# INTEGRATING DATA SOURCES TO OPTIMIZE LONG-TERM MONITORING, OPERATION, AND STEWARDSHIP

***William J. Michael*** (wmichael@uiuc.edu), David K. Tcheng, Barbara S. Minsker, andAlbert J.Valocchi (University of Illinois, Urbana, Illinois)
John J. Quinn, and
Gustavious P. Williams (Argonne National Laboratory, Argonne, Illinois)

**ABSTRACT**: Due to technical limitations and the high cost of hazardous waste site clean up, there has been a shift toward risk-based long-term management of sites, where contamination is left in place. This study demonstrates how integrating all available site data can improve LTMOS decision making and provide cost savings. A learning machine is used to integrate historic and current data from the 317/319 Area phytoremediation site at Argonne National Lab-East (ANL-E). The learning machine uses these data and daily weather data to build a model to forecast groundwater head levels. Development of the learning machine framework provides a method for integrating the diverse data sources available at the site and using that information to determine the importance of each data source in achieving monitoring objectives. Future work will determine how long the historical record will retain its accurate predictive capability and whether the value of the surrogate data (continuous samples and rainfall) increases over time. In this preliminary study, the entire historical quarterly dataset was shown to be the most important data source, which could be used to predict future water levels with far more accuracy than the most recent quarterly dataset alone.

## INTRODUCTION

Due to technical limitations and the high cost of hazardous waste site clean up, there has been a shift toward risk-based long-term management of sites, where some contamination is left in place (NRC 1994, 1999). A recent DOE report identifies the need for managing existing restoration sites for periods of 70 years or longer (NRC 2000). As the groundwater remediation field matures and the installation of remediation systems are completed, it is becoming clear that long-term monitoring, operation, and stewardship (LTMOS) of these systems will comprise a significant portion of future expenditures. LTMOS data collection objectives are typically not well defined and only a small portion of the data currently collected is used to assess the remediation progress.

To enable improved use of data and identification of data needs for LTMOS, this project seeks to lay the groundwork for a hierarchical framework that will optimize the knowledge and data stored in scattered data sets and resources through a simple and efficient system. The framework makes use of several methods, both novel and standard to the industry, and evaluates which one is best for a given problem. Ideally, larger problems will be solved on different levels by a variety of methods to achieve a much better result that any single method alone – and faster than any expert could put the methods together. Analysis of the best methods will provide insights into which data are most valuable to LTMOS objectives and which data are not. Data collection can then

focus on the most valuable data, ultimately reducing the long-term costs of monitoring while providing improved understanding of remediation performance.

## BACKGROUND

This study examines LTMOS at the 317/319 Area at Argonne National Laboratory. The 317/319 Area was used for disposal of solvents in the 1950s, leading to VOC and tritium contamination of the local groundwater aquifer. During the summer of 1999, soil mixing was performed in the source area and nearly one thousand trees were planted to provide hydraulic containment and to extract and transpire contaminants (Quinn, et al, 2001). As the trees mature, this phytoremediation system is designed to ultimately replace the pump and treat system currently in place. Because of the tremendous changes in the subsurface due to the soil mixing and tree planting in the summer of 1999, our study is limited to data since that time. The site offers the challenges of a typical groundwater remediation site – uncertain waste-treatment practice history and incomplete or sparse historical data sets.

A major objective of the remediation system at this site is to provide hydraulic control of the groundwater flows. To accomplish this objective, much of the data collection effort has focused on obtaining good estimates of water levels. For this reason, this initial study solely examines data related to water levels as a prototype for a more extensive system that would also include data related to concentrations of contaminants of concern. Three data sources are used in this study: traditional quarterly water level measurements from monitoring wells collected between November and March 2001, continuous water level readings during May 2001, and rainfall data just prior to each water level reading. To demonstrate the capabilities of the framework for integrating different datasources and improving predictions of remediation performance, these historical data are then used to predict the quarterly water levels that were measured in May 2001. Each source is described in more detail below.

