sequence are subtracted to remove ambient light from the data. After subtraction, the resulting 256 data points from each of the 25 Hadamard coded light sequences are then averaged to obtain 25 16 bit intensity values. The final 25 16 bit resulting values are exported to Matlab via the serial connection to a graphical user interface where data undergoes a reverse Hadamard transform to obtain intensity values for each of the 25 diodes.

A single scan with MatLab processing takes less than 300 ms. The switching speed of our transistors within the SSSI is significantly slow that this prototype requires a 5 μs delay before each datum reading for signal stabilization after the lights have switched states. Both times will be significantly speeded up if we move to a commercial quallity device.

We can use the data to improve our prediction of the contaminant transport by updating the initial conditions. Here, initial condition refers to the concentration distribution at some previous time step. This update reduces the computational errors associated with incorrect initial data and improves the predictions. We consider contaminant transport described by (1)-(3). Initial data is sought in a finite dimensional space. Using the first set of measurements, the approximation of the initial data is recovered. As new data are incorporated into the simulator, the initial data is updated using an objective function. We note that the formulated problem is ill-posed because there are fewer sensors than the finite dimensional space describing the initial data. Consequently, the objective function is set up based on both a measurement error as well as a penalization term that depends on the prior knowledge about the solution at previous time steps (or initial data). The prior information is refreshed using the updated initial data. The penalization constants depend on time of update and can be associated with the relative difference between simulated and measured values. In the simulations, both the prior and penalization constants change in time.

To account for the errors (uncertainties) associated with sensor measurements, we consider an initial data update 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 could use a Metropolis-Hasting Markov chain Monte Carlo (MCMC) method to generate samples from the posterior distributions. However, a sampling with MCMC is expensive since it requires iterative steps and the acceptance rate is typically low. We developed an approach that combines least squares with a Bayesian approach 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 to an initial data update. These issues will be discussed in detail elsewhere.

3 Chemical identification process

A programmable, networked, portable low-cost mil-spec sensor and network for DDDAS in extreme aqueous environments must be able to perform chemical analyses to be effective in terrorist attack and accident scenarios. Most oil sensing in the oceans is done by remote sensor systems [4].

A network of sensors immersed in the ocean water (either on fixed buoys or as roving sensors) eliminates many of the problems with remote sensing. Bad weather does not affect immersed sensors. A roving sensor can be programmed to investigate beaches, weeds or debris. The SSSI is laser fluorosensor when a filter is placed over the detector, so it can positively discriminate oil on most backgrounds. Light scattering measurements reveal droplet size, and spectral transmission and reflectance reveal droplet chemistry.

Once the spectrum of a sample has been collected, it must be classified to determine the substance present. The Bootstrap Error-adjusted Single-sample Technique (BEST) [5] is the analytical basis of SSSI, and the foundation for the chemical library. Spectra recorded at n wavelengths are represented as single points in a n-dimensional hyperspace. In this scheme, similar samples produce similar spectra that project as "probability orbitals" or "clusters" into similar regions of hyperspace. The BEST metric is a clustering technique for exploring these distributions of spectra in hyperspace.

A sample spectrum is compared to each substance in a biogeochemical and industrial library based on its direction and distance, measured in standard deviation (SD) units, from the known substances. BEST handles asymmetric standard deviations surrounding each substance nonparametrically, allowing very precise discrimination. A sample within 3 SD units of a substance is considered to be composed of the matching substance. Any substance more than 3 SD units away from any known substance is considered an unknown substance.

For a given library entry, the BEST algorithm can be suitably approximated using multiple linear regression (MLR) to substantially reduce computational requirements (see Fig. 4). In this implementation, BEST SD units are precalculated before the SSSI is deployed in a large number of directions from the population means, and MLR is used to fit the standard deviation contours as a function of direction. With a sufficient number of terms (in the example, 10th order with cross terms), the MLR version of the algorithm can predict BEST distances to within 5% of the true value.

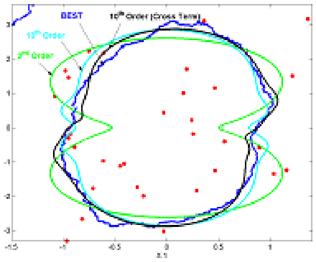

Fig. 4. Identification

Oil droplets can travel nearly anywhere in the ocean. The droplet size exerts a major effect on droplet motion [6]. The rise velocity of oil droplets extends from about 2.5×10^{-7} m/s for a diameter of 2 μm to 4.3×10^{-3} m/s for a diameter of 260 μm. Droplets traveling at 2.5×10^{-7} m/s will ascend only

0.001 m and 0.02 m, over periods of 1 hour and 24 hours, while over equivalent periods, droplets ascending at 4.3×10^{-3} m/s will climb 15 m and 370 m. In the meantime, a vertical diffusivity of 51 cm^2/s will distribute oil droplets (equally upward and downward) about 6 m and 30 m over the same time. Therefore, the smallest oil droplets act as though they are neutrally buoyant (transported only by diffusion), while the largest droplets are advected largely by their buoyancy.

