INTEGRATING DATA SOURCES TO IMPROVE LONG-TERM MONITORING AND MANAGEMENT: A HIERARCHICAL MACHINE LEARNING APPROACH

BY

WILLIAM JOSEPH MICHAEL

B.S., University of Illinois at Urbana-Champaign, 2000

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Environmental Engineering in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2002

Urbana, Illinois

ABSTRACT

As the groundwater remediation field matures and the installation of remediation systems is completed, it is becoming clear that long-term monitoring and management of these systems will comprise a significant portion of future expenditures. The purpose of this paper is to demonstrate how integrating all available site data using a hierarchical machine learning approach can improve decision-making and provide cost savings.

This study proposes a new hierarchical modeling framework and successfully tests that framework for integrating historical and current data from the 317/319 Area phytoremediation site at Argonne National Laboratory-East (ANL-E). Site data used in this study include: hourly head measurements from a subset of seven monitoring wells and quarterly head measurements from the remaining wells. The learning machine uses these data and daily weather data (to account for the impact of recharge on the groundwater regime) to forecast groundwater head levels. The performance of the model built by the learning machine is assessed by comparing predictions from previous monitoring periods to the most recently available data. By comparing predictive performance using different combinations of datasets, the most relevant data are identified to guide future sampling efforts.

The best data-driven predictions resulted in an average error of 1.5 feet, compared with average errors of 7.5 feet from an existing Modflow numerical flow model. The data-driven predictions were obtained from all of the historical quarterly data; the hourly head measurements were not as useful for prediction, most likely because of their poor spatial coverage. The Modflow model was also incorporated into the framework to test the framework's capabilities for updating numerical models to improve predictions as new data are collected. A combined data-driven and Modflow model had an average error 75% lower than the Modflow model, even when the Modflow model was updated to reflect more recent flow conditions. These results demonstrate that the proposed framework holds substantial promise for improving predictive performance of existing numerical models in areas with good data coverage.

ACKNOWLEDGMENTS

I would like to thank Barbara Minsker, for giving me more than my share of guidance throughout my pre-graduate and graduate careers, for providing a spark of motivation when I had none, and for having faith in me and providing the countless opportunities to challenge myself in more ways than I ever signed up for.

David Tcheng at the National Center for Supercomputing Applications for dragging us into the world of the learning machines and for demanding that we see things on a much larger scale. Loretta Auvil was a big help at NCSA when David wasn't available.

Al Valocchi for his expertise in groundwater modeling and for being a patient editor.

John Quinn and Gus Williams at Argonne National Laboratory-East for their hospitality during my summer at ANL-E and the quick replies to my e-mails since.

The members of the research group and others graduate students who took time to answer or ask questions and for providing insight into graduate work – in no particular order: Felipe Espinoza, Pat Reed, Gayathri Gopalakrishnan, Yong Liu, Kerry Howe, Abhishek Singh, Shengquan Yan, Meghna Babbar and Rachel Arst.

This material is based upon work supported by the Department of Energy Argonne National Laboratory under grant number IF-01628, by the National Center for Supercomputing Applications and by the Walter E. Deuchler Fellowship. Without their support I would not have had the opportunity to further my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Motivation

Due to technical limitations and the high cost of remediation 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 and management ("stewardship") of these systems will comprise a significant portion of future expenditures. In a 2001 report to Congress, DOE estimates that it will spend $5.5 billion on long-term site management between 2000 and 2006 and more than $100 million per year over the next 70 years [*DOE*, 2001]. Long-term data collection objectives are often not well defined and only a small portion of the data currently collected is typically used to assess remediation progress. To enable improved use of data and identification of data needs for long-term monitoring, this paper lays 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 system will provide straightforward mechanisms to improve site-wide predictions and allow existing models to be updated quickly and easily with the most recent data.

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 long-term monitoring 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.

This new hierarchical modeling framework's performance is successfully tested by predicting measured hydraulic heads during a routine sampling event at the 317/319 Area at the ANL-E site. The framework is then used to combine historical data with an existing Modflow numerical simulation model of the site and the results show that these novel techniques for combining data-driven and physics-based models hold promise for the future by updating existing site models and improving spatial coverage of models built from data alone. Eventually, the framework will have the ability to incorporate data at multiple scales and to expand and update all types of numerical models. Combining these diverse data sources will allow for better use of existing data to guide future data collection and improve overall predictive accuracy at groundwater remediation sites.

1.2 Site Background

This study examines long-term monitoring and management at the 317/319 Area at ANL-E. The 317/319 Area was used for disposal of solvents in the 1950s, leading to VOC and tritium contamination of the local groundwater aquifer. The aquifer of concern is a shallow, unconfined aquifer with a complex hydrogeology – a framework of glacial tills with extensive interfingerings of sands, gravels, and silts [*Quinn et al*, 2001] and a general direction of groundwater flow from the northwest to the southeast. During the summer of 1999, soil mixing was performed to remove the source contamination and nearly one thousand trees were planted to provide hydraulic containment and to extract and transpire the remaining contaminants

[*Quinn, et al*, 2001]. As the trees mature, this phytoremediation system was 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 initial 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 the first set of experiments in this study: traditional quarterly water level measurements taken at monitoring wells between November 1999 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 data sources and improving predictions of remediation performance, these historical data are then used to predict the quarterly water levels that were measured in May 2001 in the first set of experiments. For the second set of experiments, the outputs of a Modflow simulation model of site were added as a data source and the measurements of a non-routinely sampled set of wells from March 2001 were used as an additional testing set. Each source is described in more detail below.

1.3 Data Description

Currently water levels at twenty-two monitoring wells are measured routinely each quarter as part of a groundwater sampling program. In addition to the quarterly measurements, seven wells are sampled for water levels hourly and henceforth will be referred to as continuous. All of the quarterly data are included in the study because they are more spatially complete than the continuous well samples. Figure 1 shows the spatial coverage of the quarterly and continuously sampled wells relative to one another. 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 the 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 data source considered in this part of the 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 at several of the wells, as shown in Figure 2. Figure 2 shows the head levels (in feet above mean sea level – AMSL) and the 24-hr rainfall amounts from November 1999 to May 2001. Several of the wells show an almost immediate response to rainfall that tapers off over time. For this reason, rainfall amounts were included in the study 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

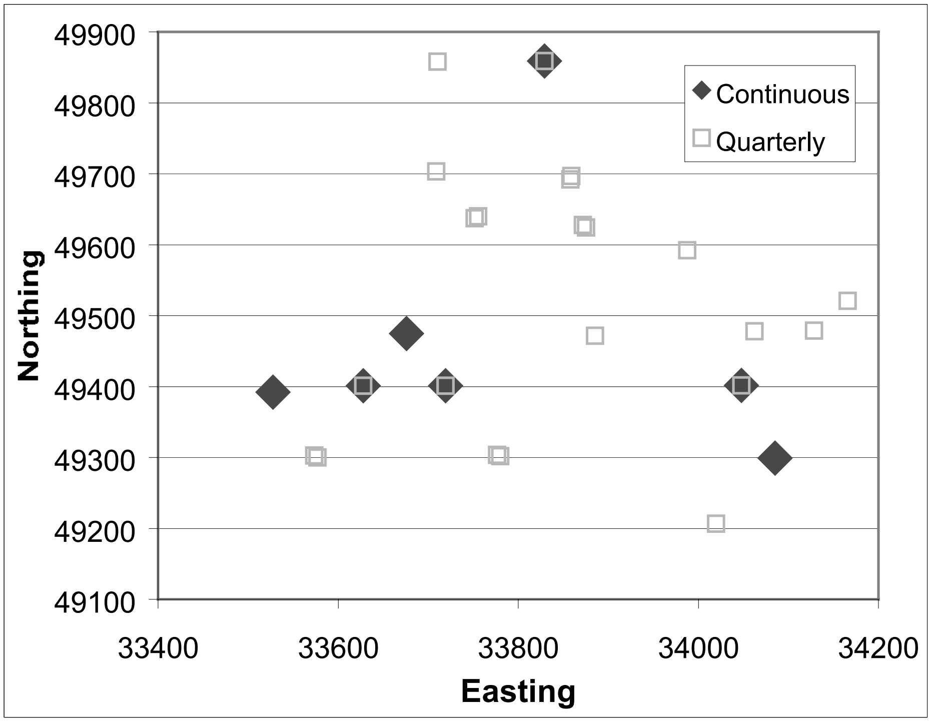

Figure 1. Quarterly and continuous water-level measurement locations at the 317/319 Area.

