# Near Optimal Sensor Selection in The COlumbia RIvEr (CORIE) Observation Network for Data Assimilation Using Genetic Algorithms

Thanh Dang[1], Sergey Frolov[2], Nirupama Bulusu[1], Wu-chi Feng[1], and António Baptista[2]

[1] Portland State University, Portland OR, USA,
dangtx, nbulusu, wuchi@cs.pdx.edu
[2] Oregon Health and Science University, Beaverton, OR, USA,
frolovs, baptista@ccalmr.ogi.edu

**Abstract.** CORIE is a pilot environmental observation and forecasting system (EOFS) for the Columbia River. The goal of CORIE is to characterize and predict complex circulation and mixing processes in a system encompassing the lower river, the estuary, and the near-ocean using a multi-scale data assimilation model.

The challenge for scientists is to maintain the accuracy of their modeling system while minimizing resource usage. In this paper, we first propose a metric for characterizing the error in the CORIE data assimilation model and study the impact of the number of sensors on the error reduction. Second, we propose a genetic algorithm to compute the optimal configuration of sensors that reduces the number of sensors to the minimum required while maintaining a similar level of error in the data assimilation model. We verify the results of our algorithm with 30 runs of the data assimilation model. Each run uses data collected and estimated over a two-day period. We can reduce the sensing resource usage by 26.5% while achieving comparable error in data assimilation. As a result, we can potentially save 40 thousand dollars in initial expenses and 10 thousand dollars in maintenance expense per year.

This algorithm can be used to guide operation of the existing observation network, as well as to guide deployment of future sensor stations. The novelty of our approach is that our problem formulation of network configuration is influenced by the data assimilation framework which is more meaningful to domain scientists, rather than using abstract sensing models.

**Key words:** Sensor selection, network configuration, coastal monitoring, data assimilation, genetic algorithm

## 1 Introduction

Earth and ocean sciences confront great opportunities and challenges in understanding the complex behaviors of large-scale physical systems with next genera-

tion sensing systems [6]. Modeling the behavior of the oceans and river estuaries is a challenging but important research field. In order to understand the state of the physical process, sensors are deployed in the environment to collect data for the modeling process. Ideally, a highly dense network of sensors will enable the collection of fine-grained information about the physical system under observation. However, for systems that operate over a large geographical region, such a deployment of sensors is infeasible. Hence, most of the existing large-scale sensing systems only deploy a sparse network of sensors and use advanced numerical methods in estimating and modeling the physical processes.

Figure 1 shows CORIE, an observation network that monitors the Columbia River estuary and the Eastern North Pacific ocean. CORIE integrates a real-time sensor network, a data management system and advanced numerical models. The goal of CORIE is to characterize and predict complex circulation and mixing processes in a system encompassing the lower river, the estuary and the near-ocean. The CORIE observation network includes an extensive array of 24 stations in the Columbia River estuary and the nearby coastal ocean. At each station, variable combinations of in-situ sensors measure one or more physical properties of water or atmosphere. Water temperature, salinity, and water levels are measured at most stations. Profiles of velocity and acoustic backscatter are measured at three stations.

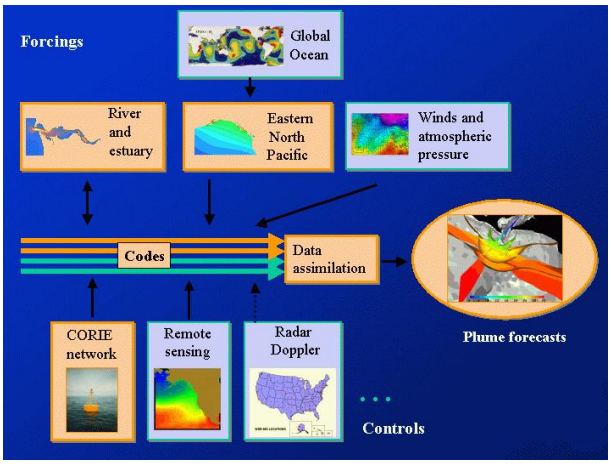


**Fig. 1.** CORIE Data Assimilation Architecture. Source: The CORIE project website

*Data assimilation* combines observational data with numerical data models to produce an estimated system state for the physical process. Data assimilation plays an important role in predicting the state of the dynamic physical process such as estuary circulation, weather and climate changes. Unlike low-powered wireless sensors such as the popular Crossbow motes which are tiny and cheap, the sensor stations in ocean monitoring are usually very expensive to deploy and

operate. Such stations typically have a number of sensors between the surface and the anchor, providing a vertical array of sensors. Measurement data from the observation network directly impacts the accuracy of the estimated system. Hence, finding a suitable network configuration is an important problem in deploying and operating an observation network because it can help reduce the resource usage while maintaining or improving the estimation accuracy.