Currently water levels at twenty-two monitoring wells are measured routinely each quarter as part of a groundwater sampling program. All of the quarterly data are included in the study because they are more spatially complete than the continuous well samples, as shown in Figure 1. In addition to the quarterly measurements, seven wells are sampled for water levels hourly and henceforth will be referred to as continuous. To ensure that the data are representative of current conditions, the continuous data are limited to the thirty days prior to the sampling dates in May 2001 that are the focus of this study. Continuous measurements taken on May sampling days were also included to test whether the data from the continuous wells could be used as surrogates for determining the levels of the other wells. Such an approach would reduce costs because the continuous wells are automatically sampled and the data collection is ongoing, whereas the quarterly sampling is an additional expense that requires labor to sample the wells, sometimes over several days for each quarter. The final dataset considered in this study is rainfall data. The addition of the rainfall data was an obvious choice from the relationship between the water levels and rainfall shown in the continuous data in Figure 2. Rainfalls were included for the day prior to each water level measurement for the quarterly samples, and for the hour prior for the continuous samples. This rainfall data comes from two different monitoring stations. One station is located at the site of concern, less than 20 meters north of the wells; however, the rainfall data collection at

this location was not complete. The main Argonne meteorological station is located about 1 kilometer directly west of the 317/319 area and any gaps in the near station were filled with data from the main Argonne site. Although the rainfall data for both sites is very similar, their proximity does not insure that they are the same.

