Interactive Multi-Objective Inverse Groundwater Modeling – Formulation and Addressing User Fatigue

Abhishek Singh[1], Barbara Minsker[2]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana Champaign, Urbana, USA, PH (217) 333-6979, email: asingh8@uiuc.edu
[2]Department of Civil and Environmental Engineering, University of Illinois at Urbana Champaign, Urbana, USA, PH (217) 333-9017, email: minsker@uiuc.edu

Abstract

This paper builds on work done on using interactive multi-objective genetic algorithms (IMOGA) to solve the groundwater inverse problem (*Singh & Minsker,* 2005) by searching for optimal hydraulic conductivity fields conditioned on field measurements of hydraulic heads and conductivities. The biggest challenge faced when using such interactive systems is that of user fatigue because the user is expected to evaluate many solutions during the search process. This paper discusses a two-step approach to reduce user fatigue. First the user is shown only a fraction of the total population in every generation. To ensure minimum redundancy during evaluation, the solutions are clustered using unsupervised clustering and the expert is shown unique samples from distinct clusters. Next the unranked solutions are ranked using a surrogate model that 'learns' from the user preferences. This is implemented using a supervised classification algorithm to cluster the solutions based on the 2-D images of hydraulic conductivity. We test 'content-based' and 'spectral' algorithms for the clustering and classification as these have been shown to be similar to how humans process images. The work on applying and testing these algorithms is on-going and this paper discusses some preliminary results. Complete results will be shown at the EWRI conference.

Introduction

Many studies such as *Yakowitz and Duckstein* (1980), *Carrera and Neuman* (1986), *Sun* (1995), and *Zimmerman et al.* (1998) have shown that the inverse problem for hydraulic conductivity is ill posed, i.e. it has non-unique solutions, which are unstable for small variations in the observation data. To deal with these problems modelers have to reduce the dimensionality of the parameter space (referred to as parameterization - see *McLaughlin & Townley*, 1996 and *Carrera et al*, 2005) and also incorporate as much "secondary" information and "prior knowledge" as possible (*Yeh*, 1986; and *McLaughlin & Townley*, 1996). Expert judgment and knowledge are critical in both these aspects. The parameterization is decided by the modeler and the inversion results are highly dependent on the type and level of parameterization chosen. It is thus important for the modeler to explore the relationship between parameterization and model accuracy during the inversion process. Moreover, the expert also has a lot of 'soft' or qualitative knowledge about the hydrogeology of the site which is often not incorporated in automated calibration methodologies. This knowledge is useful secondary information that can improve the ill-posedness of the inverse problem and may lead to more reasonable solutions. To deal with both these aspects of user knowledge, this research is exploring a novel methodology for inverse modeling using interactive multi-objective genetic algorithms (IMOGAs).

IMOGAs are an extension of the interactive genetic algorithm (*Takagi*, 2001). Interactive genetic algorithms, as the name suggests, are population based evolutionary optimization algorithms that use evaluations and feedback from users as objectives and constraints in the optimization and search process. In our research we pose inverse modeling as a multi-objective problem, with the numerical objectives being prediction accuracy, regularization level, and qualitative expert ranking of the solutions found during the search process. The basic IMOGA framework has been presented at a previous EWRI conference (*Singh & Minsker,* 2005). This framework requires the evaluation of all objectives for all individuals in the IMOGA population for every generation. While this may be easy for the numerical objectives, getting feedback from the expert for hundreds of solutions is clearly not practical. This paper thus explores several approaches to reducing user fatigue while still maintaining the solution quality of the IMOGA.

This paper is divided into four parts. First we give a brief overview of a hypothetical case study and the formulation of the multiple objectives for that problem. We next discuss the interactive multi-objective methodology used to solve the inverse problem for this case. This is followed by a discussion of the user fatigue problem and some methods to deal with it. Finally some preliminary results are discussed along with future directions for research. As this work is on-going this paper does not contain all the results for the proposed methodologies, which will be shown at the conference.