Using multiple linear regression the BEST classification algorithm can be performed in situ, allowing a rover to classify many samples, only notifying the simulation when an interesting substance is found. An initial library can be computed based on substances likely to be found in the target environment. When a substance unknown to the BEST library is found, the sensor can sample nearby points with similar spectra to create a new library entry for the new substance. Scientists can determine the type of substance present by further analyzing raw spectra of the substance provided by SSSI and by using data from their other instruments, apply these data to update the simulation. The SSSI chemical library will comprise substances expected to be in the environment in which the SSSI operates.

4 Matlab and SCIRun environments

The SCIRun-Matlab interface is designed such that SCIRun [7] detects at runtime whether Matlab is available. Hence, SCIRun does not have to be linked against any of the Matlab libraries. The way the interfacing is accomplished is thorugh a virtual shell. SCIRun accesses Matlab through stdin and stdout and files that are written to a temporary directory. The whole process of translating, saving, and opening files is hidden from the user and is initiated automatically. Data translation is also seamless for things like 0 or 1 based indices.

The module is designed in such a way that once the Module is executed it will keep Matlab running in its internal engine, hence re-executing the module will allow to access the variables that were left by a previous execution cycle. Hence the Matlab Engine can be used as well for iterative processes. An example of the matlabinterface module is depicted in Fig. 5. The figure shows how matlab is integrated into the dataflow structure of SCIRun. Once the "Matlab" module receives all the dataflow objects that are connected to it, the specific code in Matlab is executed and now dataflow objects are created for the dataflow downstream.

Hence, moving the current Matlab interface for the SSSI sensor to a problem solving environment like SCIRun is trivial.

5 Conclusions

We described how we plan to convert a traditional data collection sensor and ocean model into a DDDAS enabled system for identifying contaminants and then reacting with different models, simulations, and sensing strategies in a symbiotic manner. A drone is being built so that the SSSI will be mobile. We

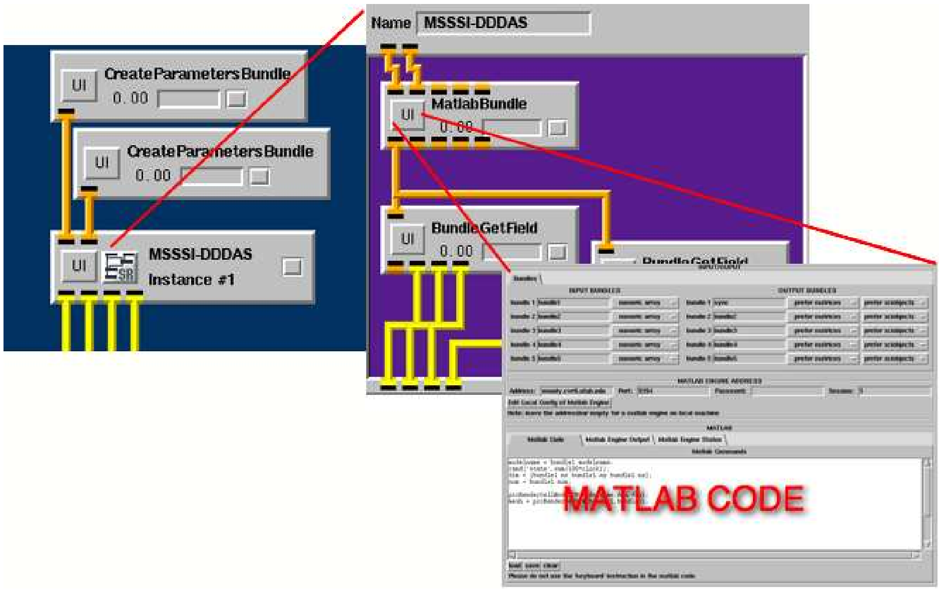

Fig. 5. SCIRun running the SSSI-cid module.

are already able to make measurements and are proceeding to program the system for remote sensing and steering. Libraries will be created for interesting contaminants during the coming year that we will use to reprogram the SSSI dynamically while we switch to an appropriate simulation for the contaminants identified to explore what else might be in the vicinity of the SSSI.

References

1. Lowell, A., Ho, K.S., Lodder, R.A.: Hyperspectral imaging of endolithic biofilms using a robotic probe. Contact in Context **1** (2002) 1–10
2. 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
3. Silva, H.E.B.D., Pasquini, C.: Dual-beam near-infrared Hadamard. Spectrophotometer Appl. Spectrosc. **55** (2001) 715–721
4. Fingas, M.F., Brown, C.E.: Review of oil spill remote sensing. In: Spillcon 2000, Darwin, Australia (2000)
5. Dieter, W., Lodder, R.A., James E. Lumpp, J.: Scanning for extinct astrobiological residues and current habitats (SEARCH) using integrated computational imaging. IEEE Aerospace and Electronic Systems (2006) (in press)
6. (OSB), O.S.B.: Oil Spill Dispersants: Efficacy and Effects (2005). The National Academies Press, Washington, DC (2005)
7. 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