© Copyright by Abhishek Singh, 2007

AN INTERACTIVE MULTI-OBJECTIVE FRAMEWORK FOR GROUNDWATER INVERSE MODELING

BY

ABHISHEK SINGH

B.E., Birla Institute of Technology and Science, Pilani, 2001
M.S., University of Illinois at Urbana-Champaign, 2003

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Environmental Engineering in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2007

Urbana, Illinois

ABSTRACT

This work presents a novel interactive multi-objective framework to solve the groundwater inverse problem - focusing on finding the optimal conductivity fields given measurements of aquifer response (such as hydraulic heads). The framework is based on an interactive multi-objective genetic algorithm (IMOGA) that considers model calibration as a multi-objective optimization process with user preference (expressing the plausibility of parameter fields) as an additional objective along with quantitative calibration measures such as prediction error and regularization. Given this information, the IMOGA converges to a set of Pareto optimal solutions representing the best trade-off among all (qualitative as well as quantitative) objectives. Results for the IMOGA show incorporating the site expert's subjective knowledge about the hydro-geology of the modeled aquifer leads to more plausible and reliable calibration results.

Since the IMOGA is a population-based iterative search it requires the user to evaluate hundreds of solutions leading to the problem of 'user fatigue'. A two-step methodology is proposed to combat user fatigue. First the user is shown only a fraction of the total population in every generation by clustering potential solutions based on spatial similarity and only showing unique samples from distinct clusters. Next the unranked solutions are ranked using a surrogate model that 'learns' from the user preferences. Research from image processing is used to build clustering and learning algorithms based on the 2-D images of hydraulic conductivity to closely mimic the human's visual evaluation of the parameter field. Such an approach is shown to reduce user fatigue by up to 50% without compromising the solution quality of the IMOGA.

An important part of groundwater inversion is assessing parameter uncertainty and its effect upon model predictions. To assess the uncertainty in prediction it is necessary to generate multiple realizations and test the prediction for each realization. This work uses a multi-level sampling approach to incorporate uncertainty in both large-scale trends and the small-scale stochastic variability. The large-scale uncertainty is addressed using a model-averaging approach considering both calibration error and regularization. A geostatistical approach is adopted for the small-scale uncertainty, which is added to the large-scale conductivity to give the combined conductivity fields. The prediction model is then run using the simulated fields to get the distribution of predictions.

These approaches are developed and tested on a hypothetical groundwater aquifer as well as a field-scale application based on the well known waste isolation pilot plant (WIPP) site.

To Sunayana…my 'True Music'

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and mentor – Dr. Minsker. She has guided me through this journey of growth and discovery for the last six years - motivating me by her example and nudging me on with her advice and counsel. Her faith in my abilities and patience with my mistakes have been instrumental in the completion of my learning years at the U of I. More than anything else, I would like to thank her for the trust she has shown in me by giving me the freedom to find my own path and explore varied research interests.

Dr. Valocchi, my co-advisor, has been a constant source of inspiration for me and I am deeply thankful for the opportunity this research has given me to work closely with him. I thank him for all that he taught me - through courses, research and personal interaction. I hope that I can emulate his example in the future.

I wish to thank all my committee members – Dr. Walker, Dr. Ellsworth, Dr. Bajcsy, and Dr. Kirlik. To Dr. Walker for his expert comments on all matters related to groundwater modeling (as well as the 'discussions on life' that we had in the midst of all the shop talk!). To Dr. Ellsworth for his openness and infinite patience with my questions on (geo)statistical matters. To Dr. Bajcsy for his guidance on the image-processing part of this research and Dr. Kirlik for his advice on human-computer interaction. I would also like to thank Dr. Sean McKenna, Dr. David Hart, and Dr. Richard Beauheim from the Sandia National Laboratory for their inputs on the WIPP model.

I wish to acknowledge my funding agency - the Department of Energy, under grant no. DE-FG07-02ER635302.

My sincere gratitude goes out to the faculty at the Civil and Environmental Engineering Department at the University of Illinois, Urbana-Champaign, especially Dr. Snoeyink, Dr. Cai, Dr. Eheart, and Dr. Kumar. Many of the ideas that are part of this research crystallized during course-work, projects, and discussions with all these people. I also wish to express my appreciation for all present and past EMSA research group members.

I think it would be an understatement to say that I have been truly blessed in my friends (my family by choice!). To Meghna, Amlan, Subhamoy, Santosh, Swagatam, Smitha, Namitha, Peter, Silvia, Marcia, Madhu, Parag, Rahul, and Animesh – you have all enriched the past so many years of my life by just being 'you'. I feel incredibly fortunate to have you all in my life.

Finally, my deepest sense of gratitude goes out to my family. To my parents who have given me everything a son needs or wants – love, support, advice and when necessary criticism. To my sister who has always turned up with a bright face and cheerful smile to light up my darkest days. To my grandmother whose loving memory seeps through every part of my life. And finally to my wife, Sunayana. Nothing I can say or do can equal her strength, tranquility, love, and caring. She is the true music of my life!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Reality is the leading cause of stress amongst those in touch with it

~Jane Wagner

As environmental scientists and engineers we often build models to explain and simulate the behavior of natural systems, and use these models to come up with solutions to real-world environmental problems. Use of such a modeling/decision-making approach has become standard practice for addressing complex water resources problems. At each stage of this process there are not only considerable uncertainties but also significant subjective decisions that need to be made. The formulation and structure of the model itself is based on the hydrologist's understanding of the processes involved leaving open the critical question of the impact such subjective modeling choices have on predictions and subsequent decisions. In addition, environmental models have numerous spatio-temporal parameters - such as precipitation, soil-type, hydraulic conductivities, aquifer storage, land-use types, etc - that need to be estimated or inferred in the absence of sufficient direct measurements. Most environmental problems do not have sufficient field data to estimate the model parameters uniquely and the estimation process is known to suffer from the problems of ill-posedness and 'equifinality' [*Beven and Freer*, 2001]. In such cases it becomes critical for the modeler to stabilize the calibration process by giving 'priors' for the parameters based on his or her understanding of the site. A critical research problem is how to best represent and express this prior information.

Moreover, while much work has been undertaken on characterizing the uncertainty in model parameters and predictions, the important issue of epistemic (related to knowledge or cognitive understanding) uncertainty in the conceptual model and 'prior' knowledge of the modeler remains to be investigated in a rigorous fashion. Finally, the decisions obtained from the model need to be translated to the real-world environmental problem. In reality, there are many complex and qualitative criteria such as regulatory requirements, socio-economic factors, and ecological or aesthetic considerations that are not easy to express within a purely quantitative mathematical optimization formulations. The challenge, then, lies in integrating these critical yet 'unquantifiable' (or difficult to quantify) criteria within the decision making process. This research aims to build a framework to incorporate such subjective criteria in model building and engineering decision making. It also addresses the issue of epistemic uncertainty and investigates its effect on the predictive abilities of the model.

More specifically, this work focuses on the groundwater inverse problem – the task of finding 'realistic' groundwater model parameters to best match the response of the real groundwater system. The groundwater inverse problem is known to suffer from the problems of non-uniqueness and instability that are further exacerbated by the paucity of field data at most real-world sites. In the past few years, there has been a shift towards using Bayesian inverse approaches [*McLaughlin and Townley*, 1997] that require prior parameter distributions to be specified by the modeler, thus compensating for the lack of data by providing 'expert knowledge' about the site. However, the specification of these priors is a challenge for most practitioners. In actuality, much of the expert's knowledge

is *cognitive* and is often difficult to express quantitatively as a 'prior'. This research aims to improve the inverse estimation of uncertain spatio-temporal parameters (specifically hydraulic conductivity or transmissivity) for groundwater models by incorporating 'subjective' or 'soft' knowledge of field experts with quantitative calibration criteria within an interactive and adaptive optimization framework.

One of the major contributions of this work is that it considers groundwater model inversion as a multi-objective problem comprising of both quantitative *and qualitative* criteria. To the best of our knowledge, this is the first time such an interactive framework has been proposed to solve the groundwater inverse problem.

The remainder of this chapter discusses the primary objectives and scope of this work (Section 1.2), and provides an overview of each succeeding thesis chapter (Section 1.2).

1.1 Objectives and Scope

As stated earlier, the objective of this research is to incorporate subjective judgment and qualitative criteria along with quantitative objectives to improve calibration for groundwater models. An iterative, population-based, multi-objective optimization approach is used to search for groundwater parameters that satisfy all the quantitative and qualitative objectives. This work uses genetic algorithms as the optimization technique to solve this problem (see *Wang* [1991], *Ritzel et al.* [1994], *Banzhaf* [1997], *Wang and Zheng* [1997], *Cho et al.* [1998], *Aly and Peralta* [1999], *Aksoy and Culver* [2000], *Hilton et al.* [2000], *Parmee et al.* [2000], *Reed et al.* [2001], *Takagi* [2001], *Kamalian et*

al. [2004], *Babbar et al.* [2006], etc. among many others for the successful applications of genetic algorithms to a variety of problem types and domains).

The specific objectives of this doctoral research are to:

- Formulate, design, and implement a model calibration framework based on the 'interactive multi-objective genetic algorithm' (IMOGA) that can be used to solve the groundwater inverse problem interactively.
- Address the issue of user-fatigue in such systems, arising due to the demands of user interaction required to evaluate multiple alternative solutions.
- Develop a framework to address both conceptual and stochastic model uncertainty for solutions found by the IMOGA.
- Test these frameworks on a challenging real-world application to demonstrate the utility of expert involvement in model calibration.

1.2 Summary of Research Approach

To achieve the objectives outlined above, this thesis addresses a number of relevant issues pertaining to the design and implementation of the interactive multi-objective framework. Chapter 2 provides a background to the terminology and theoretical underpinnings for this work, pointing the interested reader towards important literature in each field. Chapter 3 presents two case studies, the first of which is used to develop and test the IMOGA, while the second forms the real-world test bed for this approach. Chapter 4 introduces the methodology for the interactive multi-objective framework, and demonstrates that such systems can lead to important improvements in the plausibility and predictive ability of groundwater models. Chapter 5 addresses the important problem

of user fatigue, proposing novel clustering and machine-learning techniques to reduce the amount of user interaction required from the expert. Finally, Chapter 6 proposes a framework to address both conceptual and stochastic uncertainty to make robust predictions from the calibrated models identified by the IMOGA. It also applies the efficient IMOGA and the uncertainty framework to the field-scale application with a real site expert. Chapter 7 presents a summary of this work along with salient findings and future direction that this research is expected to take. A more detailed summary of the chapters is presented in the following sections.

1.2.1 Chapter 2: Background and Literature Review

Chapter 2 presents work from different research areas that are relevant to this research. This research is based on five broad areas – groundwater inverse modeling, evolutionary optimization, machine learning, human-computer collaboration, and model uncertainty analysis. Each of these areas is briefly addressed in this chapter and the salient literature is cited. The first sub-section describes the theory behind groundwater inverse modeling, and provides information about the prevalent approaches to solve this problem. The chapter then goes on to review evolutionary optimization, specifically genetic algorithms, discussing the work that has been undertaken in the areas of interactive genetic algorithms and multi-objective genetic algorithms. A broad overview of machine learning is provided next. The two categories of learning – supervised and unsupervised - that are used in this work are discussed. Next, the chapter gives an historical perspective on the development of human-computer collaborative systems, thus giving the IMOGA context within this fast-emerging field. Finally, the salient issues with model uncertainty analysis and especially model averaging and combination are outlined in this chapter.

1.2.2 Chapter 3: Case Studies

Chapter 3 presents two case studies for developing and testing the IMOGA framework. The first case study is based on the *Freyberg* [1988] experiment, which demonstrated the inherent ill-posedness and complexity of solving the groundwater inverse problem for even simple cases. The next case study is a field-scale application based on the Waste Isolation Pilot Plant (WIPP) site, situated near Carlsbad, NM. This has been site for extensive research on model calibration and aquifer characterization. For this case study it was necessary to consider uncertainty in the groundwater model predictions, thus motivating the third and final phase of this research (Chapter 6).

1.2.3 Chapter 4: Developing an Interactive and Multi-Objective Framework for Groundwater Inverse Modeling

Chapter 4 introduces the interactive multi-objective framework to solve the groundwater inverse problem. The formulation proposed in this chapter considers different sources of quantitative data as well as qualitative expert knowledge about the site. Groundwater model calibration is considered as a multi-objective problem consisting of quantitative objectives - calibration error and regularization - and a 'qualitative' objective based on the preference of the site expert for different spatial characteristics of the conductivity field. All these objectives are included within a multi-objective genetic algorithm to find multiple alternative groundwater parameter fields that represent the best combination of all quantitative and qualitative objectives. A hypothetical aquifer (based on the *Freyberg* [1988]) case study, for which the 'true' parameter values are known, is used as a test case to test the applicability of this method. Results show that automated calibration techniques *without* expert interaction can lead to parameter values that are not consistent

with site knowledge. Adding expert interaction is shown to not only improve the plausibility of the estimated conductivity fields but also the predictive accuracy of the calibrated model. This value of expert interaction is shown to become more significant as data from the site becomes sparser, indicating that this methodology enables the expert to compensate for the lack of data in this case.

1.2.4 Chapter 5: Addressing User Fatigue in the Interactive Framework

One of the major challenges in using the IMOGA framework for real-world applications is the burden it imposes on the site expert that interacts with the system. Since this approach aims at allowing the expert to represent his or her knowledge dynamically within an interactive system, *some* interaction is necessary for the IMOGA to function. A two-step methodology is proposed to reduce the number of user interactions required to identify reliable solutions. The first step is choosing a few highly representative solutions to be shown to the expert for ranking. For this unsupervised clustering approaches are investigated that group the candidate solutions based on spatial similarity of the conductivity fields. A sampling scheme is proposed to improve the diversity of solutions chosen from each. Once the expert has ranked representative solutions from each cluster, a machine learning model is trained to learn user preferences and predict rankings for the solutions not ranked by the expert. Since the expert is essentially processing visual information, algorithms from the field of image processing are used to improve both the clustering and machine learning algorithms by providing information about the spatial characteristics of the hydraulic conductivity field. To the best of our knowledge, this is the first time such image-based techniques have been used in the field of water resources model calibration. This image-based approach to clustering and machine learning is

shown to lead to significant improvements, reducing the amount of user interaction by as much as 50% without compromising the solution quality of the IMOGA.

1.2.5 Chapter 6: Predictive Uncertainty Analysis for the Interactive Framework

The first part of Chapter 6 presents the formulation and results of the IMOGA for the WIPP site. The IMOGA converges to a set of deterministic parameter fields representing the best trade-off among all (qualitative as well as quantitative) objectives. These multiple solutions represent alternative models of the large-scale structure of the parameter field. In reality, there is also uncertainty at scales smaller than the large scales identified by the IMOGA. The second part of this chapter addresses the important question of uncertainty assessment for the predictions obtained from the IMOGA solutions. A multi-level sampling approach is proposed that incorporates uncertainty in both large-scale trends and the small-scale geostatistical stochasticity. The large-scale uncertainty is modeled using a Bayesian approach where calibration error, regularization, and the expert's subjective preferences for different parameter fields are incorporated in the likelihood of a given transmissivity field. Small scale uncertainty is considered to be conditioned on the large-scale trend and correlated with a specified covariance structure. A state-of-the-art stochastic simulation algorithm called 'direct sequential simulation' is used to generate the small-scale realizations. This methodology preserves the local mean structure of the large-scale trend (thus maintaining the solutions' calibration errors within bounds). Finally the small-scale variability is added to the large-scale conductivity to give the combined conductivity fields. The prediction model is run using all simulated fields to obtain the distribution of predictions. The predictions considered for the WIPP site are the travel paths and travel times taken by a conservative (non-decaying) tracer in

the aquifer. Results, with and without expert interaction, are analyzed and the impact expert judgment has on predictive uncertainty at the WIPP site are also discussed. It is shown that for this case expert interaction leads to more conservative solutions as the expert compensates for some of the lack of data and modeling approximations introduced in the formulation of the problem.

2 BACKGROUND AND LITERATURE REVIEW

This chapter presents relevant background material as well as a review of the salient literature in each field that contributes to this thesis. The research presented in this thesis is multi-disciplinary - tying together strands from diverse fields. Each distinct research area is provided a brief overview in the following sub-sections, including groundwater inverse modeling (Section 2.1), interactive and evolutionary optimization (Section 2.2), machine-learning (Section 2.3), human-computer collaboration (Section 2.4), and model uncertainty assessment (Section 2.5). The goal is to acquaint the interested reader with the important terms and methodologies, and point him or her to the salient literature in each field for further details.

2.1 Groundwater Inverse Modeling

Inverse modeling is one of the most extensively researched topics in the groundwater literature. Thorough reviews of inverse methods have been published by *Yeh* [1984], *Sun* [1995], *Mclaughlin and Townley* [1996], *Zimmerman et al*. [1998], *de Marsily et al.* [2000], and *Carrera et al.* [2005]. This section highlights some important issues related to the science (and art) of groundwater inverse modeling.

In its most general form, for groundwater flow in porous media (in our case soil) the hydraulic head h (pressure exerted by underground water) can be written as a function of the location ($\vec{x}$), time (t), the hydraulic conductivity (K), sources and sinks such as pumping wells and recharge (Q), boundary and initial conditions (C), and the storage in the porous media (S) as:

$$h = f(\vec{x}, t, K, S, Q, C) \quad (2.1)$$

For steady state flow the head does not depend on time (t) or storage (S). The first stage of groundwater modeling consists of formulating a conceptual model, choosing the right equations, initial and boundary conditions, and environmental interactions for the above equation. This results in a (usually numerical) 'forward' model that has the unknown parameters and boundary conditions (K, Q, and C) shown in equation 2.1. The source and sink terms and the boundary conditions can be either measured directly (from rainfall gauges, pumping logs, etc) or estimated indirectly through inversion. Finding the right boundary conditions and recharge rates is in itself a challenging task. However, much of the focus of inverse modeling has been on estimating the conductivity (or the closely related parameter for confined aquifers - transmissivity, the product of conductivity and aquifer thickness) given measurements of the hydraulic heads at various locations. Conductivity is a spatial field in three dimensions that is highly heterogeneous and, being a sub-surface property, difficult to measure directly. The inverse problem consists of estimating the distribution of this parameter conditioned on field data and a forward model (in the form of Equation 2.1). The field data available at most groundwater sites typically consists of measurements of prevailing hydraulic heads (h) and effective conductivity (or transmissivity) in the aquifer at certain locations.

Different approaches to solving this inverse problem can be classified based on how they address the following four issues:

1) Parameterization: Reduction of the dimensionality of parameter space.
2) Formulation: Mathematical formulation of the objective or estimator.

3) Regularization: Constraining the parameter values to reduce ill-posedness.

4) Optimization: Search and optimization of the optimal parameter values.

These are discussed in further detail in the following sections.

2.1.1 Parameterization

Since the conductivity field is known to be continuous and highly heterogeneous it is impossible to estimate the true value of the parameter at every location in the field. Thus one needs to represent the continuous field of the parameters by a finite number of 'hyper-parameters'. It is these hyper-parameters that are then estimated by the inverse method. *McLaughlin and Townley* [1996] give a thorough review of different parameterization schemes. Hyper-parameters can be considered as different basis functions, including zones, splines, interpolation functions, influence functions, and pilot points, which can be used to represent the continuous field. If a statistical model is being used for the spatial field, hyper-parameters would also include parameters like the mean and covariance of the field. The challenge is to choose a parameterization that has low dimensionality while also being flexible enough to adequately represent highly heterogeneous parameter fields. The earliest and simplest form of parameterization was zones with constant conductivities [*Cooley*, 1977; *Carrera and Neuman*, 1986]. These zones had to be chosen so that they were less than the number of measurements (making the problem well-conditioned) while also being geologically consistent. Zonation has often been criticized as being too rigid, resulting in discrete conductivity values whereas the original field is continuous.

Continuous parameterization of the field can be achieved by using interpolation functions. These include finite elements [*Yeh and Yoon,* 1981], ridge functions [*Mantoglou*, 2003], interpolation functions [*Yeh,* 1986*; Hill et al.*, 1998], and a combination of zonation and continuous interpolation [*Clifton and Neuman*, 1989; *Tsai et al.,* 2003].

Another popular continuous parameterization technique used for non-linear inverse modeling is the pilot point methodology, introduced by *de Marsily* [1984]. Pilot point inversion has become increasingly popular because of its flexibility in representing the heterogeneity of the field without too many restrictive assumptions. Pilot points have been applied to a vast domain of problems [*Hernandez et al*, 2003; *Vesselinov et al*, 2001] and have been shown to work for complex and highly heterogeneous conductivity fields [*Doherty*, 2003].

Pilot point inversion starts with a given set of control (pilot) points over the spatial domain. The values at the control locations are interpolated (using an interpolation algorithm such as kriging) over the entire spatial domain, giving a parameter field. The values at these control locations (and consequently the parameter field) are then optimized by minimizing the calibration error of the resulting field.

Two approaches exist when locating pilot points. The first locates a pilot point based on the sensitivity of the calibration target for those locations [*Lavenue and Pickens*, 1992; *RamaRao et al.,* 1995; and *Lavenue et al.,* 1995]. Whereas the original pilot point method

proposed by *de Marsily* [1984] located the pilot points empirically, later versions by *Lavenue and Pickens* [1992], *RamaRao et al.* [1995], and *Lavenue et al.* [1995] used adjoint sensitivity analysis to search for optimal pilot point locations.

An alternative approach (proposed by *Gomez-Hernandez et al* [1997] and *Doherty* [2003]) specifies a large number of pilot points over the entire domain (guidelines for choosing pilot points are given by *Doherty* [2003]). Work by *Gómez-Hernández et al.* [1997] and *Doherty* [2003] has shown this approach is better suited to calibrate highly complex conductivity fields with high degrees of non-stationarity in the conductivity values. The latter approach is more powerful in capturing spatial heterogeneities as there are more degrees of freedom for the parameter field. However, when using a large number of pilot points additional 'regularization' criteria need to be used to make the problem stable [*Doherty*, 2003; *Moore and Doherty*, 2006; *Alcolea et al.*, 2006]. Since the problems addressed in this work are known to have highly heterogeneous conductivity fields, pilot points based non-linear calibration is chosen for this research.

2.1.2 Formulation

Once the parameterization has been decided, the next step in the inversion process is formulating the objective (or objectives) to estimate the hyper-parameters (zone values, pilot points, or interpolation functions). Different optimization objectives have been proposed based on least squares, maximum likelihood, or a Bayesian framework [*Press,* 1989]. Least-squares or maximum likelihood methods optimize the weighted fit between model predictions and field measurements, while Bayesian methodologies depend on maximum a posteriori estimation of the field (Bayesian formulations for the inverse

problem can be found in papers by *de Marsily* [1984], *Rama Rao* et al. [1995], *Gomez-Hernandez et al.* [1997], *Kitanidis* [1997], *McLaughlin and Townley* [1997], [*Zimmerman et al.*, 1998] and *Woodbury and Ulrych* [2000]). One advantage of the least square methods is that they are more flexible and it is fairly easy to impose additional constraints and objectives regarding the structure of the estimated field and prediction errors.

The theoretical underpinning to different calibration formulations can be expressed within the Bayesian context. The posterior conditional probability distribution function (PDF) of the parameters can be represented as Equation 2.2 [*McLaughlin and Townley*, 1997; *Woodbury and Ulrych*, 2000]:

$$p(K \mid d, I) = \frac{p(d \mid K, I)\, p(K \mid I)}{\int p(d \mid K, I)\, p(K \mid I) dK} = \frac{p(d - f(K))\, p(K \mid I)}{p(d \mid I)} \tag{2.2}$$

where K is the parameter vector (in this case conductivity values), d is the observation vector, $f(K)$ is the forward model (2.1) with the parameter values K, and I is some prior information regarding the model. The term $p(K|I)$ is often called the prior distribution of the parameters and represents some preferred values of the parameter given whatever is known about the site and the processes. The second equality holds if the errors and the parameters are uncorrelated and states that the distribution of the data given the parameters $p(d|K,I)$ is the same as the distribution of the errors between measurements and the model with the parameters K, while the denominator of the first equality is the pdf of observing the data ($p(d|I)$) with the uncertainty in the model parameters taken into account.

If the error distribution *p(d|K,I)* and the parameter distribution *p(K|I)* are assumed to be normal with covariances C_v and C_K, then the posterior distribution is also Gaussian and given by:

$$p(K \mid d, I) = c(d) \exp\left\{-\frac{1}{2}\left[d - f(K)\right]^T C_v^{-1}\left[d - f(K)\right]\right\} \\ .\exp\left\{-\frac{1}{2}\left[\overline{K} - K\right]^T C_K^{-1}\left[\overline{K} - K\right]\right\} \tag{2.3}$$

where *c(d)* is a normalization factor that only depends on the data [*McLaughlin and Townley*, 1997], and $\overline{K}$ is the prior mean of the conductivity distribution. The maximum a posteriori estimate is obtained by minimizing $-2Ln\{p(K \mid d, I)\}$ with respect to *K*. This results in the following objective function:

$$\underset{K}{Min} \quad \left[d - f(K)\right]^T C_v^{-1}\left[d - f(K)\right] + \left[\overline{K} - K\right]^T C_K^{-1}\left[\overline{K} - K\right] \tag{2.4}$$

There are many similarities between this objective and the least squares objective that is more commonly used in inverse problems. Note that the first term is simply the weighted sum of squares of the residuals between model predictions (*f(K)*) and the data (*d*). Minimizing just the first term would essentially lead one to the calibration least squares result. The second term is a weighted 'prior' term that is not often seen in least squares. It consists of minimizing the difference between the parameter values and some prior value. These prior values can be simply based on direct measurements of K or could also include geological knowledge of the expert. Thus, the Bayesian and least squares/maximum likelihood methods yield essentially the same results under similar statistical assumptions.

The most obvious difference between the Bayesian and least squares approaches is philosophical – the least squares/maximum likelihood methods consider the estimation of conductivities as deterministic, while the Bayesian approaches consider it as a statistical process. From the practical point of view, the important difference between the Bayesian and the least squares or maximum likelihood approaches is how they handle prior information. Traditionally, least squares and maximum likelihood were based only on minimizing prediction errors. *Neuman* [1973] was the first to recognize that there needs to be an additional 'plausibility' term included in the least square estimation. In essence this made the inverse problem multi-objective, with one objective for minimizing the prediction errors and the other to minimize deviations from some preferential or prior field. On the other hand, prior preferences are implicitly part of the Bayesian method since the a posteriori distribution of the estimates is conditioned on some prior field.

2.1.3 Regularization

Although one can parameterize the field to make solution to the inverse problem unique, this can often be overly restrictive because for this the number of hyper-parameters have to necessarily be less than the number of measurements. In general, as the number of hyper-parameters increases, the resulting field becomes more complex. Many researchers, such as *Gómez-Hernández et al.* [1997] and *Doherty* [2003], propose that parameterization should be flexible enough to accommodate complex fields that better represent (or at least approximate) the true structure of the parameter field. To find a solution to such a problem it then becomes necessary to choose the right level of complexity. Regularization is one way of constraining the spatial variability (complexity) of the estimated field so as to make the inverse problem well posed. The simplest way to

do this is to constrain the hyper-parameters within specified bounds (examples include *Lavenue and Pickens* [1992], *RamaRao et al.* [1995], and *Lavenue et al.* [1995]). Another popular method, called 'Tikhonov regularization' [*Tikhonov*, 1963], is to minimize the variance in the values for the hyper-parameters [*Doherty*, 2003] resulting in the smoothest possible field that gives desirable fit with measurements. Yet another type of regularization is what is referred to as 'preference-based regularization'. This is based on the Bayesian formulation given in the previous section and uses a prior field to define the default state of the conductivity field. The estimated field deviates from this default only as much as is required to reduce the prediction errors of the field. Preference-based regularization is similar to the 'plausibility' objective first propounded by *Neuman* [1973]. *Doherty* [2003], *Kowalski et al.* [2004] and *Alcolea et al.* [2006] were among the first to use this type of regularization with pilot points. When using regularization in the inverse problem, the level of regularization is typically included as an additional weighted objective in addition to the error minimization objective. *Doherty* [2003] and *Moore and Doherty* [2006] used the regularization objective within a constrained optimization framework, maximizing the regularization while maintaining some maximum prediction error constraint. This work uses a combination of preference-based and Tikhonov regularization that allows the parameter field to fit the prior data (wherever it is available) and be as smooth as possible in areas with little (or no) data support.

2.1.4 Optimization

Early inverse methods [*Nelson*, 1960] assumed a linear relationship between the measurements (heads) and conductivity and directly inverted the forward equation (relating the head to the conductivity) to find the conductivity matrix. Even though this

method is simple to understand it has some very obvious problems. First, it needs to know the head values at every point in the grid, while head measurements are available only at select locations. Second, it is highly sensitive to small changes and errors in head values making the algorithm rather unstable.

Another form of direct inverse method [*Dagan*, 1985; *Rubin and Dagan*, 1987; *Sun and Yeh*, 1992; *Woodbury and Ulrych*, 2000] is based on linearizing the relationship between head and conductivity and finding their cross-covariance to directly solve for the conductivity. It has been shown that this approach is essentially the same as co-kriging the conductivity with linear cross-covariance between head errors and conductivity [*Kitanidis and Vomvoris*, 1983; *Hoeksema and Kitanidis,* 1984]. However, linearization can introduce errors, especially if the conductivity values are far from their true or optimal values. Thus, non-linear indirect approaches have become the standard in the inverse literature.

For indirect methods, inversion is posed as an optimization problem where a conductivity field is iteratively optimized for a given formulation. Many inverse software packages such as MODFLOWP [*Hill*, 1992], UCODE [*Poeter and Hill*, 1998], and PEST [*Doherty*, 2004] use gradient-based non-linear search for iterative optimization. Depending on the formulation, the search space for inversion can be highly non-linear, multi-modal, and non-convex. This research uses pilot points as the parameterization for the groundwater inverse problem. Previous work [*Rama Rao et al.,* 1995; *Doherty*, 2003] that has used gradient based approaches to optimize pilot points have shown that the

initialization of the pilot point values can have a significant effect on the calibration results. Since this optimization problem is highly non-linear and multi-modal, multiple optima exist and gradient based approaches have the propensity of converging to local optima (making them highly sensitive to the initialization of the search process). Global approaches such as genetic algorithms are more likely to find the global optima and can avoid local convergence [*Goldberg*, 1989] thus making them less sensitive to the pilot point initialization. Moreover, due to non-uniqueness and parameter uncertainty it is often desirable to generate multiple solutions for the inverse problem [*Beven and Freer*, 2001]. This has led many researchers to apply heuristic global optimization techniques such as simulated annealing and tabu search [*Rao et al.*, 2003; *Zheng and Wang*, 1996], shuffled complex algorithm [*Madsen*, 2003], genetic algorithms [*Karpouzos et al.,* 2001], global-local optimization [*Tsai et al.*, 2003], and other stochastic global optimization algorithms [*Solomatine et al.*, 1999] to search for optimal conductivity fields. Due to these reasons, this research uses genetic algorithms [*Goldberg,* 1989], which have been shown to perform well for non-linear and multi-modal inverse problems. Further background on this approach is given in the next section.

2.2 Genetic Algorithms –Multi-Objective, and Interactive

Genetic algorithms (GAs), first introduced by *Holland* [1975], are population-based stochastic search and optimization algorithms based on the Darwinian principle of 'survival of the fittest'. GAs start with a 'population" of random initial solutions to the optimization problem. The decision variables are encoded as 'chromosomes' and the objective function is often called the 'fitness function' in GA literature. GAs use the operators of selection, crossover, and mutation to search for progressively better (more

optimal) solutions [*Goldberg*, 1989]. *Goldberg* [2002] has shown that well-designed or 'competent' GAs can efficiently and reliably solve difficult problems that are non-linear, non-convex, discrete, discontinuous, noisy and/or multi-modal. With the growth of computing power, GAs have begun to be used to solve the inverse problem – examples include *Solomatine et al.* [1999], *Karpouzos et al.* [20001], *Giacobbo et al.* [2002], and *Tsai et al*. [2003]. The advantage of using GAs is that they can handle the multi-modality and non-linearity that are inherent in the inverse problem. In addition, GAs do not depend on gradient information for convergence so they can be used to solve for discrete parameters.

While many different kinds of GAs have been proposed in the literature, this work focuses on two types of GAs (which until very recently have been separate fields) – multi-objective GAs and interactive GAs. The following two sections discuss these two types of GAs in more detail.

2.2.1 Multi-Objective Genetic Algorithms

Multiple objectives can be optimized either by weighting or constraining the objectives (essentially converting the multi-objective problem to a single-objective one), or by considering all the objectives together and finding the 'Pareto-optimal' set of solutions that represent the optimal tradeoff between the objectives. Mathematically, for a minimization problems with n objectives, a vector of decision variables $\vec{x}^* \in F$ is considered Pareto optimal if there does not exist another $\vec{x} \in F$ such that $f_i(\vec{x}) \leq f_i(\vec{x}^*)$ for i = 1, 2 … n and $f_j(\vec{x}) \leq f_j(\vec{x}^*)$ for at least one j (where $f_i(\vec{x})$ is the ith objective

function for the decision variable vector $\vec{x}$. In other words, the Pareto or non-dominated front consists of solutions that cannot be improved in one objective without worsening at least one other objective.