Synthetic Groundwater Aquifer Case Study

The case study for this work is based on the aquifer discussed by *Freyberg* (1988) and shown in Figure 1. This test case was chosen because previous studies (*Freyberg,* 1988) have shown that the inversion of conductivity based on head observations is an ill-posed problem with little or no correlation either between calibration error and root mean square error (RMSE) of hydraulic conductivity, or accuracy of post-calibration predictions. The aquifer is considered to be the ground truth for the parameterization, and all observations are taken from this aquifer.

The hypothetical aquifer is a 2-D, shallow, unconfined, isotropic, and vertically homogenous aquifer with steady state flow. The aquifer is discretized with 20 columns and 40 rows with each cell measuring 250 meters. There are no flow boundaries along the north, east, and west boundaries, and a constant head boundary along the south. The hydraulic conductivities as well as the bottom elevations are highest in the northwest, with a decreasing trend to the southeast. There are a total of 22 observations for head and conductivity. The hydraulic heads in this hypothetical aquifer were then solved using MODFLOW (*McDonald and Harbaugh*, 1988).

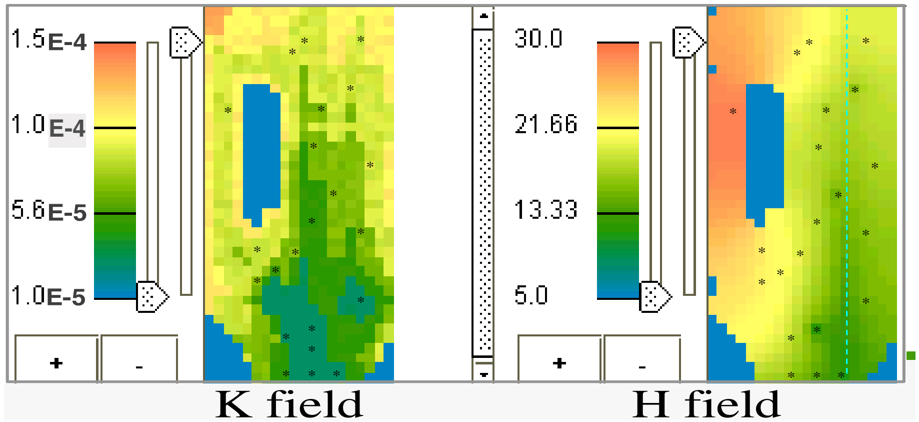

Figure 1. Conductivities (K in m^2/sec) and heads (H in m) in the original aquifer. The black dots are observation wells, and the areas in blue are the boundaries and outcrop.

Once the hydraulic heads have been calculated for the 'true case', the values at the observation points can be used as calibration targets for the inverse method. For this project only the hydraulic heads were considered unknown, while all other parameters, boundary conditions, and observations were assumed to be perfectly known. The hydraulic heads and conductivity measurements at the observation points were taken as the calibration data for the inverse model.

For this work we use non-linear inversion based on the concept of 'pilot-points' (*de Marsily*, 1984 *Rama Rao et al.*, 1995 *Doherty*, 2003). Pilot points are spatial control points spread across the modeling domain where the parameters of interest need to be estimated. The values at the pilot points are then interpolated (in space or time) to give the continuous field that one is interested in. For our problem the pilot points represent point values of conductivities (K) that are interpolated (using kriging) to give a particular conductivity field. A total of 30 pilot point locations were chosen so that they had good coverage over the entire field. The values at these 30 pilot points are decision variables for the IMOGA. The 30 pilot point values and the 22 K measurements were kriged using KT3D (*Deutsch and Journel,* 1998) to give the conductivity field to be used in the MODFLOW flow model. The optimization problem is to find the best values of these pilot points so that the groundwater model using the interpolated field has minimum prediction error with respect to measurements of hydraulic heads. The objective function is given by the following equation:

$$\underset{PP \in R^p}{Min} \quad H_{err} = \frac{1}{n}\sum_{i=1}^{n}\left[GW\left(K(PP), H_0, Q\right)_i - h_i^{obs}\right]^2 \qquad (1)$$

Where *PP* is a vector of *p* pilot points, H_{err} is the head prediction error, h_i^{obs} is the head measurement from the field at location *i*, *GW()* is the groundwater model that predicts the head values at the measurement locations *i* using the Kriged field *K(PP)* from the pilot points, the initial conditions H_0, and the source and sink terms *Q*.