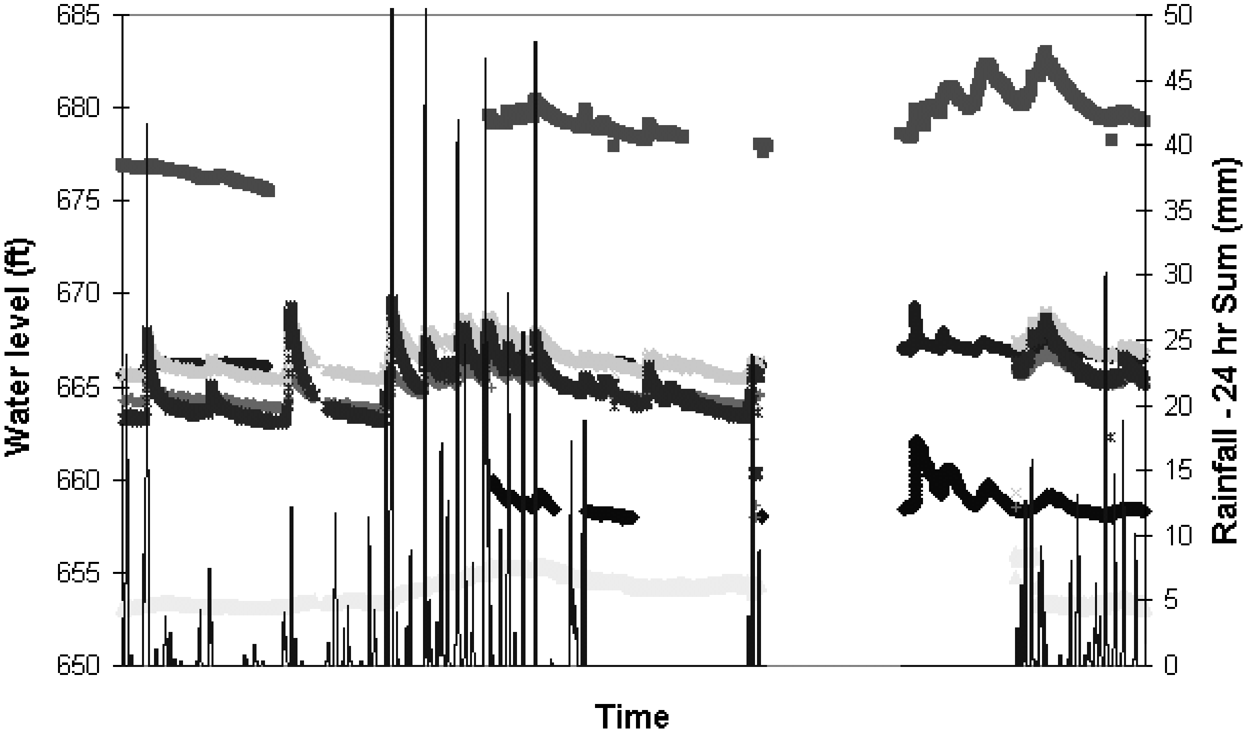

Figure 2. Correlation of the continuous head measurements and the rainfall data from November 1999 to May 2001.

rainfall data collection at this location was not complete. The main ANL-E 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 ANL-E meteorological site. Although the rainfall data for both sites are very similar, their proximity does not insure that they are exactly the same.

The Modflow simulation model, a six-year simulation model that was previously built to show the effects of the phytoremediation design [*Quinn, et al*, 2001], was also used in the study, both for comparison of predictive performance and as an additional data source. The length of the time step used in the model is monthly and hydraulic head data can be generated for any time step at any spatial location on the 3 by 3 meter numerical grid shown in the modeling domain depicted in Figure 3. Figure 3 shows that the spatial coverage of the Modflow model is much

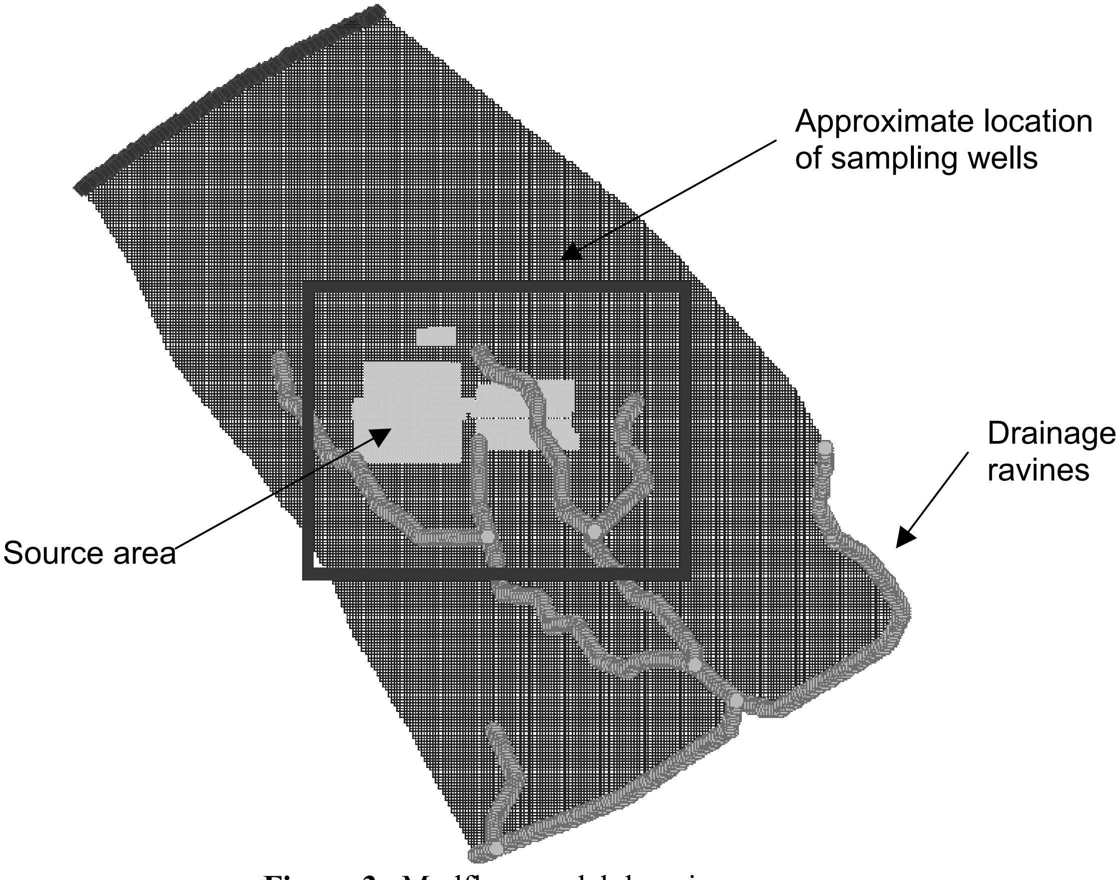

Figure 3. Modflow model domain

larger than the area where the routinely-sampled wells are located. The influence of the trees on flows are included in the Modflow model by modeling the trees as pumping wells. The amount of pumping from the trees varies, with a growing stage during the first two years and increased pumping and maturity for the remaining time. Because of the poor growing conditions in the first two years after planting, the trees had not yet reached maturity and only the Modflow predictions of the head levels during the period of interest from the first simulated year were used in the study, which fit the existing data best.

A trial and error approach was used to modify some parameters to improve the May 2001 predictions of the Modflow model. The initial head levels were updated using an inverse distance weighting interpolation of the November 1999 hydraulic head measurements, constant head boundary was increased to match head measurements more closely, and the recharge, which was originally modeled as varying seasonally, was changed to the actual monthly rainfall amounts. The hydraulic conductivity remained the same as the original model because no new data were available for updating that parameter. The original Modflow simulation model and the updated May Modflow simulation model were used as different data sources in the final experiment.

For testing the framework, these datasets are further broken down into training and testing sets, as shown in Table 1. The training sets contain the historical data that are assumed to be known at the time of the analysis and are used to train the learning machines that perform the predictions. The testing set includes all of the quarterly data collected at the time of the May 2001 sampling. Fifty measurements taken at non-routinely sampled wells in March 2001, which cover a much larger area and extend into the forest preserve south of the ANL-E site, were used to evaluate the models' ability to predict beyond the wells where the models were trained. See Figure 4 for the locations of the non-routinely sampled wells. The data collected in March 2001 at these wells represent the only complete sampling of these wells available to date, and because many of these wells were sampled only this one time since the soil mixing in the summer of 1999, training with earlier data for these wells was not possible.

Table 1. Summary of the data sets.

Water Level Training Sets	Number of Measurements
Historical quarterly data	147
March 2001 - Most recent quarterly data	16
Continuous data from the 30 days prior to the May testing dates	5,047
Testing Sets	
May 2001 – routinely sampled area	22
March 2001 – non-routine measurements & forest preserve area	50

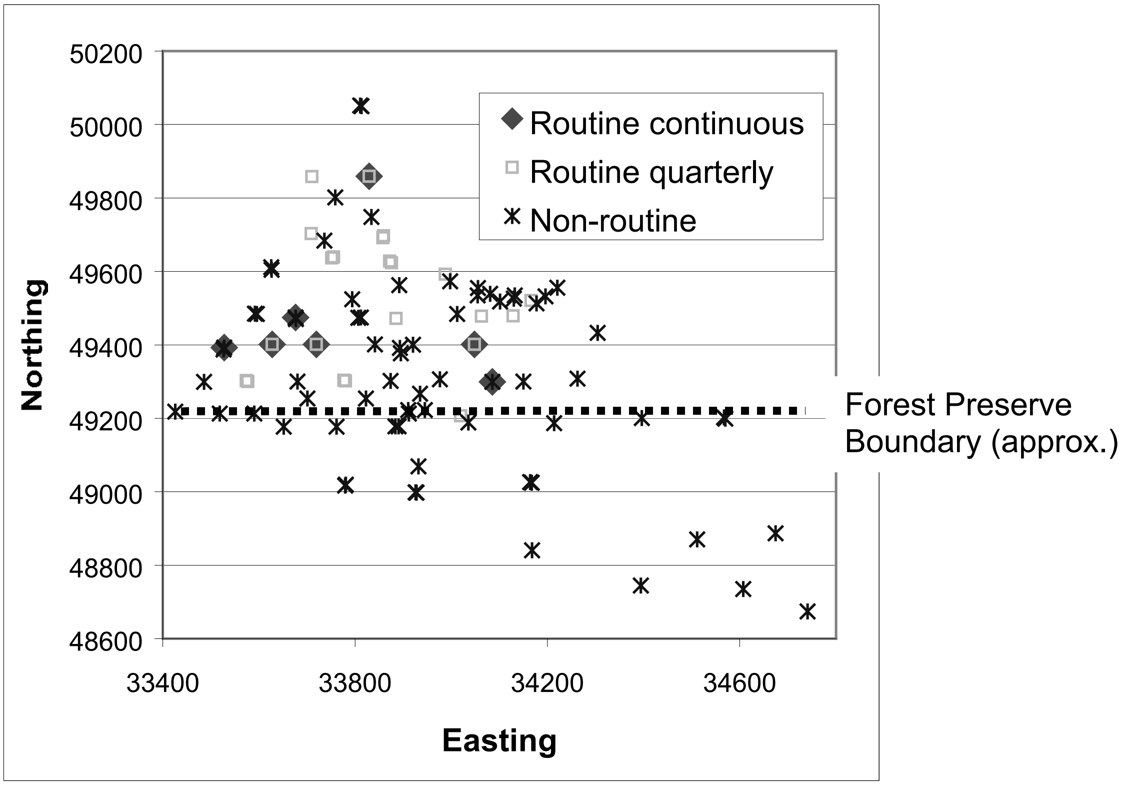

Figure 4. Locations of the non-routine well measurements collected in March 2001 relative to the routine quarterly and continuous wells.