Pareto optimization results in multiple optimal solutions; population-based approaches like GAs can solve for the entire set of Pareto optimal solution *simultaneously*, making them ideal for multi-objective optimization. Evolutionary multi-objective (EMO) methods were first introduced through the work by *Schaffer* [1984]. More work in this area was undertaken by *Fonseca and Fleming* [1993], *Srinivas and Deb* [1994], and *Coello* [1999] among others. Among others, *Cieniawski* [1993], *Ritzel et al* [1994], and *Reed and Minsker* [2004] have applied multi-objective GAs to groundwater-related problems. The application of multi-objective optimization to inverse problems has been mostly restricted to hydrological modeling (examples include *Yapo et al.*, 1998; *Gupta et al*., 1999; and *Madsen*, 2003), where multiple data sources and processes need to be incorporated in the inversion process. Multi-objective GAs were also used in a related problem of finding the best sampling design for calibrating a hydrological model [*Knopman and Voss*, 1989]. However, to the best of our knowledge this study is the first to solve the inverse problem (consisting of the plausibility or the prior term for least square optimization) formulated by *Neuman* [1973] and *Doherty* [2003] using multi-objective optimization.

This study uses an efficient and robust multi-objective genetic algorithm (MOGA) called the elitist non-dominated sorting genetic algorithm (NSGA-II) [*Deb et al,* 2000]. The

NSGA-II identifies Pareto-optimal solutions using a layer-wise ranking scheme. Each individual is compared to all others in the population and those that are strictly non-dominated are marked as the locally Pareto (or rank 1) solutions. These are then removed from the list of considered solutions and rank 2 solutions are defined as those that are strictly non-dominant with respect to the remaining solutions, and so on until all solutions in the population are ranked. Once all the individuals in a population have a rank, the objective of the NSGA-II is to minimize the rank. To ensure that the NSGA-II converges to a set of solutions that adequately cover the entire Pareto front, a metric called the 'crowding distance' is used to favor solutions that are more isolated in the decision space. The crowding distance for a particular solution is defined as the Euclidean distance (in objective or decision space) between its two adjacent neighbors on the same Pareto front as the solution. During the selection process, the NSGA-II first uses the rank of the solution to select one individual over another. For solutions with the same rank, the solution with the larger crowding distance is preferred. In addition, to preserve good solutions found by the GA, the NSGA-II utilizes an elitist selection scheme where the parent and children populations are compared and only the best solutions are advanced to future generations. Apart from using the rank and the crowding distance for selection, the NSGA-II uses the usual GA operators of crossover and mutation [*Goldberg*, 1989].

2.2.2 Interactive Genetic Algorithms

Since GAs do not require information about the gradient of the objective function, they are well suited for incorporating subjective or qualitative objectives through interactive optimization [*Banzhaf*, 1997]. Interactive optimization proceeds in the same manner as traditional optimization except that one or more of the optimization operators is provided

by a human expert. The most general case of interactive optimization is what is known as human-based genetic algorithms (HBGAs) [*Kosorukoff*, 2001] where any or all of the four GA operators of initialization, selection, crossover, and mutation are delegated to humans using appropriate interfaces. A more specific case of HBGAs is the interactive genetic algorithms (IGAs) [*Takagi*, 2001], where the fitness evaluation is subjective and is provided by the user. Because humans are better at evaluating relative differences than absolute values [*Stone and Sidel*, 1993], most IGAs employ a relative rank or preference based on the user's subjective criteria [*Takagi*, 2001; *Corney*, 2002]. Ranking allows solutions to be classified into groups according to user preference. This becomes an indirect way of understanding what types of solutions are desirable to the user's cognitive perceptions. Ranking can be as coarse as labeling solutions merely 'good' or 'bad', or it can have more classification types, e.g. 5 levels progressing from very good to very bad [*Takagi*, 2001].

IGAs have been applied to many fields where expert input is critical to finding good solutions to real-world problems such as hearing-aid fitting, lighting design, face image generation [*Takagi*, 2001], music and image retrieval [*Cho et al.*, 2002], traveling salesman optimization [*Louis and Tang*, 1999] and even creating jazz music [*Biles*, 1994]. In a more allied field, IGAs have been applied to solve inverse problems arising in geology by *Wijns et al*. [2001]. While *Wijns et al.* [2001] considered the inverse problem to have a single overriding human objective, this work considers model calibration to be a multi-objective problem where subjective information needs to be incorporated with more quantitative model performance criteria.

2.3 Machine Learning – Unsupervised and Supervised

Machine learning, as the name implies, is concerned with the design and development of algorithms and techniques that allow computers to "learn". At a general level, there are two types of learning – unsupervised and supervised learning. These two types of machine learning are discussed in the following sections.

2.3.1 Unsupervised Clustering

Unsupervised learning (see *Ghahramani* [2004] for a recent review) is a method of machine learning where a model is fit to observations in the absence of any *a priori* input or training data. Clustering [*Jain et al*, 1999] is perhaps the most common form of unsupervised learning and consists of partitioning data into subsets (or 'clusters') such that the data in the same cluster share more common features than members of other clusters. The 'commonality' of the data points is defined using a user-defined measure of similarity or proximity between data points. Clustering is a form of 'unsupervised' learning [*Jain and Dubes*, 1988] because the data are classified based purely on a similarity measure within the data set, without any prior knowledge about the true distributions that they come from. *Jain and Dubes* [1988] identify three primary steps in clustering:

1. **Data representation** (could also include feature extraction and selection) - This is the process of choosing the number, type, and scales of features of the data that are appropriate for clustering. In addition, it may be necessary to focus the clustering only on a subset of the features (feature selection), or transform the data to produce new more discriminating features (feature extraction). The distance or

proximity between different data points is then calculated as a function of the selected feature set.

2. **Clustering**: This is the actual process of finding the subset of data points with the maximum similarity. Numerous clustering algorithms have been proposed. Broadly speaking, clustering algorithms can be divided into two categories: discrete or exclusive clustering methods - such as 'K-means clustering' [*MacQueen*, 1967] - that lead to disjoint or 'flat' clusters, and hierarchical or 'nested' clustering techniques [*Johnson*, 1967] that essentially build nested, tree-like clusters based on the concept of 'linkage' between one data point and another. Hierarchical clustering groups data-points based on their immediate vicinity – typically linking data points that are closest to each other in feature space. While such an approach is good for cases with distinct clusters, it often fails when the dataset has a more complicated configuration, outliers, or noise. In recent years, a separate class of clustering algorithms, called 'spectral clustering' algorithms [*Cristianini et al*., 2002; *Ng et al.*, 2002], have emerged as a powerful and efficient tool for clustering complex and multi-dimensional datasets. Spectral clustering techniques have been shown to work well for datasets with non-linear, non-convex, and even inter-twining clusters [*Shi and Malik*, 2000; *Ng et al.*, 2002]. In addition, they do not require any prior assumptions about the form of the clusters (and are thus applicable to a wide range of problem types), are robust to noisy data, scale favorably with dimensionality of data, and are easy and efficient to implement. Reviews of spectral clustering have been provided by

Verma and Meila [2003], *Dhillon el al.* [2004], and *Luxburg* [2006] among others.

3. **Data abstraction**: This is the process of selecting or extracting representative descriptions for each clustered data set. Typical approaches consist of averaging the features to find the 'average representative' or selecting the centroid from the cluster as a prototype for that cluster [*Diday and Simon*, 1976]. Another approach is to use feature extraction methods to abstract the most predominant features within a particular cluster.

2.3.2 Supervised Learning

Supervised learning is a broad category of techniques that learn or approximate some unknown or partially known function ($y = f(x)$) from a set of 'training' data consisting of instances of inputs and outputs for the function (x, y). The output (y) can be discrete classes or labels (in which case the problem is one of learning classifications) or can be continuous values (in which case the problem is one of learning a regression function). Irrespective of the type of algorithm, a supervised algorithm begins with some training data, which it uses to learn the function by minimizing the classification error within the training data set. Once the supervised learner has been adequately trained, it is used to predict the value of the function for any valid input object. Reviews for different types of learning algorithms have been given by *Russell and Norvig* [1995], *Mitchell* [1997], and *Duda et al.* [2001]. For this study, two different types of supervised learning algorithms were used – decision trees based learning [*Quinlan*, 1986] and naïve Bayes learning [*Domingos and Pazzani*, 1996; *Michie et al*., 1994]. Details on both of these algorithms are given in Section 5.2.2.

2.4 Human-Computer Collaboration

The motivation to couple computers with human beings has a long history. In its earliest form, human-computer collaboration was as achieved through 'decision support systems' (DSS) [*Keen and Scott Morton*, 1978] – loosely defined as information systems that help the user best utilize data and models to solve semi-structured problems. Decision support systems encompass a vast suite of information systems that help decision makers in fields ranging from portfolio management to medicine. More recent artificial intelligence (AI) based 'expert systems' aim at mimicking human cognition and decision making. Early in the development of expert systems, there was a recognition that these systems were severely restricted in solving ill-posed or incomplete problems where expert knowledge was difficult to reduce to a set of rules. This led to an impetus for including interaction in expert systems [*Winograd and Flores*, 1986].

The step towards collaborative human-computer systems was taken by *Woods et al.* [1990], who coined the term 'joint cognitive systems' (JCS). According to cognitive theory, humans form plans and decisions based on their needs, motives, values, and beliefs; act on these; get feedback about the effects or consequences; and then actively modify their perceptions, plans, and behavior accordingly [*Hendry*, 1996]. Joint cognitive systems can be defined as systems that allow users to combine this process of understanding and cognition with the efficient problem-solving skills of computers to effectively solve real world complex problems. *Brezillon and Pomerol* [1997] were among the first to realize that both DSS and JCS address the same issue - how a computer system and human being can, jointly, in real time, cooperate in the achievement of a task,

so that the resulting achievement is better than if it was carried out by the system or the user alone. The difference between DSS and JCS is that while DSS tend to focus on providing 'assistance' to users to solve problems and make decisions, JCS provide the user with a collaborative environment where the computer and user can *collectively* participate in decision-making. Even early on there was a realization that joint cognitive systems are ideal for solving complex, real-world problems that could not be adequately solved either by the computer or the human expert alone.

Following the JCS paradigm, *Brill et al.* [1979] proposed the 'modeling to generate alternatives' (MGA) methodology to combine experts and mathematical models to solve incompletely defined problems. *Brill* [1979] proposed the hop-skip-jump heuristic, which can be applied to linear, mixed-integer, or non-linear optimization methods to search for different model alternatives. These model alternatives were generated so that they were maximally different from each other yet satisfied some performance constraint. This led to the discovery of multiple diverse yet satisfactory models that could be potential solutions to the ill-posed problem. The user then evaluated these alternatives based on his or her additional knowledge and found solutions that better fit qualitative and quantitative criteria. The MGA approach has been applied to many areas including land use planning [*Brill et al*, 1982], airport networking [*Brill et al*, 1990], and water resources management [*Kao and Liebman*, 1991]. *Xiao et al* [2002] were the first to recognize that population-based multi-objective optimization was essentially the same as MGA, except that in this case the solutions not only satisfied some performance constraints but also represented the optimal tradeoff between different objectives. Interestingly, *Xiao et al*

[2002] did not include expert interaction or qualitative objectives. Thus, even though their work utilized the MGA paradigm it was not really a joint cognitive system.

Finally, even though they have not been recognized as such, interactive optimization approaches such as IGAs and HBGAs are also types of joint cognitive systems. Just like JCS, these approaches require the computer and the human to work together to search for optimal solutions that could not have been found by either alone.

2.5 Model Uncertainty

Characterization of model and predictive uncertainty is increasingly becoming the norm in inverse modeling. This section discusses the predominant strands in research on both epistemic and aleatory (stochastic) model uncertainty. With regards to epistemic or conceptual model uncertainty, the need to move away from one 'optimal' model to a set of multiple models for predictions was identified early on by *Delhomme* [1979], *Neuman* [1982], *Hoeksema and Kitanidis* [1989], *Wagner and Gorelick* [1989], *Beven* [1993], and *Neuman and Wierenga* [2003] among others. *Beven* [1993; 2000] laid out the argument for considering multiple alternative models and model structures due to the problem of 'equifinality' – the notion that a unique model with an 'optimal' set of parameters is inherently unknowable. Instead, they argued for a set of acceptable and realistic model representations that are consistent with the data. Work such as *National Research Council* [2001], *Neuman and Wierenga* [2003], *Carrera and Neuman* [1986], and *Samper and Neuman* [1989] have also shown that considering only one conceptual model for a particular site can lead to biased and erroneous results that can have adverse environmental, economic, and political impacts. Given these multiple models it becomes

essential to assess the likelihood or probabilities of each model. Without such likelihood measures all transmissivity fields would be assumed to be equally likely. Different sources of information (such as direct and indirect field measurements) are typically used to assess the conditional probabilities of the alternative models given such data. This conditioning of the models on the data is important as it allows one to estimate the posterior probabilities of the models, in effect reducing the *a priori* uncertainty of the model.

The problem of model uncertainty has been addresses within two major frameworks – GLUE or generalized likelihood uncertainty estimation, and BMA or Bayesian model averaging. While there are similarities between these approaches, the major difference lies in the way they ascribe likelihood (or probability) to the different models being considered.

The GLUE methodology proposed by *Beven and Binley* [1992] rejects the idea of a single optimal model, instead starting off with a (large) set of (randomly generated) alternative model structures. A subset from this is then selected based on some performance thresholds or (possibly subjective) acceptance criteria. Each selected model is then weighted using a normalized likelihood measure related to its goodness of fit with the data available. The predictions from such models are then combined and updated using a standard Bayesian framework. While the GLUE framework *is* highly generalizable and applicable to almost all types of problems, certain aspects of the methodology have generated controversies in recent years [*Mantovan and Todini*, 2006]. These include - the

lack of statistical basis for the likelihood and threshold measures used for model selection and weighting, the computational burden required due to extensive Monte Carlo simulations, and the fact that since the model structures and parameters are not really optimized the link between GLUE and model calibration is tenuous leading to overestimation of the prediction uncertainty. However, *Beven* [2006] has answered these criticisms by contending that a) formal Bayes is a special case of GLUE and is applicable under certain strong assumptions, and b) optimization or model selection can be used within the GLUE framework to reduce unnecessary uncertainty. In recent years, the link between GLUE and optimization has become stronger with the work of *Mugunthan and Shoemaker* [2006] who showed that optimization can in fact be used to generate alternative models for GLUE. This has led to important efficiency enhancements for the GLUE framework by eliminating the need for Monte Carlo trials to generate model alternatives. *Vogel et al.* [2007] have also investigated different likelihood measures and have shown that with the right likelihood the GLUE methodology does indeed yield statistically significant results.

The other contemporaneous approach to addressing model uncertainty is the more formal Bayesian approach, which has been propounded by *Draper* [1995], *Kass and Raftery* [1995], *Hoeting et al.* [1999], *Woodbury and Ulrych* [2000], *Neuman* [2003], *Neuman and Wierenga* [2003], and *Ye et al.* [2004] among others. This approach, called Bayesian model averaging - or more specifically maximum likelihood Bayesian model averaging (MLBMA), in the case of *Neuman* [2003] - uses a more statistically consistent methodology to assess the Bayesian posterior probabilities for a given conceptual model.

While approaches such as *Draper* [1995], *Kass and Raftery* [1995], and *Hoeting et al.* [1999] rely on extensive Monte Carlo simulations to calculate the probabilities, *Neuman* [2003] proposed using likelihood measures such as KIC (Kashyap information criterion, *Kashyap* [1982]) and BIC (Bayesian information criterion, *Schwarz* [1978]) to weight different models, thus obviating the need for extensive simulations. However, MLBMA is not without its share of controversies. Since this methodology is based on the concept of a 'maximum likelihood' for a given model, it tends to give a very high weight to models which have the best goodness of fit measure. *Domingos* [1997] have argued that model combination by its very nature works by enriching the space of model hypotheses not by approximating a Bayesian distribution function. Moreover, *Domingos* [2000] compared BMA with other model averaging techniques and showed that BMA tends to underestimate the predictive uncertainty. *Minka* [2000] contended that this is hardly surprising because by definition techniques like BMA and especially MLBMA are built on the intrinsic assumption that there is *only one* model of reality. This is borne out in the original MLBMA paper by *Neuman* [2003] where he lays out the fundamental assumption for this technique - *'only one of the* (alternative) *models is correct even in the event that some yield similar predictions for a given set of data'*. In other words, MLBMA is more of a model selection technique than a true model combination methodology. In this work both methodologies are tested for the WIPP test case.

Once the uncertainty in different conceptual models has been characterized, the next step is to undertake stochastic predictive uncertainty analysis for each of the models. Stochastic modeling is a vast field in itself and it is beyond the scope of this work to give all the details. However, broadly speaking predictive uncertainty can either be estimated

using linear approximations, non-linear approximations, or Monte Carlo methods. Linear methods [*Rubin*, 1991; *Hill*, 1998] are based on assuming a linear relationship between the predictions and the first order derivative (also called the Jacobian) that measures the sensitivity of the prediction to the parameters. Assuming linearity allows one to establish confidence bounds on the predictions with respect to the uncertainty in calibration. It has been shown [*Vecchia and Cooley*, 1987] that in general linear confidence intervals underestimate the true uncertainty bounds for the predictions. Although they can be more difficult to apply, non-linear methods give a better approximation to the true confidence bounds for the predictions. Non-linear methods can either be analytical or sampling-based. Analytical non-linear methods have also been proposed by *Vecchia and Cooley* [1987], *Christensen and Cooley* [1999], and *Moore and Doherty* [2005]. Sampling-based methods that sample the distribution of the parameters are easier to implement (though computationally more expensive) and thus more popular. Most of the work conducted for the WIPP site used the sampling approach. This research uses the sampling approach, as well, as it is more conducive to the IMOGA setup, which generates multiple conductivity scenarios.

As an alternative to the two-step approach described above, which in essence addresses conceptual and stochastic uncertainty separately, another approach that has been used predominantly in the field of groundwater calibration is the concept of 'warping' stochastic realizations to fit some calibration target. This 'non-Bayesian' approach has been followed by *Rama Rao et al.* [1995], *LaVenue et al.* [1995], *Gomez-Hernandez et al.* [2003], *Doherty* [2003], and most recently by *DOE/WIPP* [2004] (interestingly, most

of these approaches were applied to the WIPP site). The methodology followed here is to create stochastic realizations of the parameter field (for example conductivity or transmissivity) conditioned on direct measurements from the field. These realizations are then 'warped' or transformed using pilot points so that each realization fits a given calibration target. These calibrated stochastic fields are then used within a standard Monte Carlo sampling scheme to assess the uncertainty in predictions. Thus, this method relies on the generation of multiple (possibly hundreds) stochastic fields, and optimizing each one of these separately. The advantage of this methodology is that different stochastic structures at different scales can be incorporated in the initial set of realizations (in the WIPP case for example, large stochastic 'indicator zones' of fractured media were included in the realizations). While, the warping approach has been applied to the WIPP test case [*DOE-WIPP*, 2004], there are some significant drawbacks to using it with the IMOGA framework. Most importantly, such an approach can be undertaken almost exclusively for fully automated calibration exercises; with interactive optimization the task of potentially ranking hundreds of parameter fields can be simply too overwhelming for the expert. A statistical drawback to such a technique, which is often overlooked, is the fact that while the initial set of stochastic fields are created to be 'equally likely' realizations, this no longer holds true for the post-optimization 'warped' fields. This is especially true because of the non-uniqueness inherent to the calibration problem. Optimization can lead to different parameter realizations (initially different from each other) converging to 'warped' fields that are similar to each other (Figure 2.1 shows this biasing effect for a uniformly generated set of 2-D data). Thus, the distribution of the parameter fields is biased by the optimization and unless likelihood weights (similar to

GLUE or BMA) are used to account for this biasing effect of optimization, the predictive probability distribution obtained from the warped fields is likely to be biased. It is for these reasons that the uncertainty analysis of the IMOGA results is based on the Bayesian framework discussed earlier.

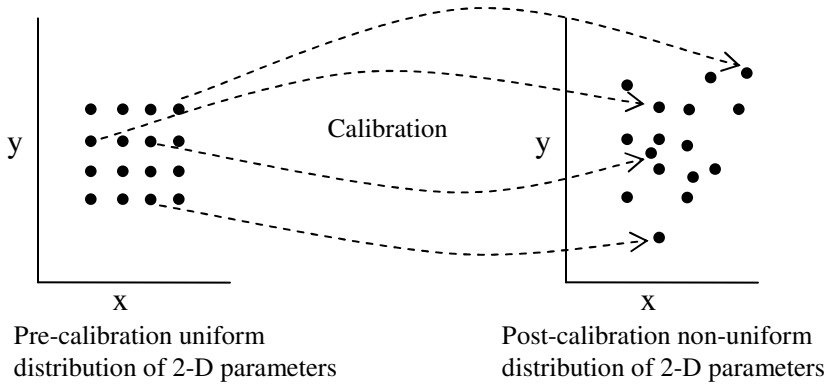


Figure 2.1 Biasing effect of optimization on uniformly generated data

3 CASE STUDIES

If you can't solve it, it's not a problem – it's reality

~ Barbara Colorose

Two case studies have been used in this research – a computationally tractable hypothetical groundwater aquifer and a more intensive field-scale application that forms the test bed for the most promising methodology identified with the hypothetical case. Chapters 4 utilizes only the hypothetical case study, Chapter 5 presents results for both the hypothetical and field-scale case study, while Chapter 6 addresses only the field-scale application. The rest of this chapter discusses details of both these case studies.

3.1 Hypothetical Groundwater Aquifer Case Study

The hypothetical case study is based on the example presented by *Freyberg* [1988] and shown in Figure 3.1. This case study is used to demonstrate the applicability and utility of the IMOGA and rapidly test different algorithmic approaches. Moreover, *Freyberg,* [1988] has shown the inversion of conductivity based on head observations is an ill-posed problem with head calibration error having little correlation with the accuracy of the calibrated hydraulic conductivity field or with the accuracy of post-calibration predictions. Since numerical calibration led to unsatisfactory results, this test case is suitable for testing the importance of expert interaction in improving the calibration process. The hypothetical aquifer is considered as 'ground truth' and all observations are taken from this aquifer.

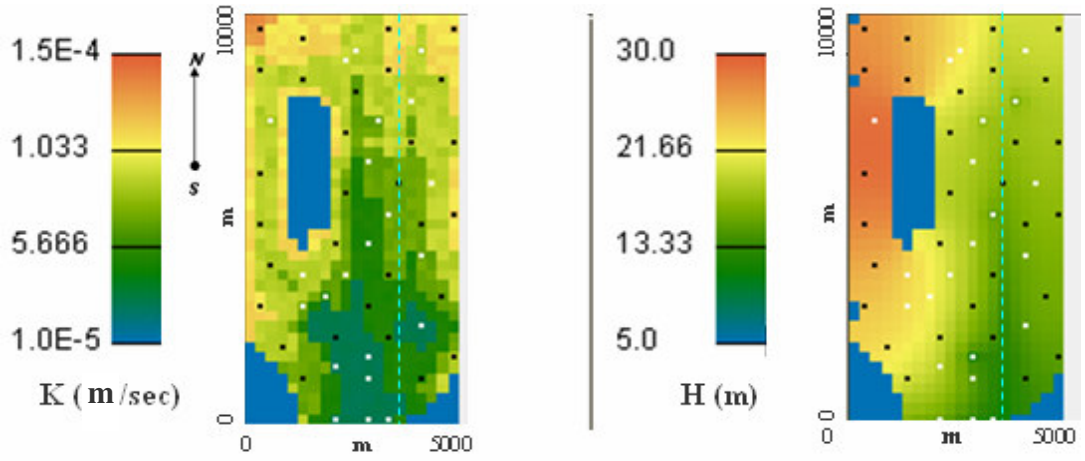


Figure 3.1 True conductivities (K) and hydraulic heads (H) for the Freyberg case

The hypothetical aquifer is a shallow, 2-D, unconfined aquifer 10 km long and 5 km wide aquifer underlain by an impervious stratum with no leakage. The aquifer is discretized into 20 columns and 40 rows, with each cell measuring 250x250 square meters. The aquifer is assumed to have isotropic hydraulic conductivities (in the two dimensions) and steady-state hydraulic heads. There are no flow boundaries along the north, east, and west boundaries, and a constant head boundary along the south. In addition, there is an outcrop in the middle, which introduces an internal no-flow region. The boundaries and outcrop are shown by the blue areas in Figure 3.1. Uniform and constant recharge is assumed. There is a river with a steady, non-uniform stage that runs north to south across the aquifer (shown by the dashed lines in Figure 3.1). The hydraulic conductivities and bottom elevations for the modeled aquifer were taken from the *Freyberg* [1988] paper. As can be seen in the figure, the hydraulic conductivities are highest in the northwest, with a decreasing trend to the southeast.

There are 6 wells pumping at constant rates, and 16 observation wells spread across the aquifer, giving a total of 22 observation points, shown by the white dots in Figure 3.1. The locations and values for the conductivities and heads are shown in Table 3.1. The hydraulic heads in this hypothetical aquifer were calculated using MODFLOW [*McDonald and Harbaugh*, 1988]. 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 research, only the hydraulic conductivities were considered unknown, while other parameters, boundary conditions, and observations were assumed to be perfectly known (this is slightly different from *Freyberg* [1988] where the uniform recharge rate, and bottom elevations were also considered unknown). The hydraulic heads and conductivity measurements at the observation points were taken as the calibration data for the inverse model. To make the problem more realistic, the head measurements were contaminated with noise from a normal distribution with zero mean and variance of 0.25 (less than 1% measurement error on average). The conductivity values were log transformed before being used for inverse modeling. A total of 30 pilot point locations (shown by the black dots in Figure 3.1) were chosen so that they had good coverage over the entire field.

As with other pilot point approaches, the values at the 30 pilot points are decision variables for optimization. A critical decision when using pilot points (or any geostatistical model) is the choice of the kriging parameters used for spatial interpolation.

These parameters should also be evaluated by the expert during the interactive optimization process. Few studies have looked at optimizing both pilot point values and kriging parameters at the same time. Thus, in addition to the values of the pilot points, the kriging parameters are also included as decision variables. These consist of the size, shape and orientation of the kriging window, and the maximum and minimum number of points used for kriging, giving 5 additional decision variables and a total of 35 decision variables. The limits of these decision variables are given in Table 3.2. The variogram, chosen by optimizing the fit between a model variogram and the weighted empirical variogram calculated from the data (shown in Figure 3.2), was a spherical model with a 0 nugget, range of 8,000 m and a sill of 0.40. It is important to note that the variogram model can have a strong influence on the final parameter field, especially when creating conditional realizations (used in uncertainty analysis, to be discussed in Chapter 6). Thus, in addition to including some of the kriging and variogram parameters in the optimization search, preliminary trial-and-error experiments were also conducted (with different variogram types) and expert judgment was used to select the appropriate (spherical) variogram model for this case study.

The ultimate goal of inverse modeling is not just to minimize calibration error, but also to find realistic parameters so that the model has better predictive capabilities. To test and validate the calibrated model, a predictive scenario was defined similar to the one used by *Freyberg* [1988]. This consisted of predicting the aquifer response after reducing the conductance of the river bed by two orders of magnitude (this can be thought of as lining the river bed for a real site). Comparing the calibrated model's hydraulic head response

with the 'true' response for this scenario provides information about the predictive capabilities of the calibrated model. Further details of the parameterization and formulation for the optimization are given in the methodology section (Sections 4.2.1 and 4.2.2) of Chapter 4.

Table 3.1 Locations and values of H and K measurements

Index	x (m)	y (m)	True H (m)	Measured H (m)	K (m/sec)	Ln(K)
1	2625	9125	22.22158	22.12778	8.99625E-05	-9.31612
2	4125	9125	19.69107	19.74297	8.70057E-05	-9.34954
3	2375	8875	22.66976	22.47836	9.22049E-05	-9.29150
4	3875	7875	16.29162	16.26502	8.98073E-05	-9.31784
5	625	7375	28.83823	28.92293	8.56501E-05	-9.36524
6	3125	7375	17.50686	17.76526	6.83011E-05	-9.59158
7	2875	6375	18.44698	18.09578	6.93262E-05	-9.57669
8	4375	5875	17.51641	17.25881	9.25927E-05	-9.28730
9	3375	5125	15.22509	15.06429	7.02220E-05	-9.56385
10	2875	4375	18.20994	18.25264	5.03435E-05	-9.89664
11	4125	4125	16.07255	16.40875	6.83458E-05	-9.59093
12	1375	3625	23.31628	23.80038	8.75577E-05	-9.34321
13	2375	3625	20.02407	20.20937	7.33922E-05	-9.51969
14	1875	3125	21.77313	21.97613	5.24470E-05	-9.85571
15	1375	2875	22.78357	22.74787	7.24584E-05	-9.53250
16	4125	2375	14.95987	14.93487	2.93851E-05	-10.43502
17	2875	1625	11.85154	11.65154	2.88240E-05	-10.45430
18	2125	1375	18.93550	19.05880	2.88149E-05	-10.45462
19	2875	1125	14.96599	15.27539	3.07929E-05	-10.38823
20	2125	125	15.60000	15.92400	2.94271E-05	-10.43359
21	2875	125	13.00000	12.93050	3.05971E-05	-10.39461
22	3375	125	12.00000	12.05430	2.99442E-05	-10.41617

Table 3.2 Decision variables for Freyberg case study

Decision Variable	Range
Minimum number of points used for kriging	1 to 20
Maximum number of points used for kriging	1 to 20
Major search radius for kriging	2500 to 8000 m
Minor search radius for kriging	2500 to 8000 m
Orientation of search window	0^o to 90^o
K values for Pilot Points (Ln(K) value for Pilot Points)	$3.35x10^{-4}$ to $4.5x10^{-5}$ m^2/sec (-8 to -11)

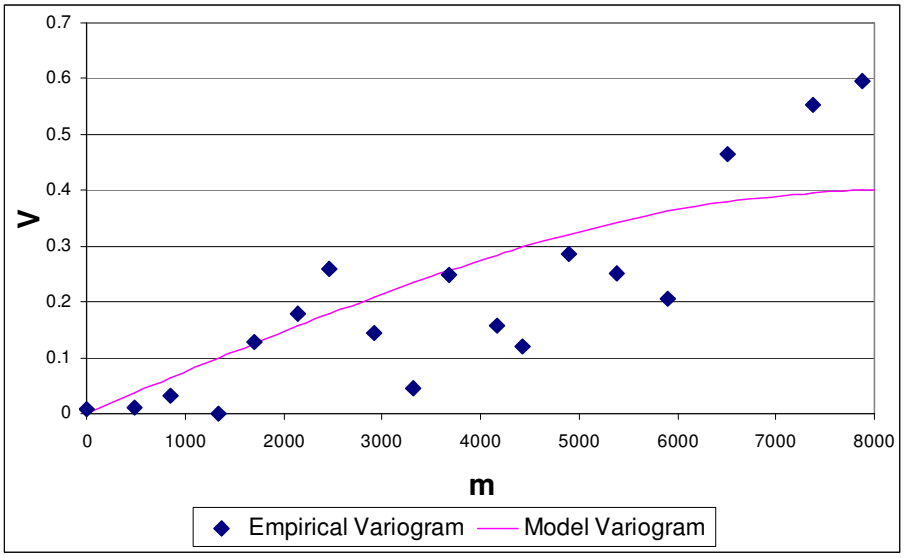


Figure 3.2 Variogram of log conductivity data for the Freyberg case

3.2 Field Scale Application – The WIPP Site

The Waste Isolate Pilot Plant (WIPP), located about 26 miles from Carslbad, NM, is the world's first underground repository used to dispose of radioactive, transuranic (TRU) waste from the research and production of nuclear weapons within the United States of America (see Figure 3.3). The WIPP has also served as a pilot facility to demonstrate the safe management, storage, and disposal of radioactive wastes. Researchers at Sandia National Laboratory, the WIPP's scientific advisors, have been primarily responsible for site selection, characterization, modeling, and experimentation to understand the interaction of the TRU wastes with the disposal environment. After almost 20 years of extensive scientific studies, public input, and regulatory struggles, the WIPP began operations on March 26, 1999 and is expected to operate until 2070 with active long-term monitoring for an additional hundred years. The concern at the WIPP site is if the radioactive wastes stored in the repository might leak and contaminate the Culebra dolomite aquifer, which is a highly permeable formation about 1000 feet above the

repository. Thus, most inverse modeling techniques have focused on characterizing the Culebra aquifer, which could prove to be an important pathway for contaminants in the case of a leak. Similar to the hypothetical Freyberg case study (Section 3.1), the objective for the WIPP case study is to identify transmissivities (product of the aquifer thickness and its saturated conductivity) for the Culebra dolomite aquifer given field measurements of transmissivities and steady state hydraulic heads.

This rest of this section is divided into two parts. The first section gives a brief overview of the prevalent hydrogeology at the WIPP site. The second section describes the groundwater model that is used in this study. Additional information about the data and the statistical assumptions for this model are provided in Appendix A.