In addition to measurements of hydraulic heads there are also measurements of conductivity at some locations. By including these K measurements in kriging (which is an exact interpolator) we force each conductivity field to match the measurements exactly. It has been shown (*Doherty*, 2003) that using a large number of pilot points leads to instability and non-uniqueness in the search process. To deal with this problem *Doherty* (2003) proposes an additional 'regularization' objective. The regularization shown below, is formulated to minimize the difference of the pilot points and the values at those locations from kriging with the observations alone:

$$\underset{PP \in R^p}{Min} \quad \Phi_R = (PP - K_0)^T C_P^{-1} (PP - K_0) \qquad (2)$$

Where *PP* is a vector with all the pilot point values, K_0 is a vector with 'prior' or expected values at the pilot point locations, and C_P is the spatial covariance matrix for pilot points. Since the pilot points are not co-located with the K measurements the K measurements are interpolated using kriging at the pilot point locations to give K_0. The covariance matrix can be either be estimated from the conductivity measurements or given based on knowledge about the true field.

A critical decision when using pilot points (or any geostatistical model) is the choice of the variogram and kriging parameters used for spatial interpolation. It has been demonstrated before (*Singh & Minsker, 2005*) that including the kriging and variogram parameters in the search process leads to K fields that have better prediction error. Moreover, these parameters should also be evaluated by the expert during the interactive optimization process. Thus in addition to the values of the pilot points, the kriging and variogram Parameters (kriging window size, variogram type, sill, and range) are also included as decision variables, giving a total of 44 decision variables. To ensure that the variogram used for kriging is close to the empirical variogram (constructed from the K measurements from the field) an additional objective of variogram fit is introduced:

$$\underset{\gamma}{Min} \quad \Phi_V = \frac{(\gamma - \gamma_0)^T W (\gamma - \gamma_0)}{n_{lags}} \qquad (3)$$

Where γ is the model semi-variogram value for a particular lag, γ_o is the observed semi-variogram value for the lag, n_{lags} are the number of measured lags, W is the weight for the lag measurement – taken as the inverse of the number of pairs available while calculating γ_o for that lag.

Interactive Multi-Objective Inversion

We use the three objectives discussed before within a multi-objective framework that identifies solutions with the optimum trade-off between the three objectives of model accuracy, regularization, and variogram fit. These solutions form what is called the 'Pareto front,' where every solution is such that one objective can not be improved without simultaneously worsening the other.

The IMOGA framework is shown in figure 2. Under an interactive approach, the GA performs the usual operations of selection, crossover, and mutation (*Goldberg*, 1989), but the user evaluates the suitability ('fitness') of candidate solutions, enabling objectives that cannot be quantified to be included in the search process. The IMOGA uses the elitist non-dominated sorting genetic algorithm (NSGA-II) (*Deb et al, 2000*), which has been shown to be a reliable and efficient algorithm for solving multi-objective problems. The quantitative objectives are calculated for every solution in the GA as discussed above. These solutions then need to be shown to the expert for subjective evaluation. At the moment the conductivity and predicted head fields are shown to the user for all of the individuals. The user then gives them a rank from 1 (best) to 5 (worst). The challenge at this step is that the user can not be expected to evaluate every solution in the GA population, especially if the population size is large. We thus need to select some representative solutions that need to be ranked by the user. Once the user has ranked these solutions the unranked solutions also need to be given an objective that reflects user preference. This requires a model to be built of user preference. This is accomplished by archiving the solutions ranked by the user and using them as training data to build a decision model to evaluate the unranked solutions. The last two steps are critical components in reducing user fatigue and are discussed in more detail in subsequent sections.