2. METHODOLOGY

2.1 General

A general hierarchical framework for combining multiple knowledge sources is proposed in Figure 5. Low-level specialty tools and models are combined by higher level "experts" that learn how to best use the specialty tools, as suggested by *Nilsson* [1995]. Each of the data sources can be used to train a specialty machine learning model that will make predictions based on the input data. Physics-based models can also be used as specialty models, as we will demonstrate in Experiment 2. *Buchanan et al*. [1978] and *Rendell et al*. [1982] have found such tiered machine learning frameworks to be highly stable and accurate for commercial business data, but to our knowledge this type of approach has not yet been tested for water resources data.

In this study, the hierarchical framework proposed in Figure 5 is tested by using the historical data to train a variety of learning machines to predict water levels at each monitoring

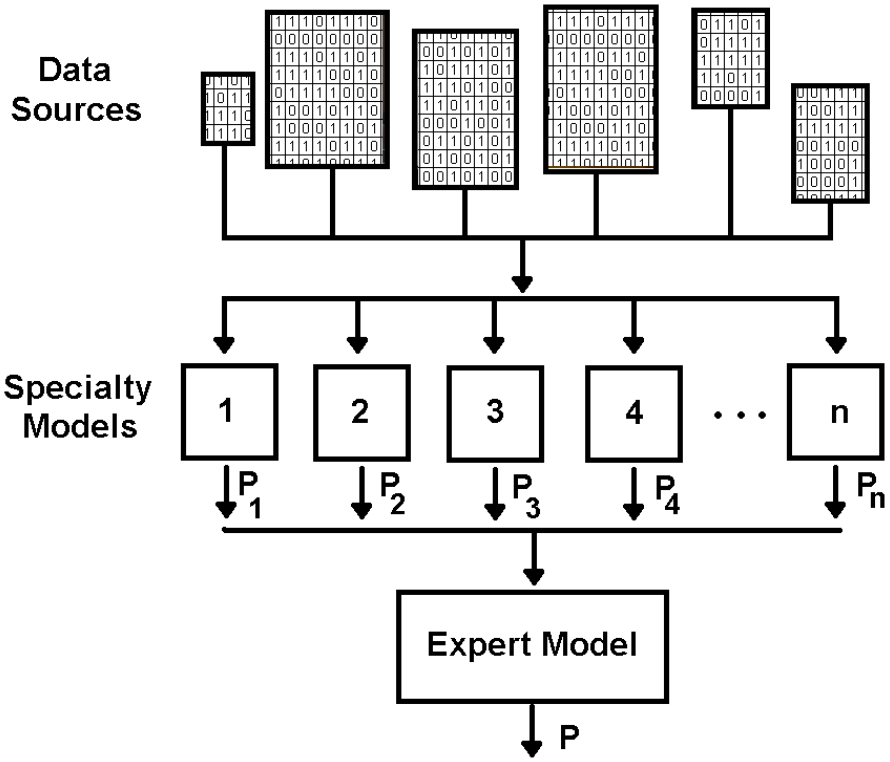

Figure 5. The proposed general hierarchical model framework.

well. Learning machines are a class of methods for automatically building computer models to predict or classify data using only a training data set. These methods can range from simple regression to sophisticated artificial neural networks.

2.2 Learning Machine Methods

For this paper, we use decision trees [*Quinlan*, 1986], instance-based weighting, inverse distance weighting, and neural network 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. Each of these approaches is briefly summarized below.

Decision Trees. Decision trees are constructed by recursively selecting the most predictive features and splitting the training sets into subsets. Splitting continues until the information in the inputs is exhausted and the terminal nodes are the classification of the final instances [*Matheus*, 1990]. In the decision tree each node represents an input and each branch a possible value of that input. The terminal nodes (or leaves) on the tree specify the output value for the combination of input values that prescribe the path to that terminal node [*UGAI*, 1995]. The maximum number of representative examples at the end of each path (leaf) was the only decision tree parameter that the learning machine adjusted during training.

Figure 6 shows a simplified example of a decision tree used to predict hydraulic heads. Each branch of the tree represents a path that is followed to make a prediction given a measured set of features. For example, a prediction with a Northing coordinate less than 0.62 (all values shown have been scaled between 0.1 and 0.9) will first take the path to the left of the root node at the top. At each successive node, the appropriate path is chosen by comparing the value of the

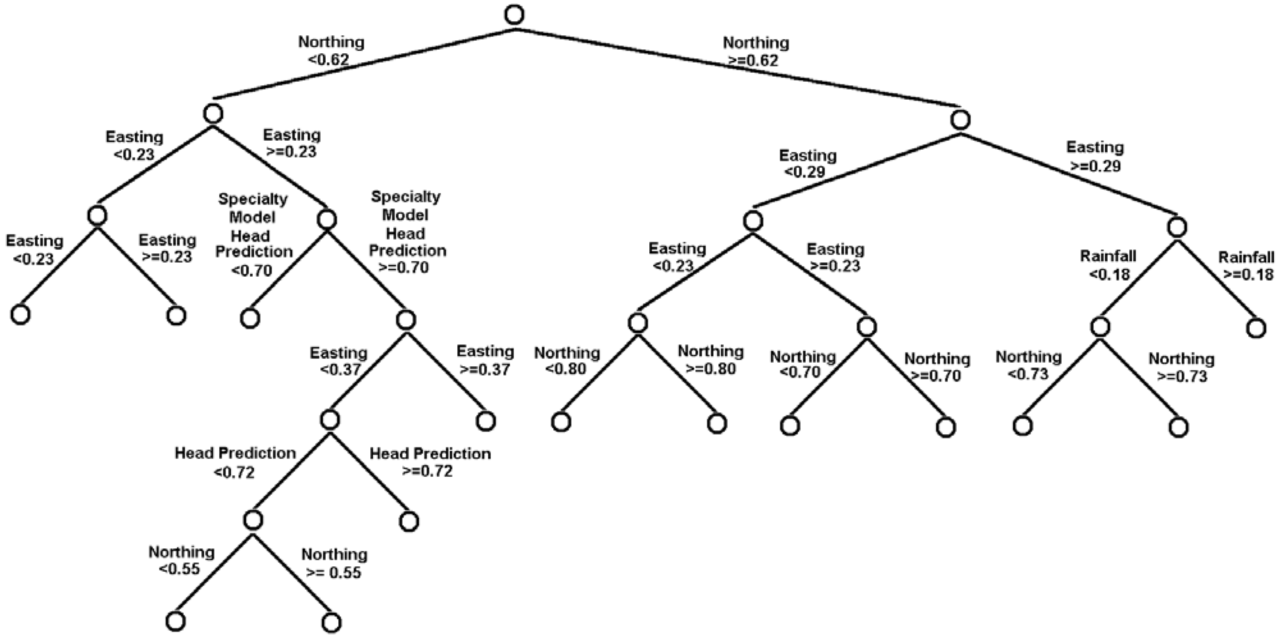

Figure 6. Graphical representation of a decision tree model used to predict hydraulic head levels.

decision input at that particular node to the threshold values shown, and the end of each path (leaf) is a hydraulic head prediction. This model branches extensively on the coordinates (Northing and Easting), a clear indication of strong spatial correlation. Other branches include the "Specialty Model Head Prediction" from a model trained on the continuous data and the "Rainfall" data.

Instance-Based Weighting. Instance-based weighting is a numerically simple yet powerful tool for estimating any unsampled point within a domain. The procedure used is a multidimensional weighted average

$$h_c^* = \left(\sum_{i=1}^{n} w_{c,i} h_{c,i} \right) \Bigg/ \left(\sum_{i=1}^{n} w_{c,i} \right) \qquad (1)$$

where h_c^* is the head prediction at location *c*, *n* is the number of nearest neighbors, $h_{c,i}$ is the head measurement at the neighboring point, *i,* and $w_{c,i}$ is the weighting factor. The weighting factors are calculated as

$$w_{c,i} = \frac{1}{d_{c,i}{}^{p}} \tag{2}$$

where p is the weighting factor and the distance $d_{c,i}$ is the Euclidean distance between the desired point and the ith closest measured point.

$$d_{c,i} = \left(\frac{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_m^2}{m} \right)^{\frac{1}{2}} \tag{3}$$

where m is the number of inputs (in this case study: northing, easting, rainfall or time) and $\Delta x_m = (x_{c,m} - \mathrm{x}_{i,m})$ is the difference between the values of the mth input at well location c and its nearest neighbor i [*Frink*, 1994]. The number of neighbors, n, and the weighting factor p, are selected by the learning machine to minimize the error on the training data. Inverse-distance weighting, which has been used extensively for interpolating spatial remediation data, is a special case of instance-based weighting where $p = 2$.