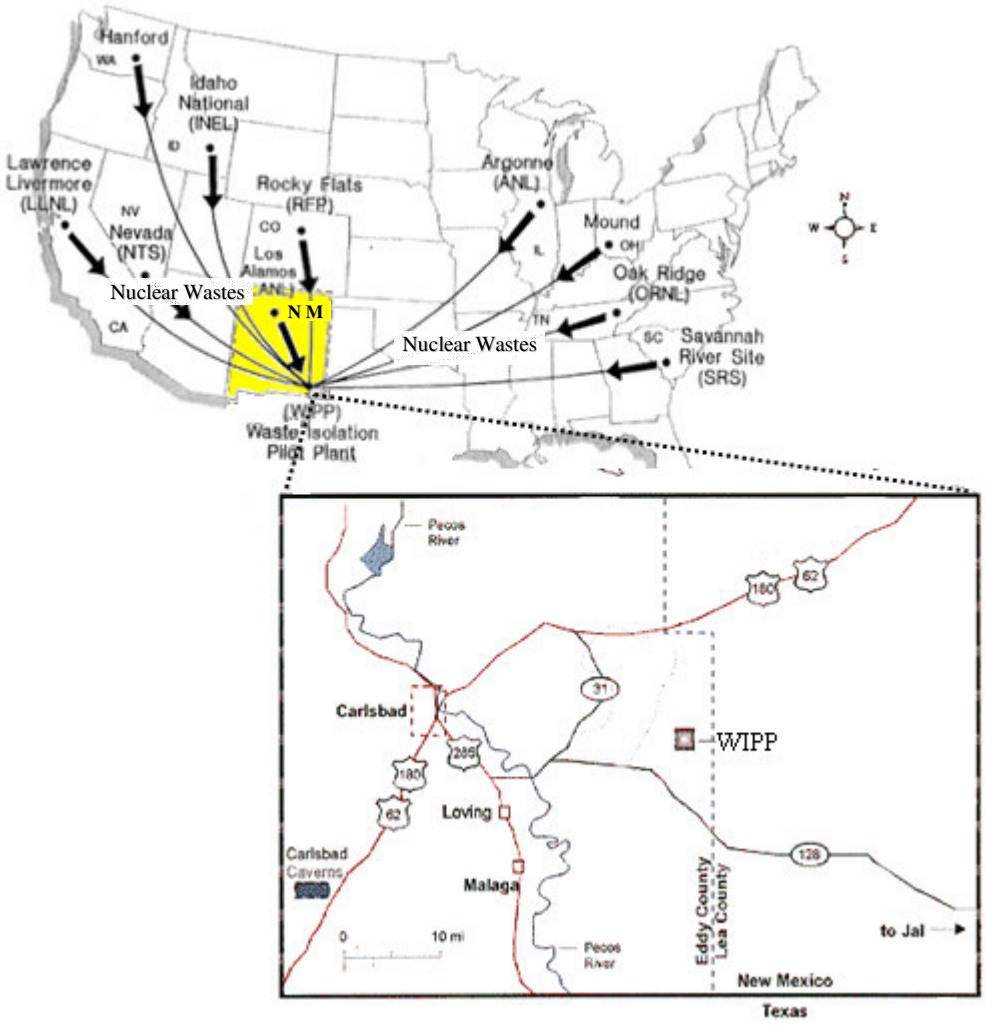


Figure 3.3 Location of the WIPP facility (arrows indicate nuclear wastes transported from different defense laboratories around the country)

3.2.1 Hydrogeology of WIPP Site

The WIPP site is located in evaporite-bearing sedimentary Delaware basin in south-eastern New Mexico. The waste repository itself is located about 650 m below the land surface in the lower part of the predominantly halite Permian Salado formation (Figure 3.4). At the WIPP site, the primary hydrologic unit of importance is the Rustler formation (directly overlying the Salado Formation, and approximately 300 m above the repository)

because it contains the most transmissive units above the repository. Of the five geological units of the Rustler at the WIPP, the Culebra and the Magenta are considered conductive while the others are considered confining units. The transmissivity of the Culebra member ($1x10^{-3}$ to $1x10^{-9}$ m^2/s) is 2 to 3 times more than the Magenta unit ($4x10^{-4}$ to $1x10^{-9}$ m^2/s), and is thus treated as the primary pathway for any contaminant leaking from the WIPP repository. Most modeling exercises [*LaVenue and Pickens*, 1992; *RamaRao et al*, 1995; *LaVenue et al*, 1995; *Capilla et al*, 1998; *Zimmerman et al*, 1998; *DOE/WIPP*, 2004] have focused on modeling the fate and transport of contaminants in the Culebra layer.

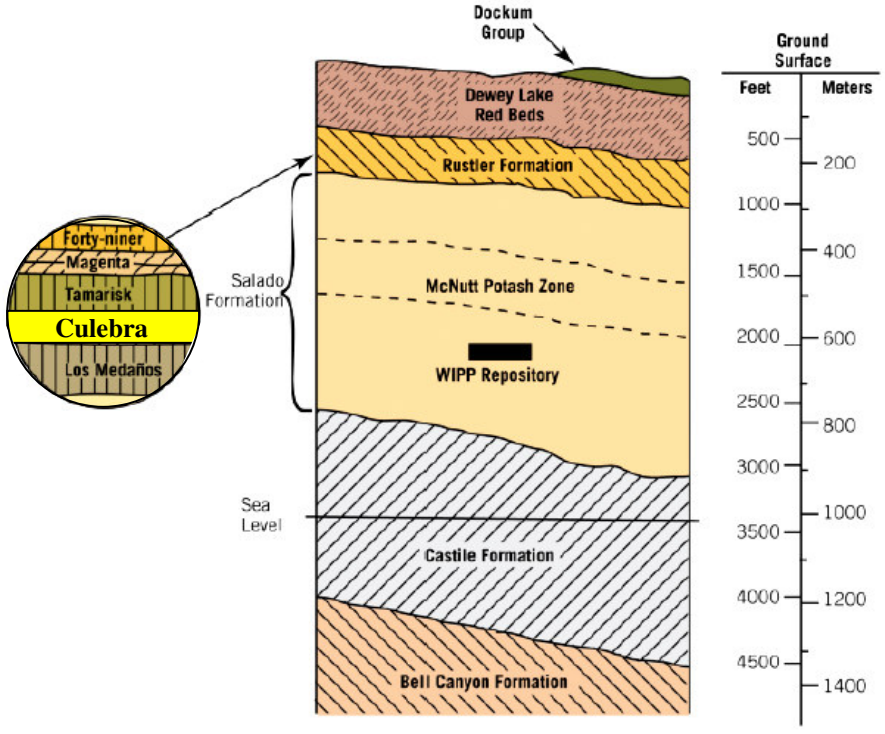


Figure 3.4 WIPP Repository and Geological Strata at the Site

3.2.2 The WIPP Model

This work considers the confined, single-layer, steady state groundwater flow model of the Culebra aquifer developed for the WIPP compliance recertification application [*DOE/WIPP*, 2004]. The model domain is 30.7 x 22.4 km^2 consisting of a total of 68,768 cells (307 north-south by 224 east-west), each being 100 x 100 m^2. Fixed head boundaries exist north, east, and south of the model domain, and a no-flow boundary exists on the west (these boundary conditions are shown by the purple lines in Figure 3.5). The no-flow boundary approximately passes through the middle of the 'Nash Draw' (a huge depression in the land surface, west of which Karst topology is thought to exist). The entire exercise is based on the same assumption given in DOE/WIPP [2004] – the "*…permeability of the Culebra can be adequately modeled as a continuum at the 100 m (328 ft) × 100 m (328 ft) grid block scale and the measured T values used to condition the model are representative of the T in the 100 m (328 ft) × 100 m (328 ft) grid block in which the well test was performed. Implicit in this assumption is the prior assumption that the hydraulic test interpretations were done correctly and used the correct conceptual model.*"

The Culebra transmissivities are thought to exist in three geological zones, with the highest transmissivities on the western edge of the site and the lowest transmissivities on the eastern edge. A conceptual model of the transmissivity zones based on the hydrogeology of the site was developed in *DOE/WIPP* [2004]. This conceptual model was based on a linear regression model that related transmissivity to the a) thickness of overburden above Culebra, b) dissolution of the upper Salado formation, c) spatial

distribution of Halite in the Rustler formation above and below the Culebra aquifer, and d) fracture interconnectivity within the Culebra aquifer. Of these, the first three factors are treated deterministically while the forth is treated stochastically (multiple equally likely indicator simulations of fracture zones are created since there is little data to assess the true fracture interconnectivity in the aquifer – see *DOE/WIPP* [2004] for details). Based on these factors, 'base fields' were created encapsulating all these relationships (see *DOE/WIPP* [2004] and *Holt et al.* [2005] for further details on how the base fields were created). In the *DOE/WIPP* [2004] study multiple base fields were used, however only the average of all of these base fields was made available for this study.

The average base field, which has three distinct transmissivity zones, is shown in Figure 3.5. The zone marked in red on the west corresponds to what is known as the 'Salado dissolution zone', where the Salado layer has dissolved, leading to the subsidence and fracturing of the Culebra and, consequently, higher transmissivities. The zone on the east marked in blue is bounded by the presence of halite found in cores taken from the Rustler formation (both above and below the Culebra). High transmissivities in the Culebra would presumably lead to more dissolution of Halite in and around the Culebra and thus the presence of halite is thought to be inversely related to Culebra transmissivities (for more details on the formulation and calibration of this linear model for the base field see *DOE/WIPP* [2004]).

The data available for this case study consists of 35 equilibrated head (H) measurements (from the year 2000) and 46 transmissivity (T) measurements based on pumping tests

conducted in the Culebra aquifer (of these, 28 measurements are co-located). These measurements are shown as the yellow triangles (head measurements) and blue squares (transmissivity measurements) in Figure 3.5. The data are also tabulated in Table A.1 shown in Appendix A). The object of calibration for the WIPP site is to 'realistically' modify the base field such that it is conditioned both on a) direct measurements of T values and b) head measurements. All other model parameters and boundary conditions are considered 'known' and are fixed at the values used in the original *DOE/WIPP* [2004] model.

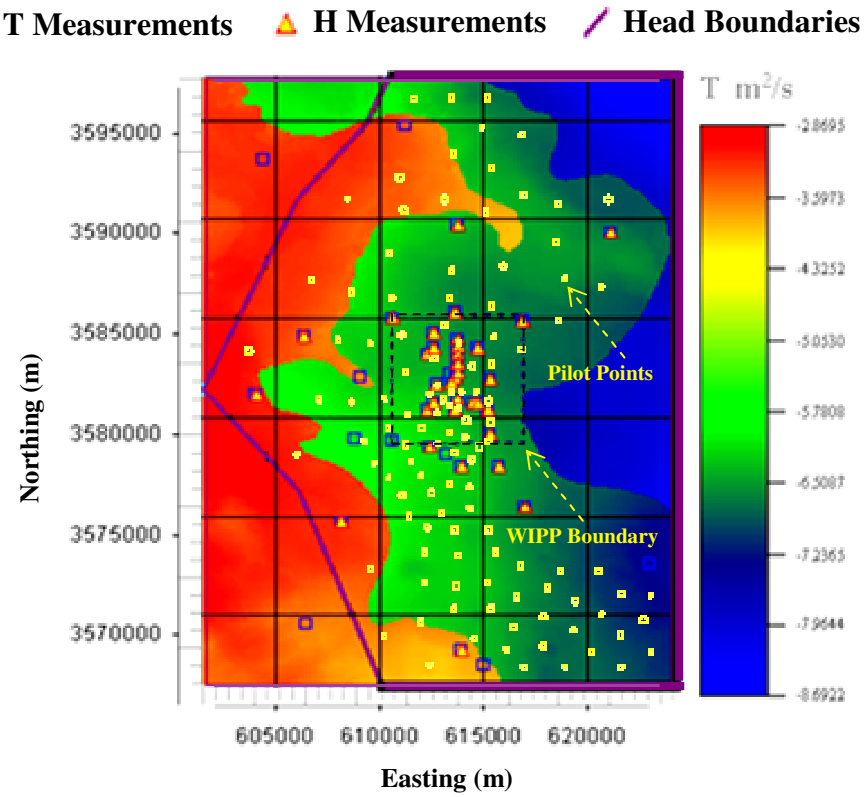


Figure 3.5 Groundwater Model for Culebra Aquifer

As with the hypothetical case study, pilot points [*Doherty*, 2003] are used to calibrate the Culebra transmissivities (pilot points were also used in *DOE/WIPP* [2004]). However,

unlike the *Freyberg* [1988] case where the pilot points represented log-conductivities directly, here they represent *deviations* from the given base field. The goal of the optimization algorithm is to find the best way to perturb the base field at the pilot point locations, while keeping all the objectives and constraints in mind. Details about the parameterization and formulation for the optimization are given in the methodology section (Section 6.2) of Chapter 6.

A total of 115 pilot points are used for this case (see Figure 3.5). The number and locations of these pilot points is kept the same as that given by *DOE/WIPP* [2004]. In general, the pilot points were located along a regular grid approach. However, the geometry of specific pilot points was changed to accommodate pumping and observation wells so that a pilot point existed between each pair of observation and pumping well (this is in keeping with recommendations given in *Doherty* [2003] and laid forth in *McKenna and Hart* [2003]).

3.2.3 Predictive Uncertainty Analysis for the WIPP site

Once the transmissivity fields for the WIPP model have been calibrated, uncertainty analysis consists of predicting the path and travel time of a conservative solute to assess the impact transmissivity has on contaminant pathways within the Culebra aquifer (in the case of a leak from the WIPP repository). For this purpose a particle tracking model is used to simulate the path and travel time of a particle released approximately above the center of the WIPP waste panels (the exact coordinates of the release location being UTM (Universal Transverse Mercator) X = 613, 597.5 and Y = 3,581,385.2; model grid X = 120, Y =158). *DOE/WIPP* [2004] used DTRKMF, a semi-analytical particle tracking

software created by *Rudeen* [2003] for this purpose; however, this model is available only to the Sandia National Laboratories so MODPATH [*Pollock*, 1994] – a public domain software that computes three-dimensional flow paths based on output from MODFLOW – is used in this work, instead. The particle tracking model used a constant advective porosity of 0.16 - the same value used in *DOE/WIPP* [2004] calculations. As in *DOE/WIPP* [2004], the final goal is to identify a cumulative distribution function for the travel times to assess the effect uncertainty in the transmissivity fields has on particle paths and travel times. Note that, as observed by *DOE/WIPP* [2004], the travel times are primarily for the comparison of different transmissivity fields and establishment of relationships between spatial trends in the transmissivity fields and travel time for particles, not for estimating the true travel times for contaminants in the field.

It is interesting to note that historically it was with the emergence of the WIPP project that predictive uncertainty analysis for contaminant transport became of practical importance. By Federal law, the WIPP project had to undertake stochastic analysis and consider alternate models to determine if the nuclear wastes stored at the site could accidentally pollute the groundwater. Inverse modeling techniques undertaken for the WIPP site all addressed predictive uncertainty [*LaVenue and Pickens*, 1992; *RamaRao et al*, 1995; *LaVenue et al*, 1995; *Capilla et al*, 1998]. *Zimmerman et al* [1998] compared the predictive performance of seven geostatistical inverse approaches for a test case that was based on the WIPP site. Subsequently, many other studies (such as *Rubin* [1991], *Medina and Carrera* [1996], *Hill* [1998], *Woodbury and Ulrych* [2000], *Moore and Doherty* [2005], and *McKenna et al* [2003]) were also undertaken to address this issue.

4 DEVELOPING AN INTERACTIVE MULTI-OBJECTIVE FRAMEWORK FOR GROUNDWATER INVERSE MODELING

We will surely get to our destination if we join hands

~Aung San Suu Kyi

4.1 Introduction

Mathematically well-posed problems have a solution that a) exists, b) is unique, and c) is stable with respect to small variations in the data. In contrast, the groundwater inverse problem is known to be an 'ill-posed' problem [*Yakowitz and Duckstein*, 1980; *Carrera and Neuman*, 1986; *Sun*, 1995; and *Zimmerman et al.*, 1998]. In effect, this means that the estimated conductivity field is often non-unique and highly sensitive to small changes in head measurements. In such cases minimizing the difference between the predicted aquifer response and the measured aquifer response is not sufficient to guarantee a unique and stable conductivity estimate. Due to this non-uniqueness there is no guarantee that the estimated conductivity field estimate is close to the 'true' conductivity, leading to unreliable and potentially erroneous model predictions. In other words, reducing the calibration error in itself is not sufficient and it is necessary to define an additional objective that ensures the conductivity estimate is close to reality. *Neuman* [1973] defines this additional 'plausibility' objective as an '*objective or subjective estimate of the expected (in a statistical sense) parameter values as obtained from field test and other sources*'. Of course, since the true parameter field is unknown, the challenge is to define

an appropriate plausibility heuristic that measures how far the parameter is from reality. It is in this context that 'prior information' becomes of utmost importance to inverse modeling.

Studies, including *Woodbury and Ulrych* [2000], *Marsily et al* [2000], *Kowalski et al* [2004], and *Alcolea et al* [2006], have discussed and demonstrated the importance of prior information for inverse modeling. Prior information itself can be of three types. The first type consists of direct measurements of the parameters to be estimated. Such 'hard' data provides a direct measure of the true conductivity field and should be incorporated in inverse modeling. The second type of prior information - known as 'soft' or 'indirect' data - consists of indirect indicators such as topology, land-use, and geophysical images that can be used to give some information about the parameter of interest. In this study the primary concern is a third type of prior information – expert knowledge about the hydrogeology of the site. Due to sparse and noisy measurement data, expert knowledge is often critical in model calibration and selection. Experts may have knowledge about the geological and sub-surface characteristics of the site (such as predominant rock or soil types, geological history of the site, existence of high conductivity flow channels, etc) that may not be reflected in the field data. Direct and indirect prior information have been incorporated into inverse modeling using Bayesian approaches [*McLaughlin and Townley*, 1996; *Woodbury and Ulrych*, 2000; *Kowalski et al*, 2004] or by imposing preferred value 'regularization' constraints and objectives [*Alcolea et al*, 2006; *Moore and Doherty*, 2006]. Though both these approaches provide frameworks to include expert knowledge, this requires the expert to express essentially qualitative knowledge in purely

quantitative terms. Bayesian techniques require the expert to represent their knowledge of the site in terms of a 'prior' distribution of conductivity fields – something that the experts may not feel comfortable or confident doing. Regularization requires the expert to express their knowledge in terms of low-order metrics (such as room mean square error), objectives, and constraints. Such representation will always suffer from what is known as the 'blindness of norms' - the loss of information when multi-dimensional information is represented by a low-dimensional statistic or norm. In fact, *Brill et al* [1990] have demonstrated that optimizing a multi-dimensional problem in a lower dimensional space leads to loss of solutions that would be optimal in the higher-dimensional space. They considered a hypothetical three-objective problem and solved it successively using one and two objectives. Most of the optimal solutions for the three-objective problem projected into the inferior region for the two-dimensional objective space (see Figure 4.1). This was the motivation for their proposal of a 'joint cognitive' approach where computers were used to generate multiple optimal (albeit in the low-dimensional objective space) alternative solutions that human experts could search to find solutions that are optimal in the higher dimensional domain. In this case, the 'higher-dimensions' that are difficult to represent using low-order metrics are the spatial characteristics of the parameter field that is being estimated. Without this information, it is possible that the optimal solution found by optimizing low-order metrics would lead to solutions that do not display desirable spatial characteristics for the calibrated parameter field.

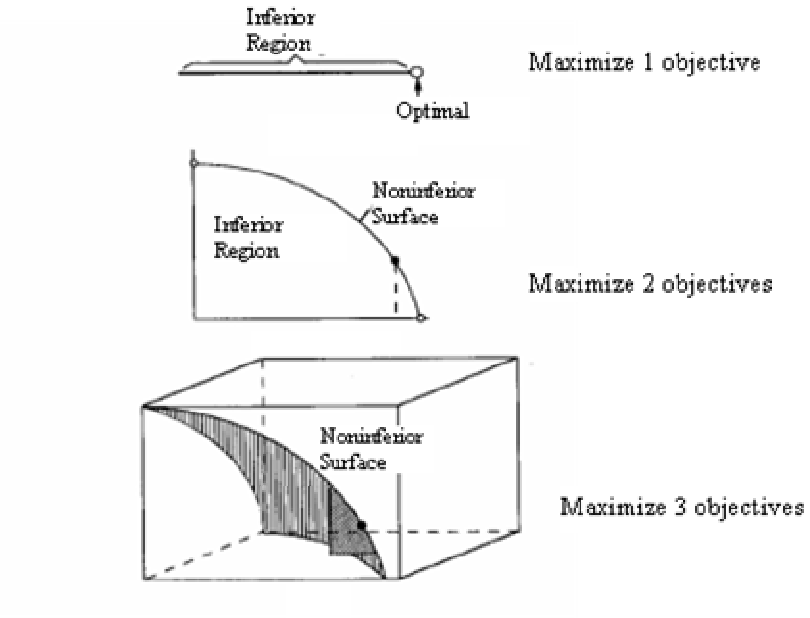


Figure 4.1 Optimal solutions in high dimensions project to inferior regions in lower dimensions (adapted from *Brill et al.* [1990])

Finally, expert knowledge is dynamic and adaptive and can evolve as the expert is allowed to interact and learn from the model. Thus, there is a need for a qualitative and adaptive environment that allows site experts to give maximum information during inverse modeling, while simultaneously evaluating and updating their own understanding of the model.

In this chapter, a novel interactive framework is proposed to adaptively incorporate qualitative expert knowledge in the inversion process. The proposed methodology is based on the concept of 'interactive optimization' [*Takagi*, 2001], where responses from users are used to drive the optimization search. Such an interactive framework is essentially similar to the 'joint cognitive systems (JCS)' proposed by *Woods et al* [1990] and *Brill et al* [1990] (see Section 2.4), combining the cognitive abilities of humans with

the computational powers of the computer to solve complex real-world problems. In the case of the groundwater inverse problem, expert knowledge is used to assess plausibility of conductivity fields, while the computational powers of the computer are used to optimize the quantitative calibration objectives and generate the best alternatives for the expert to evaluate.

To the best of our knowledge, this work represents the first attempt to investigate an interactive optimization framework for groundwater inverse modeling. Moreover, this approach introduces an important distinction from most previous interactive optimization approaches: instead of an over-riding 'human objective' (as used by *Wijns et al*, 2001 and *Takagi*, 2001), this work treats quantitative and qualitative criteria as multiple objectives within a Pareto optimization approach (this is similar to the multi-objective approach used by *Babbar et al.* [2006]). This is important for the inverse problem where quantitative objectives, such as calibration error, and agreement with measured prior data can be as important as qualitative expert preferences. Another advantage of such a multi-objective approach is that it allows one to evaluate the tradeoff between different kinds of information that feed into the inverse problem. With this methodology, which is called the interactive multi-objective genetic algorithm (IMOGA), it is possible to generate multiple alternatives for the parameter field, each representing an optimal tradeoff between different quantitative and qualitative criteria.

The goal of this chapter is to formulate, implement, and demonstrate the applicability and utility of the interactive optimization framework for model calibration. Since the goal at

this stage is a proof of concept, the IMOGA is tested on the hypothetical Freyberg [1988] case presented in Section 3.1. The hypothesis is that given user knowledge about the plausibility of the groundwater model, it is possible to incorporate this subjective knowledge into the inverse modeling process and identify improved (more plausible) solutions. It is also demonstrated that such knowledge can be useful for improving the generalizability and predictive power of the groundwater model.

4.2 Methodology

The different stages of model calibration – parameterization, formulation (including regularization), and optimization - have been discussed in Section 2.1. Each of these stages is discussed with respect to the Freyberg case study in the following paragraphs.

The parameterization chosen for this work is based on the concept of 'pilot-points'. Details of the pilot point implementation are given in Section 4.2.1. Once the parameterization has been selected, the next step is to formulate the objectives for the inverse problem. For this work, the quantitative objectives are given by the calibration error – a measure of the difference between the model predictions and measurements of aquifer response – and a physical plausibility objective – a measure of how far the conductivity field deviates from direct field measurements of conductivity. Finally, the user's subjective judgment of the plausibility of conductivity fields is treated as a third *qualitative* objective. Details on the multi-objective formulation are given in Section 4.2.2.

Given the decision variables and the quantitative and qualitative objectives, a two-stage optimization strategy is used to find the best conductivity fields. In the first stage, the inverse problem is solved using only the quantitative objectives (calibration error and regularization) using an efficient optimization strategy, ensuring that these objectives are adequately optimized. The interactive phase then starts when the non-interactive stage ends and allows the expert to further optimize the solutions by incorporating his or her subjective knowledge, while still ensuring good performance on the quantitative objectives. Details about the interactive optimization framework are given in Section 4.2.3.

4.2.1 Parameterization

The parameterization used for this study is based on the pilot point methodology proposed by *Doherty* [2003] and uses a large number of pilot points. The conductivity field is assumed to be log-normal and consequently all conductivity measurements are log-transformed (even though the methodology is presented for log-conductivity estimates it applies without any loss of generality to other transformations). Ordinary kriging [*Deutsch and Journel,* 1998] is used to interpolate the pilot point values *and the conductivity field measurements* to get the final conductivity field, ensuring that the resulting field honors the conductivity measurements exactly but deviates from the kriged field in different ways at the pilot point locations (see Figure 4.2). Details of the kriging methodology are given in Appendix B.

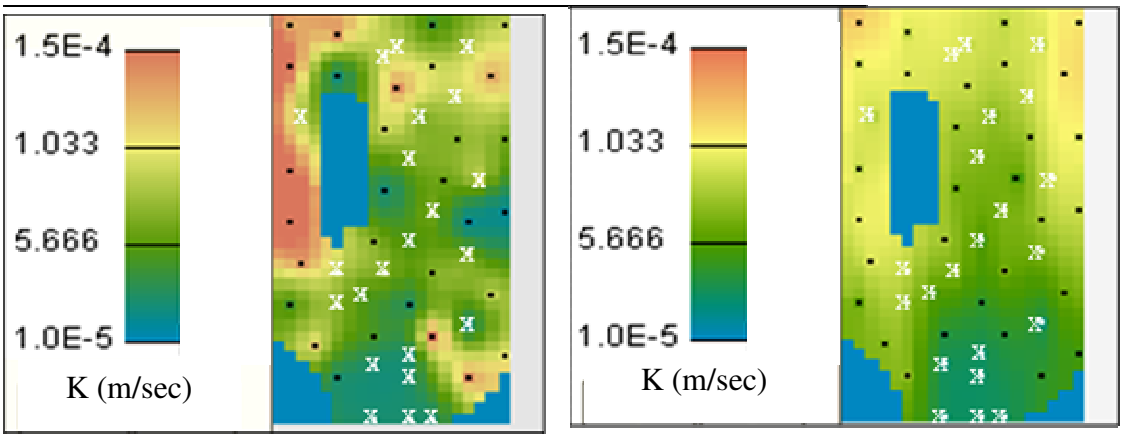


Figure 4.2 Two conductivity fields that fit the measured values (white crosses) exactly, but have different values for the pilot points (black dots) resulting in different spatial fields

The conductivity field obtained from kriging the data and the pilot points is then used to predict hydraulic heads using a groundwater flow model (MODFLOW [*McDonald and Harbaugh*, 1988] in this case). These predictions are then used to calculate calibration errors for inverse modeling. Different values for pilot points give different conductivity fields, all of which match the field measurements but give different calibration errors and deviate from the 'ground-truth-data' conductivity field (as given by the best estimate from the prior conductivity data – see Section 3.1 for details) to different degrees. A major goal of the calibration can be defined as finding 'plausible' conductivity field(s) with the best calibration error and minimum deviation from the prior field.

4.2.2 Calibration Formulation

This work follows the multi-objective formulation proposed by *Neuman* [1973] to optimize the conductivity field using *both* the quantitative objectives of calibration error

and plausibility. A third qualitative objective is used during the interactive phase to include input from the site expert.

The calibration objective consists of minimizing the calibration error (H_{err}) with respect to field measurements of hydraulic heads. This is given by the weighted root mean square error (RMSE) between measured field data and the prediction of the groundwater flow model for a given conductivity field (Equation 4.1):

$$\underset{PP \in R^P}{Min} \quad H_{err} = \left(\frac{\left[GW\left\langle K_{PP} \right\rangle - H^{obs} \right]^T C_{err}^{-1} \left[GW\left\langle K_{PP} \right\rangle - H^{obs} \right]}{n_{obs}} \right)^{1/2} \tag{4.1}$$

where PP is a vector of p pilot points, H^{obs} is the head measurement vector, C_{err} is the head measurement error covariance matrix, and $GW\left\langle K_{PP} \right\rangle$ is a vector with groundwater model predictions of the head values using the kriged field K_{PP} from the pilot points PP that are considered as decision variables. It is typical to assume that C_{err} is dependent on the measurement accuracy of the instruments used for the head measurements and can thus be assumed to uncorrelated and unbiased (leading to a diagonal matrix for C_{err} with measurement error variance along the diagonal). Since the measurement error is known for the Freyberg case this matrix is known for this problem.

As discussed earlier, when using a large number of pilot points it is necessary to impose some form of 'regularization' [*Moore and Doherty*, 2006; *Alcolea et al*, 2006] to reduce unwarranted variations in the pilot point values. *Moore and Doherty* [2006] and *Alcolea et al* [2006] have proposed a regularization scheme that aims at minimizing the deviations of the pilot point values from some prior conductivity field (that can either be

specified by the expert or be obtained from existing field data). This regularization is essentially a *quantitative* measure of the plausibility of the pilot point values with respect to direct field measurements. For the regularization objective it is first necessary to come up with a prior conductivity field as given by the measurement data. To do this, the conductivity measurements are kriged using ordinary kriging (using only the direct field measurements of conductivity) at every pilot point location. This gives the best linear unbiased estimate of the log conductivity $K_{pp}^{'}$ (note that here as elsewhere the notation has been simplified by using K to be synonymous with Ln(K) - the conductivity is assumed to belong to a log-normal distribution and is kriged in log space) at each pilot point location. The regularization objective (K_{err}) is given by the weighted root mean square of the difference between the pilot point values and the kriging estimate at those locations (Equation 4.2):

$$Min \quad K_{err} = \left(\frac{\left[K_{pp} - K_{pp}^{'}\right]^T C_{pp}^{-1} \left[K_{pp} - K_{pp}^{'}\right]}{n_{pp}} \right)^{1/2} \tag{4.2}$$

where K_{pp} is the log conductivity value at the pilot points, $K_{pp}^{'}$ is the kriging estimate at the pilot point locations based on the conductivity data, n_{pp} is the number of pilot points, and C_{pp} is the estimation error covariance for the pilot point locations. An important distinction between this study and previous work (such as *Doherty* [2003]) is that instead of the typical diagonal matrix (with the diagonal elements specifying the kriging error variance) C_{pp} is considered to be a full matrix specifying the *estimation error covariance* – thus ensuring that the pilot point values fit the specified covariance structure to the extent possible. The covariance between the estimation error at pilot point i and pilot point j given by C_{pp}^{ij} is calculated by the equation below:

$$C_{pp}^{ij} = \sum_{\alpha=1}^{n_i}\sum_{\beta=1}^{n_j}\lambda_\alpha\lambda_\beta Cov\left(u_o^\alpha - u_o^\beta\right) + Cov\left(u_{pp}^i - u_{pp}^j\right) - \sum_{\alpha=1}^{n_i}\lambda_\alpha Cov\left(u_o^\alpha - u_{pp}^j\right) - \sum_{\alpha=1}^{n_i}\lambda_\alpha Cov\left(u_o^\alpha - u_{pp}^j\right) \quad (4.3)$$

where C_{pp}^{ij} is the covariance between the estimation error at the location of pilot point i (located at u_{pp}^i) and the estimate at pilot point j (at u_{pp}^j), *Cov* is the model covariance function used for kriging, λ_α is the kriging weight and u_o^α the location of the α^{th} data point used for kriging at pilot point *i*, λ_β is the kriging weight and u_o^α the location of the β^{th} data point used for kriging at pilot point *j*, and n_i and n_j are the number of data points used for kriging at pilot point locations *i* and *j*. The derivation of this covariance equation is given in Appendix B. The first term in equation 4.3 gives the covariance between all of the data points used for kriging the priors for the two pilot points, the second term gives the expected covariance for the distance between the two pilot points, and the third and the fourth terms give the covariance of the data point used for kriging one pilot point with the location of the second pilot point. Note that this formula reduces to the familiar kriging estimation variance for $i = j$ [*Deutsch and Journel,* 1998]. Weighting the objective function with such a covariance matrix ensures that a) pilot points located close to data points are highly correlated to those values and b) pilot points are more correlated with nearby pilot points than with more distant points. Higher values of the regularization objective typically mean more deviation from the field data, and more heterogeneity in the conductivity field.

In addition to the two quantitative objectives (given by Equations 4.1 and 4.2), there is an additional qualitative objective that is given by the human expert. The qualitative objective is obtained by showing the kriged conductivity fields and the predicted head field to the user along with information about the numerical objectives. The expert then ranks different solutions based on how realistic they seem based on his or her knowledge about the site. This process is discussed in further detail in the next section.

4.2.3 The IMOGA

To solve the formulation given in Section 4.2.2, the optimization process consists of two phases: non-interactive and interactive multi-objective optimization. The non-interactive phase is similar to other groundwater inverse methodologies such as *Alcolea et al* [2006] and *Moore and Doherty* [2006] except for one important difference. Previous approaches have either combined the calibration objective with the regularization objective using a weighting scheme [*Alcolea et al*, 2006], or specified one as a constraint and used the other as an objective for optimization [*Moore and Doherty*, 2006]. The constrained optimization problem is then solved using a non-linear optimization algorithm. In this framework, the two objectives of regularization and calibration are treated *independently* within a Pareto-optimization framework. Such an approach converges to the non-dominated solution set (solutions that can not be improved on one objective without worsening the other), allowing the expert to evaluate and select the best tradeoff between calibration error and physical plausibility of the conductivity fields. An added advantage of this multi-objective approach is that it does not require the prior specification of weights, penalties, or constraint levels for optimization.

For this framework an efficient multi-objective genetic algorithm called the elitist non-dominated sorting genetic algorithm (NSGA-II) [*Deb et al*, 2000] is used to find the Pareto optimal set of solutions (see Section 2.2.1 for details on the NSGA-II). During this first non-interactive optimization phase, a large population size and number of generations are be used to effectively solve for the quantitative objectives.