In this paper, we first propose a metric for characterizing the error in the CORIE data assimilation model and study the impact of the number of sensors on the error reduction. Second, we propose a genetic algorithm to compute the near optimal configuration of sensors that reduces the number of sensors to the minimum required while maintaining a similar level of error in the data assimilation model.

This problem is relevant to the sensor network research community because the deployment of the existing observation network was based on an intuition of the underlying physical process, with little knowledge about how sensor placement would affect the resulting data models. Hence, a solution to the problem will help conserve resources by using fewer sensors.

The problems in oceanography have their own distinct challenges. First, the physical model is extremely complex. Unlike models in previous work [1, 8, 20, 11, 12], we address a complex 3D circulation and mixing processes in a system encompassing the lower river and the estuary. This has been formally recognized as a challenging task in ocean modeling [10]. Second, the computation is very expensive due to its large state space size. For example, the state space size of the CORIE model is 878,850. Given larger memory and computing platforms, this number would increase with new resources by increasing spatial and temporal resolution. Third, the observation model incorporates multiple sensing modalities such as salinity, temperature, elevation, and velocities. Each sensing modality provides different information about the observed environment. Therefore, the solution must take into account not only the correct set of sensors but also the correct type of sensors to ensure good estimation results. In addition, the solution must be model independent so that it can be used in other environmental monitoring deployments provided that they use the same data assimilation framework. Fortunately, the framework we use is a state-of-the-art data assimilation and estimation system [10]. Finally, selecting an optimal sequence of sets has already been shown to be NP-hard in many settings [1]. Therefore, we must consider not just only polynomial class solutions but also how much time it actually takes to converge to an acceptable result.

We present a method to partially address the problem of finding near optimal network configuration for an observation network, which uses a data assimilation framework based on a sigma-point Kalman filter [10]. The main contributions of this paper are:

- We formulate the problem of optimizing network configuration based upon data assimilation (Section 3) and apply it to an ocean modeling application.
- We propose a framework that uses genetic algorithms to partially address the problem of selecting a suitable subset of sensors (Section 4).

– We evaluate the approach on data from the CORIE observation network (Section 5) and demonstrate that we can reduce the use of sensing resources by 26.5% and operating expenses by $10,000 a year while maintaining a similar level of estimation accuracy.

## 2 Data Assimilation Overview

This section provides a brief overview of the CORIE data assimilation framework used in the formulation of the sensor selection problem. While we describe only the data assimilation framework used in the CORIE project [10], we do not imply that the problem is only applicable for this specific framework. In fact, the problem is suitable for any situation provided that the error of the estimation can be calculated.

### 2.1 CORIE Data Assimilation Framework

The complete data assimilation framework, proposed and implemented by Frolov *et al.* [10], is complex and draws upon several disciplines including numerical analysis, machine learning, and estimation theory. We provide a brief overview of CORIE here and refer readers to [10] for more details. Figure 1 shows the high level components of the CORIE modeling system. It integrates model and field controls. The main purpose of CORIE is to simulate 3D circulation in the region that lies between the Columbia River estuary and near ocean but also extends further inland in Oregon, USA to the Eastern North Pacific. CORIE performs multiple tasks and provides the following: short term forecasts, actual past conditions, characteristic climatology conditions, and scenario conditions.