Artificial Neural Networks. "ANNs are distributed, adaptive, generally nonlinear learning machines built from many different processing elements (PEs)" [*Principe*, 1999]. Each PE is connected to other PEs or itself. The manner in which the PEs are connected define the type of ANN. The information passed between the PEs is scaled by adjustable weights. The PEs sum the contributions of their connections and produce an output by applying a non-linear static function to the sum. These outputs can be system outputs or outputs to other PEs. The ANN builds functions from the PEs with the number and shape of functions being determined by the type of ANN used. The input-output function is defined by the weights, and these weights are adjusted directly from the training examples without any assumptions about the distribution of the data. Thus, a central issue of ANN design is the training of the network. See *Principe* [1999] for more background on ANNs.

ANNs have been used extensively in water resources [*ASCE,* 2000], with several different types of ANNs. Feedforward backpropagation ANNs have been used most extensively, but recurrent and radial basis functions have been shown to be more accurate for some applications [*Coulibaly et al,* 2001a; *Coulibaly et al* 2001b]. For this initial study, we have used only the simple feedforward backpropagation ANN as a preliminary exploration into the mechanics of ANNs and a coarse comparison with the other methods readily available within our current software package. The exploration of the parameters that define neural networks in this study were limited to: number of layers, nodes per layer, and the number of weight updates.

2.3 Data to Knowledge (D2K) Environment

All of the framework development and machine learning takes place in the Data to Knowledge (D2K) "graphic data flow" environment – a Java-based data mining tool from the National Center for Supercomputing 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 "data flow" program - and the itinerary automatically builds a model that is used to make the prediction.

For each type of learning machine and model in D2K, a ten-fold cross validation is performed on the example set in each experiment and twenty experiments are performed (automatically in the D2K itinerary using random search within a user-defined range) to identify good model parameters (i.e. the ones that yield the lowest cross-validation error).

In the tenfold cross-validation the training set is divided into ten “testing” sets; for each of these ten sets, the learning machine is trained on the remaining data to predict the head values in the testing set. The cross-validation error is the difference between the predicted values and the actual values in each testing set. Then, in another itinerary, the entire training set and its optimized learning model parameters are fed into a model builder and that resulting model is used to make the necessary predictions.

3. EXPERIMENTS AND RESULTS

Using the methodology described in the previous section, two major experiments were performed to demonstrate the capabilities of the proposed hierarchical framework. The experiments and their results are described separately below.

3.1 Experiment 1: *Performance of Data-Driven Models for Predicting Water Levels at Routinely Sampled Wells*

The first experiment tests the capabilities of the data-driven machine learning models alone (decision trees, instance-based weighting, inverse distance weighting, and artificial neural networks) for predicting heads at routinely sampled wells in May 2001.

Figure 7 shows the hierarchical framework developed for this experiment as an itinerary within D2K. In the first level of the hierarchy, (labeled "Learning Machine #1" in Figure 7) a specialty model is built to predict the water levels at each of the twenty-two routinely sampled wells using the continuous and rainfall data. The spatial coordinates of each well (x & y), the rainfall (r), and the continuous head data are used to train the model to predict water levels. This water level prediction is then added to the quarterly data set as another input into the second ("expert") level of the hierarchy (labeled Learning Machine #2 in Figure 7), 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 in terms of number of examples. 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.

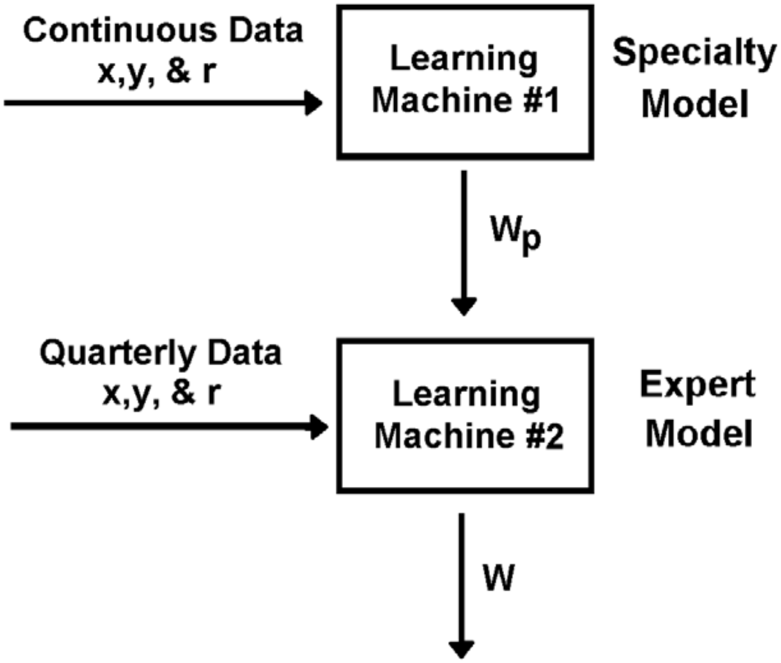

Figure 7. Example hierarchical framework diagram for experiment 1.

Table 2 shows the results of various combinations of the data sources and learning machine algorithms tested in Experiment 1. Each of the four methods (decision trees, instance-based weighting, inverse distance weighting, and artificial neural networks) were trained on five different sets of data: all historical quarterly measurements, the March 2001 measurements alone (the most recent quarterly data to the date of interest), the continuous head data for May, and two combined data sets – all the historical and continuous data combined and the March 2001 and continuous data combined. Further, for each of those data sets the learning machines were given different sets of inputs to determine which data had the most value. The inputs were either the spatial coordinates only (x & y), the spatial coordinates and the rainfall data (x,y, & r), the spatial coordinates with time, (x,y, & t) or all of the inputs (x, y, r, & t).

The framework performs extremely well at predicting the water levels in May 2001, as shown in Table 2. Note that the prediction errors shown are simply the spatially-averaged

Table 2. Data-driven model summary of results.

Predicting May Errors in ft		**Historical**	**Continuous May**	**Historical + Continuous May**	**March**	**March & Continuous**
Decision Trees	x & y	1.66	6.42	1.66	7.04	7.04
	x, y, & r	1.66	6.45	3.27	7.04	7.04
	x, y, & t	1.50	6.30	3.08	7.04	7.04
	x, y, r & t	1.50	6.30	3.08	7.04	7.04
Inverse Distance	x & y	1.80	8.61	1.74	5.21	5.19
	x, y, & r	1.72	8.68	1.83	5.21	5.19
	x, y, & t	4.79	8.87	4.29	5.21	6.65
	x, y, r & t	4.79	6.91	4.61	5.21	6.39
Instance Based	x & y	1.80	8.49	1.74	5.25	5.25
	x, y, & r	2.74	8.64	2.76	5.25	5.25
	x, y, & t	5.97	8.87	5.99	5.25	6.86
	x, y, r & t	5.91	8.87	6.81	5.25	5.99
Neural Networks	x & y	6.01	7.06	5.98	6.60	6.56
	x, y, & r	5.99	7.10	5.98	6.69	6.57
	x, y, & t	6.23	6.95	6.23	6.60	6.50
	x, y, r & t	6.24	6.95	6.24	6.42	6.50

differences between the predicted and actual values in feet at each well. The prediction errors remain low relative to the ranges of the values of the head levels in the aquifer – from about 620 to 685 ft above mean sea level. 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 from March 2001. In particular, inverse-distance weighting, which is widely used for spatial interpolation of data in a single time period, shows substantial errors when only the March data were used for prediction. However, when all of the historical data are used, Table 2 shows that inverse distance weighting has reasonable performance, almost as good as the best method.

The rainfall information improves the predictions from the historical quarterly data only in the case of inverse distance weighting, while the addition of time improves the predictions for

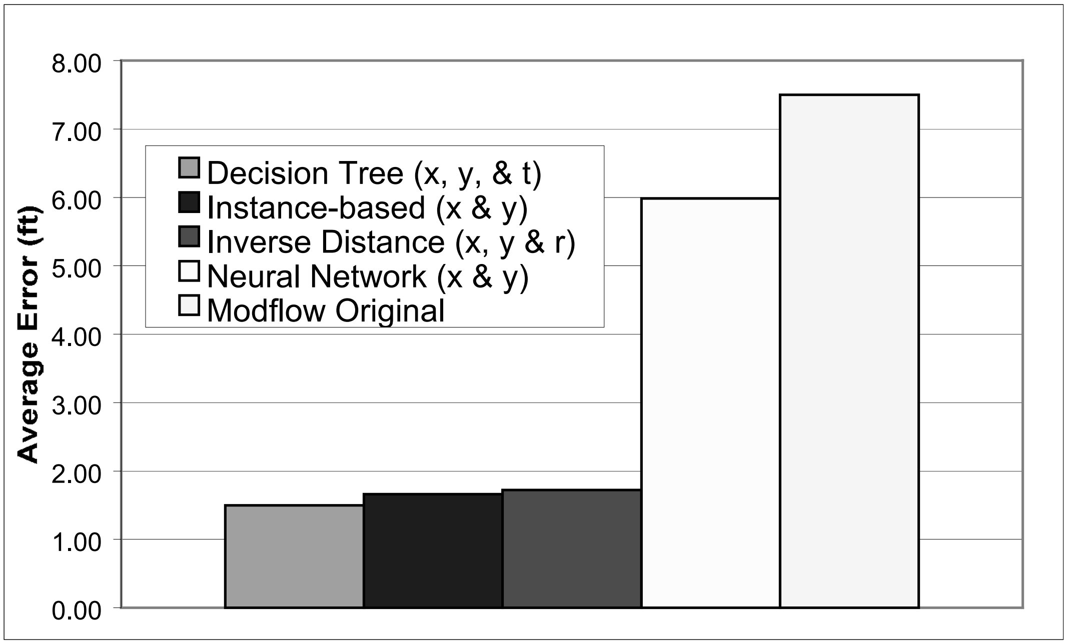

Figure 8. Summary of the best results by method.

the decision trees. Figure 8 shows a summary comparison of the trials with the lowest spatially-averaged errors for each of the methods. Overall, the complete historical quarterly dataset proved to be the most valuable data source, most likely due to its good spatial coverage. Decision trees provided the most accurate predictions, but all of the methods outperformed the existing Modflow model. (Detailed histogram comparisons of the individual methods can be found in the Appendix.)