Once this non-interactive session has converged, the expert then examines the estimated conductivity fields and the calibration targets for the optimal solutions. If the expert believes there are certain aspects of the estimated parameter fields that have not been captured or have been misrepresented, he or she initiates the next phase – the interactive multi-objective optimization. Figure 4.3 gives an overview of the IMOGA framework. The starting population of the IMOGA is initialized with solutions from the optimal tradeoff front of the non-interactive run, as well as other randomly generated conductivity fields. This ensures that the expert has a good starting set of solutions to search from, while the GA has sufficient diversity to effectively explore the region around the quantitatively optimal solution space.

In each generation of the IMOGA, the decision variables (consisting of the pilot points and kriging parameters) are varied to create different 'images' of the hydraulic conductivity field. A particular conductivity field leads to a unique hydraulic head field. Both of these fields are then evaluated for the quantitative objectives of calibration error and regularization. The qualitative objective is supplied by the expert after visual inspection of the conductivity and head fields through a graphical user interface - see

Figure 4.4. As can be seen from Figure 4.4, the kriged conductivity and predicted hydraulic head field (corresponding to that conductivity field) are shown side by side. Additional information such as the numerical objectives (labeled as H Err and K Err in the grey panels above the images for conductivities and head fields), previous human rank of the solution (set to 0 if the solution has not been previously ranked), and other case-specific details are also shown to the expert. The user gives these solutions a rank from 1 (best) to 5 (worst). It is noteworthy that this two-phase approach is consistent with the 'joint cognitive' philosophy of *Brill et al* [1990] who reasoned that *'if the most important dimension or dimensions* (as expressed by the expert criteria) *are included in the model* (or mathematical formulation)*, then the problem solution*(s) *is likely to be in the region of the optimal solution'* of the quantitative formulation [*Brill et al*, 1990].

Given these three objectives, the IMOGA uses the same operators (crossover, mutation, Pareto selection) as the NSGA-II to find the best tradeoff between both quantitative and qualitative criteria. The generations are repeated either until a maximum generation (specified by the expert depending on how much time he or she wants to spend on the interaction) or until the expert is satisfied with the solutions found by the IMOGA.

It is worth noting that this multi-objective approach is different from earlier interactive optimization approaches used in other fields, which either considered only the human objective [*Wijns et al.*, 2001] or used it as an over-riding objective to select solutions from the quantitative tradeoff curve [*Parmee et al.*, 2000]. Both these approaches place

the entire onus of the optimization search on the user, neglecting the quantitative objectives, which for the inverse problem are also very important.

The IMOGA framework has been built within the Data 2 Knowledge (D2K) system from the National Center for Supercomputing Applications (NCSA), IL. D2K is a visual programming system that supplies a core set of data mining and machine learning modules, application templates, and a standard API (application programming interface) for further software component development. All D2K components are written in Java for maximum flexibility and portability. See *Welge et al.* [2003] for more details on the D2K software.

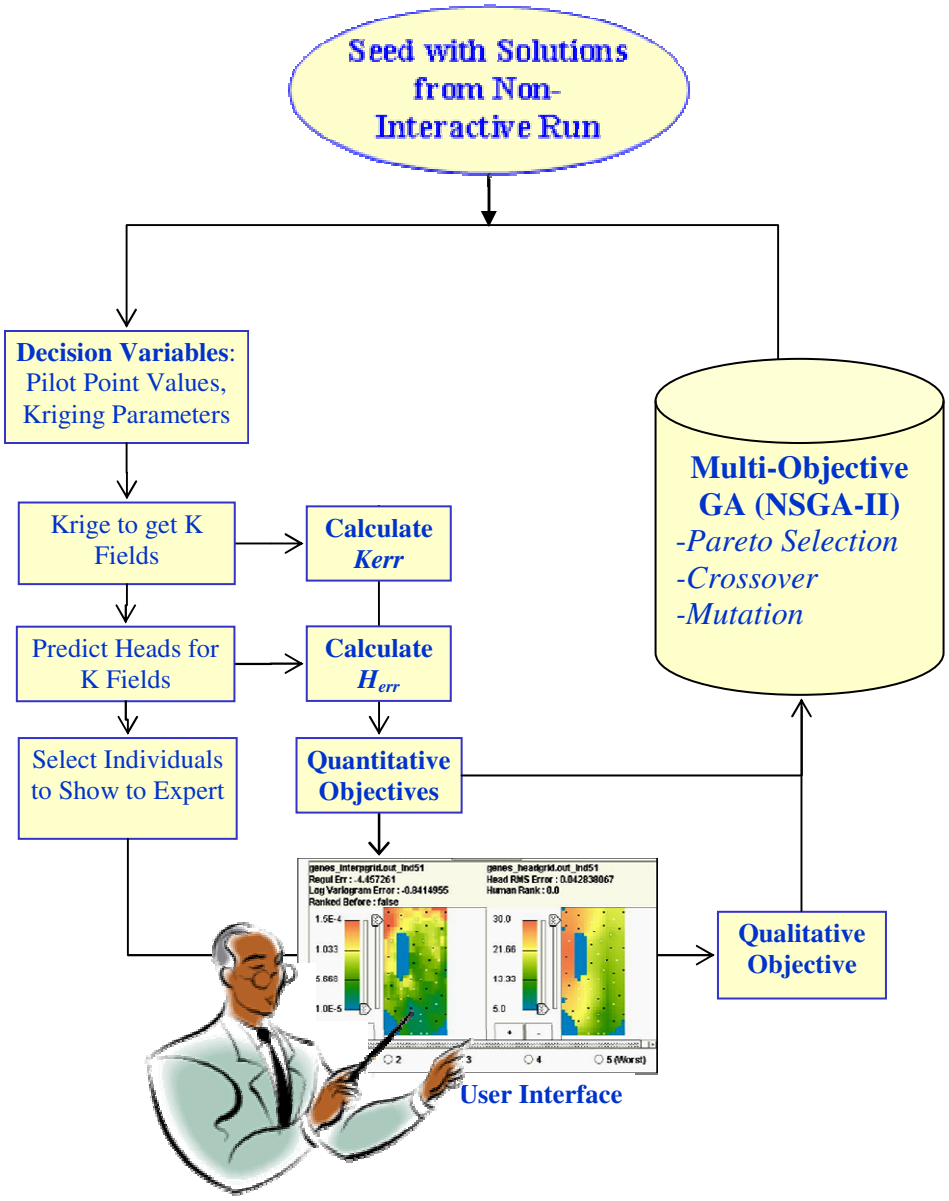


Figure 4.3 IMOGA framework for interactive multi-objective groundwater inversion

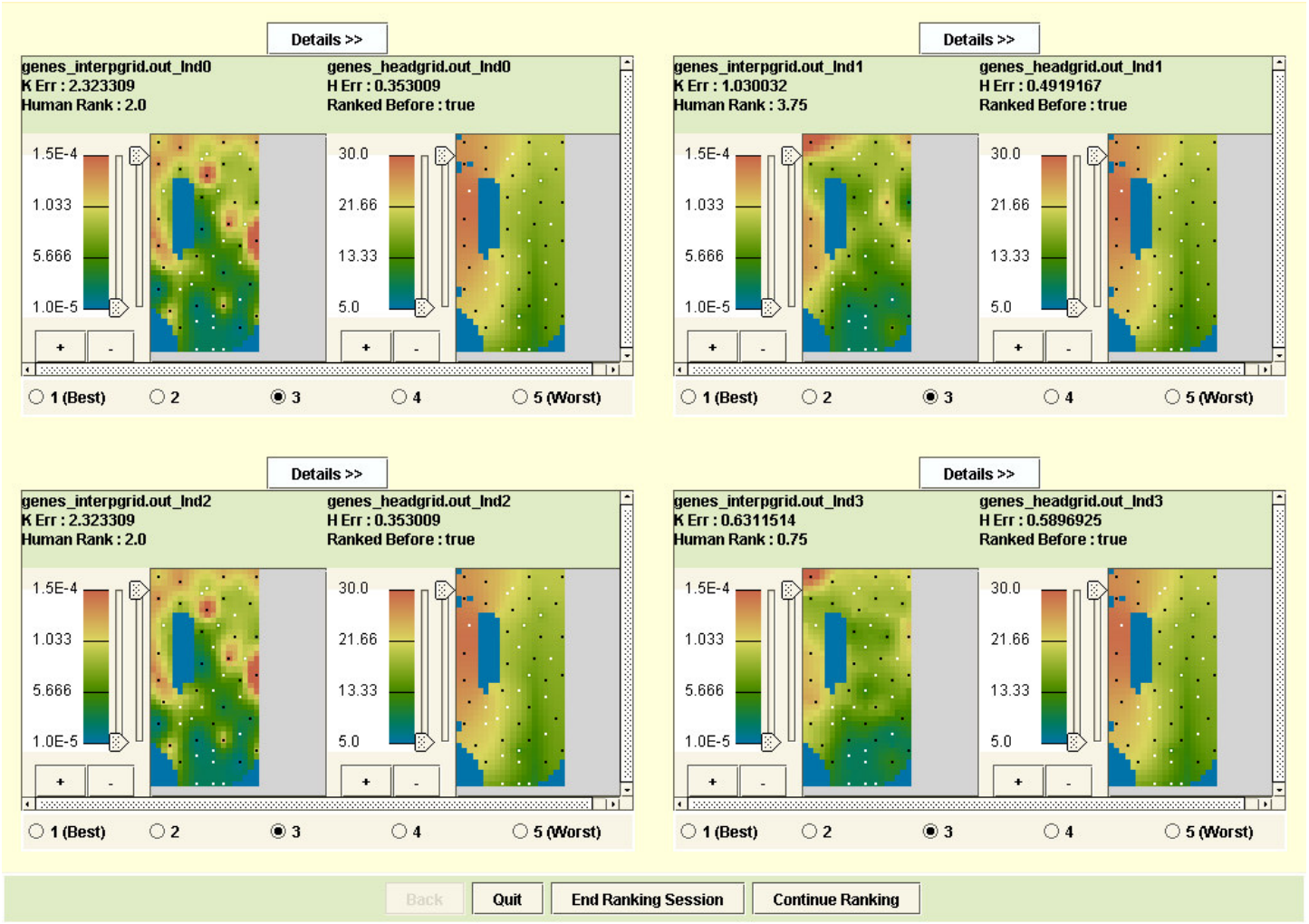


Figure 4.4 Ranking panel for the IMOGA (each panel displays 4 unique IMOGA solutions with conductivity on the left and hydraulic head field on the right)

4.3 Results and Discussion

The Freyberg case (discussed in Section 3.1) was solved using the two-phase multi-objective interactive approach discussed in section 4.2. For the initial non-interactive NSGA-II, a population size of 250, tournament size of 2, crossover probability of 0.5, and mutation rate of 0.004 were used. The population size for this problem is set using the iterative framework proposed by *Reed et al.* [2003], where small population sizes are iteratively increased till there is no further addition of information for the multi-objective Pareto front. The other parameters are set using a trial-and-error approach.

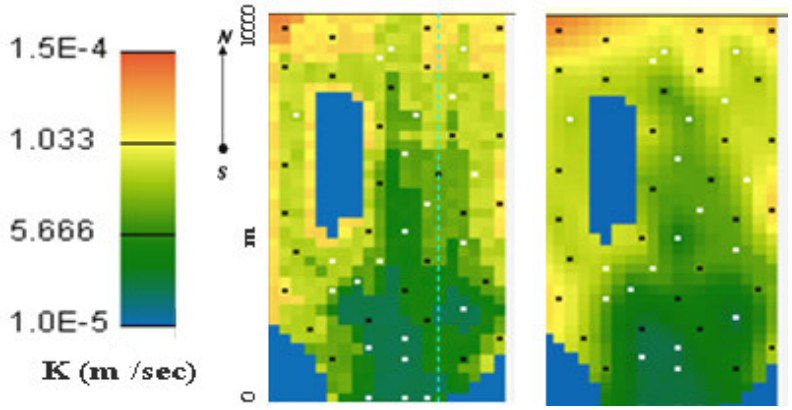
The interactive phase provides the expert an adaptive framework to incorporate his or her preferences and site-knowledge to improve these results. Since this was a hypothetical aquifer the expert (the author) had knowledge of 'reality'. To make the human responses more realistic, the true conductivity field was only shown to the expert in advance (before both the non-interactive and interactive phases). The expert was not allowed to make direct comparisons between the true field and the conductivity fields shown by the IMOGA but rather was forced to work from memory. This is similar to real applications where experts may have secondary soft information in the form of maps and/or satellite images and also their own field experience to help them in decision making. Despite the fact that the case-study and expert-knowledge were 'synthetic,' this experiment serves as a proof of concept that subjective expert knowledge can be incorporated effectively with quantitative objectives. It also provides a standardized test-bed where different algorithmic approaches can be compared.

For the interactive phase, small population sizes are necessary so that the user could evaluate the solutions without being overwhelmed by the amount of interaction. For this reason, the population size for the IMOGA was set at 20 individuals with 25 generations. Half the IMOGA population was initialized from the results of the non-interactive run and the other half was randomly initiated. This selection strategy was chosen based on recommendations given by *Babbar* [2006] for initializing interactive genetic algorithms. To better explore the search space, the crossover and mutation rates were kept relatively high at 0.9 and 0.5, respectively.

The results for these experiments are presented in four parts. Section 4.3.1 discusses the results obtained from the non-interactive multi-objective optimization. Section 4.3.2 then compares results from the interactive approach to the non-interactive results. Section 4.3.3 presents results for the predictive scenario (defined in Section 3.1) for both the non-interactive and interactive conductivity field estimates. Finally, to further explore the importance of user interaction for problems with sparse data, the number of observations was reduced (from 22 to 17). The model was then calibrated with this reduced data set using both the interactive and non-interactive approach. These results are presented in Section 4.3.4.

Before presenting these results it is important to consider the best inverse modeling solution possible for the case study considered in this paper. This solution, representing the closest approximation to the real conductivity field given the number and locations of pilot points available, is obtained from the true conductivity values at all of the pilot

points. In essence this represents the solution achieved with perfect knowledge of all of the decision variables for this optimization problem, which is the 'true' optimal solution to this groundwater calibration problem. The difference between the ground-truth reality (shown in Figure 3.1) and the true-optimal case can be attributed to 'model structure' errors [*Doherty and Moore*, 2005] that arise due to the parameterization used and approximations in the conceptual model itself. Of course, this solution is only known for this hypothetical case study where the ground truth is known. The optimal pilot-point conductivity field is compared with the true conductivity field in Figure 4.5. As expected, compared to the 'true' conductivity field, the optimal pilot-point field is smoother due to the effect kriging has on the interpolated field (this smoothening is the source of the so-called structural error for this case). However, the overall spatial characteristics of the true conductivity field are preserved. Since the optimal conductivity field in Figure 4.5.b is the ideal solution that can be obtained with the current parameterization, it serves as a standard with which to compare the inverse modeling results.



(a) **(b)**

Figure 4.5 (a) True conductivity field and (b) optimal pilot-point conductivity field

4.3.1 Results for Non-Interactive Optimization

Figure 4.6 shows the optimal tradeoff between the calibration error (as measured by H_{err}) and the regularization/plausibility objective (as measured by K_{err}) for the non-interactive optimization. It is worth noting that constrained single-objective calibration approaches using pilot points (such as PEST [*Doherty*, 2003]) would converge to one point on this tradeoff front, depending on the constraint level set for one of the objectives. The graph also shows the calibration error and regularization objective for the optimal pilot-point solution shown in Figure 4.5. The results shown in Figure 4.6 give interesting insights into the tradeoff between model accuracy and model complexity (as measured by the regularization term). In general it is seen that highly regularized conductivity fields (that are more homogenous and fit the K data better) lead to higher calibration errors, while fields that are under-regularized have lower calibration errors (among others [*McKenna et al.*, 2003] have discussed this problem of over-fitting for groundwater model calibration as well as the role regularization has in reducing it). Interestingly, the optimal pilot point solution is not part of the final Pareto front, and is in fact dominated by the solutions found by the NSGA-II. This finding is not as surprising as it may seem at first and, in fact, forms the basis for the interactive approach. The dominance of the optimal pilot-point solution indicates that there are other solutions with the same level of regularization that have lower calibration errors. This is primarily due to the ill-posedness of the groundwater inverse problem where small variations in head can lead to very divergent conductivity estimates (recall that the head measurements had a certain amount of noise). In regions with good data support regularization constraints the conductivity values at their optimal levels, but in areas with little to no data support the optimization

algorithm changes the conductivity field to better fit the noise in the head measurements, leading to (implausible) conductivity fields with lower calibration errors. Thus, with respect to the quantitative objectives of calibration error and regularization, the optimal pilot point solution is perceived as sub-optimal and is not part of the final Pareto front. In such a case, optimizing the conductivity field based on the quantitative objective alone would not find the true optimal solution, thus forming the motivation for the IMOGA's 'joint cognitive approach' [*Brill et al.*, 1996].

Figure 4.7 compares conductivity fields for three solutions taken from the non-interactive tradeoff front with the optimal pilot point conductivity field (Figure 4.6). These solutions represent conductivity fields with low K_{err} and high H_{err} (high plausibility and high calibration error), medium K_{err} and H_{err}, and high K_{err} and low H_{err} (low plausibility and low calibration error). Some important issues related to inverse modeling are evident in Figures 4.6 and 4.7. Notice that in terms of the spatial structure of the hydraulic heads, very different conductivity fields lead to almost identical head fields (the total range of head residuals is just 0.3 m from 0.25 to 0.55 m in Figure 6) indicating the problem of non-uniqueness for this example. It can also be seen from Figure 4.7 that as K_{err} increases the conductivity fields become more heterogeneous because the pilot points are less constrained. However, the more 'calibrated' conductivity fields (i.e. solutions with low calibration error) are clearly different from the true conductivity field, indicating that in this case regularization and not calibration error is the more important objective. In this way, comparing the quality of different solutions along the tradeoff front allows one to make decisions regarding the compromise between different objectives (something that

would not be possible with the standard approach of using single-objective optimization with a fixed regularization or calibration error constraint).

The solution closest to the true optimal field (Figure 4.7.a) is the most regularized solution with least K_{err} (Figure 4.7.b). However, even this conductivity field suffers from unrealistically high conductivities in the western portion of the site and low conductivities in the north-west portion of the aquifer. Unreasonably high conductivity values are observed in the west and north-west portions of the site for all other solutions along the non-interactive tradeoff front. These are the areas with the least data support and thus the quantitative plausibility objective (as measured by K_{err}) is insufficient to ensure realistic conductivity values in these regions. The other difficulty is that the plausibility of the conductivity field is based on a two-dimensional spatial field. Any one-dimensional metric, such as the regularization term used here, is bound to lead to some loss of information.

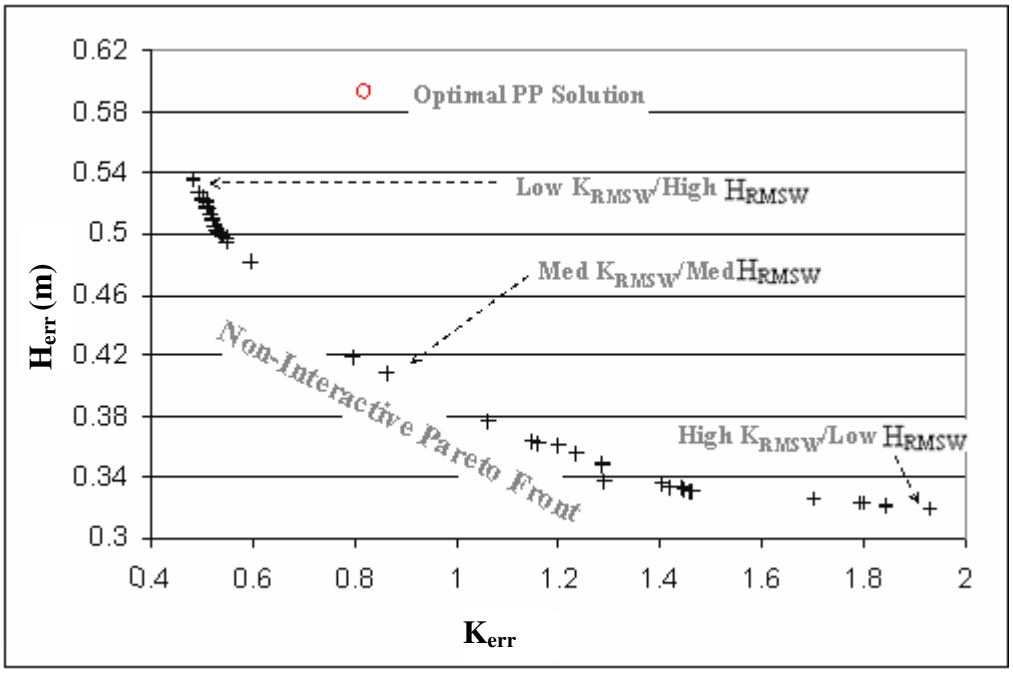
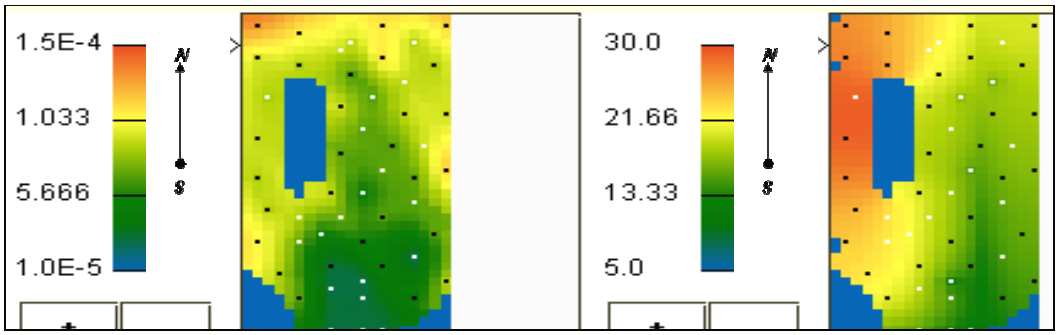
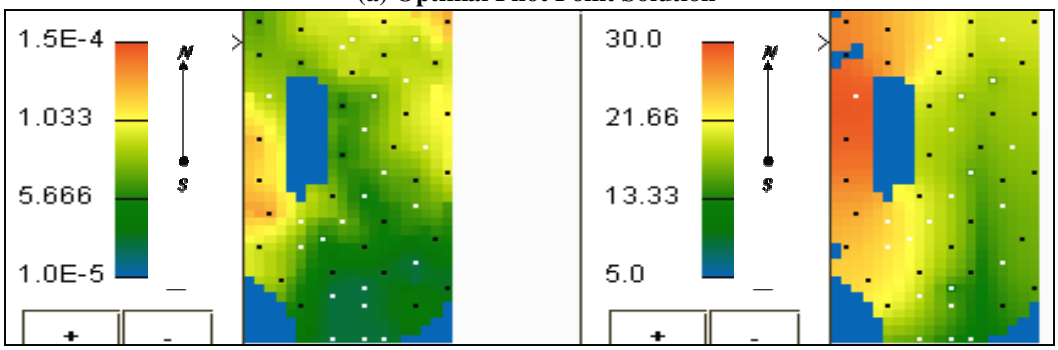


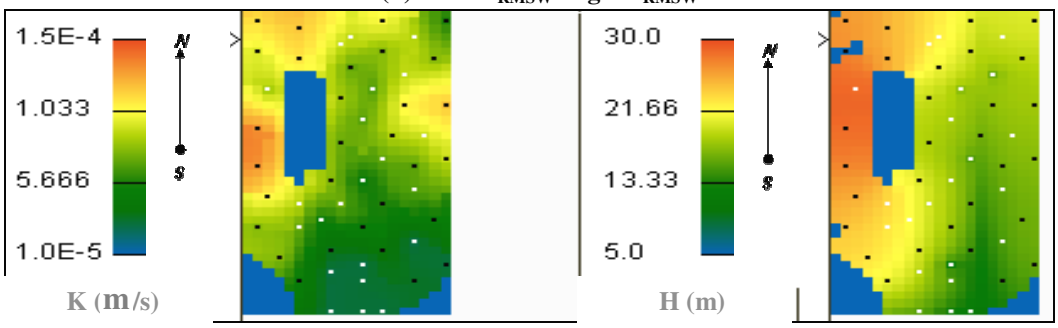
Figure 4.6 Tradeoff between calibration error and regularization objectives (for non-interactive optimization)



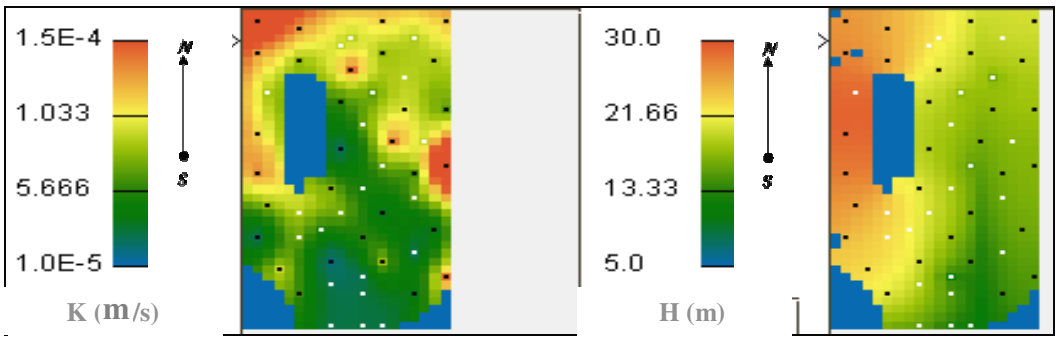
(a) Optimal Pilot Point Solution



(b) Low K_{RMSW}/High H_{RMSW}



(c) Medium K_{RMSW}/Medium H_{RMSW}



(d) High K_{RMSW}/Low H_{RMSW}

Figure 4.7 Comparison of conductivity (left) and head (right) fields found by non-interactive optimization (the units are m/s for conductivity and m for heads)

4.3.2 Results for Interactive Optimization

The results from the non-interactive phase indicate that quantitative objectives can not guarantee sufficiently plausible conductivity fields. When the non-interactive results are fed into the IMOGA, Figure 4.8 shows the resulting tradeoff curve. In this case there are three objectives – the first two, K_{err} and H_{err}, being the same as before, and a human rank (from 1 to 5). The first two objectives are displayed on the x and y axis of the graph and the human rank is shown by different colored icons on the graph. Because IMOGA works with a much smaller population, the tradeoff curve is sparser than the non-interactive results (Figure 4.6). The graph also shows that there is some tradeoff between the expert rankings and quantitative objectives. Thus, some low human rank (rank 3) solutions are on the quantitative Pareto front, and higher human-ranked solutions (ranks 2 and 1) are diverging from the quantitative Pareto front. In fact, rank 1 solutions (indicated by diamonds on the graph) are farthest from the quantitative Pareto front, and more interestingly the closest to the 'true' optimal solution, indicating that these conductivity fields are closest to 'reality'.

As mentioned earlier, this issue has been pointed out by *Brill et al* [1990] when discussing the need for interactive systems. When solving multi-criteria problems, many solutions that are optimal in higher dimensional space are perceived as sub-optimal in low-dimension objective space and are lost when optimizing with the low-order objectives. It is important to note that the multi-objective approach considers both quantitative and qualitative objectives and thus some solutions on the quantitative Pareto front are also retained.

These differences in objective space are also reflected in the spatial features of the conductivity fields founds by the IMOGA. Figures 4.9 a, b, and c show three solutions from the interactive Pareto front with different human ranks (these are shown on the Pareto front in Figure 4.8). It can be seen that while rank 1 solutions have the spatial structure preferred by the expert, rank 2 and rank 3 solutions deviate from the expert-preferred structure and are closer to the solutions found only by the non-interactive run. Thus, in one step the IMOGA was able to find solutions that represent different levels of compromise (trade-off) between spatial structures found plausible by the expert and by the quantitative plausibility measure as well as the calibration error. This trade-off is important information as it gives the expert the relative value of the field data with respect to their own understanding of the site and model. It also safe-guards this approach from possible biases of the expert by establishing conflicts between expert knowledge and field data. In a situation such as this one, the expert is forced to make a conscious and considered decision about the type of conductivity structure appropriate for the model.

Finally, Figure 4.10 compares the conductivity field of a Rank 1 solution (the solution found most plausible by the expert shown by Figure 4.10 c), the conductivity field with the lowest regularization/plausibility error from the non-interactive phase (Figure 4.10 b), and the true optimal pilot-point solution (shown earlier in Figure 4.5.b and shown again in Figure 4.10.a). Note that both (4.10 b and 4.10 c) these solutions have very similar calibration errors (0.58 m for the interactive solution and 0.54 m for the non-interactive solution). The interactive solution does not suffer from the high conductivity values seen

in the west portion of the non-interactive solution. Moreover, unlike the non-interactive solution, interactive inversion is also able to capture the high conductivity values in the north-west region seen in the original conductivity field.

To compare the results from the interactive and non-interactive approaches more thoroughly, two additional metrics were calculated. The first was a measure of the root mean square difference between the 'true' conductivity field and the estimated conductivity field. The second was the root mean square difference between the 'true' hydraulic head field and the head field as predicted by the calibrated models. Note that unlike the calibration error objective that only used head data at measurement locations, this head difference is calculated for the entire spatial domain of the model. Figure 4.11 shows the difference between the true conductivity field and the most plausible non-interactive (Figure 4.10 b) and interactive conductivity fields (Figure 4.10 c), respectively. As expected the figure shows the non-interactive conductivity field underestimates the conductivity (negative differences are shown in blue) in the north-west region and over-estimates (positive differences are shown in red) it in the west portion of the site. The interactive conductivity field, on the other hand, seems to underestimate the conductivity field in the south-west portion of the site but has some local high values (seen by the red and orange dots in Figure 4.11 b). Both fields match the true field exactly on the boundaries, outcrop, and the conductivity measurement locations (shown in white). Interestingly, the root mean square (RMS) differences for the two fields were found to be almost identical - equal to 1.6×10^{-5} m/s. Thus, even though the overall spatial trend of the interactive estimate is more consistent with the real

conductivity field, in terms of the RMS metric the two fields are equivalent. This again demonstrates the fact that low-order summary statistics like RMS may not fully capture two-dimensional (spatial) information.

The next metric compared is the difference in head predictions. For this metric, the known hydraulic heads from the true case are subtracted from the predicted heads for each of the two calibrated models, and displayed in Figure 4.12. Figure 4.12 shows that the non-interactive head field has higher errors (shown by the red regions) in the north-west and south-west portions of the model, while the interactive solution has more errors (shown by the blue region) in the middle portion of the model. The root mean square error (RMSE) between the head fields was found to be 0.300 and 0.324 for the interactive and non-interactive solutions, respectively. Thus, on average the interactive solution does 8% better in predicting the head field compared to the non-interactive solution.

4.3.3 Results for Predictive Scenario

Next, the performance of the non-interactive and interactive calibrated models is compared for the predictive scenario. The predictive scenario consisted of reducing the bottom conductance of the river bed by two orders of magnitude and recalculating the hydraulic head field with the calibrated groundwater model [*Freyberg*, 1988]. Figure 4.13 shows the difference between the head field of the calibrated models (same as in Figures 4.10 b and c) and the head field from the true model, both with lined river beds. This figure shows that while both cases seem to over-predict the heads (as seen by the mostly positive differences in heads in Figure 4.13), the magnitude of error for the non-interactive field is *greater* than that for the interactive solution. In terms of the root mean

square difference between the two calibrated models and the true model, the non-interactive field leads to a predictive RMSE of 0.377 while the interactive field has an RMSE of 0.334 – representing about 13% improvement for the latter.

4.3.4 Results with Reduced Data

The previous results showed that the plausibility of the estimated conductivity and the predictive performance of the calibrated model both improved when allowing the expert to incorporate his/her knowledge about the spatial characteristics of the site. The improvements were seen in regions with the least data support, indicating that the expert can compensate for sparse data with his/her knowledge about the site. To explore this idea further, data were selectively removed from the hypothetical aquifer to assess the importance of user interaction, especially in the face of sparse and skewed data sets. Figure 4.14 shows one such case where the five measurement locations from the middle of the site (where earlier there was maximum data support) have been removed.

The non-interactive and interactive optimization approaches were then tested with this reduced data case, with locations of the removed measurements used as pilot points. Figure 4.15 shows the difference between the most plausible conductivity fields with and without interaction for the reduced data case and the true conductivity field. The non-interactive conductivity field (Figure 4.15 a) has higher errors than the interactive conductivity field. Compared to the results with more data (Figures 4.11 a and b), the errors around the edges of the model domain have been exacerbated and the non-interactive solution tends to underestimate the conductivity along the east and south-west edges of the site. Moreover, the conductivity errors in the middle are also higher (as seen

by the patch of orange – positive errors) compared to the results with more data (Figure 4.11 a). In terms of the root mean square error of the difference between the estimated and true conductivity field, the interactive approach leads to an RMSE of 1.43×10^{-5} m/s, while the non-interactive solution has an error of 1.84×10^{-5} m/s. This represents a 29% mean improvement for the interactive approach compared to the non-interactive solution.

The root mean square error for the hydraulic head field (compared with the true head field) for the interactive solution is 0.188 m compared to 0.244 m for the non-interactive solution. This represents a 30% mean improvement for the interactive solution. Finally, the predictive scenario with the lined river bed was also run with and without interaction. The root mean square difference between the calibrated fields and the true field (both with lined river beds) was found to be 0.307 and 0.354 for the interactive and non-interactive approaches, respectively. This represents a 15% improvement of the interactive solution compared to the non-interactive solution for the predictive scenario. Subsequent experiments with reduced data from different locations showed similar results. In general, upon removal of data points, non-interactive inversion performed poorly. The IMOGA was able to partially compensate for the lack of data, improving results both in terms of the plausibility of the field and the predictive performance of the model.