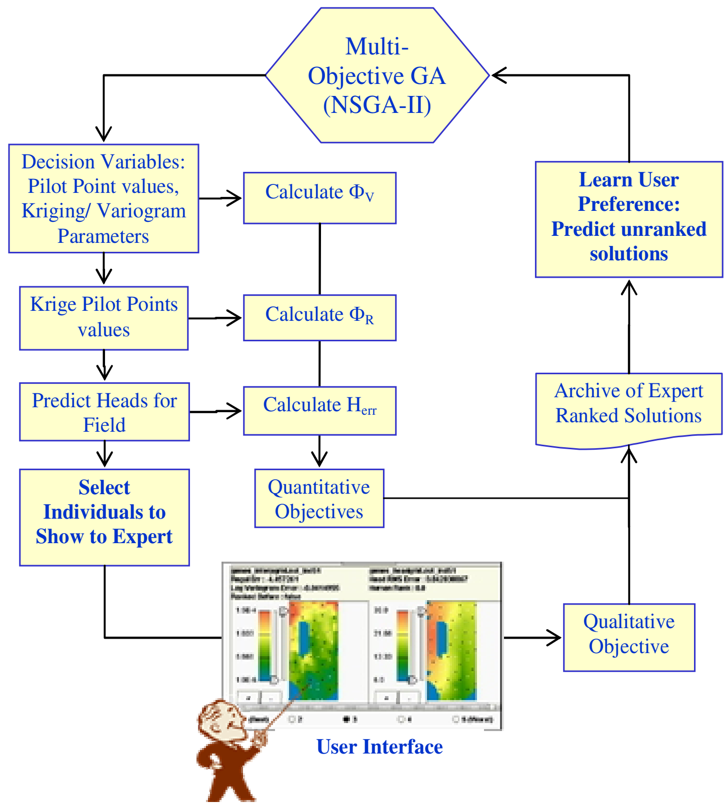

Figure 2: IMOGA Framework

Reducing User Fatigue

As discussed before there are two stages to reducing the number of solutions that need to be ranked by the user - the appropriate selection of representative solutions from the current population and the extrapolation of user rank for the unranked solutions. Given a population of n individuals, we want to select k (k<n) individuals that have best coverage over the entire set. By best coverage we mean that the k individuals should capture the maximum variability (information) of the n individuals. This is akin to dividing the n individuals into k distinct clusters such that solutions in each cluster are most similar to each other and most dissimilar to members of other clusters. Picking one representative from each cluster would result

in the selection of individuals that are most distinct from each other and are representative of the entire population. This can be accomplished using existing 'unsupervised' (meaning that they do not have any training data and are based only on the variability of the dataset) clustering algorithms like K-means clustering. K-means clustering is an iterative algorithm that starts off with a random initialization of the centroids for the *k* clusters. Each individual is then categorized with one of the centroids based on which centroid it is closest to. The centroids are then recalculated and the process is repeated until the centroid values do not change over successive iterations.

An important issue in using the K-means (or any distance based) clustering algorithm is defining a metric for the difference or distance between two points. In our case the points to be clustered correspond to two-dimensional conductivity fields that are based on the 30 pilot point values and the 14 kriging and variogram parameters. We can thus use the Euclidian distance between these 44 decision variables as a measure of distance between two individuals. However there is one major problem with this approach. The Euclidean metric used in K-means clustering assumes that every variable for a particular individual is equally important. This is obviously not true in this case where small changes in kriging parameters may lead to very different K fields even with similar pilot point values. Thus the 44 parameters can not be equally weighted when calculating the distance metric between individuals. One way to deal with this problem is to use the actual 2-D conductivity fields to calculate the distance metric instead of the decision variables. There are two advantages to this approach – first, each location in the field is equally important, so the problem of weighting the distance metric is circumvented; second, the expert actually ranks the solutions based on these fields so it makes more sense to cluster the individuals based on the corresponding conductivity fields. The disadvantage of using this approach is that depending on the problem the conductivity fields can be very large and the K-means algorithm is known to become unstable with very high dimensional data. It is thus necessary to optimally reduce the dimensionality of the K fields. Much research has been conducted in the field of image classification and retrieval on efficient algorithms that cluster images (or any 2 dimensional field) based on the content of the image. These 'content-based' algorithms use what are known as 'spectral' clustering methods that reduce the dimensionality of the images using spectral techniques that are based on the eigen-components of the images. An efficient and effective spectral clustering algorithm is what is known as the N-cuts algorithm given by *Ng et al* (2002). In terms of dimensionality reduction, N-cuts is the extremely efficient as it reduces the n dimensions into k (the number of clusters). The clustering dimension is thus only dependent on the number of clusters and not on the dimension of the data to be clustered. Moreover N-cuts clustering has been shown to handle non-convex and non-linear proximal relationships that can not be handled by other linear clustering techniques. The proof and details of this algorithm are beyond the scope of this paper and can be found in *Ng et al* (2002). In brief, this algorithm is based on spectral graph partitioning theory which uses the Laplacian of a connected graph to find approximately optimal cuts (partitions) for the graph (giving the method it's name – normalized cuts based clustering). The salient aspect of this theory is the use of eigenvectors of the Laplacian L instead of the covariance matrix to reduce the dimensions of the data optimally.