The addition of the continuous data to the quarterly measurements in the combined datasets provides improvement only in the case where inverse distance is used with spatial coordinates only. This indicates that the continuous data are not as useful as the quarterly data, probably because of their poor spatial coverage (see Figure 1). Figure 9 shows the distribution of errors at the individual wells for the best historical model, which was the decision tree model using the spatial coordinates and time as inputs (x, y, & t). Most of the wells have low prediction errors, less than one foot. Note the poor performance in the center of the graph where there is a

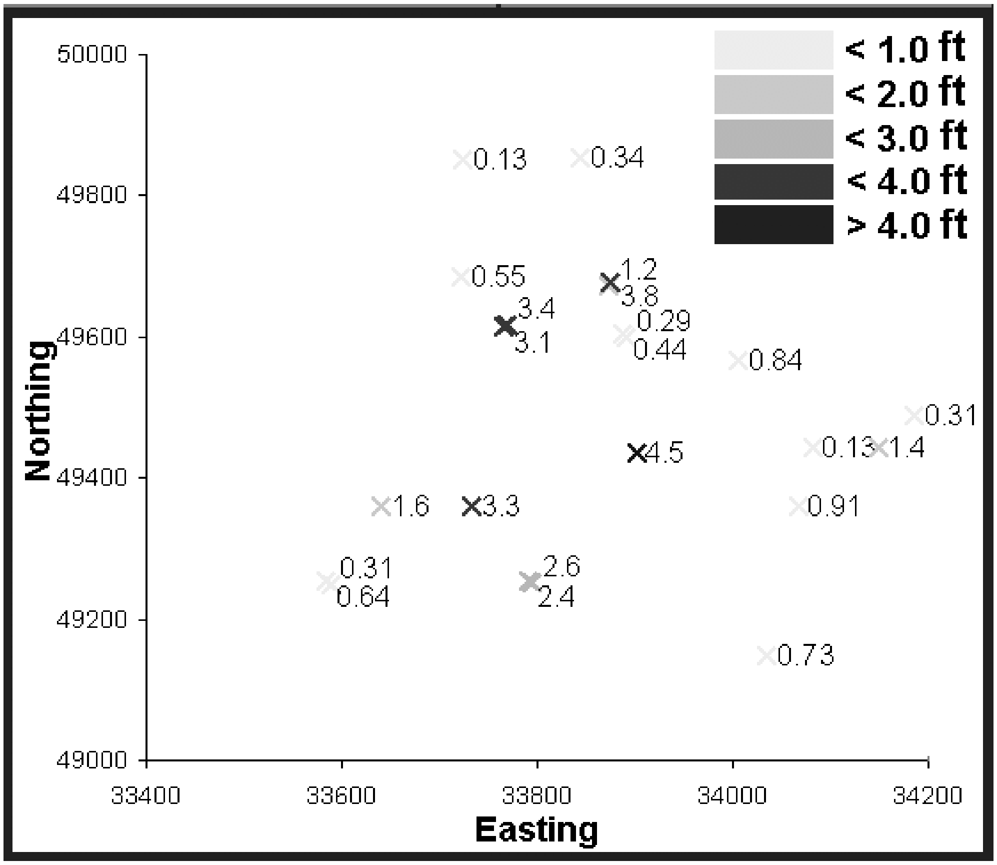

Figure 9. Error of the best data-driven historical model prediction for each well in May 2001**.**

"lone" well. Whether there is some interesting phenomena happening in the subsurface that is unique to this area or the spatial coverage is simply insufficient, it is clear that this area is a good candidate for further sampling.

3.2 Experiment 2: *Performance of combined data-driven and physics-based models for predicting water levels at routinely and non-routinely sampled wells*

Numerical modeling of contaminated groundwater sites is a common step in remediation design, particularly for flows. These models are built using hydrogeologic expertise and often contain intricate and specific details about an individual site. As the groundwater remediation field matures, it is becoming clear that it is not cost effective to refit these models whenever new data become available. While some automated updating methods such as Kalman filters [*Gelb*, 1974; *Graham and Tankersly*, 1993; *Graham*, 2001] are available, such methods are not widely used because of their complexity and computational burden [*Eppstein and Dougherty*, 1996]. In this experiment, we test the potential for machine learning methods to automatically update existing numerical models by using the Modflow model as a specialty model in the hierarchical framework proposed previously.

As shown in Figure 10, the errors in the Modflow prediction alone are much worse than the data-driven model and there is not any clear pattern of systematic error across the site. However, the Modflow model does perform well in the center of the routinely sampled well area, a location

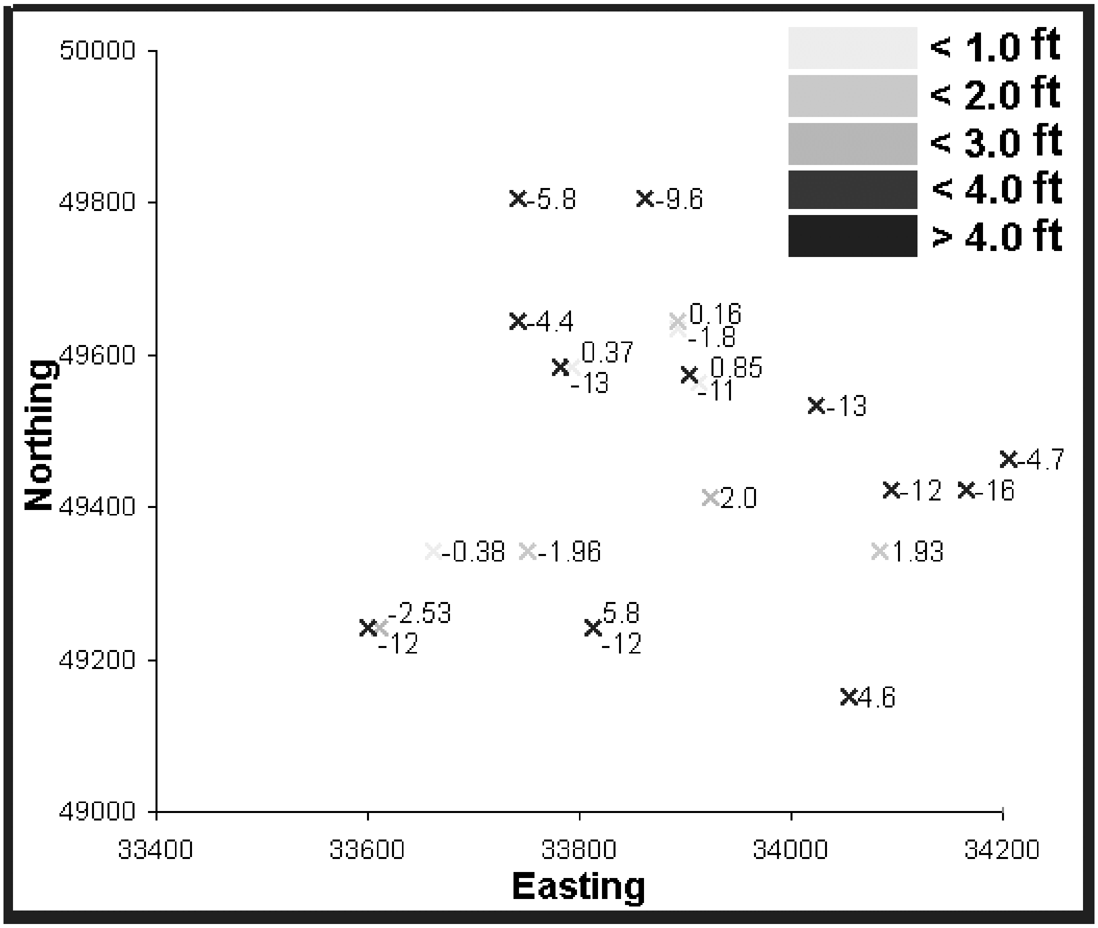

Figure 10. Modflow updated individual well errors in predicting the May 2001 data

where the data-driven model did not perform well. Other advantages of the Modflow model are the larger area of coverage – such as in the forest preserve where it is more difficult to collect data – and more measurements over a longer time period.

Figure 11 shows the structure of the combined model tested, with the Modflow predictions at the 22 routinely sampled wells fed into the expert model with the quarterly data. For comparison, the input data (the spatial coordinates and time or x,y, & t) and the decision tree learning algorithm used in the combined model are the same as those from the best data-driven model that was trained on the historical quarterly data.