In order to accomplish these tasks, one of the key components is data assimilation which integrates observational data from sensors into a non-linear ocean model. The model integrates information from the CORIE network, Doppler radar and remote sensing with forcings from the river, estuary, winds, atmosphere, and ocean to predict the behavior of the underlying physical processes. The work of Frolov *et al.* [10] proposes and implements a fast framework with model surrogates for data assimilation, illustrated in Figure. 2. The data assimilation framework includes two main components. The first component is *off-line learning* illustrated on the left block in Figure. 2. Its main purpose is to train a model surrogate, which is an equivalent model in the reduced space. In order to do that, the original system state of 878,850 variables is reduced to 60 variables using principle component analysis, a popular method for extracting patterns and compressing data [17], based on the singular value decomposition (SVD) algorithm. The model surrogate is trained using a recurrent neural network [15]. All training is carried out off-line using an existing database of model hindcasts generated by the traditional circulation model [18]. The second component is *data assimilation* illustrated on the right block in Figure. 2. The core of the assimilation algorithm is the *sigma-point kalman filter* [19]. The filter estimates the state of the dynamic system using the model surrogate and measurements from sensors. An existing framework such as data assimilation using ensemble Kalman filter [9] is computationally very expensive, limiting its use. In contrast,

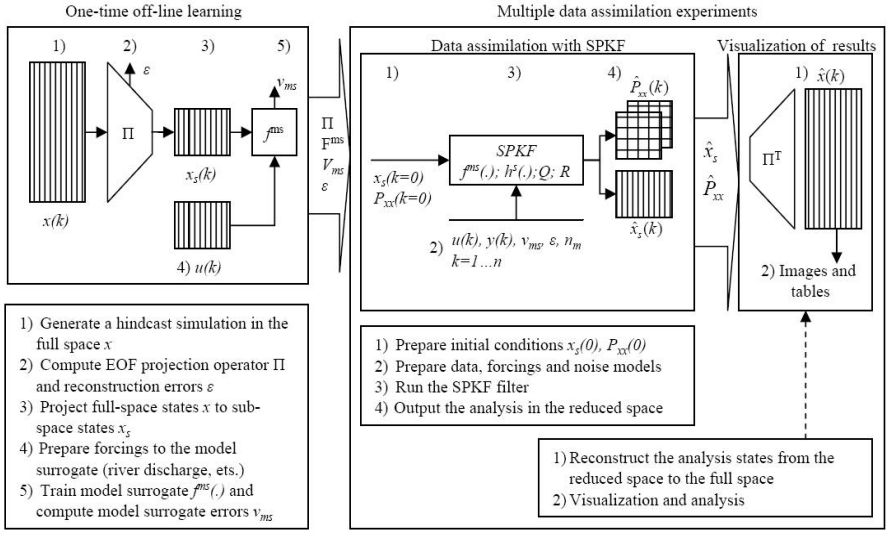


**Fig. 2.** CORIE Data assimilation framework — Reproduced with permission from Frolov *et al.* [10]

the model surrogate framework performs 1000 times faster than existing frameworks and can significantly increase the estimation error reduction (defined in Equation 2) on the given measurement set [10].

Clearly, this framework provides a significant contribution in improving the estimation of the ocean model. Nevertheless, a trivial observation is that the configuration of sensors including not only sensor location but also sensor type plays a very important role in providing better estimation. Hence, there are two problems:

- What is the configuration of sensors that would maximize the estimation accuracy given an observation network?
- What is a subset of sensors to achieve or maintain a certain estimation accuracy?

These have been proven to be hard problems [1]. In our work, we only address a part of the latter problem, which we describe formally in the next section.

## 3 Problem Formulation

Network configuration refers to the number of sensors, their type and their locations. The configuration of sensors plays an important role in estimation in general. For example, an object's location in two dimensional space can be better estimated from range measurements with a triplet of non-collinear sensors

than with a triplet of collinear sensors. The temperature in a room can be better characterized from sensors spread throughout the room rather than sensors concentrated in one specific area.