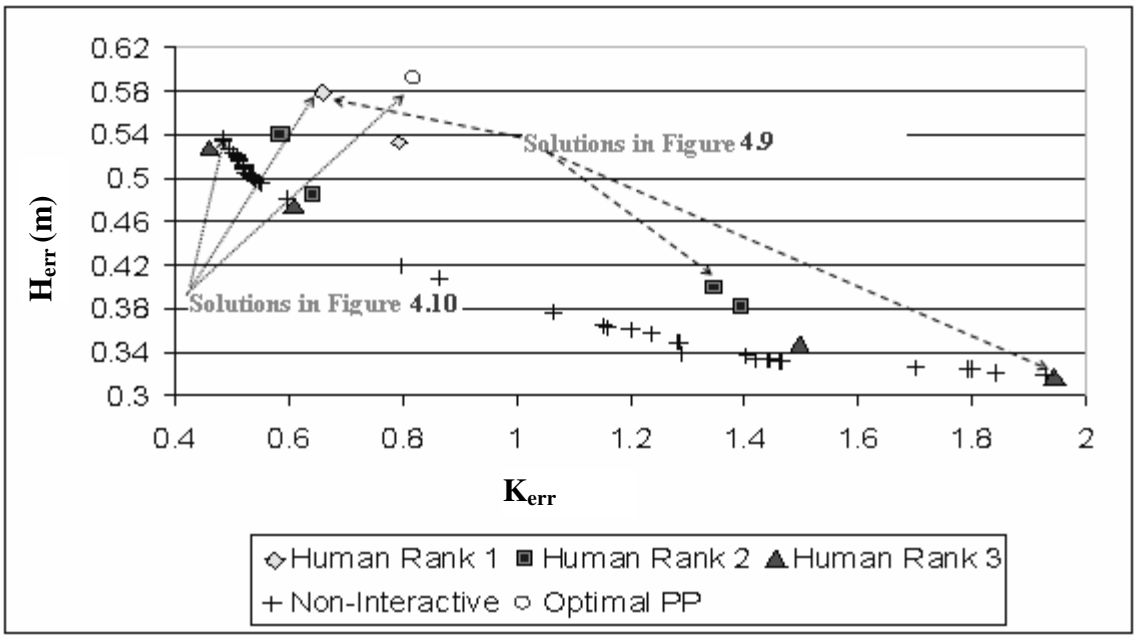
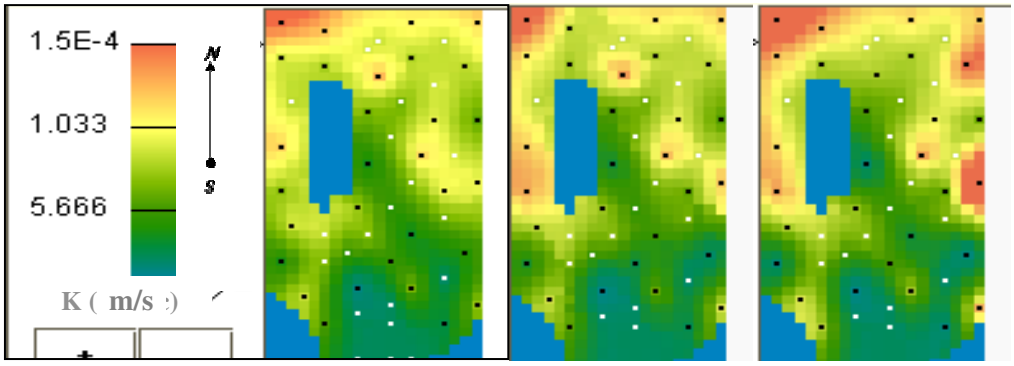
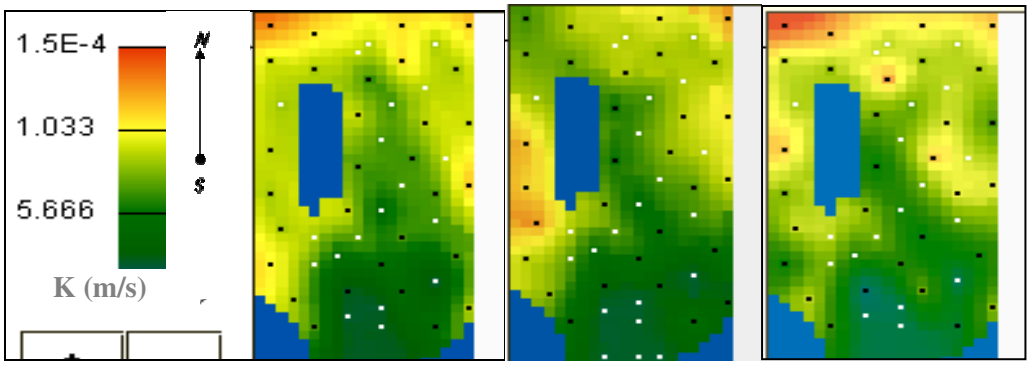


Figure 4.8 Tradeoff curve for IMOGA and non-interactive optimization



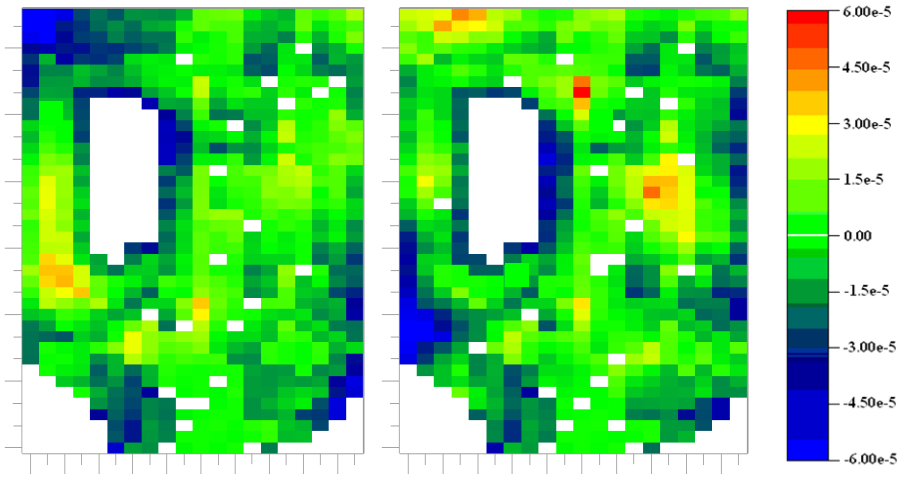
(a) **(b)** **(c)**

Figure 4.9 (a) Rank 1, (b) Rank 2, (c) and Rank 3 solutions from interactive Pareto front Figure 4.8



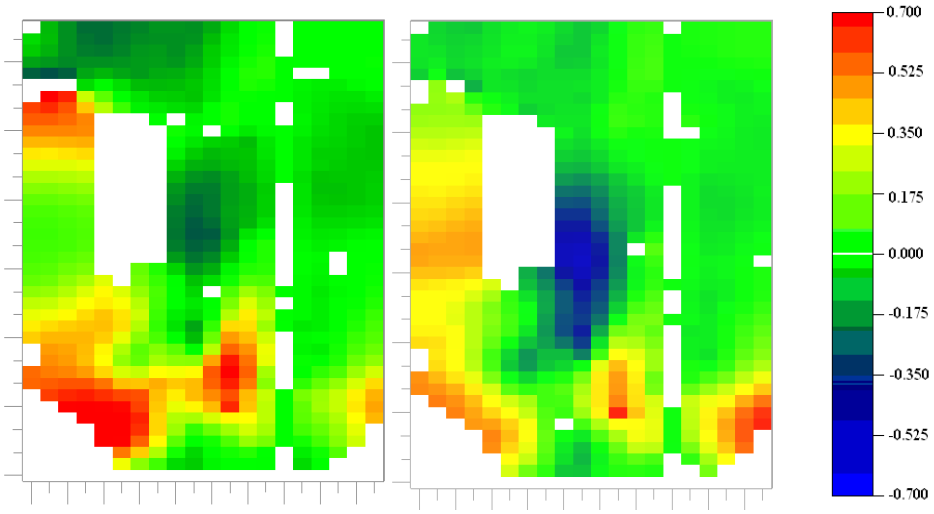
(a) (b) (c)

Figure 4.10 Conductivity fields of (a) true optimal pilot point field, and most plausible (b) non-interactive conductivity estimate, and (c) interactive conductivity estimate



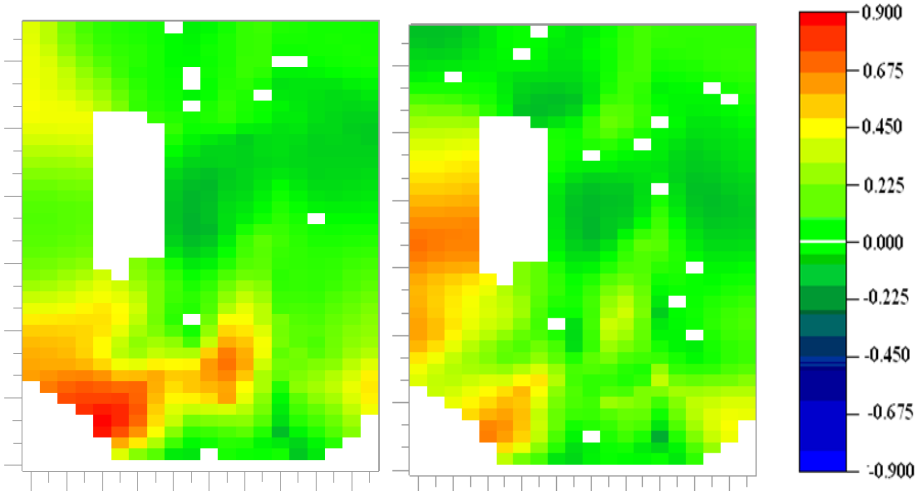
(a) (b)

Figure 4.11 Differences (in m/s) between a) the non-interactive conductivity field and true conductivity field and b) the interactive conductivity field and true conductivity field



(a) (b)

Figure 4.12 Differences (in m) between the a) the non-interactive solution's head field and true head field and b) the interactive solution's head field and true head field



(a) (b)

Figure 4.13 Differences (in m) between the a) the non-interactive solution's head field and true head field with lined river bed and b) the interactive solution's head field and true head field with lined river bed

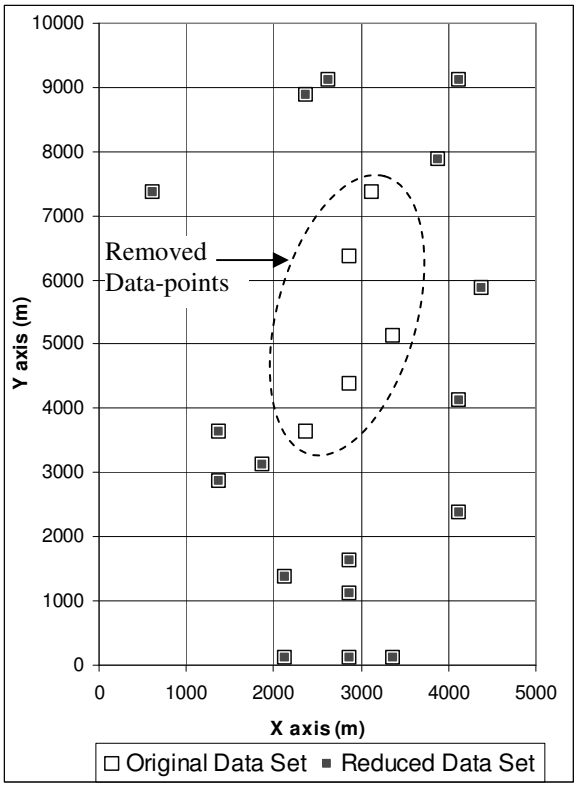


Figure 4.14 X and Y locations of original data set compared to reduced data set

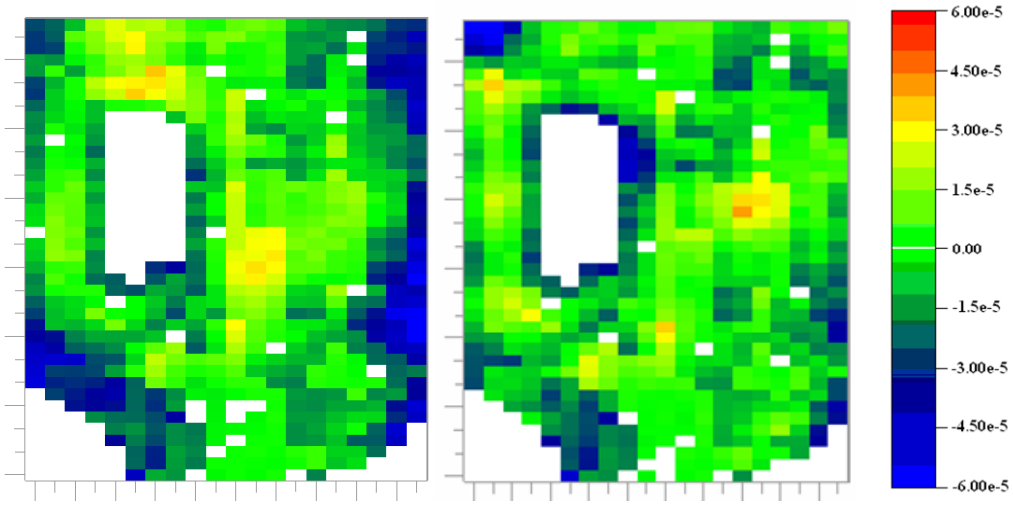


Figure 4.15 Differences (in m/s) between a) the non-interactive conductivity field and true conductivity field and b) the interactive conductivity field and true conductivity field with reduced data

4.4 Summary and Conclusions

The goal of the research presented in this chapter was to build a framework to systematically integrate expert knowledge, secondary information, and hard data in the search for optimal parameters for groundwater models. This chapter introduced a novel interactive multi-objective calibration framework based on the IMOGA that allows the user to adaptively interact with the search process and identify solutions with different levels of tradeoffs between expert knowledge, field data, and calibration objectives.

To test this approach, the IMOGA was applied to the hypothetical Freyberg aquifer using pilot point based calibration with the two quantitative objectives of head prediction error and regularization level, and an additional qualitative objective of 'user preference'. The results indicated that even without user interaction there is a significant tradeoff between the two quantitative objectives of accuracy (as measured by the calibration error) and plausibility. The most accurate solutions were in fact those that differed the most from the prior field (in other words, were most implausible with respect to field data) and this was ascribed to over-fitting of the measurement errors in the head data. It was seen that the objective values for the 'true optimal pilot point' solution (which in a sense is the best approximation of reality possible with the current parameterization) was perceived as sub-optimal when only considering these quantitative objectives.

Moreover, the conductivity fields found with non-interactive optimization suffered from unrealistic conductivity values in regions with low data support. If the expert has some knowledge about these regions with low data support, then interactive optimization

allows him/her to include this knowledge while still maintaining the (quantitative) optimality of the solutions in areas with high data support. Comparison of the most plausible interactive conductivity field with the most plausible non-interactive conductivity showed that the former had a spatial structure closer to the true optimal conductivity field. The IMOGA solution was seen to perform better than the non-interactive solution both in terms of the head response for the calibration scenario and the head response for a predictive scenario that served as validation for both models. Moreover, the improvements seen in the IMOGA increased as the data became sparser and less uniformly spread. Interaction with the expert could in fact provide information about areas where there was a lack of direct field measurements constraining the conductivity values, leading to more plausible fields. These findings indicate that the IMOGA is able to effectively incorporate expert knowledge in the search for optimal conductivity fields.

This chapter presented the first application of interactive approaches to groundwater model calibration and a number of important issues remain to be addressed for improving this methodology. The greatest challenge to be addressed is referred to as 'user-fatigue' (the burden imposed on the human user from interacting with such systems) in the interactive optimization literature. *Takagi* [2001] has discussed some of the issues of user fatigue and has pointed out how it can lead to deterioration of the solution quality of the interactive optimization by introducing errors and biases from the user. *Takagi* also discusses certain steps for reducing user fatigue – making interfaces simple, limiting the population size and the number of generations, and using 'surrogate functions' to model

and supplement human interaction. Some of these recommendations (such as small population sizes, seeding from large population runs to reduce the computational burden, and simple interfaces and ranking scheme) have already been incorporated in the IMOGA presented in this chapter. The next chapter investigates methods to reduce user fatigue by using clustering and machine learning models to 'learn' user preferences from the human-ranked results and supplement user interaction when needed.

5 ADDRESSING USER FATIGUE IN THE INTERACTIVE SYSTEM

…recognizing a face glimpsed in a crowd across an airport lobby, two human eyes can do more image processing than all the supercomputers in the world put together

~From the 'Silicon Eye', by George Gilder

5.1 Introduction

The first phase of this research demonstrated that the IMOGA can be used effectively to combine expert knowledge with the search and optimization powers of a computer. However, since the IMOGA is a population-based iterative search it requires the user to evaluate hundreds of solutions, leading to the problem of 'user fatigue' [*Takagi*, 2001]. The second phase of this PhD research addresses this issue and proposes a three-step methodology to deal with this problem.

Takagi [2001] has discussed some of the issues related to user fatigue and has pointed out how it can lead to deterioration of the solution quality of interactive optimization by introducing errors and biases from the user. The burden of prolonged user interaction also makes such interactive frameworks difficult to be used by practitioners. User fatigue is a very critical concern for researchers working with interactive optimization. Past research efforts have focused on two main approaches to reducing user fatigue – reduction in the population size requirements of the IGA and using 'surrogate' models to partially substitute human interaction with automated response functions.

Most IGAs (e.g., *Takagi* [2001], *Cho et al.* [2002], and *Kamalian et al* [2004]) have used small population sizes and run for only a few generations to reduce user fatigue. *Takagi* [2001] argues that for most cases small populations are adequate because the user is not interested in finding one optimal solution, but is more interested in a range of desirable solutions that all satisfy some user requirements. Small populations and a few generations are enough for the IGA to converge to such solutions. However, for the calibration problem the quantitative objectives given by the calibration error and regularization are arguably as important as the additional subjective plausibility criterion (necessitating the proposed interactive multi-objective approach). Given the non-linearity and complexity of the inverse problem [*Yakowitz and Duckstein*, 1980; *Carrera and Neuman*, 1986; *Sun*, 1995; *Zimmerman et al.*, 1998], small population sizes may be insufficient to solve this problem adequately.

To address this issue, the framework presented in the previous chapter employed a two-stage optimization approach that involved seeding the IGA with solutions from the non-interactive Pareto optimal set. This approach was based on the work of *Babbar* [2006], who investigated different initialization strategies for the IGA search. However, despite the reduction in population size requirements due to effective initialization, even small IGA populations can impose a significant burden on the user. Thus, every effort needs to be made to a) efficiently display the information shown to the expert and reduce the amount of user interface interaction, and b) actively reduce the number of solutions that need to be evaluated by the expert. With respect to the former objective, *Takagi* [2001]

has pointed out that the type of evaluation scheme and user interface used for the IGA have an important effect on user fatigue. Research on improved input interfaces and evaluation schemes [*Takagi et al*, 1996; *Ohsaki et al*, 1998] has shown that discrete evaluation schemes with 5 to 7 levels are desirable in most cases. IGA practitioners also keep their input interface relatively intuitive and easy to navigate, showing only information that is cogent to the problem being solved. These recommendations have been used in the design of the IMOGA framework described in the Chapter 4.

The second issue of actively reducing the number of solutions shown to the expert has received much attention in the IGA literature (see references below). Most practitioners have recommended the use of surrogate predictive models that aim to 'learn' and mimic user preference. IGAs have used two broad categories of predictive models to learn user preferences - 'nearest-neighbor' type models and 'regression-based' models. Nearest neighbor models identify the most similar previously evaluated solution (or set of solutions), and use the labels of these 'neighborhood' solutions to classify a new (unevaluated) solution. On the other hand, regression-based methods try to build a functional form for the predictive model based on the features and parameters of the evaluated solutions. Nearest-neighbor type approaches have been applied by *Nagao et al* [1998] and *Lee and Cho* [1999] to IGAs for image retrieval systems. Examples of the use of regression models include convex functions [*Takagi*, 2001], fuzzy rule-based models [*Nishio et al.*, 1997; *Babbar et al.*, 2006], neural networks [*Tokui and Iba*, 2000], and support vector machines [*Llora et al.*, 2005]. *Llora et al.* [2005] compared these two approaches and pointed out that nearest-neighbor type approaches can be detrimental to

GA convergence and recommended that regression-based models be used for small population genetic algorithms such as IGAs. Thus, this is the approach used in this work.

The primary motivation for this phase of the PhD research is to 'intelligently' reduce the user interaction for the IMOGA so that a few carefully chosen solutions evaluated by the user can be used to drive the IGA towards the desired solution space. A two-step methodology is proposed to achieve this. The first step of this approach addresses how to best choose a few solutions from the IMOGA population to be shown to the user. Since the results of the IMOGA depend heavily on user ranking, it is imperative that the user be shown as many different kinds of solutions as possible, allowing him or her to explore the solution search effectively. To do this different clustering approaches are investigated that group the search space based on the similarity of the solutions. Solutions are then selected from different groups to improve the diversity of solutions shown to the user for ranking. Finally the ranked sample solutions have to effectively 'drive' the large population GA towards solutions with desirable properties. For this different machine learning approaches are investigated to model user preferences for unevaluated solutions. Such user-preference models are then used as surrogates to supplement (but not entirely replace) user interaction and are adaptively retrained as more user evaluations become available.

During IMOGA sessions for the case studies considered in this thesis, the information provided to the user is mainly perceptual - users provide feedback to the IMOGA based on visual inspection of the spatial characteristics of different conductivity fields. To

incorporate this spatial information, novel approaches are proposed that utilize the 'images' of the conductivity fields to drive the clustering and machine learning algorithms. The hypothesis is that such spatial information improves the performance of these algorithms and better reflects the choices and preferences of the IMOGA user. To the best of our knowledge, this work is the first to propose the use of image-processing algorithms to drive 'visual' decision making environments such as the IMOGA.

The rest of this chapter is divided into three parts. The first part (Section 5.2) presents a number of algorithmic approaches to a) cluster the conductivity fields based on the spatial content of each field, and b) use the spatial information of the conductivity fields to build learning models of user preferences. In the second part (Section 5.3) the performances of these algorithms are compared for the hypothetical test case (described in Section 3.1) using well-defined performance metrics. The algorithms with the best performance and reliability are then implemented for the field-scale application (described in Section 3.2) and these results are presented at the end of Section 5.3. Finally, Section 5.4 gives a summary of the results and conclusions for this phase of research.

5.2 Methodology

Figure 5.1 provides an overview of the strategy proposed to reduce user fatigue. The basic operations of the IMOGA remain the same, but in each generation the IMOGA population (that has not yet been evaluated by the user) are clustered to identify unique solution types to show to the expert. The expert inspects this subset of the population, resulting in a partially-ranked population. The solutions ranked by the expert are stored in

an archive. The archive is then used to train a machine-learning algorithm that predicts user preferences for the remaining unranked individuals (those not selected for user inspection by the clustering algorithm). Once the entire population has been ranked (either by the expert or by the surrogate preference model) the entire process is repeated for all subsequent IMOGA generations.

Thus, compared to the IMOGA used in the previous chapter, the new aspects of the IMOGA process introduced here are the clustering and selection of potential solutions for user evaluation and the building of machine-learning models to reflect user preferences for the unranked solutions. The following subsections present the methodologies that have been tested for each of these two processes. The first subsection (5.2.1) deals with clustering - investigating different data transformations and clustering techniques to identify the most promising alternative(s) for this application. The next subsection (5.2.2) develops machine-learning models of user choices. Since user preferences are 'perceptual' or based on visual inspection of different conductivity fields image processing tools are explored to incorporate this visual or spatial information in both the clustering and machine-learning approaches.

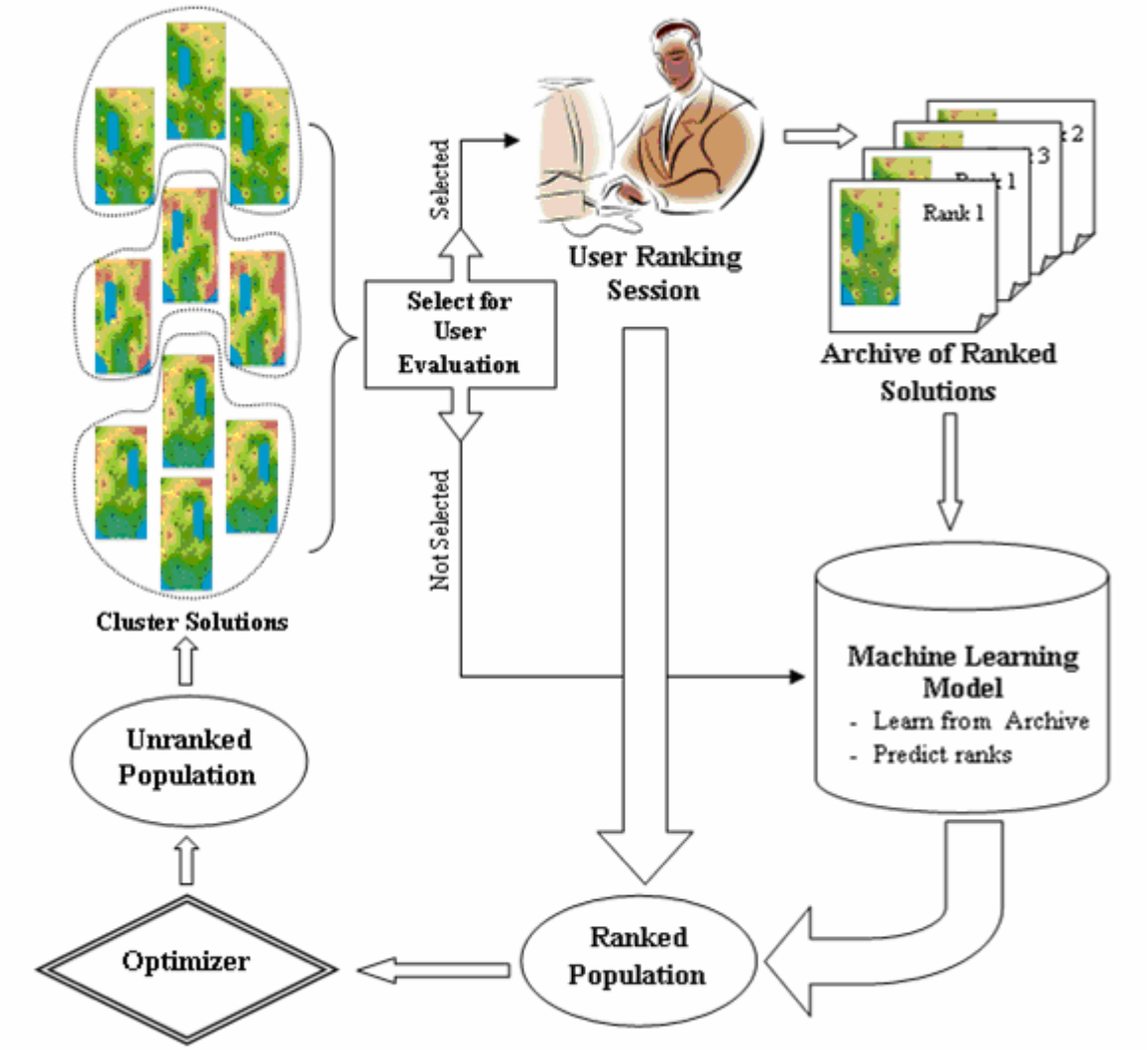


Figure 5.1 Framework to reduce user fatigue

5.2.1 Clustering Conductivity Fields

The three stages of clustering, as given by *Jain and Dubes* [1988] are data representation, clustering, and data abstraction (see Section 2.3.1 for details). An important clustering issue to consider is the optimal number of clusters to be used. The following subsections discuss each of these issues with respect to this particular application. Section 5.2.1.1 addresses the first stage of data representation by exploring different clustering

representations of the conductivity fields. Section 5.2.1.2 discusses the second stage – the actual clustering of the data – by investigating two types of clustering algorithms for grouping the conductivity fields based on spatial similarities. Section 5.2.1.3 addresses the issue of determining the optimal number of clusters present in the dataset. Finally, Section 5.2.1.4 deals with data abstraction or how to best choose representative solutions from each cluster.

5.2.1.1 Data Representation

Effective clustering is highly dependent on choosing the right features to represent high-dimensional data such as the conductivity fields that need to be clustered. The two most obvious choices to represent each conductivity field are: 1) the decision-variables (pilot point values and kriging parameters) used by the GA to create the different conductivity fields, and 2) the two-dimensional grid values for the conductivity fields (kriged from the pilot points). Clustering by decision variable values is simple, but may not perform well because different decision variables have substantially different effects on the final kriged conductivity field (for example the kriging window that defines the number of neighboring points to use in kriging has a huge impact on how smooth or irregular a particular kriged field is) . Using the 2-D kriged grid directly for comparisons avoids such complications, but this substantially increases the dimensionality of the problem. Depending on the scale and discretization level of the numerical groundwater model, the conductivity grid for groundwater models can have tens of thousands of grid values leading to prohibitively high dimensionality for the clustering algorithm. Thus it is appropriate to investigate dimension reduction strategies to best represent this spatial information in low dimensional space. In this study, two such approaches are investigated

– nested spatial moments and spatial principal components – that are discussed in more detail in the following subsections.

Nested Spatial Moments

Spatial moments have long been used for 'content-based' image classification [*Mandal and Aboulnasr*, 1996; *Vailaya et al.*, 2001]. Based on the the concept of 'recursive region splitting' of *Ohlander et al.* [1978] this work uses 'nested' spatial moments to represent the spatial variability of the data. The methodology is as follows:

1. Start with the entire image (2-D field conductivity field). Calculate the required spatial moments for the entire image.
2. Divide the image into disjoint sub-images (either uniformly or based on some prior knowledge of the 'structure' of the image). Calculate the required spatial moments for each sub-region.
3. Repeat 2, sub-dividing the regions recursively and calculating spatial moments.
4. Stop when a user-specified minimum sub-region size (typically corresponding to the smallest scale of heterogeneity of concern to the expert) is used.

Figure 5.2 provides an overview of this methodology. For this study uniform splits are used and the first two moments - mean and standard deviation - are calculated for each sub-region. When calculating the moments, cells that are not within the flow boundaries or are within outcrop regions (see Section 3.1 and Figure 3.1 for this case study) are not considered. The nested spatial moment technique represents each image as a set of average and standard deviation values. The total number of moments is given by k:

$$k = 2\left(\sum_{i=0}^{n-1} 4^i\right) \tag{5.1}$$

where n is the level of nesting. Three levels of spatial nesting were found to be sufficient for the Freyberg case study through trial-and-error experimentation (giving a total of 21 pairs of averages and standard deviations).

Once all the spatial moments have been calculated for all the conductivity fields in a dataset, the difference between 2-D conductivity fields is then calculated as the root mean square difference between corresponding spatial moments.

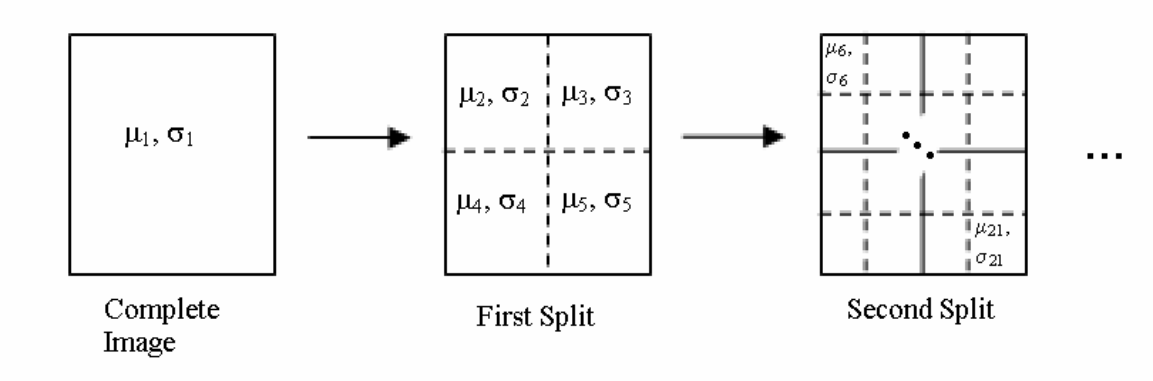


Figure 5.2 Calculating nested spatial moments (μ and σ are the mean and standard deviation of the nested blocks)

Spatial Principal Components

The principal components of a data set represent the dominant patterns that exist in the data. Given n images $I_1, I_2 \ldots I_n$ each with p pixels, relatively few (m << p) orthogonal projections $E_1, E_2 \ldots E_m$ (principal components) can be found that capture a large proportion (say 95%) of the variability in all the images. These principal components, referred to as the 'eigenimages' in the field of image processing, represent the optimal basis that can be used to represent any given image in the dataset. Eigenimages have been

used in many applications for pattern recognition and image compression for large image databases [*Kirby and Sirovich*, 1990; *Turk and Pentland*, 1991; *Yang et al*., 2004].

Once the eigenimages have been calculated for an image database, each corresponding image in the database can be reconstructed as a unique linear combination of the dominant eigenimages (i.e. $I_j = y_{1,j} E_1 + y_{2,j} E_2 + \ldots y_{m,j} E_m$). The m scaling factors y (called eigenscores) in this reconstruction represent the projections of that particular image in the hyperspace defined by the eigenimages. Thus, whereas earlier each image had a dimensionality of p, after projection in the eigenimage space each image can be uniquely identified by its set of only m eigenscores. Eigenimage analysis thus provides a very useful tool to represent higher-dimensional data such as the conductivity fields in only a few dimensions. The n images in the database can now be clustered using the m eigenscores for each image. Details about the algorithms to calculate the eigenimages and eigenscores for an image database are given in Appendix C. In this case the distance metric between two images is simply the root mean square difference between the corresponding eigenscores for the two images. Another attractive aspect of such principal component analysis is that it is completely non-parametric and does not require any restrictive assumptions about the dataset.