The errors in the best combined models are slightly worse than the best data-driven model alone for predicting head levels in the routinely sampled wells in May 2001 (see Figure 12) with an average error of 1.97 ft. The errors differ from either the data-driven model or Modflow alone, so the model is learning a compromise between the two sources. Figure 13 shows the changes in the error predictions for each individual well from the separate data-driven

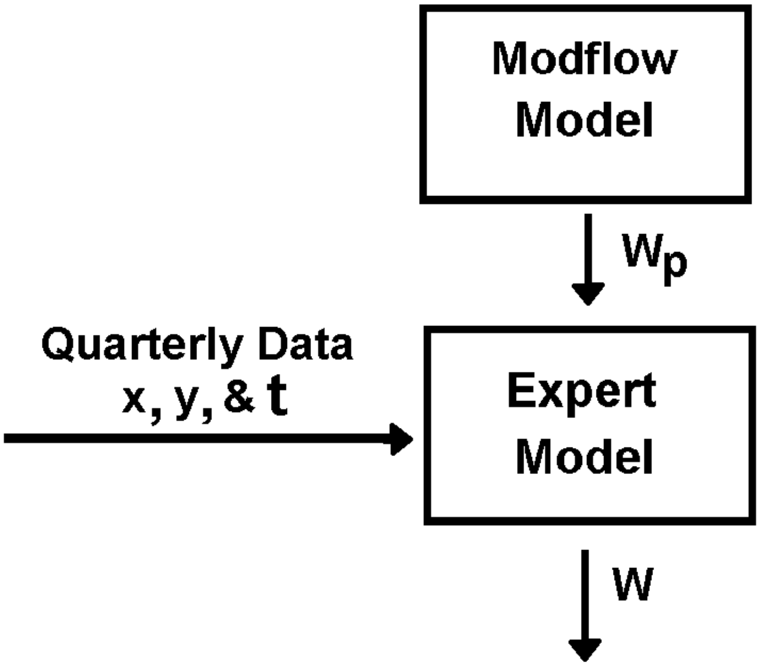

Figure 11. Hierarchical model framework for experiment 2.

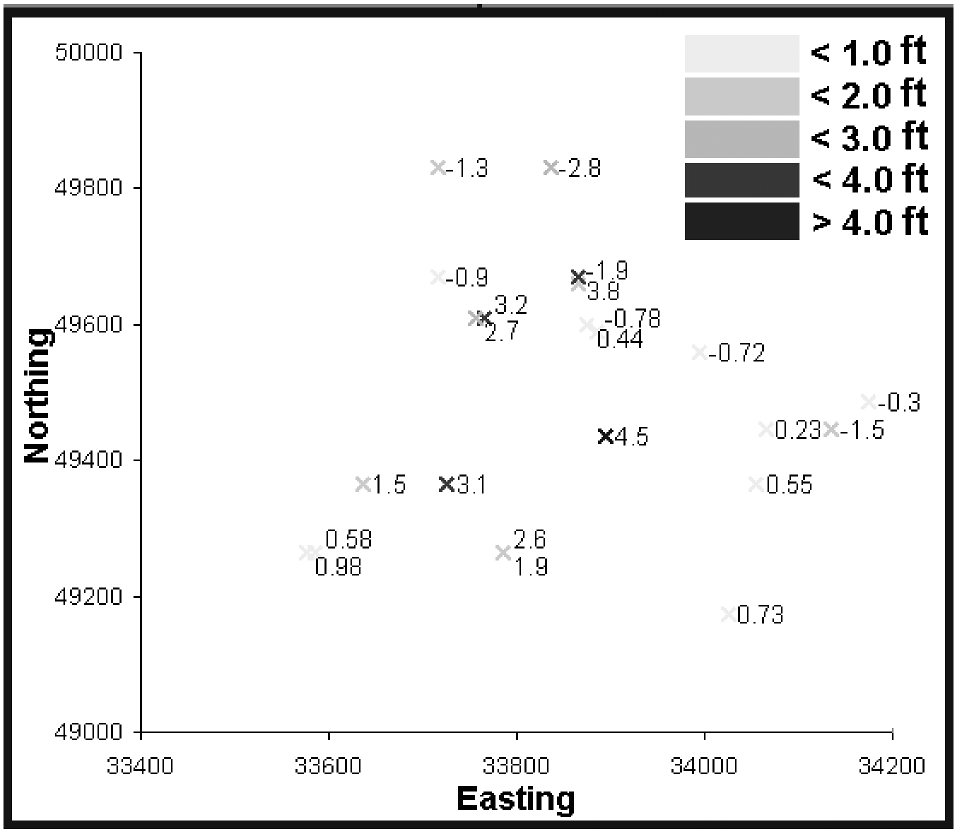

Figure 12. Individual well errors for the prediction of the May 2001 data using the combined Modflow and data-driven model.

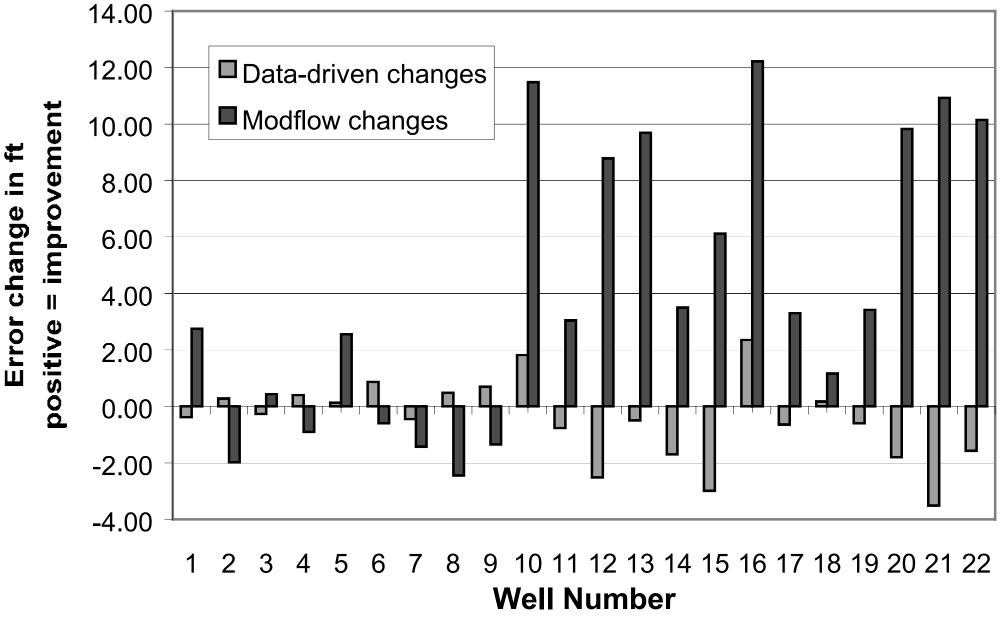

Figure 13. Well by well changes in errors from combining models.

model (historical data only) and the physics-based model (Modflow) to the new combined model, where a positive value indicates an improvement in the prediction error at that well. For example, for the first well shown, the prediction from the combined model is 2.5 feet better than the Modflow prediction and .43 feet worse than the data-driven model alone.

Taking a broader look at Figure 13, the bias in the Modflow model is greatly reduced in the combined model. Comparing Figures 11 and 12, the combined model has somewhat higher and lower errors in different areas than the data-driven model alone, with improvements from the data-driven model generally localized in the center and slightly increased errors in the wells around the edges of the domain. Unfortunately, the combined model did not reduce the error at the "lone" well in the center relative to the data driven-model, despite the Modflow model's improved prediction at this location (see Figure 11). This result may indicate problems from the relatively small historical quarterly dataset (only 147 data points, of which only seven are at that well) or a need for a more sophisticated hierarchical model (e. g, one that allows the user to provide weights on data sources in different areas).

Next, the combined model was tested at the non-routinely sampled wells (see Figure 4) in March 2001 (the only period in which most of those were sampled) to assess its ability to extrapolate beyond the training data. The results for the non-routinely sampled wells were expected to be worse due to lack of training data from that area and the differences in the values of the measurements relative to the routinely sampled area, especially within the forest preserve. (The forest preserve head levels are much lower than the other measurements.) The results, shown in Figure 14, confirmed that expectation. The best data-driven model using all of the historical data has the best performance in the routinely sampled area, but it has the worst

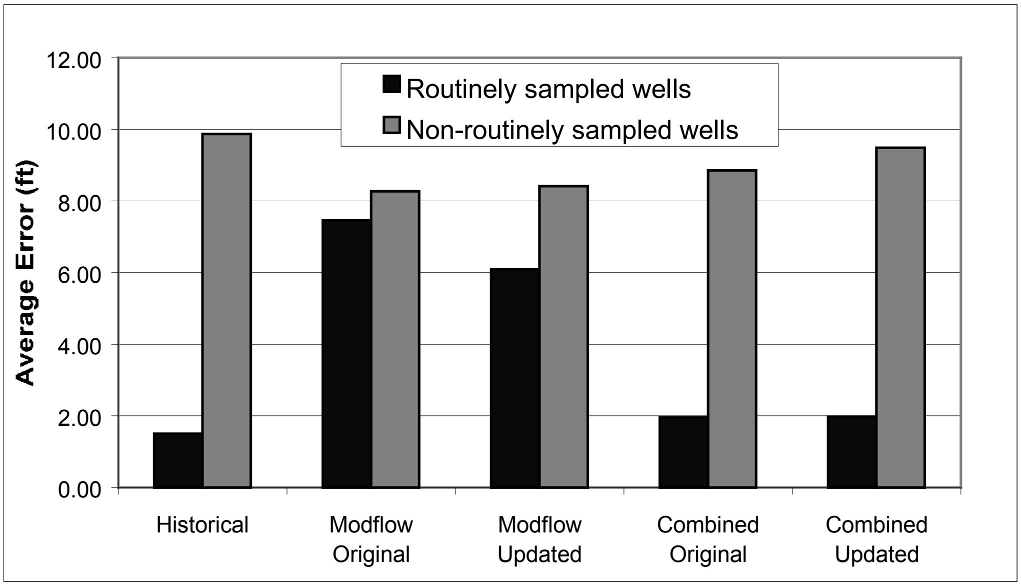

Figure 14. Comparison of performance of the combined models on the routinely and non-routinely sampled wells.

performance in the non-routinely sampled area. The opposite is true for the original Modflow model – it is best for the non-routinely sampled area and worst for the routinely sampled area. Combining the models makes use of the strengths of both, with the larger-scale physics of the site captured in the original Modflow model and the smaller-scale local trends in the historical dataset. The result is a good compromise between the two models, with only slightly higher errors in the routinely sampled wells and lower errors in the non-routinely sampled wells. These results show that the use of physics-based models within the hierarchical framework will be important for more accurate extrapolation beyond the historical data. However, the errors in the non-routinely sampled wells are still relatively high, even for the best model, indicating a need for more data from those wells to improve predictions.