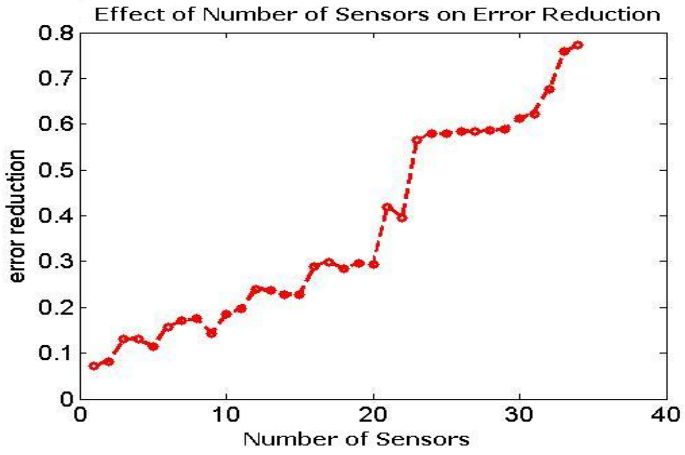


**Fig. 3.** Number of sensors versus estimation error reduction

For example, we have run data assimilations on an increasing number of sensors in CORIE network. Figure 3 shows the estimation error reduction (defined in Equation 2) versus the number of sensors used in CORIE network. As we can see, some sensors have more impact on error reduction than others [3]. In addition, some configurations of sensors might have lower error reduction even though they have more sensors than other configurations. Therefore, finding a minimum set of sensors that can provide the most information is an interesting problem, the *network configuration problem*. Formally, it can be stated as the following optimization problem:

$$min|S| \; subject \; to \; D(S) \leq \varepsilon \; and \; S \subseteq A \tag{1}$$

where $A$ is the set of all sensor information $A = \{s_1, s_2, ..., s_n\}$ and $s_i = (type, x, y, z, \delta)$ in which *type* is the sensor type, which can be temperature, salinity, or elevation. ($x$, $y$, $z$) is the sensor's location. $\delta$ is the standard deviation in the sensor reading obtained by calibration. $\varepsilon$ is the threshold error. $D(S)$ is the simplified form of the function of error reduction of the data assimilation. In other words, it is the cost function to be optimized. $D(S)$ can be calculated as the estimation error reduction as follows:

$$error\_reduction = 1 - \frac{sum[(xs\_twin - xs\_data).^2]}{sum[(xs\_twin - xs\_free).^2]} \tag{2}$$

[3] We compare the error reduction in data assimilation using the observational data relative to relying on the numerical model alone.

where $xs\_twin$ is the true system state, $xs\_data$ is the estimated system state, $xs\_free$ is the simulated system state, i.e., the estimated system state without considering sensor measurements. The notation $(.)^{.2}$ denotes the vector of squared elements. We can consider $sum[(xs\_twin - xs\_data)^{.2}]$ as the squared error of the data assimilation using sensor measurements and $sum[(xs\_twin - xs\_free)^{.2}]$ as the squared error of the data assimilation without sensor measurements. Hence, Equation 2 shows how much error the data assimilation can reduce when it uses additional sensor measurements.

A similar derived optimization problem can be formulated as follows:

$$max|D(S)| \; subject \; to \; S \subseteq A \; and \; |S| = n \; (n \leq |A|) \tag{3}$$

to find a configuration of the network that maximizes the error reduction $D(S)$ of the data assimilation.

There are several parameters to be considered here. The first parameter is a sensor's *type*. Intuitively, sensors of different types may provide a better data set for data assimilation than sensors of a single modality e.g. temperature sensors. The second parameter is the sensor *location*. It is important that sensors should be deployed in critical locations such that they together report data representing the underlying physical process. The final parameter is the *number* of sensors which is our optimization objective. Unfortunately, the complete problem is very difficult to solve due to the fact that selecting an optimal sequence of sets is NP-hard [1] and the behavior of function $D(S)$ is unknown. Therefore, our work can only address a part of the problem where the sensor locations are fixed. Hence, the problem becomes a *sensor selection* problem. The next section presents our approach to solve this problem.

## 4 Sensor Selection Using Genetic Algorithm

The key idea in our proposed solution is to apply genetic algorithms to search for an acceptable sensor set. We consider genetic algorithms (GAs) for this problem because they have been applied successfully to a variety of optimization problems, and especially for optimizing the topology and learning parameters for artificial neural networks [15]. GAs can search for the optimal solution by observing the behavior of the system without actually knowing how the system works. GAs can optimize cost functions with multiple minima without numerical gradients for the cost functions. Hence, it is well suited for our sensor selection problem because we have little prior knowledge about the relationship between the error reduction and the configuration of sensors. For a complete discussion on genetic algorithms, please see [15].

The search for an appropriate configuration begins with a collection of initial configurations. Members of the current population are used to generate the next generation population by means of operations such as random mutation and crossover, which are patterned after processes in biological evolution. At each step, the configurations in the current population are evaluated by the reduction of error after data assimilation. Those with the highest error reduction are selected probabilistically as seeds to produce the next set of configurations.