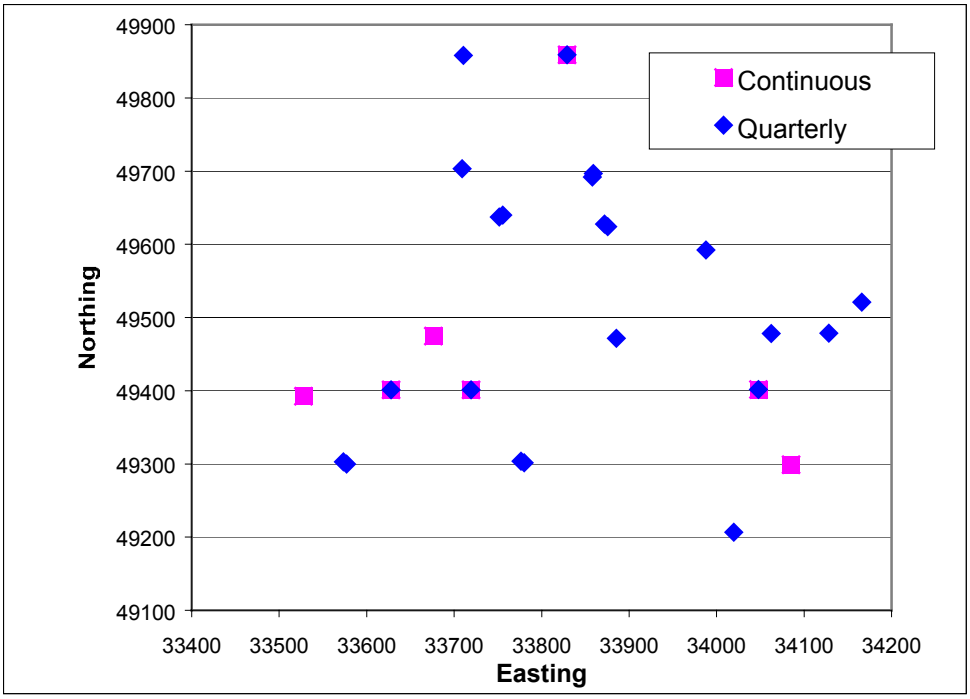


**FIGURE 1 – Well Locations at the 317/319 Area.**

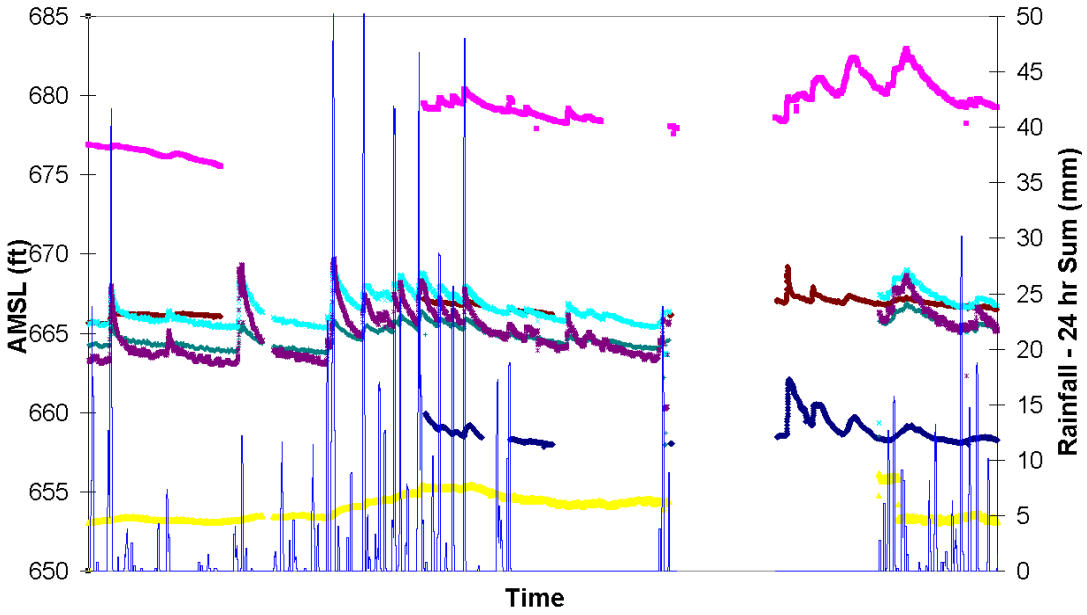


**FIGURE 2 – Continuous head measurements and precipitation data.**

For testing the framework, these datasets are further broken down into training and testing sets. The training sets contain the historical data that are assumed to be known at the time of the analysis and is used to train the learning machine that performs the predictions. The testing set includes all of the quarterly data collected at the time of the May 2001 sampling.

**TABLE 1 – Summary of the data sets.**

| **Water Level Training Sets** | **Number of Measurements** |
|---|---|
| Historical Quarterly Data | 147 |
| Most Recent Quarterly Data | 16 |
| Continuous data from the last 30 days prior to the May testing dates | 5047 |
| | |
| **Testing Set** | |
| May 2001 | 22 |

## METHODOLOGY

The hierarchical framework developed in this study uses the historical data to train a variety of learning machines to predict water levels at each monitoring well. Learning machines are a class of methods for training computers to predict or classify data using a training data set. These methods can range from simple regression to sophisticated artificial neural networks. For this demonstration, we compare decision trees (Quinlan, 1986) and inverse distance weighting approaches. Decision trees have been used extensively for data mining of commercial datasets, while inverse distance weighting has been a standard approach in the remediation industry for interpolating spatial data.All of the framework development and machine learning takes place in the D2K environment – a Java based data mining tool from the National Center for Super-Computing Applications (see http://www.ncsa.uiuc/D2K). Each data set is read into a D2K itinerary - a collection of prewritten modules that essentially make up a program- and the itinerary automatically builds a model that is used to predict the May 2001 data. Training sets with and without the rainfall data were evaluated.

Figure 3 shows the hierarchical framework developed as an itinerary within D2K. In the first level of the hierarchy, a model is built to predict the water levels using the spatial coordinates of each well (x,y) and the rainfall (r), training on the continuous data alone. The water level prediction from this model is then added to the quarterly data set as another input into the second level of the hierarchy, which automatically weights the relative importance of each data set in making a final prediction. This hierarchical structure is necessary because of the vastly different scales of the continuous and quarterly datasets. If the quarterly data were simply combined with the continuous data into a single learning machine, our initial trials indicated that important information contained in the historical quarterly data would be overwhelmed by the volume of the continuous readings. The hierarchical arrangement avoids this difficulty.