The N-cuts clustering process is carried out before each ranking session to partition the *n* individuals in the population into *k* distinct groups and one individual is chosen from each cluster to be shown to the user. Once the user has evaluated these *k* solutions, the remaining *n-k* individuals need to be given an estimated human rank. This is the same as supervised classification of the *n-k* individuals using the previously ranked individuals as training data. Our approach uses 'content-based' classification techniques similar to those employed for clustering the conductivity fields. Specifically, we use the spectral or eigenvector (also called 'eigen-images' in the image processing community) information from the training set as attributes to train decision trees (*Quinlan*, 1986) to predict human rank. Decision trees recursively select the most predictive feature based on information gain theory and split 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. In our case the final instances that need to be predicted correspond to the ranks (from 1 to 5) and the training is performed based on the already ranked individuals stored in the archive.

To select the spectral information to be used as features in the decision trees, the principal components of the archived images are found and those which capture the maximum variability (have the largest eigenvalues) are chosen. Initial tests have shown that these 15 to 20 eigen-images can capture as much as 90% of the variability of conductivity fields. Another popular methodology uses wavelets to classify images (*Wang et al.,* 1997). Wavelet analysis allows one to extract information about the dominant scales of heterogeneity in the conductivity field. This information is often as important as the actual pattern of conductivity and can be used to classify images. We are still developing a framework to apply and test these content-based classification techniques for the IMOGA and results from these experiments will be presented at the conference.

Results

To show the usefulness of the interactive approach, initial experiments were conducted with small population sizes (20 individuals) so that the user could rank every solution in every generation. To compare solutions of the IMOGA with the non-interactive multi-objective GA (MOGA), the final results were evaluated by the user with respect to the original (known in this case) aquifer structure. The best solutions from the IMOGA and the non-interactive approach are shown in figure 3. Comparing these results to the original conductivity field in figure 1, it can be seen that the IMOGA solution is closer to the true conductivity scenario, indicating that the IMOGA was able to effectively incorporate user knowledge in the estimation process.