5.2.1.2 Clustering Algorithms

Next, two clustering techniques - hierarchical agglomerative clustering [*Duda et al.*, 2001] and a spectral clustering algorithm called 'N-cuts' clustering [*Ng et al.*, 2002] - are investigated to cluster the data processed using the above approaches. The common K-means clustering algorithm was not chosen for this work because this type of clustering

tends to perform well only when the data are naturally clustered in roughly 'globular' or 'spherical' clusters. Moreover, K-means clustering is highly sensitive to the specification of an initial set of 'centroids' and different initializations can lead to very different clustering configurations.

Hierarchical Clustering

The first clustering technique applied in this work is hierarchical clustering. Hierarchical clustering can be of two types: agglomerative (also known as the bottom-up or 'clumping' approach) – where clusters are built by iteratively combining data points, and divisive – where clustering begins with the assumption that all data points are in one cluster and then the clusters are built by splitting the dataset successively. Agglomerative clustering is the one that is most often used, and is thus used for this study. Details about agglomerative hierarchical clustering are given in Appendix D.

The results of this clustering algorithm depend heavily on the similarity metric used to compare clusters during the clustering process. *Duda et al.* [2001] define three major types of metrics that can be used for this purpose – nearest neighbor (also called single-linkage), farthest neighbor (also called complete linkage), and average linkage. *Duda et al.* [2001] discuss some of the drawbacks with using the first two types of metric - high sensitivity to outliers for the nearest neighbor metric and spurious clustering for the farthest neighbor metric. Consequently, this work uses the 'average' linkage similarity metric which has been shown to ameliorate some of the problems with the first two types of metrics. As the name implies, 'average' linkage uses the average distance between members in two clusters as a measure of the similarity between the two clusters.

Other similarity metrics exist, such as 'centroid linkage', 'median linkage', and 'Ward's linkage' [*Ward*, 1963], however these were not investigated for this study. In general, the choice of the type of linkage is generally based on assumptions about the structure of the natural clusters in the dataset. When such structure does not exist, the clustering algorithm can fail to find meaningful groups. The average linkage has been shown to be the most general type of linkage, working well for most types of data configurations [*Duda et al.*, 2001].

Spectral Clustering

As mentioned earlier, spectral clustering algorithms do not require assumptions about the structure of the data and are thus a natural alternative for a framework like the IMOGA that can be applied to a wide range of problems. There are several spectral algorithms ([*Luxburg*, 2006] provides a good overview of major spectral clustering approaches) but 'normalized cuts' (commonly referred to as 'N-cuts') clustering [*Ng et al.*, 2002] has been shown to be a stable and robust algorithm, working equally well for many different types of data configurations that can not be partitioned by other clustering approaches [*Verma and Meila*, 2003]. For this reason N-cuts clustering is chosen for this study.

Details of the N-Cuts clustering algorithm are given in Appendix D. In brief, this algorithm is based on spectral graph partitioning theory. The data to be clustered are represented as a connected graph, and the Laplacian of this graph is used to find the optimal partitions (or 'cuts' - giving the method its name 'normalized cuts clustering') for this graph based on the proximity of different graph nodes to each other. Compared to other clustering techniques, the salient aspect of this methodology is the use of the

eigenvectors of the Laplacian matrix (instead of the more commonly used covariance matrix) to reduce the dimensions of the data optimally. In terms of dimensionality reduction, N-cuts is extremely efficient as it reduces the dimensions of the data to k - the number of clusters. The clustering dimension is thus only dependent on the number of clusters and not on the dimensionality of the data to be clustered.

5.2.1.3 Finding Optimal Number of Clusters

A critical issue when clustering data is the number of clusters into which the data set will be partitioned. Since clustering is by definition unsupervised, there are no validation data that can be used to test the accuracy of classification. In the literature various metrics have been proposed to test the separability of the clusters (*Milligan and Cooper* [1985] have provided a good comparison of different alternatives). Typically the clustering algorithm is run for multiple cluster numbers and the cluster number corresponding to the optimal metric is chosen as the best cluster size. This study uses the average 'silhouette width' [*Kaufman and Rousseeuw*, 1990], which has been shown to work well for image databases [*Ng and Han*, 1994], for the clustering performance metric. The silhouette value for each point is a dimensionless quantity that is a measure of how similar that point is to points in its own cluster, as compared to points in other clusters, and ranges from -1 to +1. It is defined as:

$$S_i = \frac{\min_m(D_{i,m}^{Inter}) - D_i^{Intra}}{\max\left\{D_i^{Intra}, \min_m(D_{i,m}^{Inter})\right\}} \tag{5.4}$$

where S_i is the silhouette value for point i, $D_{i,m}{}^{inter}$ is the average distance between point i and all points in a different cluster m, D_i^{intra} is the average distance between point i and

all points in the same cluster as *i*. The silhouette value is 1 for data points that are well identified within their clusters (i.e. the inter-cluster distance between these data points and data points in other clusters is much larger than the intra-cluster distance between the data points and other members of the same cluster), and it is low (or negative) for 'outliers' (with the intra-cluster distance being more than the inter-cluster distance).

The silhouette width for a particular cluster is simply the average of the silhouette values for all of its members. Based on extensive testing, *Kaufman and Rousseeuw* [1990] have proposed the following interpretation of Silhouette values:

Table 5.1 Interpretation of silhouette scores

Silhouette Width	**Interpretation**
0.71 – 1.0	Strong cluster
0.51 – 0.71	Reasonable cluster
0.26 – 0.5	Weak or artificial cluster
≤ 0.25	No cluster found

Ng and Han [1994] proposed using the average silhouette width (the average across all clusters of the silhouette widths) to find the optimal cluster sizes. Starting from a cluster number of 2 they iterate through higher cluster numbers as long as the average silhouette width keeps increasing. Whenever a higher cluster number leads to a loss in average silhouette width it indicates that the clusters are being artificially split and the optimal cluster number was identified in the previous iteration. This approach was used in this study, but the number of clusters was also limited to be no more than the number of samples to be shown to the expert, since the expert is shown at least one member from each cluster.

5.2.1.4 Selection from Clusters

Once all of the data have been clustered into the optimal number of clusters, the final stage is the selection of conductivity fields from each cluster to show to the expert. To allow the expert to evaluate the widest range of solutions possible, the following selection strategy is proposed that takes into consideration both the cluster size and the silhouette scores for the individuals. To choose *m* members from a total of *c* clusters with n_1, $n_2 \ldots n_c$ members:

1. Calculate the number of members to be selected from each cluster, proportional to the cluster size:

$$p_i = round\left(m \frac{n_i}{\sum_{j=1}^{c} n_j} \right) \tag{5.5}$$

   where p_i is the number of members to be chosen from cluster i with n_i members, and *round( )* is the mathematical rounding operator (p_i can only be an integer).

2. For each cluster calculate the range R_i of silhouette scores from highest to the lowest.

$$R_i = \underset{n_i}{Max}(S_i) - \underset{n_i}{Min}(S_i) \tag{5.6}$$

3. Divide silhouette range into p_i bins of equal size R_i/p_i.
4. Divide cluster members equally within the silhouette bins based on their silhouette scores.
5. Randomly select one member from each silhouette bin.

An example for this selection strategy is shown in Figure 5.3, where the silhouette scores are given for a two-dimensional data set with 5 members to be selected from 10 solutions that were divided into 3 clusters. With this selection strategy not only are the most representative solutions chosen from each cluster, but other members that are not as representative are also shown to the expert for evaluations, ensuring that the expert evaluates the widest range of solution types.

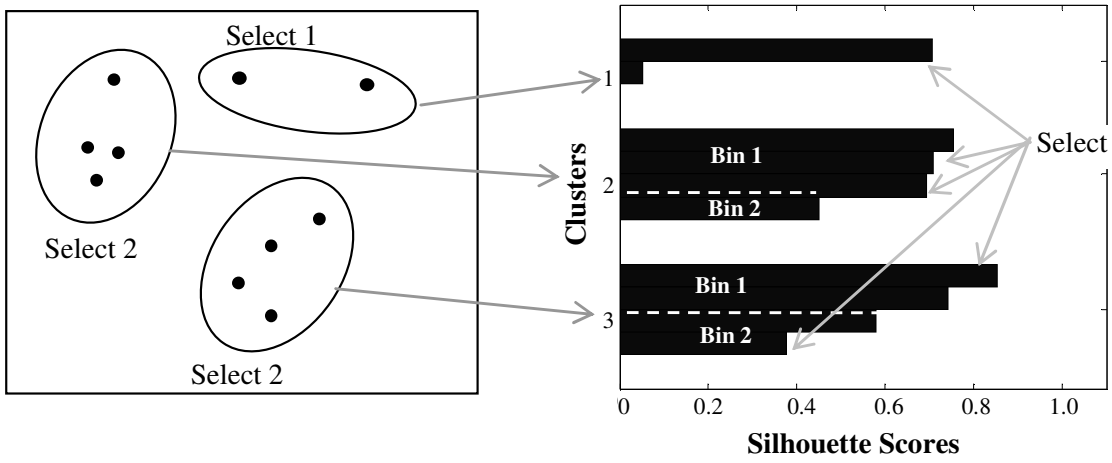


Figure 5.3 Strategy for selecting 5 solutions among three clusters, based on silhouette scores (dashed lines indicate selection bins for each cluster)

5.2.2 Building Models of User Preferences

Once the conductivity fields have been clustered and candidate solutions have been chosen from each cluster, the user is shown these conductivity fields and these are then ranked during the interactive session. All of the solutions ranked by the user are then stored in an archive (see Figure 5.1). The solutions that have not been ranked by the user in the current generation are then ranked using machine-learning algorithms.

As in the case of clustering, appropriate parameters need to be chosen to represent the conductivity fields for the learning models. While for clustering the only requirement was a measure of similarity between different conductivity fields, for supervised learning algorithms these parameters become the input to the learning model. For this reason, using low-dimensional and salient representations becomes all the more important. Therefore, three feature extraction approaches were implemented to reduce the dimensionality of the conductivity fields. These are: using the decision variables directly as inputs to the learning algorithms, using the nested spatial moments for each conductivity field (described in Section 5.2.1.1), and using the eigenimage scores (described in Section 5.2.1.1) for each image calculated with respect to the testing and training set used for the learning algorithm (note that these eigenimage scores would need to be recalculated for each image as new conductivity fields are added to the testing or training data set).

Two types of user-preference models – decision trees (*Quinlan*, 1986) and naïve Bayes (*Jensen*, 1996) learning models were compared in this work. A brief overview of the two approaches is presented in the next two sub-sections.

5.2.2.1 Decision Trees

Decision trees (DT) [*Quinlan*, 1986] are locally-linear supervised learning algorithms that build a tree function from the input data to predict outputs, where each node in the tree represents a choice between different levels of the input parameters (see Figure 5.4 for an example of a decision tree for playing tennis). Decision trees recursively select the most predictive input feature based on the information gain for different levels of the

inputs 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. For this work an implementation of decision trees based on the public domain C4.5 software [*Quinlan*, 1994] was used. C4.5 can handle both continuous and categorical attributes and also allows for the decision tree to be 'pruned' (i.e. decision tree branches that are not highly informative are removed and replaced by leaf nodes – leading to a more compact tree).

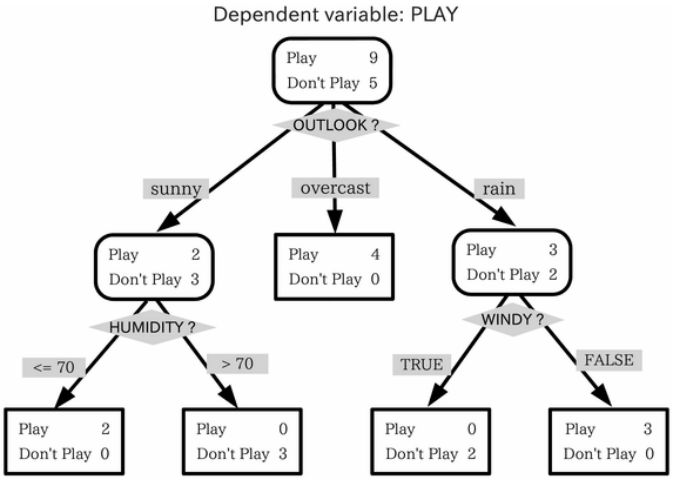


Figure 5.4 Decision tree to decide whether to play tennis or not based on outlook, humidity, and wind

5.2.2.2 Naïve Bayes

Naïve Bayes (NB) is a type of Bayesian learning method, based on a probabilistic approach to learning. In general, the aim of Bayesian learning is to estimate the posterior distribution of the outputs given the training data following the fundamental Bayes theorem:

$$P(O \mid I) = \frac{P(I \mid O)P(O)}{P(I)} \tag{5.7}$$

where $P(O|I)$ is the probability of the output given the input vector I, $P(I|O)$ is the probability of the inputs for a given output, $P(O)$ is the prior probability of getting the output, and $P(I)$ is the probability for generating the inputs. Naïve Bayes classifiers are based on a simplifying (naïve) assumption of conditional independence between all of the inputs – meaning that the probability of occurrence of an input I_i is independent of the probability of occurrence of all other input, given an outcome O. With this assumption $P(I|O)$ is simply given by the product of the probabilities $P(I_i|O)$ for each input (I_i) given the output. For a particular data set, these probabilities are calculated based on the frequency of each input variable for a certain output. For this reason, Naïve Bayes can only handle categorical data, but can still be used for continuous attributes by dividing the range of the continuous attributes into discrete bins and then counting the frequency of occurrence of each bin in the training dataset. Using bins of equal width is recommended as this can effectively approximate skewed, multimodal, and/or heavy-tailed probability distributions [*Elkan*, 1997].

Despite the strong assumption of conditional independence, naïve Bayes have been shown to perform remarkably well in practice, matching and outperforming neural networks and decision trees in performance [*Domingos and Pazzani*, 1996; *Michie et al.*, 1994].

5.3 Results

The methodologies given in Section 5.2 were first tested on the *Freyberg* [1988] case (described in Section 3.1). To consistently test the different algorithmic approaches in Sections 5.3.1 and 5.3.2, five IMOGA sessions with an expert (the author) were created for the Freyberg case from Section 3.1. For each session, the human responses represent the 'ideal' case that the algorithms should reproduce as far as possible. Testing was conducted by employing the clustering and machine learning algorithms on the pre-ranked IMOGA results and comparing the results with the true expert responses. The first subsection (5.3.1) presents results for the two clustering algorithms (hierarchical and N-cuts) outlined in Section 5.2.1.2. The second section (5.3.2) compares the different machine learning algorithms and presents results for the testing and training of these user-preference prediction models. Once the best combination of clustering and machine learning approaches was selected using this off-line approach, this was then applied to a real 'online' session where the selected methodology was run in real time with partial expert interaction for both the Freyberg case study (Section 5.3.3) and the WIPP case study (Section 5.3.4). The online runs allow one to assess the impact of this partial interaction on the calibration solution quality.

5.3.1 Results for Clustering

To test both clustering techniques using all four measures of similarity, the following eight combinations were tested: hierarchical clustering with a) decision variables, b) complete image, c) eigenimage scores, and d) nested spatial moments; and N-cuts clustering with a) decision variables, b) complete image, c) eigenimage scores, and d) nested spatial moments. To test these combinations, the IMOGA solutions in the first

generation, consisting of 20 candidate conductivity fields, was shown to the human expert, who was then asked to cluster the images based on visual perception. The clusters predicted by the clustering algorithms are then compared to the human clusters and the classification accuracy metric is calculated for each clustering algorithm. It is worth noting that the initial generation for the IMOGA has the highest variance and is thus the most difficult to cluster compared to subsequent converged populations. Thus the performance for this generation is expected to be a conservative measure of the clustering algorithm's performance for the rest of the IMOGA generations.

Testing the methods using such a 'human-based' approach is common practice in the field of image processing and image retrieval, where the perceptual and pattern-recognition skills of human beings are accepted as superior to automated techniques, and in fact provide the ultimate test for such algorithms. Moreover, this approach is particularly appropriate for IMOGA, where the clusters are being used for visual inspection – and thus the ideal clusters should correspond to spatial patterns that the human could identify and use as a basis for ranking.

To estimate the classification accuracy of the clustering algorithms, both the average silhouette width (Equation 5.4), which is a measure of how uniquely identified and distinct the clusters are from one another, and the 'Wallace index' [*Wallace*, 1983], which measures the accuracy of the clusters with respect to labeled data, are used. Wallace index has been extensively used in the literature [*Verma and Meila*, 2003; *Ruan and Zhang*, 2006] to evaluate the performance of clustering algorithms with respect to

labeled data. Following the notation of *Ruan and Zhang* [2006], the Wallace index (W) between the true (labeled) clustering Γ and the predicted clustering Γ' is defined as:

$$W(\Gamma,\Gamma') = \min\left(\frac{N_{\Gamma,\Gamma'}}{S(\Gamma)}, \frac{N_{\Gamma,\Gamma'}}{S(\Gamma')}\right) \tag{5.8}$$

where $N_{\Gamma,\Gamma'}$ is the number of pairs of members that appear in the same cluster in both clustering schemes. $S(\Gamma)$ is the total number of linkages of intra-cluster members in the labeled clustering scheme, and $S(\Gamma')$ is the total number of linkages of intra-cluster members in the predicted scheme. Details about calculating the Wallace index are given in Appendix D.

Figures 5.5 and 5.6 show the average Wallace indices and cluster silhouette widths for the eight clustering techniques. These figures show that N-cuts clustering *always* performs better than ordinary hierarchical clustering for this case study. Not only are the resulting clusters more consistent with the spatial patterns of the conductivity fields (indicated by the higher Wallace indices), but the cluster separabilities are also much better, as indicated by the higher silhouette scores. N-cuts clustering is known to be able to resolve clusters that are of complex and irregular shapes. The fact that N-cuts always leads to better clusters indicates that the conductivity image data used in this study have complex configurations that are not well resolved with ordinary clustering. On average, across the different distance metrics, using N-cuts clustering led to an improvement in cluster accuracy ranging from 6% (for the nested spatial moments) to 60% (for the direct image and eigenimage metrics), and an improvement in silhouette scores ranging from

7% (in the case of decision variables) to more than 40% (in the case of nested spatial moments).

When comparing the distance metrics, it is clear that using the decision variables alone was not sufficient - the Wallace index for the decision variables was the lowest among all of the four distance metrics. Interestingly, the silhouette scores for the clusters with decision variables were consistently high – indicating that there *were* natural clusters in the decision variable space, but these did not correspond well with the similarities in spatial patterns seen in the actual images. Thus, using spatial information, in general, leads to improved clustering performance.

Among the three image-based metrics that use spatial information, the results vary considerably. The clusters formed by the eigenimages were *exactly* the same as those formed using the complete image information (both for hierarchical and N-cuts clustering). The only benefit of eigenimages was marginal improvements in the separability of the clusters as indicated by an increase of about 2% (for hierarchical clustering) and 11% (for spectral clustering) in the overall silhouette scores. The nested moments metric performed well compared to the direct image and eigenvector (albeit only for hierarchical clustering) with an increase of 30% in the cluster accuracy (as given by the Wallace index). While, the silhouette width of the nested moments approach for hierarchical clustering was the lowest, spectral clustering with nested moments led to a significant increase (more than 40%) in the silhouette width.

Finally, N-cuts with the direct image or eigenimage information led to the most accurate clustering methodology with a very high Wallace index of 0.8 and high silhouette scores ranging from 0.8 to 0.9. Figure 5.7 shows the actual clusters found by this algorithm, which correspond well with the dominant spatial features in the conductivity fields. The N-cuts algorithm with direct image information was also the fastest among all of the other image-based algorithms (with the eigenimages-based algorithms being the slowest due to the fact that they had to solve the eigenvectors for an 800 x 800 matrix). Thus N-cuts image-based clustering was seen to be an efficient and accurate spatial clustering algorithm and was used for subsequent runs of the IMOGA.

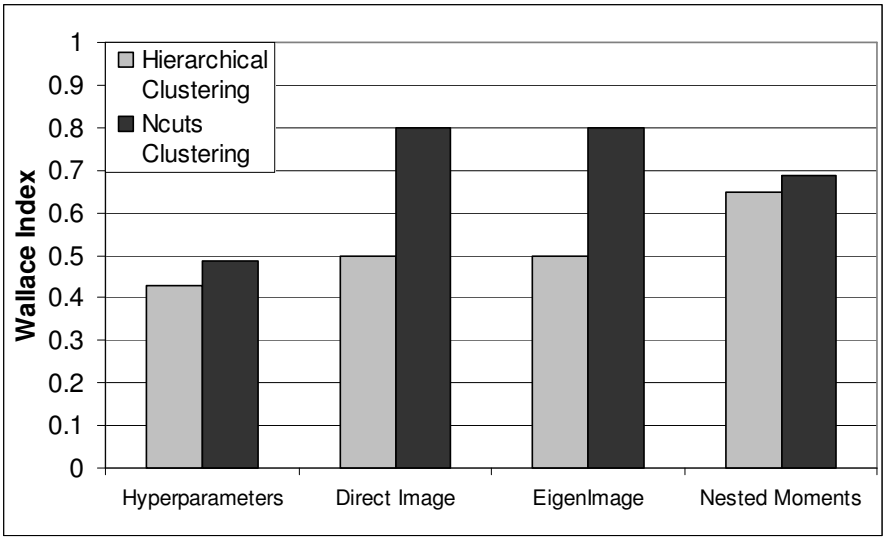


Figure 5.5 Wallace indices for hierarchical and N-cuts clustering with different distance metrics

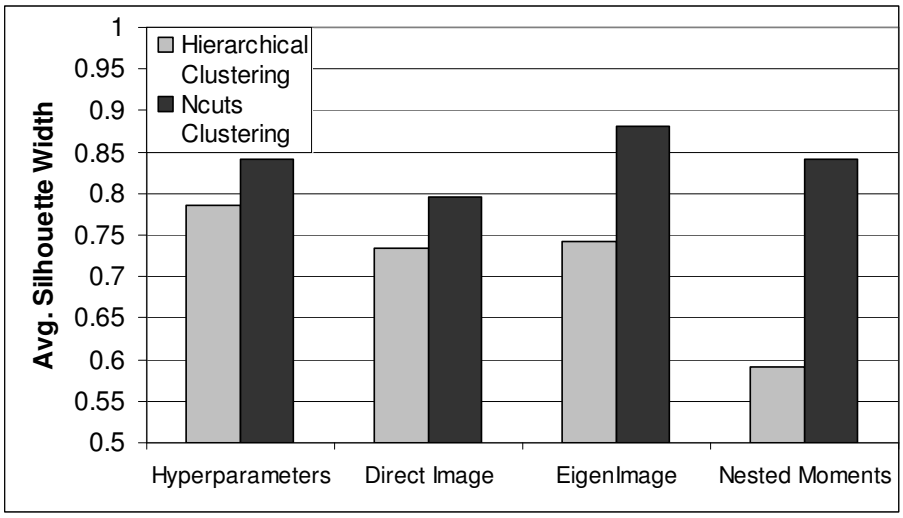


Figure 5.6 Average silhouette widths for hierarchical and N-cuts clustering with different distance metrics

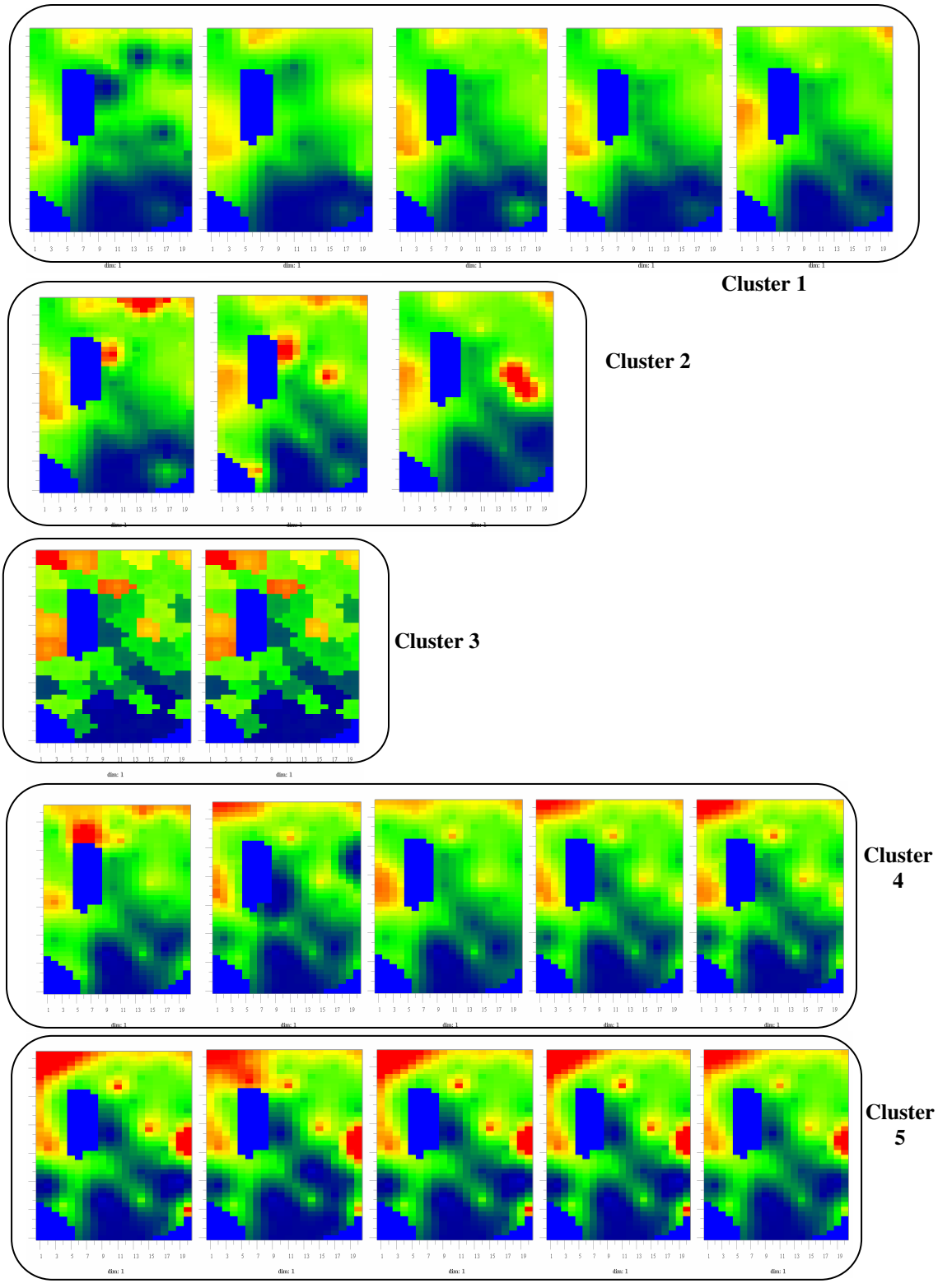


Figure 5.7 Five clusters of conductivity fields identified by N-cuts clustering

5.3.2 Results for User-Preference Learning Models

To test the decision tree and naïve Bayes models, the first 3 of the 5 offline training runs described previously were used to parameterize each model (see Appendix E for details). The models with the best parameterization were then compared in the last two trials as 'validation'. Figure 5.8 shows the testing and validation average root mean square error (RMSE) and the corresponding standard errors in rank predictions for the decision trees and naïve Bayes models without the clustering-based selection. Note that this figure shows results only when decision variables are used as the training attributes; results for the image-based approaches are shown subsequently. The average RMSE is calculated for each GA run as the average root mean square difference between the user rankings and the ranks predicted by the machine-learning algorithm for the 20 individuals over all 10 generations. The testing RMSE is then calculated as the average RMSE over the first three offline GA runs and the validation RMSE as the average over the final two validation GA runs. As can be seen from Figure 5.8, the two models perform similarly, with almost equal errors during the testing phase - the root mean square error (RMSE) in the rank prediction for naïve Bayes was 1.044 compared to 1.039 for the decision tree. However, the naïve Bayes was seen to perform marginally better for the validation dataset (with an RMSE of 1.010 compared to 1.11 for the decision tree – a 10% improvement).

To further investigate these differences, the average (across trials) root mean square error for each model was calculated for each of the 10 generations across all members of the population. Figure 5.9 plots the average *difference* in the root mean square errors (along

with the standard error ranges for the average difference) for the decision tree and the naïve Bayes (both with decision variables as input) across generations. Positive values indicate that the prediction error for the decision tree was more than that for the naïve Bayes. As can be seen from the figure the differences are significantly positive (indicating that the naïve Bayes is performing better) for the first few generations, but become closer to zero for later generations (indicating that both algorithms are performing equivalently). During the first few generations, the archive size for training the user-preference models is small because the expert has only ranked a few solutions. These results indicate that the naïve Bayes may be better at handling sparse data compared to decision trees. Both models perform almost equally well with larger datasets (in later generations). Such results have been reported in the literature for other applications with sparse data, where naïve Bayes has been shown to favorably compare to more sophisticated learning models such as decision trees and neural networks [*Domingos and Pazzani*, 1996; *Michie et al*, 1994].

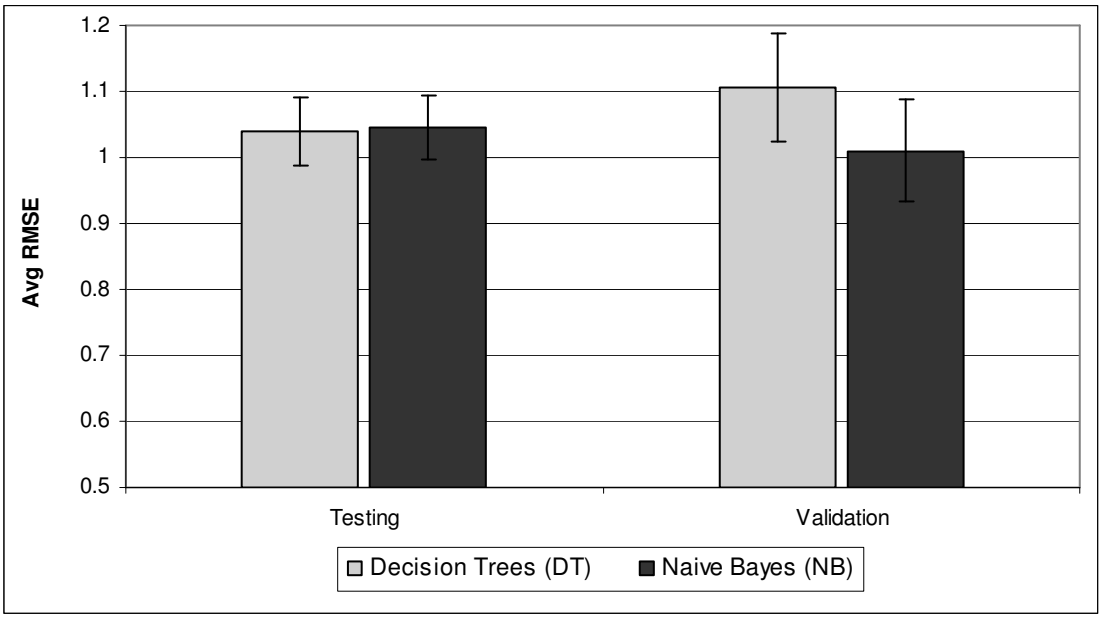


Figure 5.8 Testing and validation performance of decision tree and naïve Bayes without clustering (decision variables as inputs)

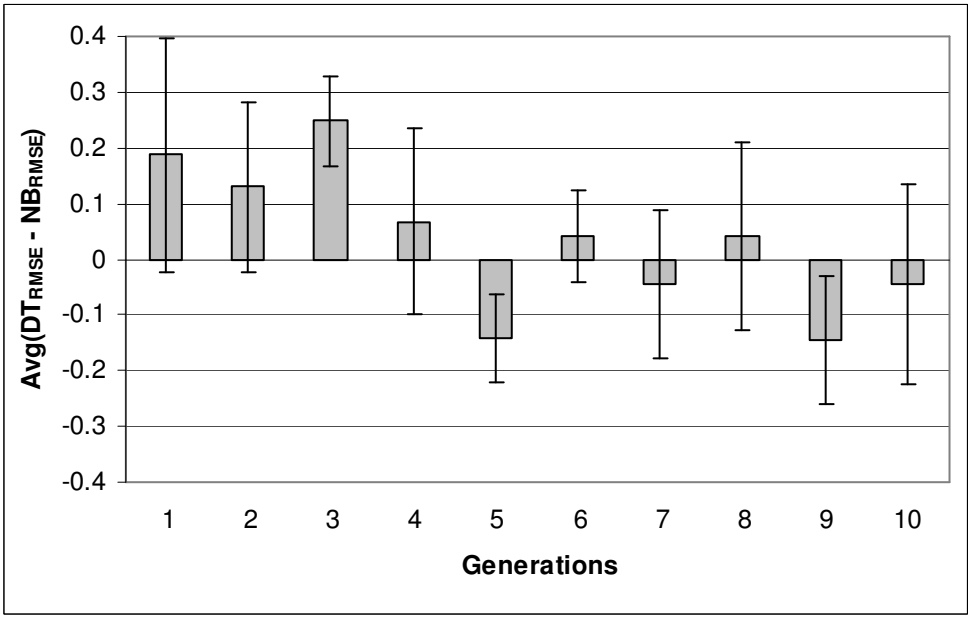


Figure 5.9 Average difference between the RMSE for decision tree and naïve Bayes plotted against generations (positive differences indicate that the decision tree performs worse than the naïve Bayes, and vice-versa)

Next, the machine-learning models were tested with the optimal clustering algorithm (N-cuts clustering) selecting conductivity fields for the user to rank, which were then used for the training data. Figure 5.10 shows the average root mean square for the decision tree and naïve Bayes models (with decision variables as input) for the testing and validation datasets. As can be seen from Figure 5.10, in general clustering led to an improvement in the prediction accuracy of the learning models. Of the 10 trials (5 each for decision trees and naïve Bayes), eight showed a decrease in the average error. The reduction in the error was an average of 8% for the naïve Bayes, while for decision trees the average error reduced by 4%. However, it is noteworthy that the clustering-based training led to an improvement of more than 20% (from an RMSE of 1.3 to 1.04) in the worst prediction error of the machine learning models.