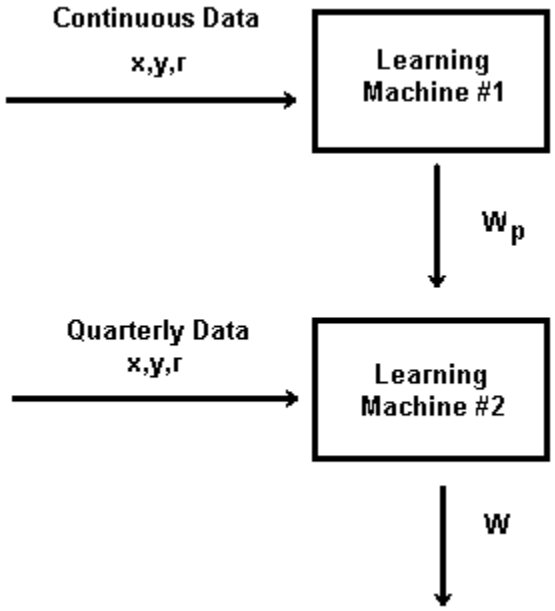


**FIGURE 3 – Hierarchical model diagram.**

**RESULTS**

The framework performs extremely well at predicting the water levels in May 2001, as shown in Table 2. Note that the prediction errors shown in Table 2 are simply the average percentage differences between the predicted and actual values. All of the training sets that include the quarterly data perform better than the continuous data alone. Moreover, using all of the historical quarterly data provides far better predictions than using only the most recent quarterly data. The addition of the rainfall information improves the predictions from the historical quarterly data

The addition of the continuous data to the quarterly measurements in the combined datasets provides improvement only in the case where decision trees are used with the rainfall data. This indicates that the continuous data are not as useful as the quarterly data, primarily because of their poor spatial coverage. (see Figure 4).

The best performance came from the inverse distance weighting model on the historical quarterly data with the rainfall data, with an average percentage error in the predicted water levels of only 2.34%. Figure 4 shows the spatial distribution of errors for this case, with an absolute range from 0.01% to 5.09%.

These results show substantial promise for this framework. Further results will be presented at the conference to compare other types of learning machines and to examine the performance of the models on more recent data to determine whether the value of the historical data set deteriorates over time. This analysis will be useful for identifying how long we can collect only the surrogate data (rainfall and continuous water level measurements) and not collect quarterly samples before our predictions will suffer.

**TABLE 2 – Summary of results.**

| Predicting May 2001 | | Testing Error | | | |
|---|---|---|---|---|---|
| | | Decision Trees | | Inverse Distance Weighting | |
| Training Sets | | w/o Rainfall | w Rainfall | w/o Rainfall | w Rainfall |
| Most Recent Well Measurements | | 11.09% | 11.09% | 7.30% | 7.78% |
| All Well Measurements | | 3.28% | 3.23% | 3.29% | 2.34% |
| Continuous Well Measurements | | 12.60% | 12.62% | 17.13% | 17.18% |
| **Training Sets - Combined Sources** | | | | | |
| Most Recent Well Measurements | Continuous Well Measurements | 11.09% | 11.09% | 7.00% | 13.06% |
| All Well Measurements | Continuous Well Measurements | 3.28% | 2.69% | 3.29% | 3.20% |