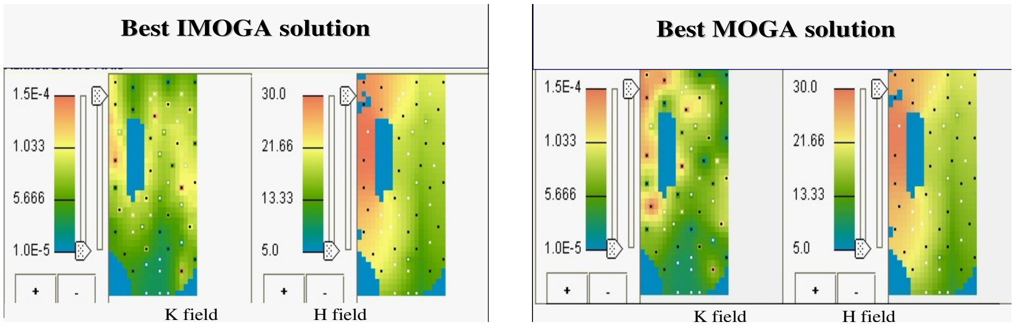

Figure 3: Comparison of best results from IMOGA and non-interactive MOGA

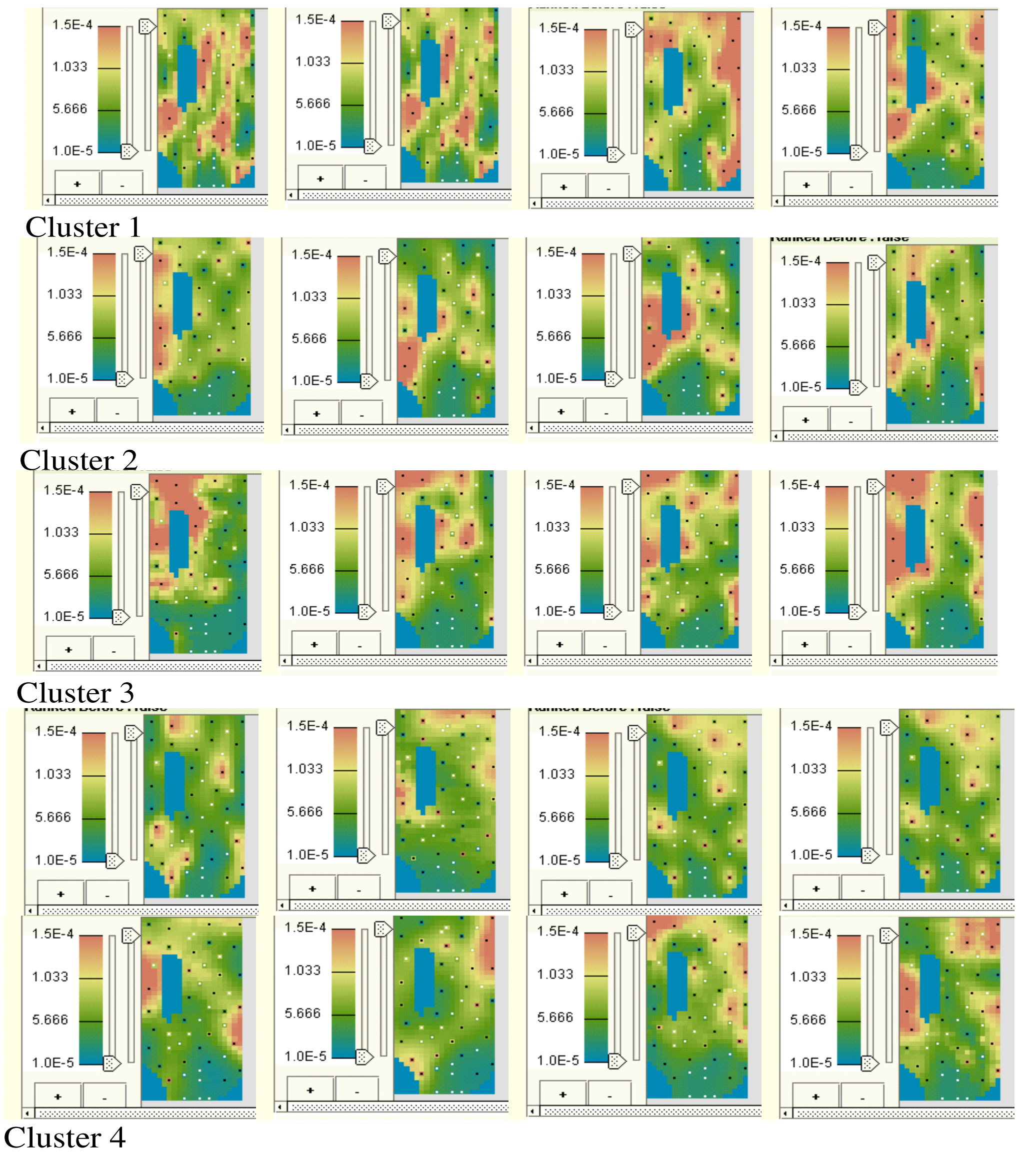

Figure 4: Clusters for 20 conductivity fields from IMOGA

To test the performance of the spectral clustering technique the first generation of 20 conductivity fields was clustered into 4 clusters. The 4 clusters are shown in figure 4. As can be seen, the members of each cluster share common perceptual features. Clusters 2 and 3 are especially well defined, while there are some discrepancies in clusters 1 and 4. An important issue with unsupervised clustering is specifying the correct number of clusters. Ideally this should correspond to the natural clusters inherent in the data. From the results in figure 4 it can be seen that cluster 1 and 4 can be further split into two clusters each. We have adapted the K-means algorithm to search for the optimal set of clusters and the results for this approach will be presented at the conference.