To investigate the effect clustering has on the learning mechanism, the variance and min-max range of the ranks in the training set were also calculated for each trial. With clustering, the variance in the training set increased by more than 16% on average, ranging from 7% to 30% across all trials. In general, a more varied training dataset leads to better learning and a more generalizable prediction model – which is reflected by the lower errors in Figure 5.10. Figure 5.11 shows the min-max range of ranks within the training set with and without N-cuts clustering. The training set corresponds to solutions that have been shown to the expert; training sets with large ranges indicate that the expert evaluated many different types of solutions, which should not only improve the performance of the prediction model but also lead to better results for the IMOGA. As can be seen from Figure 5.11, random selection leads to a small range in solution types

shown to the expert. In fact, in early generations random selection leads to an average range of just 2 ranks shown to the expert (i.e. the best solution shown to the expert is on average rank 2, while the worst solution is on average rank 4). With clustering, however, the range of solutions shown to the user is consistently wider. Another important fact that can be inferred from Figure 5.11 is that with clustering the expert is shown rank 1 solutions (solutions that are preferred the most) as early as the third generation, while with random selection it takes the IMOGA 9 generations before a rank 1 solution is shown to the expert (and thus included in the training set).

Like the previous trials without clustering, the naïve Bayes classifier performed better than the decision trees, with an average of 8% lower RMSE compared to decision trees. As before, the naïve Bayes had stronger performance in early generations with smaller datasets.

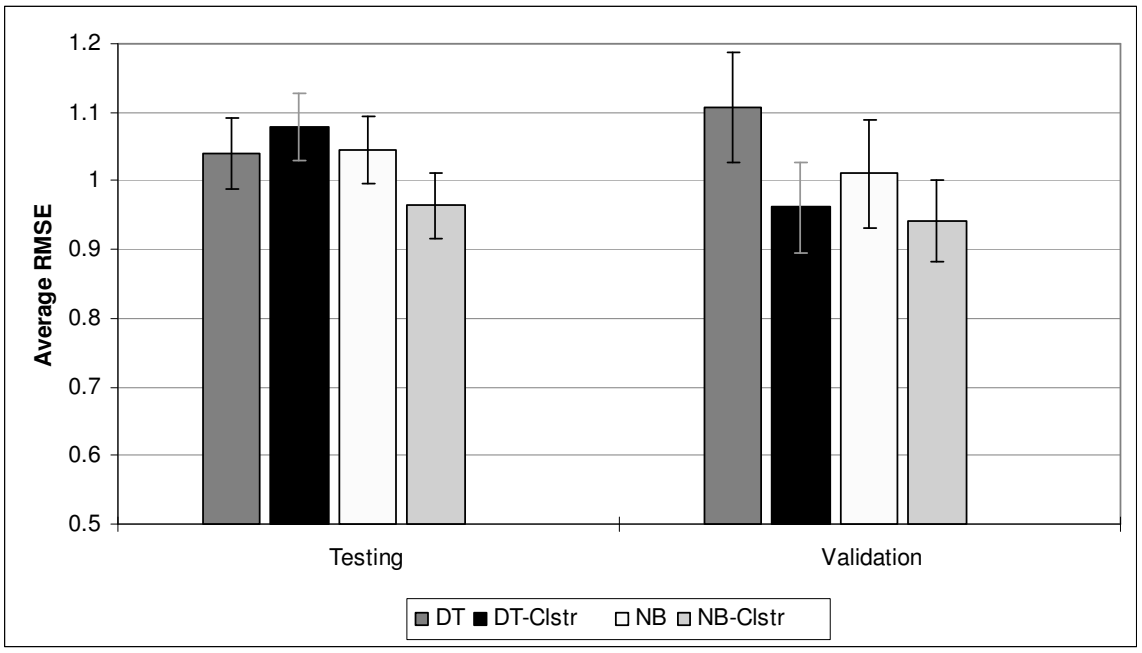


Figure 5.10 Average (across generations) RMSE for the learning models with N-cuts clustering

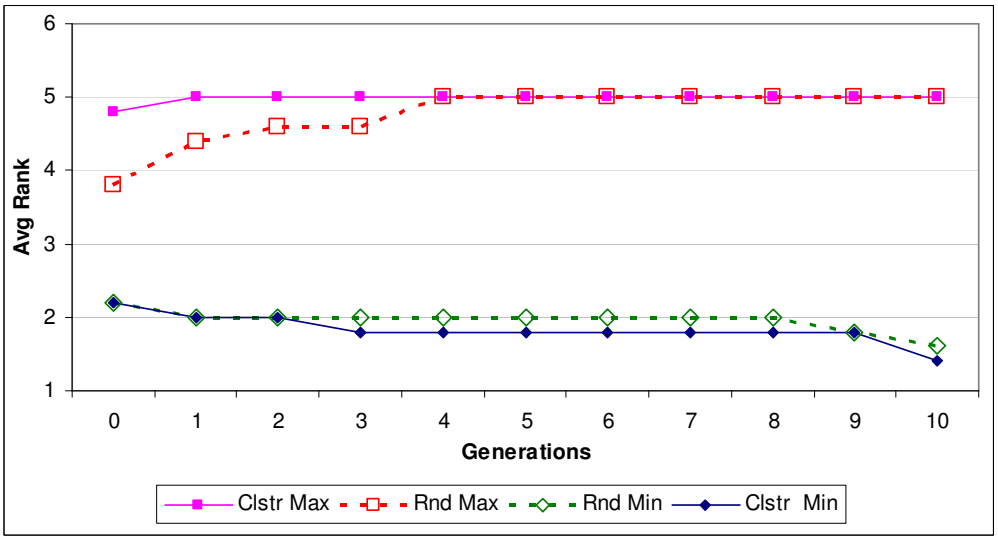


Figure 5.11 Average (across trials and populations) of minimum and maximum ranks in the training set with and without N-cuts clustering

The previous results were obtained with machine-learning models trained using only the decision variable values as input attributes. The final offline experimental results examine how the machine learning models trained with spatial information (nested spatial moments and eigenimage scores) perform. Previous results revealed that the naïve Bayes model performed favorably compared to the decision tree model. The best results were seen for the naïve Bayes model using clustering for the training data. Thus, subsequent results are presented to show the effect of spatial information on the best model found so far - naïve Bayes with N-cuts clustering. It is worth noting, that the experiments described below were also undertaken for the decision trees model, and results similar to Figures 5.8 and 5.9 were seen again, with the naïve Bayes model performing marginally better than the decision tree, especially in the early GA generations.

Figure 5.12 compares the root mean square errors for both the eigenimage scores and nested spatial moments attributes for training naïve Bayes models with clustering. Compared to the naïve Bayes model without any spatial information (given by NB+Clstr), additional spatial information in the form of both eigenimage scores and spatial moments was seen to consistently improve model predictions (although the eigenimage scores do not lead to much improvement for the testing trials). The improvement with spatial moments is the most significant, with a decrease of more than 12% in the root mean square error (compared to the original naïve Bayes model without clustering, this meant a total reduction of the RMSE by more than 20%). The average RMSE for naïve Bayes with spatial moments and with N-cuts clustering was approximately 0.86.

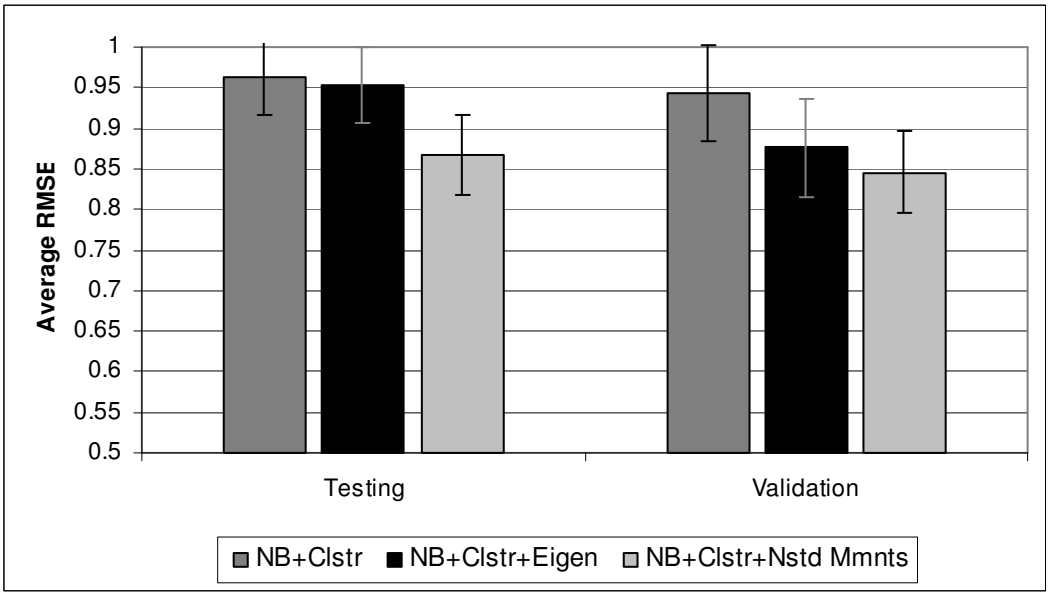


Figure 5.12 Comparison of RMSE of naïve Bayes (with clustering) with eigenimage scores and nested spatial moments as input attributes

5.3.3 Results with Online Interaction

The final set of results is for a real-time run of the IMOGA. For these results, two real-time expert sessions were carried out with the IMOGA. In the first session, the expert (the author) ranked all of the solutions in every population (i.e., 100 percent user interaction). In the second session, the user again ranked *all* of the solutions in every population but only half the ranked population was used directly in the IMOGA. This half was chosen using the optimal clustering algorithm (N-cuts clustering). The ranks of the other half were predicted by the optimal machine learning algorithm (naïve Bayes with nested spatial moments), similar to the methodology outlined in Figure 5.1 (i.e., 50 percent user interaction). Since the expert also provided ranks for the second half of the

population, the accuracy of the predictions from the naïve Bayes model could be compared with user-given ranks. Since the IMOGA uses half the ranks given by the user and half provided by the prediction model, this experiment demonstrates the effect the proposed methodology has on the solution quality of the IMOGA with only 50% user interaction. Figure 5.13 shows the online RMSE of the learning model in every generation. The RMSEs correspond to the difference between the ranks given by the expert and those predicted by the naïve Bayes model for the half of the population used to test the machine-learning algorithm. The prediction errors are higher in the initial generations, but as the archive grows the prediction errors are reduced. The average prediction error for the online run with 20 generations was found to be 0.92 (i.e., less than a prediction error of 1 rank). Finally, Figure 5.14 compares a rank 1 (as assessed by the expert) calibrated conductivity field from the run with 100% user interaction with the rank 1 conductivity field with only 50% interaction (both these solutions are almost identical in terms of the quantitative calibration objectives) . The figure shows that the overall spatial characteristics of the calibrated conductivity field are preserved with only half the user interaction.

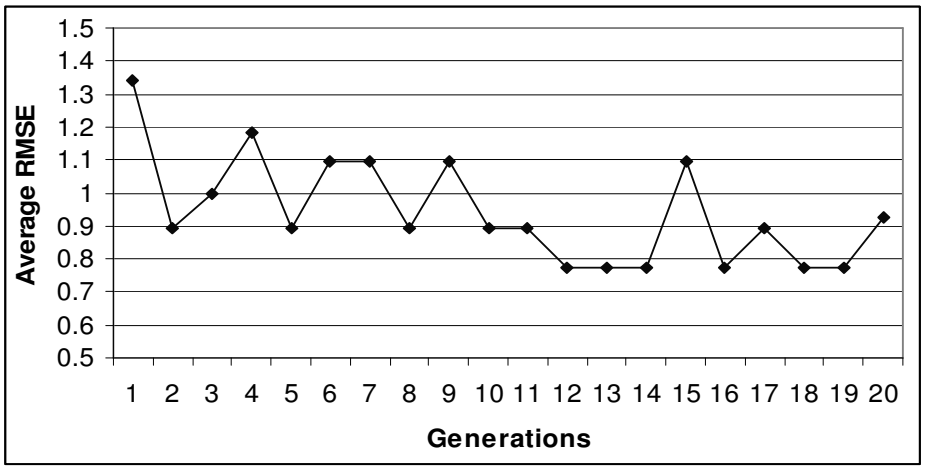


Figure 5.13 Average RMSE of selected prediction model for online user interaction

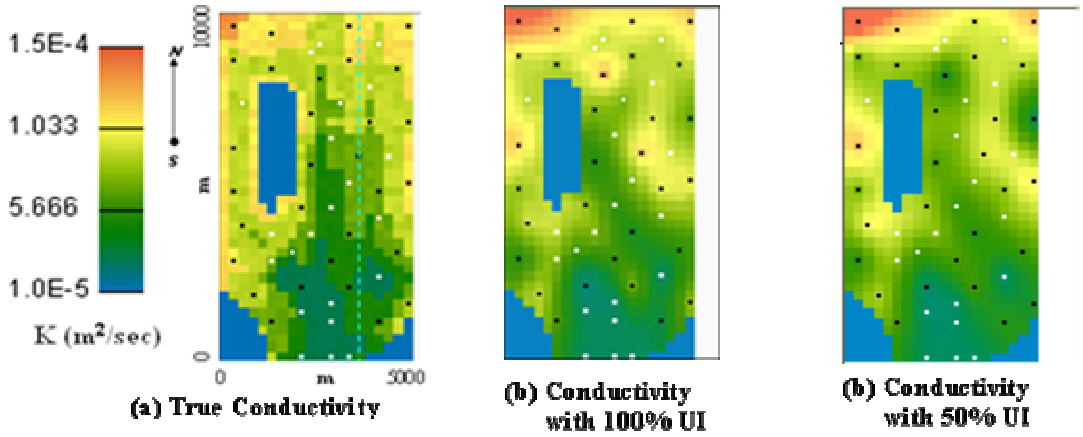


Figure 5.14 Comparison of rank 1 conductivity fields for the Freyberg case with 100% and 50% user interaction (the head error in both differs less than 10%)

5.3.4 Results for the Field-Scale WIPP Application

To test the generality of the findings of the previous sections, the best performing algorithm (naïve Bayes with nested spatial moments and N-cuts clustering) was applied to the field-scale WIPP model (see Section 3.2.2 for details). Similar to the Freyberg case, two runs were completed with the author serving as the simulated expert, with a population size of 20 for 10 generations. As before, 100% and 50% runs were undertaken to compare both the prediction accuracy and the solution quality of the IMOGA.

Figure 5.15 shows the online RMSE for the WIPP case, which has a population average RMSE between 1 and 1.5 (except for generation 10 when it falls below 1). The overall average is 1.21, which is higher than the average RMSE (0.95) for the Freyberg case. Note that an average RMSE of 1 corresponds to a difference of one rank between the user and the machine-learning model (for example, a rank 2 solution could on average be ranked as 3 by the machine-learning model). Thus, an RMSE of 1.21 is a reasonably

good result given the size and complexity of this problem. Moreover, the prediction errors of the learning model are seen to decrease with the GA generations (similar to the Freyberg case) with a final RMSE less than 1.

Figure 5.16 compares two solutions from the 100% and 50% IMOGA runs that were given the highest human rank (in this case, since the 'true' transmissivity field is unknown the ranks are given based on the amount of heterogeneity that is acceptable to the field expert for this site – highly smooth and overly heterogeneous fields tend to be given worse ranks and fields with mid-range heterogeneity are given better ranks – see Section 6.3 for more details about the preference rankings for the WIPP site). Figure 5.16 shows that the overall spatial patterns for both fields are very similar. Figure 5.16.b, the solution from the 50% IMOGA, does have more small-scale heterogeneities in the lower part of the field. This result likely occurs because such small-scale variability is averaged out when calculating the nested moments and thus the machine learning and clustering algorithms are not able to identify differences at such small scales. However, these results do show that for the scale of spatial information given to the learning model it is able to learn the user preferences for spatial characteristics of the transmissivity field.

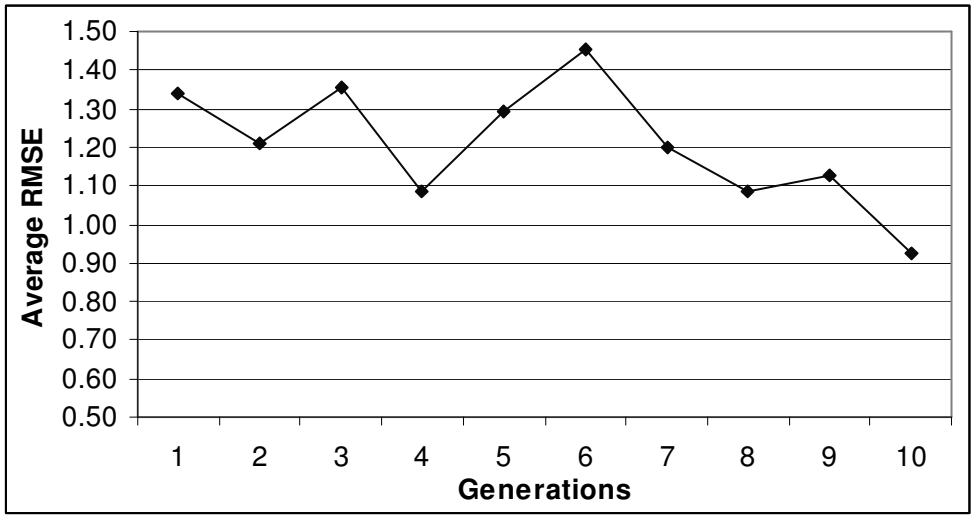
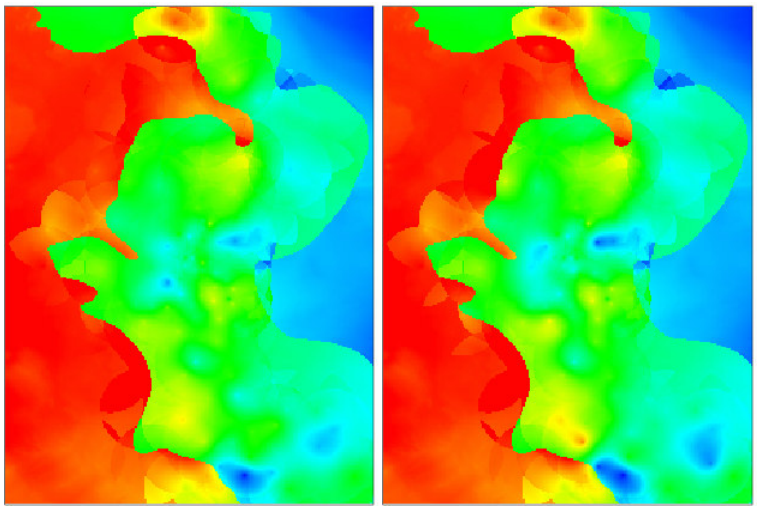


Figure 5.15 Average RMSE for the naïve Bayes model (with nested spatial moments and N-cuts clustering) for the WIPP case



(a)100% UI **(b)50% UI**

Figure 5.16 Comparison of rank 1 transmissivity fields for the WIPP site with 100% and 50% user interaction

5.4 Conclusions

User fatigue is an important concern for interactive environments such as the IMOGA. Not only can user fatigue reduce the quality of solutions found by the IMOGA (due to errors and noise in the ranking process), it also makes such an interactive framework difficult for use by experts for some real world problems. This research investigated approaches to make user ranking more efficient and effective and also reduce the burden on the user by employing machine-learning tools to learn user preference.

The first step in improving the efficiency of the ranking process is to reduce the redundancy and increase the diversity of solutions shown to the user through spatial clustering algorithms that group similar conductivity fields together and show the user samples from each group. Two types of clustering methodologies and four types of clustering attributes were tested, with the results indicating that the N-cuts algorithm with spatial information substantially improved clustering performance. N-cuts clustering with the entire 2-D matrix had the best performance overall, followed by N-cuts clustering with nested spatial moments.

The second step is to partially replace human rankings with predictions from machine learning models. Two learning algorithms - decision trees and naïve Bayes - were tested with and without clustering and with various input attributes for the prediction models. In general, the naïve Bayes model performed competitively with decision trees for large datasets but was observed to have better prediction accuracy for small datasets in the early stages of the IMOGA. Including clustering of the solutions in the user ranking

process consistently led to a significant improvement in the prediction accuracy of the learning models. This was attributed to an increase in the variance of the training data set and a wider range of solution types being shown to the expert. Including spatial information in the form of nested moments further improved the performance of the learning model, with the most significant improvement from the nested spatial moments. The effect on performance of including the eigenimage scores was at best equivocal – with the performance improving for some of the trials but deteriorating for others. Thus, using N-cuts clustering and naive Bayes learning based on nested spatial moments of conductivity fields was shown to be the most promising methodology to reduce the number of user evaluations required by the IMOGA leading to an improvement of more than 20% compared to the learning model without spatial information or clustering.

Finally the IMOGA was run with online user interaction in conjunction with the machine learning model. Results for both the hypothetical case and the field-scale application indicate that the solution quality of the IMOGA remained consistent with as little as 50% user interaction. Results for the WIPP site indicate that the proposed framework is applicable to a larger problem with more complex spatial characteristics. Over all, these results indicate that the proposed strategy is an effective tool in combating user fatigue for interactive decision making.

In conclusion, the focus of this study was the use of machine learning and image-analysis tools to improve the efficiency of the IMOGA. However, the methodologies that have been presented here have broad applicability to many different fields. The use of image-

processing tools to extract spatial information from hydrologic parameters (such as conductivity) is a particularly promising approach that can be applied to many other areas of environmental modeling. Particularly useful applications could be delineating rainfall patterns, zonation of aquifer conductivities/transmissivities, categorization of aquifers based on topology, reducing the number of stochastic realizations for uncertainty analysis (using clustering to identify predominant spatial patterns within the ensemble). Finally, other *visual* decision making frameworks and expert systems (similar to the IMOGA) could also benefit from such image-based approaches.

6 PREDICTIVE UNCERTAINTY ANALYSIS FOR THE INTERACTIVE FRAMEWORK

There are two kinds of probabilities

~From 'Degrees of Belief', by Stephen G. Vick

6.1 Introduction

The first two phases of this thesis developed the IMOGA framework (Chapter 4) and introduced efficiency enhancements (Chapter 5) to make the system more feasible. When applied to a groundwater model calibration problem this efficient IMOGA calibration framework converges to a Pareto front that represents the best tradeoff among model accuracy (calibration error), model complexity (regularization), and plausibility (assessed qualitatively by the expert). The solutions on the Pareto front correspond to different large-scale trends for the parameter fields (hydraulic conductivity in the Freyberg case study) that fit (to varying degrees) the field measurements and the experts' understanding of the site. An important use of such calibrated models is to make predictions to support environmental management decisions. The task of predictive uncertainty analysis is to characterize the uncertainty in predictions and relate it to uncertainty in parameters. The final chapter of this dissertation focuses on extending the IMOGA framework to consider both conceptual and stochastic uncertainty in the calibrated parameter fields and assess the impact these have on the predictive response of the calibrated model(s). To the best of our knowledge, this is the first time that a study has addressed both levels of uncertainty within a multi-objective and interactive groundwater calibration framework.

The two types of uncertainties considered in this research are conceptual and stochastic uncertainty. For the IMOGA calibration framework, conceptual uncertainty corresponds to the multiple alternative large-scale trends that arise due to the non-uniqueness of the inverse problem and the subjectivity in the expert's knowledge of the site (this is consistent with the 'equifinality' philosophy proposed by *Beven* [1993, 2000]). This is what is typically referred to as 'epistemic' uncertainty – resulting from inaccurate or incomplete information. Epistemic uncertainty itself can be due to lack of subjective knowledge (referred to as 'subjective' uncertainty) or due to inadequate field data (referred to as 'objective' uncertainty). This is reflected by the multiple large-scale parameter fields identified by the IMOGA that all honor the data as well as the expert's knowledge of the site. However, though the IMOGA does identify multiple possible parameter fields it does not yield the probability distribution for these large-scale trends. The challenge is to translate user preference and the likelihood metrics into a consistent 'weighting' or sampling function to express combined probabilities for these estimated trends. In this study a Bayesian framework is proposed to incorporate both objective (as given by the calibration and the regularization errors) and subjective (as given by the expert's assessment of the plausibility of different fields) likelihoods to weight the different calibrated models.

The second type of uncertainty is 'aleatory' or stochastic uncertainty – resulting from irreducible randomness in the system modeled. It is well-recognized in the groundwater literature that transmissivity/conductivity is highly variable (changing by orders of

magnitude) over very small scales. While epistemic uncertainty in the large-scale trend can be reduced by additional data, aleatory uncertainty in the small-scale fluctuations is for all practical purposes irreducible since it is virtually impossible to adequately measure aquifer properties at such small scales. For this reason, this small-scale variability is typically modeled as a stochastic process in the groundwater literature. The IMOGA is essentially a deterministic calibration framework that identifies multiple large-scale parameter fields. However, groundwater predictions related to contaminant transport are known to depend on the small-scale variability in the aquifer. Thus, the predictive uncertainty framework also needs to consider this additional source of uncertainty. This research proposes a methodology to also incorporate this small-scale variability in the IMOGA results.

This framework is developed and applied to a field-scale application based on the well-known Waste Isolation Pilot Plant (WIPP) case study (details of this case study are provided in Section 3.2) that has been the subject of extensive research on stochastic groundwater inversion [*LaVenue and Pickens*, 1992; *RamaRao et al*, 1995; *LaVenue et al*, 1995; *Capilla et al*, 1998; *Zimmerman et al*, 1998].

The rest of this chapter is organized as follows. The methodology section (Section 6.2) first presents the formulation used to solve the WIPP case study, highlighting the important differences between this case study and the Freyberg formulation presented in Section 4.2.2. It next presents the predictive uncertainty framework incorporating uncertainty in both the large-scale and small-scale transmissivity values. Section 6.3

presents the results of the IMOGA for the WIPP site. It next discusses the results for the predictive uncertainty analysis introduced in the previous section (Section 6.2). Finally Section 6.4 provides a summary of the results, concluding with the important findings of this research. Details on model averaging and model selection, topics that are pertinent to the research discussed in this chapter, are given in the background and literature review section (Section 2.5).

6.2 Methodology

This section is divided into two parts. The first sub-section (Section 6.2.1) presents an extension of the calibration formulation presented in Sections 4.2.1 and 4.2.2 that also considers a prior parameter field that is obtained through expert analysis before calibration. This section also discusses the design of the user-interface - an integral part of any interactive framework - for the WIPP case study (Section 6.2.1.2).

The second sub-section (Section 6.2.2) presents the uncertainty framework for post-calibration predictive analysis of the IMOGA results. As already discussed, two sources of uncertainty are considered in this work, and both sources of uncertainty are addressed in separate sub-sections (Section 6.2.2.1 for the large-scale uncertainty and Section 6.2.2.2 for the small-scale stochastic uncertainty).

6.2.1 Parameterization and Formulation of the Optimization Framework

The (quantitative) calibration formulation presented in Sections 4.2.1 and 4.2.2 was based solely on the head and conductivity measurements. Often, the field experts are also able to express a prior field that represents their understanding of the geology of the site. The

calibration formulation, in such cases, needs to consider not only the direct field measurements but also the prior field provided by the expert. For the WIPP case study this prior field consists of the 'base field' that represents predominant geological features that are thought to exist at the regional scale (details about the base field are given in Section 3.2.2).

A schematic for using pilot point calibration with a prior base field is shown in Figure 6.1 for a one-dimensional case. Given the base field, the data points and the pilot point values represent deviations from the base field (in terms of log T (m^2/s) values). The perturbations from the pilot points and the residuals from the measured data are interpolated, using 'Ordinary Kriging' [*Deutsch and Journel*, 998] with a variogram based on the residuals of the direct T data (see Appendix A for details), to give a 'residual field' that can then be added to the base field to yield a final transmissivity field that exactly matches the measured data but deviates from the base field wherever the pilot points have non-zero values (note that this methodology is essentially the same as kriging with an external drift [*Deutsch and Journel*, 1998]). This schematic is essentially the same as the one used in *DOE/WIPP* [2004], the difference being the regularization objective that differs from *DOE/WIPP* in that it considers both the field data and the base field in its formulation. Details of this formulation are given in Section 6.2.1.1.

Note that the all calculations for the WIPP study are carried out in the log-transmissivity space; to simplify notation 'T' is assumed to be synonymous with log T, with the implicit understanding that all calculations and interpretations are in the log T domain.

6.2.1.1 Quantitative Objectives

The pilot point 'deviations' from the base field are the decision variables for the optimization algorithm, and are optimized to minimize the calibration error for the model and the measured data. As before, the calibration fit - expressed as the weighted root mean square error between the model prediction and the head measurements (identical to Equation 4.1, in Section 4.2.2) – forms the first quantitative objective for calibration.

To control the magnitude and correlation of the perturbations, regularization needs to be used as an additional objective. The perturbations for the pilot points are controlled in two ways – first, constraints are fixed on the minimum and maximum allowable deviation from the base field. The range for the pilot points is set at ±3 (in terms of log T values) in the mid-transmissivity region (shown by green in Figure 3.5) and ±1 (also in terms of log T) in the high-transmissivity region (shown by red in Figure 3.5). These ranges are recommended by *DOE/WIPP* [2004] and *McKenna and Hart* [2003] because the base field fits the measurements in the high-transmissivity zone relatively well, but does not do so in the middle range. Moreover, the WIPP site and most of the head data are located within the middle-T region (see Figure 3.5) and thus the transmissivities in this region are thought to be more critical for the hydraulic behavior of the site. In addition to setting constraints on the range of pilot point perturbations, an additional regularization objective (T_{err}) is defined as shown in Equation 6.1, below.

$$Min \quad T_{err} = \left(\frac{\left[T_k - K\left\langle D_T \right\rangle\right]^T C_{\Delta T}^{-1} \left[T_k - K\left\langle D_T \right\rangle\right)\right]}{n_{pp}} \right)^{1/2} \tag{6.1}$$

where n_{pp} is the number of pilot points, T_k is an n_{pp} dimensional vector with the log T residual values for the pilot point locations (the decision variables for optimization) for the k^{th} field, and $C_{\Delta T}$ is the estimation covariance for the pilot point locations based on the residuals of the log T data from the base field (similar to C_{pp} used in Section 4.2.2, see Appendix A for details). The final term, $K\langle D_T \rangle$, is an n_{pp} dimensional vector which is populated with values kriged at the pilot-point locations from the log T data residuals, wherever such an estimate is available. For pilot points located beyond the range of measurement data (such as the ones near the north and south boundaries of the WIPP model – see Figure 3.5) this vector is set to zeroes (so that the interpolated residual field does not perturb the base field in these regions). In keeping with the regularization objective proposed by *Moore and Doherty* [2005], the $K\langle D_T \rangle$ vector contains 'prior' or preferred values for the pilot point perturbations - ensuring that pilot points close to measurements have perturbations close to the residuals for these data points, and pilot points in regions with little or no data are as close to the base field as possible.

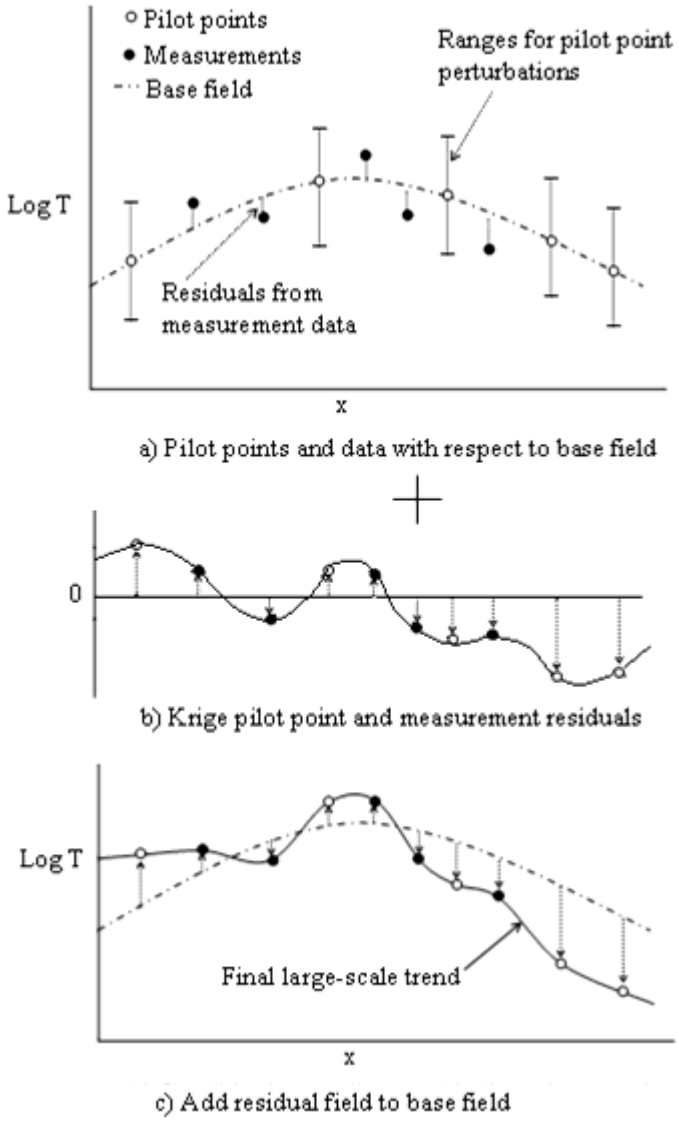


Figure 6.1 Pilot point calibration with field data and prior transmissivity field

6.2.1.2 Qualitative Objective

In addition to quantitative calibration objectives, the additional qualitative criterion is based on an expert's assessment of the plausibility of different transmissivity fields. The expert interaction for the WIPP case study was given by Dr. D. Walker, who is familiar with the hydrogeology and the geospatial characteristics of the WIPP site.