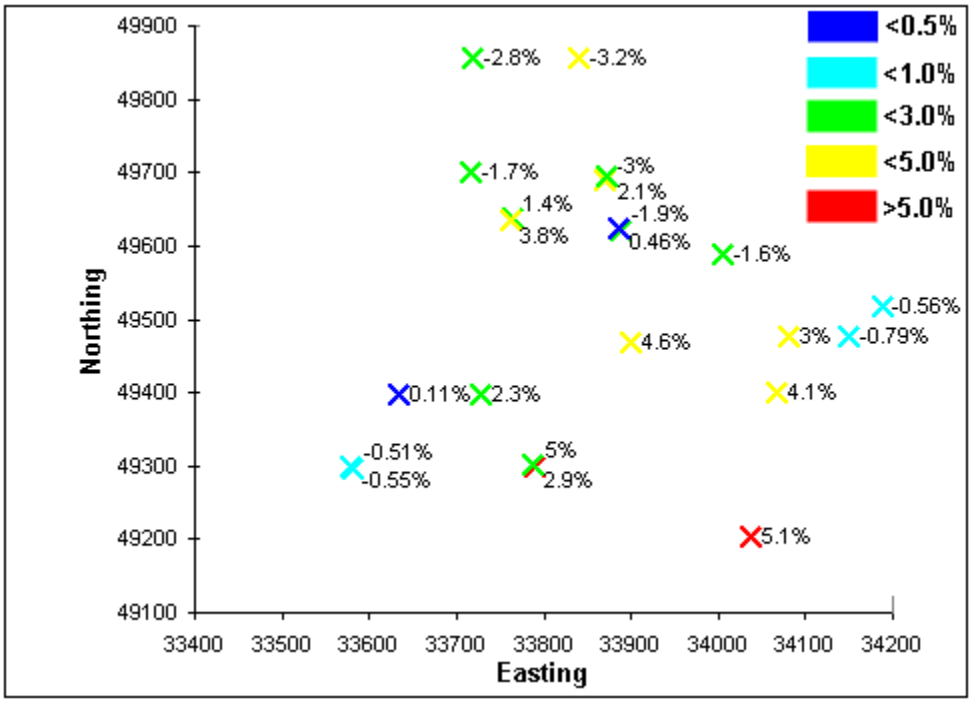


**FIGURE 4 – Error of the model prediction for each well sample in May 2001.**

## CONCLUSIONS

This paper demonstrates that knowledge integration, which combines different sources of data in a machine learning framework, holds considerable promise for the future of groundwater monitoring. Most monitoring data analysis includes only the most

recently collected data, but our results show that the historical quarterly sampling record is the most valuable source of information at this site. Further work is needed to determine how long the historical record will retain its accurate predictive capability and whether the value of the surrogate data (continuous samples and rainfall) increases over time.

Combining non-traditional data sets and optimizing the value of current data as demonstrated in this paper will improve predictions so that the most useful data can be collected efficiently and cost-effectively. This same type of approach could be applied to contaminant concentrations with other surrogate data such as dissolved oxygen or pH sensor measurements, oxidation reduction potential, and measurements from remediation systems such as off-gassing from the trees in the phytoremediation system or effluent measurements from the pump and treat system.

## ACKNOWLEDGMENTS

This material is based upon work supported by the Department of Energy Argonne National Laboratory under grant number IF-01628 and by the National Center for Supercomputing Applications.

## REFERENCES

National Research Council (2000). *Natural Attenuation for Groundwater Remediation*, National Academy Press, Washington, D. C..

National Research Council (1999). *Groundwater and Soil Cleanup: Improving Management of Persistent Contaminants*, National Academy Press, Washington, D. C..

National Research Council (1994). *Alternatives for Groundwater Cleanup*, National Academy Press, Washington, D. C..

Quinlan, J. R., Induction of decision trees. *Machine Learning*, 1:81-106, 1986

Quinn, J.J., M.C. Negri, R.R. Hinchman, L.M. Moos, J.B. Wozniak, and E.G. Gatliff, 2001, Predicting the Effect of Deep-Rooted Hybrid Poplars on the Groundwater Flow System at a Phytoremediation Site: *International Journal of Phytoremediation*, vol. 3, no. 1, p. 41-60.