Conclusions

This paper introduces an interactive inverse approach that aims at combining user knowledge with hard data. This is done within a multi-objective framework that allows a range of Pareto dominant solutions to be found. Results indicate that using the proposed interactive inverse modeling approach could lead to significant improvements in the solutions found. Automated parameterization results in parameter values that only honor the data at the observation points. The spatial structure of the parameters, which is of utmost concern in modeling groundwater flow and transport processes, can usually not be explicitly included in the solution search. Allowing the user to systematically interact with the optimization process allows 'user knowledge' to drive the search towards more 'realistic' solutions. However for this approach to be feasible, user interaction needs to be kept at a minimum. Two steps were identified to deal with user fatigue. The first step was choosing only a few representative samples to show to the user. Content-based clustering was used to cluster the conductivity field based on their spatial information. This was shown to yield clusters which shared visual patterns of the conductivity field. Content-based image supervised classification can be used to give a rank to the unranked solutions. These methodologies are being tested and will be presented in more detail at the conference.

Acknowledgements

This research was funded by the Department of Energy - Grant No.: DE-FG07-02ER635302. We also wish to thank Dr. Albert Valocchi for his help in formulating the inverse problem, and Meghna Babbar for her work on the D2K IGA setup.

References

Carrera, J., and S. P. Neuman (1986) Estimation of aquifer parameters under transient and steady state conditions, Maximum likelihood method incorporating prior information, *Water Resources Research*, 22(2), 199–210.

Carrere J., A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten (2005) Inverse problem in hydrogeology, *Hydrogeol J*. 3:206 –222.

Deb, K., S. Agrawal, A. Pratap, and T. Meyarivan. (2000) A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimisation:

NSGA-II, *Proceedings of the Parallel Problem Solving from Nature VI Conference*, 849-858.
Deutsch, C. V. and A. G. Journel (1998) *GSLIB: Geostatistical Software Library and User Guide*; Oxford University Press: New York.
Doherty, J. (2003) Groundwater model calibration using pilot points and regularization, *Ground Water*; 41(2): 170-177.
Freyberg, D. L. (1988) An exercise in Ground-Water Model Calibration and Prediction, *Ground Water*, V. 26, No. 3, May-June.
Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, Reading, MA.
McDonald, M.G., and A. W. Harbaugh (1988) *A modular three-dimensional finite-difference ground-water flow model*. Techniques of Water Resources Investigations 06 A1, United States Geological Survey.
de Marsily, G. (1984) Spatial variability of properties in porous media: A stochastic approach, in *Fundamentals of Transport in Porous Media*, edited by J. Bear and M. Y. Corapcioglu, pp. 719–769, Martinus Nijhoff, Boston.
McLaughlin, D., and L. R. Townley, (1996) A reassessment of the groundwater inverse problem, Water Resources Research, 32(5), 1131–1161.
Ng, A. Y., M. I. Jordan, and Y. Weiss (2002) On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, Cambridge, MA. MIT Press.
Quinlan, J. R. (1986), Induction of decision trees. *Machine Learning*, 1:81-106.
Ramarao B. S, A. M. Lavenue,G. H. de Marsily, and M. G. Marietta MG (1995) Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 1.Theory and computational experiments. *Water Resources Research* 31(3), 475 –493
Singh, A., B. S. Minsker, and H. Takagi, (2005) Interactive Genetic Algorithms for Inverse Groundwater Modeling, American Society of Civil Engineers (ASCE) Environmental & Water Resources Institute (EWRI) World Water & Environmental Resources Congress 2005 & Related Symposia, Anchorage, AK
Sun, N.-Z. (1995) "*Inverse Problems in Groundwater Modeling*" Kluwer Academic Publishers, the Netherlands.
Takagi, H. (2001) Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation, Proceedings of the IEEE, 89(9), 1275-1296.
Wang J,. Z., G. Wiederhold, O. Firschein, and S Xin Wei (1997) Content-based image indexing and searching using Daubechies' wavelets, *International Journal on Digital Libraries*, Springer-Verlag, 1:311-328.
Yakowitz, S., and L. Duckstein (1980) Instability in aquifer identification: Theory and case study, *Water Resources Research*, 16(6), 1045–1064.
Yeh, W. W-G. (1986) Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resources Research*, 22(1), 95–108.
Zimmerman, D. A., et al. (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resources Research*, 34(6), 1373–1413.