### 4.1 Representing The Network Configuration

We employ a standard hypothesis bit-string (or chromosome) representation that is often used in GAs. The advantage of this representation is that it can be easily manipulated by genetic operators such as crossover and mutation.

Since we are only optimizing the number of sensors in the network, the network configuration can be represented by an $n$-bit string $\underbrace{10111...1}_{n}$

where $n$ is the total number of sensors in the network.

0 means that the sensor is not used.

1 means that the sensor is used in the configuration.

For example: 10101 is a configuration of a network of 5 sensors in which the $1^{st}$, $3^{rd}$, and $5^{th}$ sensors are used while $2^{nd}$ and $4^{th}$ sensors are not used.

### 4.2 Fitness Function and Selection

The fitness function defines the criterion to rank the configurations for the purpose of selection. In our problem settings, the most appropriate criterion is the error reduction in data assimilation. Hence, the fitness function calculates the error reduction in the data assimilation using that configuration.

There are several popular selection methods such as fitness proportionate selection, tournament selection, and rank selection. Each selection method has its own advantages and disadvantages [15]. In our approach, we use the tournament selection method which runs a competition among a few individuals selected randomly and select ones with the best fitness. The tournament selection method often yields a more diverse population than other methods. Hence, a broader range of configurations can be considered during training.

### 4.3 Crossover and Mutation

We use standard settings for the crossover and mutation functions. We use scattered crossover as the operator instead of single point or intermediate crossover because it maximizes the information exchange among individuals. We use the gaussian mutation strategy because it is popular and standard in GAs.

## 5 Experimental Results

This section describes the experiments conducted to evaluate if GAs can produce a good set of sensors. The hypothesis we propose and test is that the sensor set found by a GA can save significant resources while maintaining a level of estimation accuracy similar to the current observation network.

The metric we used is the same as the cost function in Equation 2 because it is the optimization criterion. The evaluation of the sensor set is based on the error reduction in the data assimilation using the data from this configuration.

$$error\_reduction = 1 - \frac{sum[(xs\_twin - xs\_data).^2]}{sum[(xs\_twin - xs\_free).^2]} \tag{4}$$

The $error\_reduction$ lies between 0 to 1 because $sum[(xs\_twin - xs\_data).^2]$ is smaller than $sum[(xs\_twin - xs\_free).^2]$ as the measurements are incorporated in the estimation of $xs\_data$. Ideally, the higher the $error\_reduction$ is, the better the set of sensors.

Another metric that we consider is the cost of sensor equipment, deployment, and maintenance, that we can save by reducing the number of fully operational sensors in CORIE. We assume the costs for different sensors are the same. In practice, this is not true. Deployment and operation costs for sensors depend on sensor location and sensor type. However, we simplify the model to give an idea of how much money we can save by selectively reducing the number of sensors as follows:

$$cost\_reduction = (eq\_cost + dep\_cost + mnt\_cost) * num\_sensor \quad (5)$$

where $num\_sensor$ is the number of sensors we can remove from CORIE. The average equipment cost, $eq\_cost$, is approximately \$4,500 per sensor. The deployment cost, $dep\_cost$, is about \$500 per sensor. The maintenance cost, $mnt\_cost$ is about \$1000 per sensor. These are derived from actual costs in CORIE. We do not take into account the cost to deploy the station and the power and communication system because we can use one station for several sensors.

### 5.1 Experimental Design

We conduct the experiments using data from the CORIE observation network. The network consists of 23 stations with 34 sensors deployed in the Columbia river estuary.

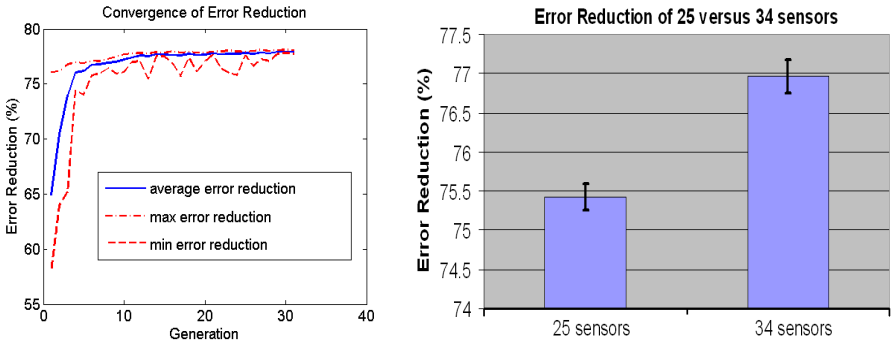


**Fig. 4.** Results: a) The error reduction converges and reaches a stable state after 10 generations. b) The error reduction of the data assimilation using 25 GA-selected sensors is only 1.55% smaller than using all 34 sensors.