Note that the updated Modflow results shown in Figure 14 might indicate overfitting – that is the model is so well fit (relatively) to the May 2001 data that it has more difficulty predicting other time periods (since the non-routinely sampled wells were sampled in March

2001). Essentially, the simulation model has been adjusted to provide the desired results at one instance at the expense of the accuracy of the predictions of other instances. These results show that the manual updating of the Modflow model was unnecessary and in fact detrimental to its general predictive ability. Using the machine learning model to update the Modflow model and capture knowledge from the historical data appears to be a more promising approach.

4. CONCLUSIONS

This paper demonstrates that knowledge integration, which combines different datasets and models in a machine learning framework, holds considerable promise for improving groundwater monitoring and management. The proposed hierarchical framework has successfully predicted head levels in a routine sampling event at the 317/319 Area at the ANL-E site with substantially greater accuracy than an existing Modflow simulation model. Most monitoring data analyses include only the most recently collected data, but our results show that the entire historical quarterly sampling record is the most valuable source of information at this site.

Decision trees were shown to be useful tools for quick evaluation and development of different hierarchical frameworks. They consistently provided the most accurate predictions for this problem and the resulting models are intuitive and easy to understand (see example in Figure 6). The training time for decision trees is very short and their simplicity makes them ideal for rapid, on-site adaptive field sampling. On-site adaptive sampling would provide a means for new data to be analyzed quickly and further data needs to be identified in the field. New data could be used to update the model online and identify predicted areas where loss of containment may be occurring, for example. These areas could then be targeted for further sampling. Given decision trees' capabilities for incorporating diverse data sources, they could also be used to combine new surrogate or indicator data (such as rainfall) with existing historical data to identify whether the new data indicate potential problems that would warrant collection of more traditional samples. Future research will investigate this possibility.

Combining non-traditional data sets and optimizing the value of current data as demonstrated in this paper should 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.

Finally, combining data-driven and physics-based models can improve accuracy of physics-based models in areas with substantial data as well as extend spatial coverage of data-driven models. When complete, such combined hierarchies could allow off-the-shelf, historical physics based models to be updated automatically, creating “living” models that are expandable and adaptable as new data, analysis, and modeling techniques become available.

While the results of this preliminary study are promising, much additional research is needed. More sophisticated hierarchical models should be developed that can further exploit knowledge about the site. For example, the continuous data could prove more valuable if used for training only in certain areas of the site. While the spatial coordinates used for training the models do allow for spatial zones to be established automatically, the quantity of spatially-distributed data available in this study may have hampered the effective use of this capability. Another example of a more sophisticated approach would be to automatically weight different data so that the influence the data has on the model would be proportional to its real world significance.

Further research should also be conducted to expand the study of machine learning methods and algorithms – particularly in the area of artificial neural networks. Recurrent and radial basis function neural networks are obvious choices, since they have performed well on other water resources problems [*ASCE*, 2000]. Other machine learning methods exist that could also be tested, such as support vector machines [*Burges*, 1998].

Another analysis that would be useful is identifying how long we could collect only surrogate data (such as rainfall and continuous water level measurements) and not collect quarterly samples before our predictions using the historical quarterly data and new surrogate data will suffer. Although the results are not presented here, the August 2001 and November 2001 prediction errors using the same historical quarterly dataset from November 1999 to March 2001 were similar to the May errors, indicating that the historical data are useful for at least one year without additional sampling. This type of analysis would help identify which data sources are the most essential and aid in the development of sampling plans. Genetic algorithms could then be used to optimize sampling plans for all data sources, in an approach similar to *Reed* [2002] and *Reed et al* [2001, 2000].

The full capabilities of D2K should also be explored to automate more processes. The ultimate goal should be to embed physics-based models in D2K with automatic updating of the hierarchical model parameters as new data become available. In addition to fast updates, the framework could also provide indicators to suggest when a new conceptual model may be necessary. If the new data being collected exceeds a range of expected values by a predetermined amount or when a formerly accurate model produces errors that are consistently above a threshold of tolerance, the hierarchy could trigger a message to review the new data and the current model.

REFERENCES

ASCE Task Subcommittee on Application of Artificial Neural Networks in Hydrology, Chair Govindaraju, R.S, Artificial Neural Networks in Hydrology. II: Hydrologic Applications, *Journal of Hydrologic Engineering*, 5(2) 125-156, 2000.

Buchanan, B.G., C. R., Johnson, T. M. Mitchell, and R.G. Smith, Models of learning systems, *Encyclopedia of Computer Science and Technology*, J. Belzer, ed. 1978.

Burges, C.J.C, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery* 2, 121-167, 1998.

Coulibaly. P., Anctil, F., Aravena, R., Bobée, B., Artificial neural network modeling of water table depth flucuations, *Water Resources Research*, 37(4), 885-896, 2001a.

Coulibaly. P., Anctil, F., Bobée, B., Multivariate Reservoir Inflow Forecasting Using Temporal Neural Networks, *Journal of Hydrologic Engineering* 6(5), 367-376, 2001b.

Eppstein, M.J. and D.E. Dougherty, Simultaneous estimation of transmissivity values and zonation, *Water Resources Research*, 32(11) 3321-3336, 1996.

Frink, N.T., "Recent Progress Toward a Three-Dimensional Unstructured Navier-Stokes Flow Solver," AIAA Paper 94-0061, 1994.

Gelb, A., Applied Optimal Estimation, MIT Press, Cambridge MA, 1974.

Graham, W., and Tankersly, C., Forecasting Piezometric Head Levels in the Floridian Aquifer: A Kalman Filtering Approach, *Water Resources Research*, 29(11), 3791-3800, 1993.

Graham, W., Optimal estimation and prediction of hydrogeochemical parameters using Kalman filters, Chap. 9 in Stochastic Methods in Subsurface Contaminant Hydrology, edited by Rao Govindaraju*, ASCE Press*, Reston, VA, 2001, in press.

Matheus, C. J. 1990. "Feature Construction: Analytical Framework and an Application to Decision Trees." Ph.D. Thesis. University of Illinois, Urbana.

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..

Nilsson, N.J., Eye on the prize, *AI Magazine* 16(2), 9-17, 1995.

Quinlan, J. R., Induction of decision trees. *Machine Learning*, 1:81-106, 1986
Principe, Jose C. and Euliano, Neil R. and Lefebvre, W. Curt. *Neural and Adaptive Systems: Fundamentals Through Simulations*. pg 101. John Wiley and Sons, New York, 1999.

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.

Reed, P. and B. S. Minsker, "Striking the Balance: Long Term Groundwater Monitoring Design for Multiple, Conflicting Objectives." *Journal of Water Resources and Planning Management*, submitted, 2002.

Reed, P. M., B. S. Minsker, and D. E. Goldberg, "A multiobjective approach to cost effective long-term groundwater monitoring using an elitist nondominated sorted genetic algorithm with historical data." Invited paper, *Journal of Hydroinformatics*, 3, 71-89, 2001.

Reed, P. M., B. S. Minsker, and A. J. Valocchi. "Cost effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation." *Water Resources Research*, 36(12), 3731-3741, 2000.

Rendell, L.A., R.M. Seshu, and D.K. Tcheng, More robust concept learning using dynamically variable bias, *Proc. of the Fourth International Workshop on Machine Learning*, 66-78, 1987.

Principe, Jose C. and Euliano, Neil R. and Lefebvre, W. Curt. *Neural and Adaptive Systems: Fundamentals Through Simulations*. pg 101. John Wiley and Sons, New York, 1999.

UGAI Lectures, Temple University, Department of Computer Science, *Building Classification Models: ID3 and C4.5*, http://yoda.cis.temple.edu:8080/UGAIWWW/lectures95/learn/C45/#8, 1995.

U.S. Department of Energy, Office of Environmental Management, January 2001, "Report to Congress on Long-term Stewardship", January 2001, Rep No DOE/EM-0653.