Due to the fact that we never know the true state of the dynamic system, we set up twin experiments that use the real data to estimate the true state of the

system and use this estimated true state to simulate the measurements for the data assimilation.

We use a separate hindcast data $xs\_twin$ and consider it as the true state. $xs\_twin$ is then used to simulate the observations from the sensor network. The measurements are used as the input to the data assimilation. The output of the data assimilation is $xs\_data$, which is the estimated system state. On the other hand, by using the model only, we also simulate $xs\_free$ as the simulated system state. $xs\_free$ is obtained without data assimilation. Readers should distinguish between the process model, which is known and used to simulate $xs\_free$ and the error reduction model, which we have little understanding about. The settings for GA are listed below.

- Hypothesis representation: 34-bit chromosome
- population size: 20
- crossover rate: 0.8
- crossover operator: scattered
- mutation strategy : gaussian
- selection method: tournament
- number of generation: 30
- fitness function: average error reduction of 5 runs
- runtime: 35 days

Due to limited processing capability, we do not set the threshold error reduction $\varepsilon$ to find the configuration. Instead, we observe the best configuration after 30 generations.

### 5.2 Results and Analysis

The experiments finished after 35 days with the error reduction convergence. One might wonder about the experiment run time. As we mentioned earlier, the 3D circulation model state size is 878,850 — 8370 grid points × [(1 salinity + 1 temperature + 2 velocities) × 26 levels + 1 elevation]. Although the data assimilation is operated in the reduced space of 60 variables, the evaluation of error reduction of individual sensing type must be done in the full space of 878,850 at each time step. Hence, one complete data assimilation alone takes 20 minutes on 2-day data. The total time to finish the GA can be estimated as $20 \times 30 \times 5 \times 0.3/24 = 37.5$ days. However, this number can be significantly reduced by leveraging the inherent parallelism in GA. For example, the total time can be reduced to one week if 5 machines are used for the experiment. However, this motivates the design of a new algorithm to make GA parallelism possible. This is not the focus of our work. However, there exist several popular ways to accomplish it [15].

The best configuration after the $30^{th}$ generation was:
11111010011111111001111001011101111.

This means that 9 sensors or 26.5% resources are not used. We verify this configuration by the second experiment, in which we run data assimilation 30

times for 2-day data. The error reduction achieved was 75.42%. This is only 1.55% lower than that using all 34 sensors as shown in Figure 4.

If the difference in the error reduction is negligible, it means that we can save 9 sensors. According to the estimated costs for initial equipment, deployment and maintenance, we can save around 40 thousand dollars of initial expense and 10 thousands dollars for maintenance per year.

## 6 Related Work

The problem of network configuration or sensor selection has attracted significant interest in the sensor networks research community. Several papers [22] [13] [20] [1] try to address the problem for varying classes of sensors, network scale and the underlying physical process that the network is monitoring.

In one of the earliest works on the sensor coverage problem, Megeurdicherian *et al.* [14] proposed a solution that given the knowledge of existing sensor positions uses Voronoi diagrams to compute the maximal breach paths in the sensor field and find gaps in coverage, where additional sensors can be deployed. Similarly, Wang *et al.* [21] proposed a solution to network coverage by integrating sensing and connectivity constraints. The limitation of their work is that they use a simple signal attenuation model for a particular sensing modality to evaluate the utility of each sensor, rather than considering the complete data assimilation process.

Willett *et al.* [22] proposed an adaptive sampling scheme called Backcasting. They try to address a similar problem to ours, which is to minimize the number of active sensors while maintaining high accuracy. However, they assume a dense uniform distribution of sensors and eliminate sensors by considering the correlation of the environment estimated from a fusion center. The context in coastal modeling is slightly different w there are only a few expensive sensor stations deployed in a very large geographical area. Hence, the assumptions are no longer valid.

One direction in solving the sensor selection problem for target tracking tries to use concepts from information theory [8] [13] [20]. Ertin *et al.* [8] and Liu *et al.* [13] consider the *mutual information* between the predicted sensor observation and the current target location distribution as the criterion for selecting sensors. This approach works because mutual information actually represents the reduction in the uncertainty of one random variable to the knowledge of the other [4]. Wang *et al.* [20] overcome the expensive computation of mutual information by introducing an *entropy-based* approach. The authors claim that the difference between the entropy of the probability distribution of the sensor view and the entropy of the sensing model for a true target is strongly related to the mutual information. Hence, this information can be used to sort sensors more quickly while still maintaining similar results. While these attempts show very interesting findings, they are formulated for target tracking and localization problems. It is unclear how the approaches can be applied to the ocean monitoring problem. In addition, the approaches implicitly assume that a greedy selection of the set

of the most informative sensors provides the most information. However, as we observed in our data assimilation problem, this is not always the case.

The work of Krause *et al.* [12] and Bian *et al.* [1] formulate the problem as a form of optimization with some cost function, utility function or sensing quality, subject to constraints such as energy consumption [1] or communication cost [12]. We found that they are very close to our problem theoretically. However, their problem context is different from ours because we do not have any constraints on energy or communication cost. All sensor stations in CORIE are wired with power and data cables. Another difference is that our problem addresses a very large and complex geographical region, the Columbia river estuary. Therefore, determining the super modular utility function [1] or predicting sensing quality [12] is infeasible.

There are various works attempting to solve related problems such as adaptive sampling for localized phenomena [7], and sensor deployments that differ from our work in that they optimize certain specific criteria [16] [2] [5]. The problems are related but different to ours. Therefore, most of their approaches are not applicable to the problem we address.

Finally, there are attempts to use genetic algorithms to select sensor parameters [3] or select noisy sensor data [11]. Our work is different in that we show that we can use genetic algorithms combined with data assimilation for applications in ocean observation and coastal monitoring.

## 7 Future Work

Genetic algorithms do not use knowledge about the relationship between sensor configuration and monitoring precision. However, detailed investigation about the physical process model may help in better explaining the relationship between sensor configuration and monitoring precision. We also would like to assess the effectiveness of the genetic algorithm results with monthly, seasonal, and yearly environmental changes.

As mentioned before, finding an optimal configuration of sensors is an unsolved hard problem. As future work, we would like to investigate optimization algorithms that take into account not only the number of sensors but also the sensor type and sensor location to determine an optimal network configuration. We would also like to try our framework with other modeling and sensing systems such as atmospheric sensing besides the Columbia river estuary system to ensure the usefulness of our approach in practice. Finally, in sparse wide-area observation networks such as CORIE, the long term data collection from static stations is often augmented with opportunistic data collection from mobile stations. In the CORIE project, Clatsop Community College's M/V Forerunner serves as a mobile station of opportunity, and several cruises have been conducted over the years. As part of the CORIE project, we are currently investigating how to guide the trajectory of these vehicle cruises to optimize the observation process.

## 8 Conclusion

CORIE is a pilot environmental observation and forecasting system for the Columbia River. The CORIE observation network differs from low-power, dense wireless sensor networks in one aspect - sensor stations are sparse and expensive to deploy and maintain. The challenge for scientists is to maintain the accuracy of their modeling system while reducing the use of expensive resources.

We showed that genetic algorithms can aid in optimizing the configuration of the CORIE observation network. Specifically, we were able to reduce the number of observation stations without compromising the accuracy of the state estimate; leading to potential savings in the deployment and maintenance cost for the observatory. The novelty of this paper is that our problem formulation of sensor selection is influenced by the data assimilation framework which is more meaningful to domain scientists, rather than by abstract sensing models. Our approach and algorithm are simple and potentially generalizable to other wide area environmental sensing systems.

**Acknowledgements** We would like to thank Michael Wilkin for providing information about the cost of operating expenses of equipments in the CORIE project. We also would like to thank Dave Maier, Todd Leen, and Eric Wan for useful discussions during the project. The research described in this paper was supported by National Science Foundation grants NSF 05-14818 and 01-21475.