APPENDIX

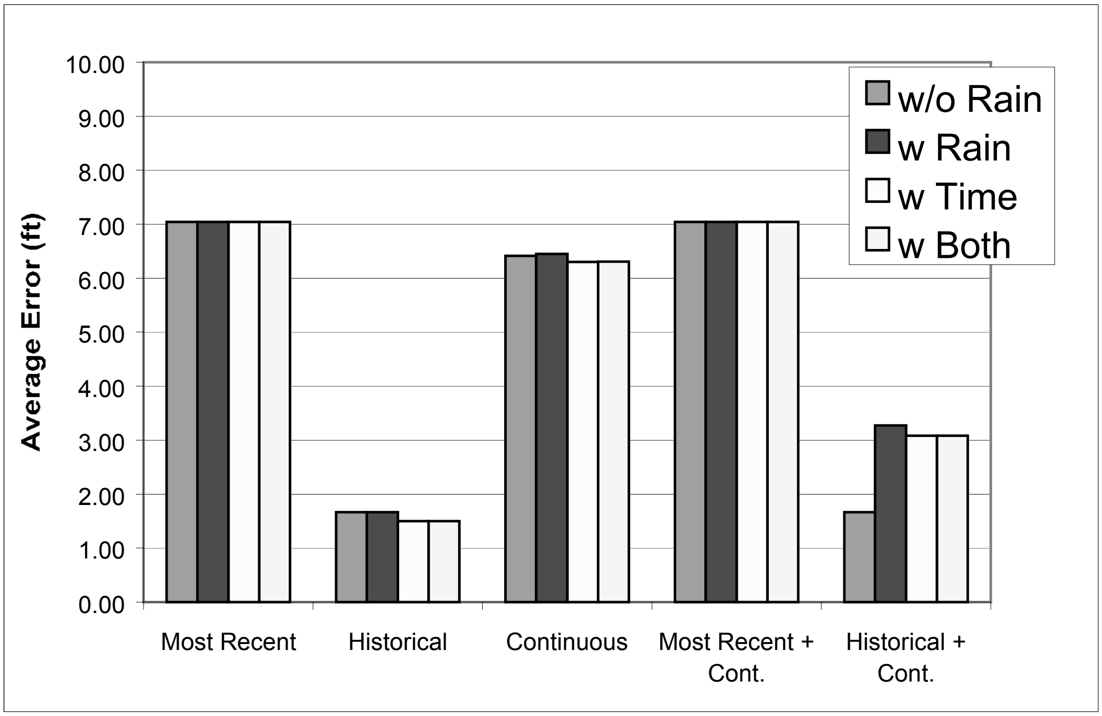

Figure 15. Decision tree results by input.

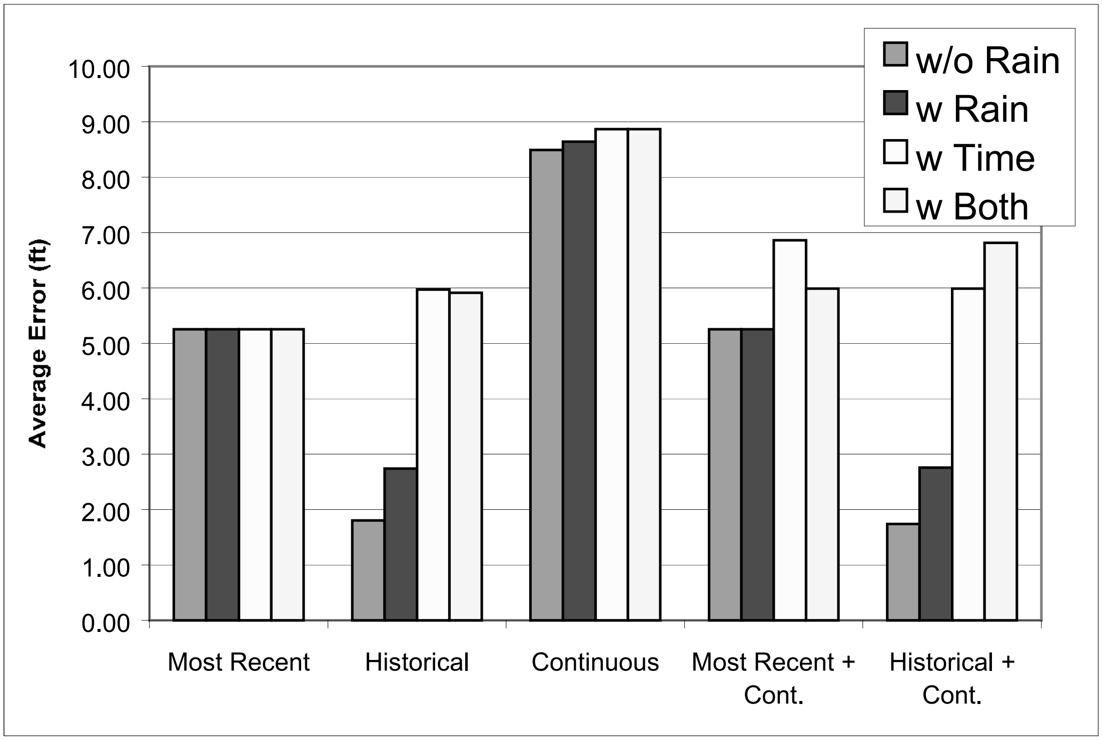

Figure 16. Instance-based results by input.

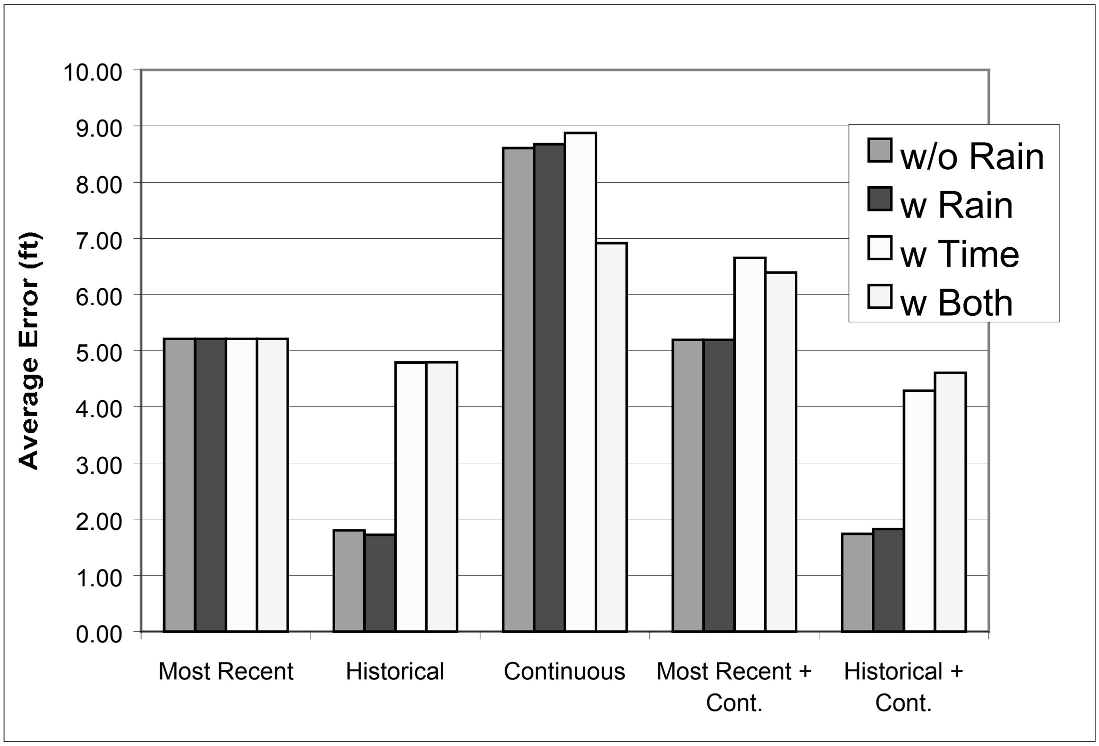

Figure 17. Inverse distance weighting results by input.

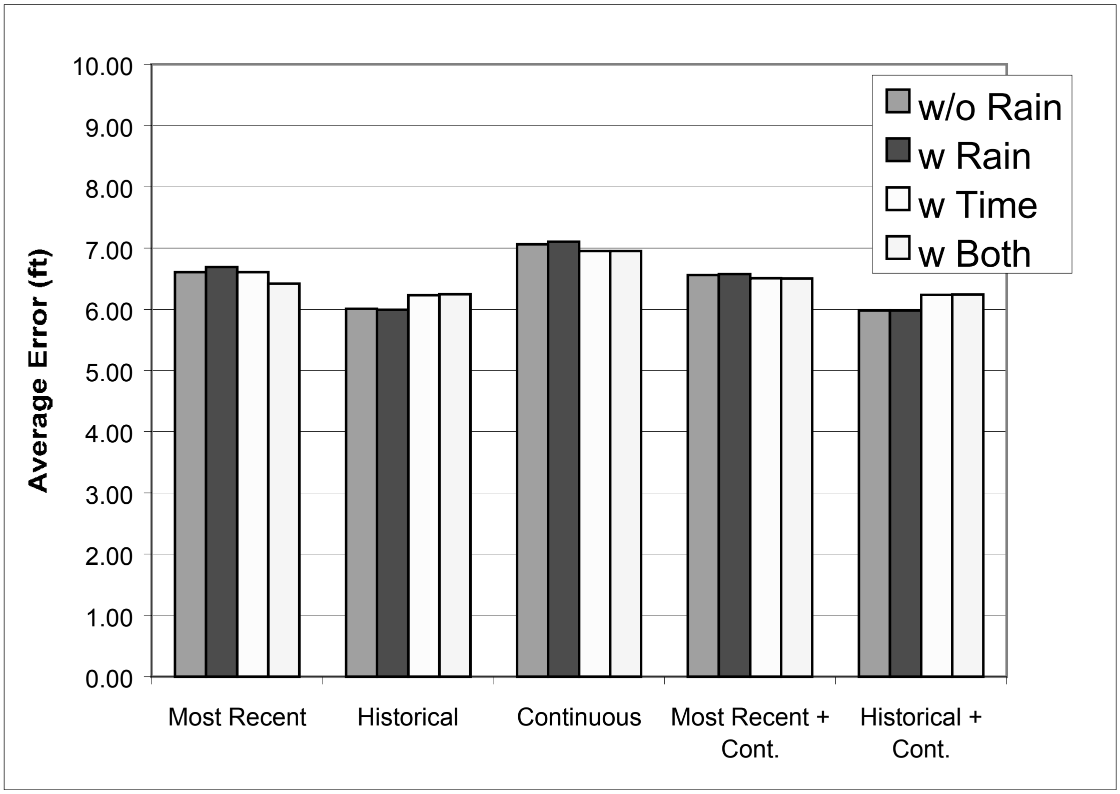

Figure 18. Artificial neural network results by input.