## References

1. F. Bian, D. Kempe, and R. Govindan. Utility based sensor selection. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN 06)*, pages 11–18, Nashville, Tennessee, April 2006.
2. J. L. Bredin, E. D. Demaine, M. Hajiaghayi, and D. Rus. Deploying sensor networks with guaranteed capacity and fault tolerance. In *Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing*, pages 309 – 319, Urbana-Champaign, Illinois, May 2005.
3. P. Corcoran, J. Anglesea, and M. Elshaw. The application of genetic algorithms to sensor parameter selection for multisensor array configuration. *Sensors and Actuators A Physical*, 76:57–66, February 1999.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc, New York, 1991.
5. R. Cristescu and M. Vetterli. On the optimal density for real-time data gathering of spatio-temporal processes in sensor networks. In *Proceedings of the Fourth International Symposium on Information Processing In Sensor Networks*, pages 159–164, Los Angeles, California, April 2005.
6. J. Delaney. Keynote: Next-generation earth and ocean sciences: Opportunities and challenges. In *Proceedings of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, San Diego, California, November 2005.
7. E. B. Ermis and V. Saligrama. Adaptive statistical sampling methods for decentralized estimation and detection of localized phenomena. In *Proceedings of the Fourth International Symposium on Information Processing in Sensor Networks (IPSN 05)*, pages 143–150, Los Angeles, California, April 2005.

8. E. Ertin, J. W. Fisher, and L. C. Potter. Maximum mutual information principle for dynamic sensor query problems. In *Proceedings of the Second International Symposium on Information Processing in Sensor Networks (IPSN 03)*, pages 405–416, Palo Alto, California, 2003.
9. G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research*, C5(10), 1999.
10. S. Frolov, A. Baptista, Z. Lu, R. van der Merwe, and T. Leen. Fast data assimilation with model surrogates: Application to circulation in a highly stratified estury. In *Submission to Ocean Modeling*.
11. A. A. Khan and M. A. Zohdy. A genetic algorithm for selection of noisy sensor data in multisensor data fusion. In *Proceedings of American Control Conference*, pages 2256–2262, Albuquerque, NM, June 1997.
12. A. Krause, C. Guestrin, A. Gupta, and J. M. Kleinberg. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN 06)*, pages 2–10, Nashville, Tennessee, April 2006.
13. J. Liu, J. Reich, and F. Zhao. Collaborative in-network processing for target tracking. *EURASIP Journal on Applied Signal Processing*, 4, 2002.
14. S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava. Coverage problems in wireless ad-hoc sensor networks. In *Proceedings of the Conference on Computer Communications 2001 (INFOCOM 2001)*, pages 1380–1387, Anchorage,Alaska, April 2001.
15. T. M. Michell. *Machine Learning*. Mc Graw Hill, New York, 1997.
16. S. Ray, W. Lai, and I. C. Paschalidis. Deployment optimization of sensornet-based stochastic location-detection systems. In *Proceedings of the Conference on Computer Communications 2005 (INFOCOM 2005)*, Miami, Florida, March.
17. L. I. Smith. A tutorial on principle component analysis. *http://kybele.psych.cornell.edu/Ẽedelman/Psych-465-Spring-2003/PCA-tutorial.pdf*, February 2007.
18. R. van der Merwe, T. Leen, Z. Lu, S. Frolov, and A. M. Baptista. Fast neural network surrogates for very high dimensional physics-based models in computational oceanography. *Neural Computation (To appear)*, 2007.
19. R. van der Merwe and E. A. Wan. Sigma-point kalman filters for probabilistic inference in dynamic state-space models. In *Proceedings of the Workshop on Advances in Machine Learning*, Montreal, Canada, June 2003.
20. H. Wang, K. Yao, G. Pottie, and D. Estrin. Entropy-based sensor selection heuristic for target localization. In *Proceedings of the third international symposium on Information processing in sensor networks (IPSN 04)*, pages 36–45, Berkeley, California, USA, April 2004.
21. X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless, and C. Gill. Integrated coverage and connectivity configuration in wireless sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems (Sensys)*, pages 28–39, New York, NY, USA, 2003. ACM Press.
22. R. willett, A. Martin, and R. Nowak. Backcasting: Adaptive sampling for sensor networks. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN 06)*, pages 36–45, Nashville, Tennessee, April 2006.