As for the Freyberg case (Section 4.2.3), the expert gave ranks from 1 (best) to 5 (worst) through a graphical user interface that displayed potential solutions. As with any interactive system, this user interface is critical in extracting meaningful information from the user. Thus, the user interface for the WIPP study was carefully designed based on consultation with the field expert involved. Figure 6.2 shows the graphical user interface for the WIPP case study. As for the Freyberg user interface (Figure 4.4), candidate transmissivity fields (far left panel in Figure 6.2) and predicted head fields (far right panel in Figure 6.2) were visualized for the expert. However, since the scale of the study was large, the expert believed that it was necessary to contextualize these fields by overlaying important geographical features and boundaries on the calibrated and predicted fields. Thus, four important geological boundaries – Salado dissolution boundary, Nash draw, and two halite detection boundaries (marked as M3/H3 and M2/H2 corresponding to the presence of Halite above and below the Culebra, respectively) – as well as the WIPP site boundary, were overlaid on the transmissivity and head fields (see the far left panel in Figure 6.2). In addition, the expert was also interested in visualizing the spatial distribution of the errors in the head predictions. To enable this, the individual measurement wells were shown with the head errors color coded based on their magnitudes (the middle panel in Figure 6.2). The sizes of the three graphic panels were fixed to minimize the amount of scrolling required by the user (which would add to user fatigue). However, sliding scales and zooming controls were provided to allow the expert to explore each panel in more detail. Additional information about the quantitative objectives was also provided for each solution (text in the horizontal strip at the top of the

interactive panel). Based on all of this information, the expert is then asked to give the solution a rank from 1 (best) to 5 (worst). A sample ranking panel for the WIPP site is shown in Figure 6.3.

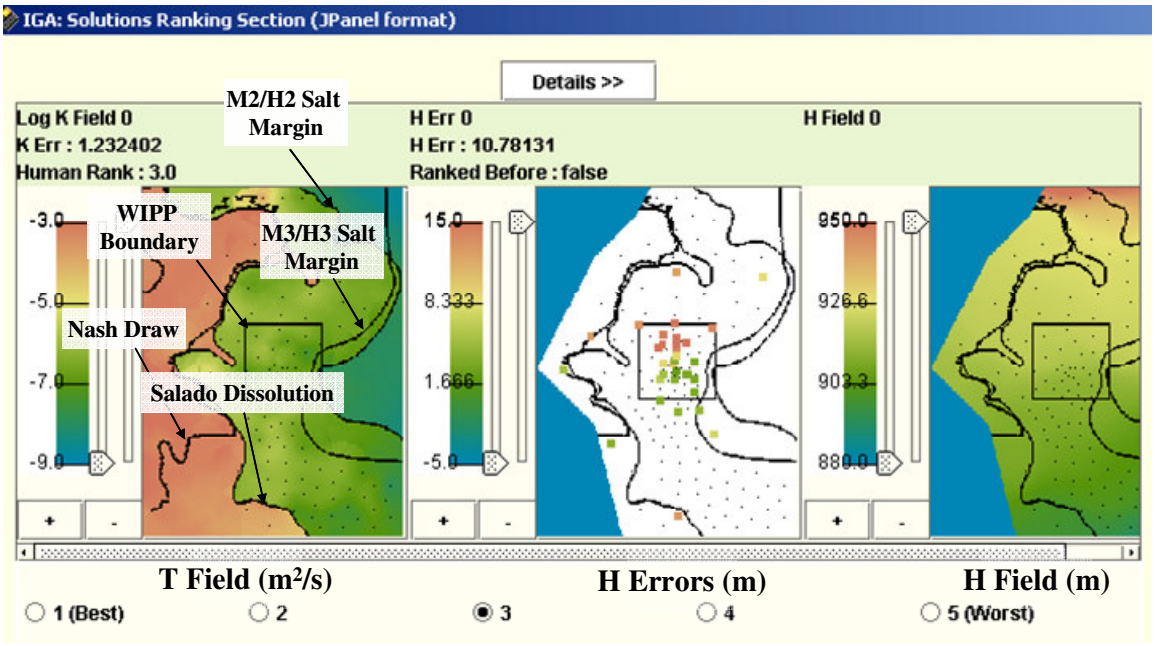


Figure 6.2 Graphical user interface for the WIPP site

6.2.2 Uncertainty Analysis of IMOGA Results

Figure 6.3 shows the overall framework for uncertainty analysis for the IMOGA results (for a one-dimensional case, similar to Figure 6.1). Once the IMOGA (developed in Chapters 4 and 5) is applied to the calibration problem it yields multiple solutions that represent the best tradeoff between calibration error, spatial heterogeneity, and expert preferences. Each solution on the Pareto front represents an alternative model for the large-scale structure of the underlying transmissivity field conditioned on the hydraulic head measurements, the direct transmissivity measurements and the experts'

understanding of the site. This study assumes that the uncertainty in the large-scale trends is reflected by the multiple transmissivity fields found by the IMOGA. While it is true that there may be many more transmissivity fields that are not found by the IMOGA (due to the restricted population size of the IMOGA and the inherent bias of the optimization approach) the fields that *are* part of the final Pareto front represent the most likely fields (as assessed by different goodness-of-fit measures and subjective plausibility) given a particular population size. It is thus assumed that a) the IMOGA solutions have good coverage across the Pareto front and thus adequately represent the diversity of solutions across the front, and b) solutions far away from the Pareto front are 'less likely' (and would thus be given very small or zero likelihood weights) and would not have significant impact on the predictive uncertainty assessment analysis. Moreover, the methodology presented in this research is based on *Mugunthan and Shoemaker* [2006], who showed that results from calibration optimization (such as the IMOGA) can be used for uncertainty assessment (with the appropriate likelihood measure to reduce the bias in the solution set).

It has been shown [*McLaughlin and Townley*, 1996] that hydraulic head fields are most sensitive to the large-scale trend in the transmissivity field; but this is not the case with contaminant transport, which is heavily dependent on small-scale variability. Since the optimal transmissivity fields found by the IMOGA do not have this small-scale variability it needs to be added on to these large-scale fields.

To do this, the log transmissivity field $T(x)$ is decomposed into a large-scale trend $T_L(x)$ and small-scale variability $T_S(x)$ [*Mclaughlin and Townley*, 1996; *Alcolea et al*, 2005] as:

$$T(x) = T_L(x) + T_S(x) \tag{6.2}$$

The concept of 'scale' is largely 'functional' here. The 'large-scale' trend ($T_L(x)$ in Equation 6.2) is the trend estimated from the given head data, while $T_S(x)$ represents the *unidentifiable* small-scale fluctuations from this large-scale trend. Before conducting any type of predictive uncertainty analysis, it is important to address uncertainty both in the estimated large-scale structures and the unidentifiable small-scale heterogeneity.

Since kriging is known to yield the best unbiased linear estimator of the data values [*Cressie*, 1993], the different calibrated fields can be thought to represent multiple alternative mean values for the large-scale trend 'T_L(x)' of the transmissivity field. However, since kriging yields an 'average' field, it tends to smooth out the local fluctuations that comprise $T_S(x)$. (In general, the scale of heterogeneity that can be represented in the calibrated field can, at most, be at the resolution of the pilot point locations.) In this work, the small-scale variability is treated within a stochastic framework and multiple (equally likely) realizations are created conditioned on each (Pareto) optimal large-scale structure found by the IMOGA (see Figure 6.3).

The prediction of concern to the decision makers - particle travel times in the case of the WIPP case study - are then simulated for each set of realizations (corresponding to a particular large-scale trend) to yield a probability distribution function for a particular large-scale trend. These probability distribution functions are then combined using a

model-averaging paradigm. This study uses both the generalized likelihood uncertainty estimation (GLUE) [*Beven and Binley,* 1992] and Bayesian model averaging (BMA) [*Hoeting et al*., 1999; *Neuman,* 2002; *Neuman*, 2003] for model averaging.

It is important to point out that the small-scale variability generated stochastically is still at the same grid size used in the groundwater model. In other words, the proposed approach does not address the sub-grid variability (i.e. variability in the transmissivity values within the 100m x 100 m grid) within a stochastic framework. However, the proposed methodology can be extended to consider this source of uncertainty in future research. One way of doing so would be to use 'random-walk' models [*Prickett et al.*, 1981] instead of the deterministic MODPATH [*Pollock*, 1994] (used for the WIPP case study predictions – see Section 3.2.3 for details) to simulate the travel times and travel paths for the stochastic simulations. Such random walk models take into consideration the effects of sub-grid dispersion, yielding a probability distribution for the contaminant transport predictions. Another approach would be to sub-divide the modeled domain into smaller grid sizes (downscale) and conduct stochastic simulation at the smaller scale while maintaining the effective transmissivity found at the larger scale (see pages 596 – 602 of *Chiles and Delfiner* [1999], for a discussion on up/downscaling of permeability).

The rest of this section is divided as follows. The first sub-section (Section 6.2.2.1) addresses uncertainty for the large-scale model and gives details about the GLUE and BMA model-averaging techniques. The section that follows (Section 6.2.2.2) implements

a methodology to incorporate small-scale heterogeneity in the transmissivity fields without disturbing the optimal large-scale structure found by the IMOGA.

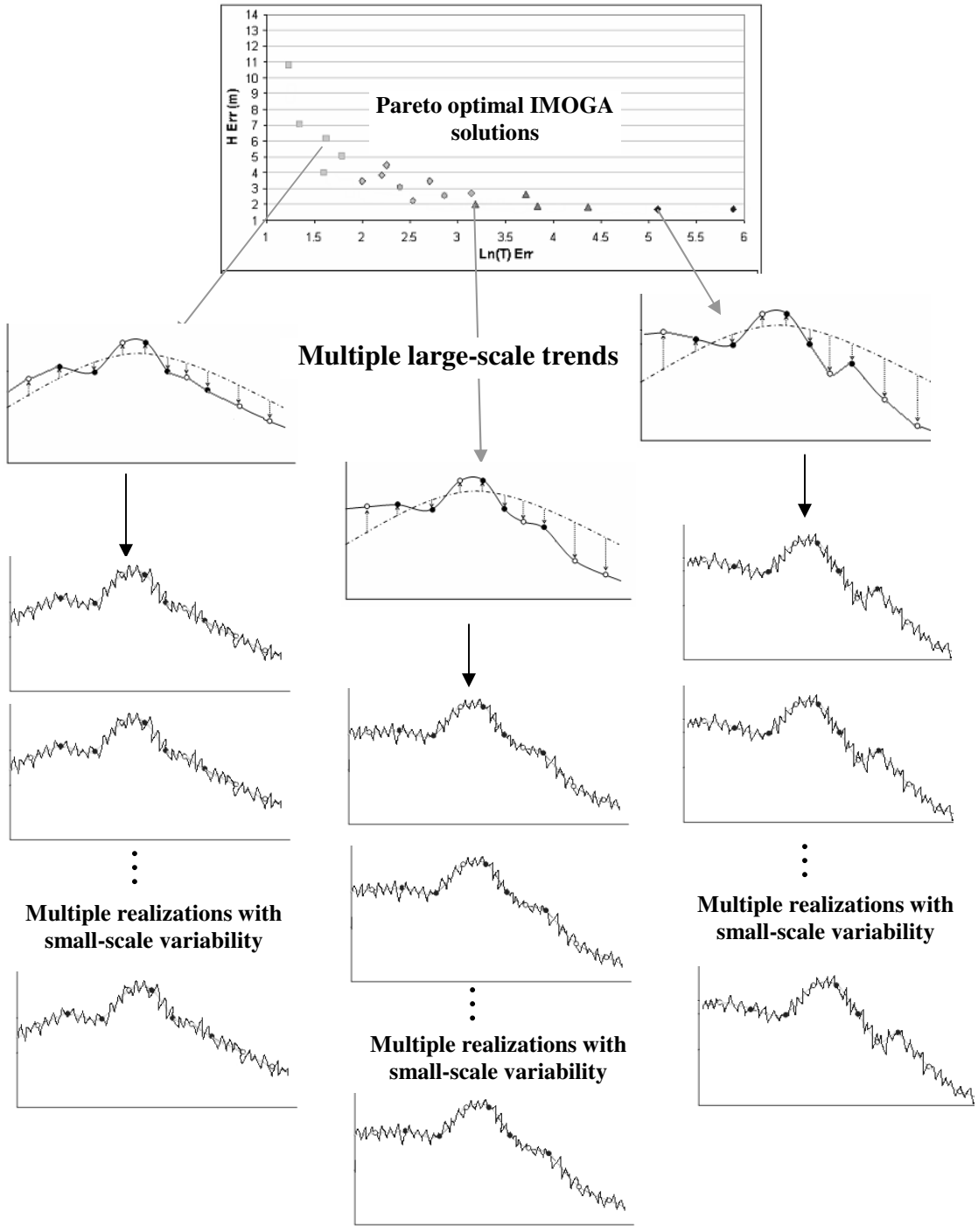


Figure 6.3 Uncertainty assessment framework for the IMOGA

6.2.2.1 Uncertainty in Large-Scale Trends

The non-uniqueness of the inverse problem and the subjective uncertainty in the expert's knowledge of the site lead to multiple transmissivity fields identified by the IMOGA. Given the large-scale transmissivity fields identified by the IMOGA, there are multiple sources of information – such as calibration error, field measurements, and/or expert judgement - that can be used to assess the likelihood for a particular calibrated field.

To use the IMOGA fields within a probabilistic framework, it is necessary to weight each trend model based on how likely it is. This weight should correspond to the 'belief', 'likelihood' or 'probability' for that transmissivity field. Each transmissivity field found by the IMOGA has a certain fit with the head data (D) and also a particular regularization level. In this case, the regularization level can be assumed to correspond to a certain large-scale heterogeneity level (S_k) with respect to the prior base field. Given D and S_k, a Bayesian framework can be used to write the probability for a particular transmissivity field T_k as:

$$P(T_k|D, S_k) = P(D|T_k, S_k)P(T_k \mid S_k)P(S_k) \quad (6.3)$$

where $P(T_k|D, S_k)$ is the probability (or likelihood) of the k^{th} transmissivity field (T_k), given the calibration data (D) and the level of heterogeneity S_k. $P(D|T_k, S_k)$ is the probability of getting the field data D given the transmissivity field and the level of heterogeneity. This is what is referred to as the 'direct PDF' [*Townley and McLaughlin*, 1996] and can be estimated from the error in predicting the field data using transmissivity field T_k. $P(T_k|S_k)$ is the likelihood for a particular transmissivity field T_k for S_k level of

heterogeneity. Finally, P(S_k) is the 'prior' likelihood for that large-scale heterogeneity as assessed (interactively) by the expert.

Different methodologies exist to calculate P($T_k|D\ S_k$). In this work two approaches are implemented based on a) the GLUE framework proposed by *Beven and Binley* [1992] and b) the MLBMA (maximum likelihood Bayesian model averaging) approach used by *Raftery et al.* [1996] and *Neuman* [2002, 2003] (see section 2.5 for a discussion on different approaches to addressing model uncertainty). Details of both these methods are given in the following paragraphs.

The GLUE framework requires likelihoods for the first two terms in Equation 6.3. This work uses a well-known metric for likelihood that has been used in various GLUE applications (*Binley and Beven*, 1989; *Beven and Binley*, 1992]. This likelihood metric is based on the variance of the predictions with the data and is given by:

$$L(\mathrm{D} \mid T_k, S_k) = (\sigma^2_{Herr,k})^{-\mathrm{N}} \tag{6.4}$$

where $\sigma^2_{Herr,k}$ is the head prediction error variance for the model given by:

$$\sigma^2_{Herr,k} = \left[\frac{\left[\mathrm{D_H} - \mathrm{GW}\langle \mathrm{T_k} \rangle\right]^T C_H^{-1} \left[\mathrm{D_H} - \mathrm{GW}\langle \mathrm{T_k} \rangle\right]}{n_{obs}} \right] \tag{6.5}$$

N in Equation 6.4 is what is called a shaping factor [*Beven and Binley*, 1992] and is normally set to 1. Note that in Equation 6.4 the notation is changed from P($D|T_k,S_k$) to L($D|T_k,S_k$) to denote the likelihood as opposed to the probability, which is more in keeping with the GLUE methodology. Also note that Equation 6.5 is simply the square of the head prediction RMSE that was used as a calibration objective (Section 6.2.1.1).

Based on the work of *Townley and McLaughlin*, 1996] and *Neuman* [2002, 2003], $P(T_k|S_k)$ can be written in a similar likelihood form to Equations 6.4 and 6.5:

$$L(T_k|S_k) = (\sigma^2_{Terr,k})^{-N} \tag{6.6}$$

where

$$\sigma^2_{Terr,k} = \left[\frac{\left[T_k - K\langle D_T \rangle\right]^T C_\Delta^{-1} \left[T_k - K\langle D_T \rangle)\right]}{n_{pp}} \right] \tag{6.7}$$

where the definitions of the terms remain the same as Equations 6.1. It is worth pointing out that the traditional GLUE approach does not take model complexity into account and only weight the models based on Equation 6.4. However, in this work model complexity, as measured by the regularization of the parameter fields, is explicitly used within the GLUE framework.

Since the sum of all probabilities should be equal to 1, Equation 6.3 can be written as:

$$L(T_k \mid D, S_k) = \frac{\dfrac{P(S_k)}{\sigma^2_{Herr,k}\sigma^2_{Terr,k}}}{\displaystyle\sum_{k=1}^{K} \frac{P(S_k)}{\sigma^2_{Herr,k}\sigma^2_{Terr,k}}} \tag{6.8}$$

The MLBMA approach does not use likelihood functions, but instead derives the probability for a particular transmissivity field (T_k) as a function of the Bayesian information criterion (see *Neuman* [2003] for the derivation). According to this $P(T_k|D,S_k)$ can be written as:

$$P(T_k \mid D, S_k) = \frac{\exp\left(-\frac{1}{2}\Delta BIC_k\right) p(S_k)}{\sum_{k=1}^{n} \exp\left(-\frac{1}{2}\Delta BIC_k\right) p(S_k)} \tag{6.9}$$

where

$$\Delta BIC_k = BIC_k - BIC_{\min} \tag{6.10}$$

P(S_k) is the subjective probability for the particular transmissivity field based on expert judgment, BIC_k is Bayesian information criterion for the transmissivity field T_k defined as below, and BIC_{min} is the minimum BIC among all competing models (alternately BIC_{min} can also be the minimum *possible* BIC measure that sets a standard to compare the different models).

$$BIC_k = NLL_k + n_k Ln(n) \tag{6.11}$$

where n is the number of calibration data, n_k is the number of parameters, and NLL_k is the negative log-likelihood estimation for T_k and is given by:

$$NLL_k = -2\ln P(D \mid T_k) - 2\ln P(T_k) \tag{6.12}$$

and

$$\begin{aligned} LnP(D \mid T_k) &\propto -\frac{1}{2}\left[\left[\mathrm{D_H} - \mathrm{GW}\langle \mathrm{T_k} \rangle\right]^T C_H^{-1}\left[\mathrm{D_H} - \mathrm{GW}\langle \mathrm{T_k} \rangle\right]\right] \\ LnP(T_k) &\propto -\frac{1}{2}\left[\left[T_k - K\langle D_T \rangle\right]^T C_\Delta^{-1}\left[T_k - K\langle D_T \rangle\right)\right] \end{aligned} \tag{6.13}$$

For this work, n_k represents the contribution from each pilot point. As such the number of pilot points is fixed; however some of them may have zero or near zero values. Thus, n_k is calculated as:

$$n_k = \sum_{p=1}^{n_{pp}} \frac{T_p}{\max_k (T_p)} \tag{6.14}$$

where T_p is the value for the pilot point for the k^{th} transmissivity field, $\max_k(T_p)$ is the maximum value (across the different transmissivity fields) for that pilot point from all the transmissivity fields (or 1 if all values for that pilot point are zero), and n_{pp} is the total number of pilot points. In other words, n_k is the sum of the ratio of each pilot point's value to the maximum value (over all the transmissivity fields) for that pilot point. The maximum value of n_k is n_{pp} (when all the pilot points are at their maximum value), and the minimum value of n_k is zero (when all the pilot points are zero).

Note that in their work *Neuman* [2002; 2003] and *Ye et al.* [2004] use the allied Kashyap information criterion (KIC) for the probabilities, which requires calculation of the second-order derivatives of the calibration measure with respect to the model parameters. For highly parameterized (having a large number of calibration parameters/pilot points) problems such as the WIPP site, this poses a computational challenge (*Neuman* [2003] and *Ye et al.* [2004] have demonstrated the use of KIC for problems with relatively small parameter dimensionality). Moreover, as the number of data points increase, KIC becomes asymptotically equivalent to BIC. On the other hand, in the absence of extensive data the estimation of the Fisher information criterion is unreliable, making the benefit of this extra term questionable [*Kashyap*, 1982; *Sclove*, 1994; *Rutledge*, 1995]. Another alternative is to use the Akaike information criterion (AIC) given by $NLL_k + 2n_k$. However, AIC does not take into consideration the amount of calibration data when penalizing model complexity (note that unlike KIC and BIC the *Ln(n)* term is not part of AIC) and tends to favor higher dimensional models compared to BIC and KIC [*Rutledge,*

1995]. For these reasons, BIC is used in the proposed Bayesian framework. This criterion has been successfully shown to work for a host of different problem types [*Raftery et al.*, 1996; *Volinsky et al.*, 1997; *Hoeting et al.*, 1999].

It is worth noting that while the GLUE methodology depends on the variance of the errors (and is thus independent of the number of data samples), the MLBMA weights depend on the weighted sum of residuals (and thus grow exponentially with the number of data points). This has an important consequence in the weightings for each of these methodologies.

Both the GLUE and MLBMA approaches have the $P(S_k)$ term in Equation 6.8 and 6.9 and allow for the expert preferences to be included in the weighting scheme. Thus, both these approaches allow for subjective likelihoods or probabilities to be included with quantifiable (based on the field data) probability measures when calculating posterior probabilities for each large-scale trend model based on both the quantitative and qualitative criteria. In traditional model-averaging frameworks the expert is expected to express $P(S_k)$ as a prior statistical distribution expressing his or her conception of the spatial distribution for the parameter field – a challenging task in most cases. In the interactive framework, this is not required from the expert, who can simply express the degree to which a particular T_k field (found by the IMOGA) satisfies his or her conception of reality with the probability $P(S_k)$. In this work, the $P(S_k)$ term is directly provided by the expert, based on visual inspection of transmissivity images, from each rank of the final IMOGA solution set. These probabilities represent the relative

probabilities for the occurrence of each type of solution based on the expert's judgment. As expected, the probabilities would be correlated with the ranks; however, the relative importance of one rank over the other can be controlled by the expert in this post-calibration phase. This is one way of 'eliciting' probabilities from the experts. In the future, more formal methods of 'probability elicitation' as proposed by *Lau et al*. [1998] and *van der Gaag et al.* [1999], among others, may be used to obtain these probabilities.

Once the likelihoods/probabilities for each transmissivity field have been calculated from Equations 6.8 or 6.9, these can be used as weights to combine the output from each of the field within a Bayesian model averaging paradigm [*Raftery et al.*, 2003] as:

$$P(y \mid D, H_k) = \sum_{k=1}^{n} P(y \mid T_K) P(T_k \mid D, S_k) \tag{6.15}$$

where $P(y|T_k)$ is the probability for the prediction one is interested in (such as contaminant travel time or breakthrough curves estimated using analytical stochastic techniques or Monte Carlo sampling) conditioned on the calibration data (and with a specified level of model heterogeneity S_k), $P(y|T_k)$ is the probability of that prediction for the k^{th} transmissivity field, $P(T_k|D,S_k)$ is the probability (or likelihood) calculated as in Equation 6.8 or 6.9, and n is the total number of large-scale transmissivity fields.

6.2.2.2 Incorporating Small-Scale Variability

The methodology discussed in the previous section addressed uncertainty in the large-scale structure for the transmissivity field. Since kriging is used to estimate these large-scale trends, the resulting transmissivity fields are smoothed out, leading to low local variance. Such smooth interpolated maps should not be used for predictions such as

contaminant transport that are known to be highly sensitive to local variability and connectivity [*Gooevaerts*, 1997]. The goal of incorporating small-scale variability is to generate multiple realizations of the transmissivity field, each conditioned on a particular optimal large-scale structure found by the IMOGA, but without the smoothing artifact of kriging.

In this work, 'conditional sequential simulations' [*Goovaerts*, 1997; *Deutsch and Journel, 1998; Chiles and Delfiner*, 1999] are used to represent the local variability in the simulated field. Since these conditional simulations need to maintain the (Pareto optimal) calibration objectives found by the IMOGA, the challenge is to create conditional realizations that simulate the small-scale variability *while still preserving the large-scale spatial trend and the covariance structure* of the transmissivity fields (found by the IMOGA). This is necessary because the head calibration errors are sensitive to these aspects of the transmissivity fields.

This work uses a new class of simulation algorithms called 'direct sequential simulation' [*Journel*, 1994; *Soares*, 2001; *Oz et al.*, 2003; *Hansen and Mosegaard*, 2007]. This methodology rests on the important theoretical result [*Journel*, 1994] that, in order to reproduce a given covariance model, the successive conditional distributions used in the sequential path can be of any type as long as they identify the kriging mean and variance. The main advantage to using this simulation paradigm is the freedom from the multi-Gaussian assumption that is required for the more popular sequential Gaussian simulation (SGS) algorithm [*Goovaerts*, 1997; *Deustch and Journel*, 1998]. Consequently the

simulation algorithm uses the original data and covariance model directly without requiring any transformations (unlike Gaussian simulation that requires the data and covariance model to be transformed into a normal distribution) – making them very convenient and efficient to use with the IMOGA. For this work the simulation data and variogram used for the direct simulations are based on the transmissivity measurement residuals (see Appendix A).

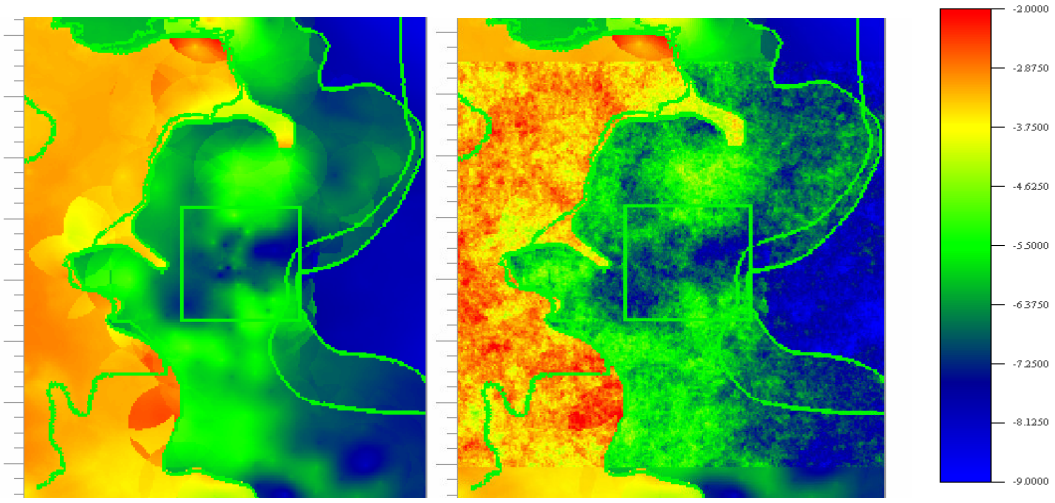
A commercially available geostatistical software package called 'VISIM' [*Hansen and Mosegaard*, 2007] was used to implement direct conditional simulation. To preserve the large-scale structure conditional simulation is carried out with a locally varying mean [*Deustch and Journel*, 1998] obtained from the kriged, optimal, large-scale residual transmissivity field. Since the current version of VISIM does not allow for specifying a locally varying mean, this feature was incorporated in the package by allowing it to read these values from the transmissivity file created during the IMOGA process.

Initial experimentation showed that direct simulation was superior in preserving the local average and the covariance structure for the large-scale transmissivity fields compared to the more popular SGS approach. This is consistent with the work done by *Journel* [1994] and *Caers* [2000] who have shown such techniques are better at preserving the local accuracy of kriging algorithms. Moreover, *Caers* [2000] also advocated using direct simulation with untransformed data because conditional simulations that work with transformed data (such as SGS) can only preserve the covariance of the transformed data, with no guarantee of preserving the original data covariance. Finally, this approach

avoids another drawback of the SGS approach – the reduction in the connectivity of high values, possibly leading to biased estimates for contaminant predictions [*Goovaerts*, 1997].

In addition to being sensitive to the large-scale structure of the transmissivity field, the head calibration error was also found to be extremely sensitive to the transmissivities close to the fixed head boundaries. Thus, to preserve the optimal calibration errors found by the IMOGA, conditional simulation was conducted away from the boundaries of the WIPP model. In essence this allowed the simulated transmissivity fields to maintain the optimal boundary transmissivities that were identified by the IMOGA.

An example of the conditional realization is shown in Figure 6.4, where the optimal kriged field and the realization conditioned on this field are shown. Figure 6.4 shows that the conditional simulation maintains the local mean structure of the kriged field while incorporating small-scale variability in the transmissivity field. Also note that the upper and lower boundaries of the WIPP model are not perturbed as part of the stochastic simulation in order to maintain the head calibration error that was most sensitive to these boundaries. Since the predictive analysis only needs to be undertaken within the WIPP site boundary (shown by the green rectangle in the middle of each Figures 6.4 a and b) this 'screening' ensures that the boundary effects are negligible on the predictive analysis for the realizations.



(a) (b)

Figure 6.4 (a) Optimal kriged transmissivity field and (b) one of the realizations conditioned on the kriged field (green lines indicate geological and site boundaries)

Once transmissivity realizations similar to Figure 6.4b have been created for each IMOGA solution, predictions can be made for each set of realizations. The cumulative distribution function (CDF) for the prediction can be calculated by Equation 6.16 as:

$$P(t \le t_i) = \sum_{k=1}^{n} P(t \le t_i \mid T_K) P(T_k \mid D, S_k) \quad \forall i \tag{6.16}$$

where $P(t \le t_i \mid T_k)$ is the conditional probability of the prediction t (in the case of the WIPP case study this is the particle travel time to the WIPP site boundary) being less than a certain amount t_i for the large-scale transmissivity structure T_k (this is essentially the CDF of the travel times for the particle for the set of realizations corresponding to T_k), $P(T_k|D, S_k)$ is the conditional probability or likelihood for the large-scale transmissivity structure T_k (this is the obtained from the weighting scheme given by Equations 6.8 or

6.9), and $P(t \leq t_i)$ is the final probability of the prediction t being less than a certain amount t_i.

As a concluding note, it is important to point out a significant distinction between the stochastic realizations generated using the above framework and those used by *DOE/WIPP* [2004]. As pointed out earlier (Section 3.2.2), the *DOE/WIPP* [2004] study created multiple base transmissivity fields incorporating both small-scale heterogeneity *and regional* heterogeneities that represented fractured media in the site. Each base field represented a stochastically generated fracture scenario (see *DOE/WIPP* [2004] for details). Since the IMOGA is essentially a deterministic optimization framework it uses a single *average* base field instead of the multiple base fields used by *DOE/WIPP* [2004].

6.3 Results

The results presented in this section are divided into two parts. The first section (Section 6.3.1) presents interactive calibration results for the WIPP case study. Section 6.3.2 presents the results of the predictive uncertainty analysis for the WIPP case study. As mentioned earlier (Section 6.2.1.2) the primary expert for this study was Dr. D. Walker and the results for Sections 6.3.1 and 6.3.2 correspond to his interactive session. In addition, two experts from the Sandia National Laboratory – Dr. S. McKenna and Dr. R. Beauheim – also evaluated the IMOGA framework and contributed short-duration (due to time constraints) interactive sessions. Preliminary results from these sessions are discussed in Section 6.3.2. These results identify some key aspects of the IMOGA and also provide motivation for future directions of research with the IMOGA.

6.3.1 IMOGA Results for the WIPP site

Both the non-interactive multi-objective genetic algorithm (MOGA) and the IMOGA were run for the WIPP case. The MOGA was run with a large population (population size = 300, number of generations = 300). The IMOGA was seeded with solutions from the MOGA and run with a population size of 20 for 10 generations, with the expert ranking 10 solutions per generation (the rest being ranked by the surrogate machine-learning model discussed in Chapter 5). The entire final generation from the IMOGA was reassessed by the expert, and subjective probabilities were provided for each rank class (i.e. rank 1, rank 2, rank 3, etc solutions were shown to the expert who was then asked to provide a probability ranging from 0 to 1 for the likelihood of each rank of solutions).

Figure 6.5 shows the Pareto front generated for the WIPP case study from the interactive optimization methodology. The crosses indicate the tradeoff curve for the calibration error (H_{err}) and the regularization term (T_{err}) found by the non-interactive MOGA (i.e., with no expert interaction). Figure 6.5 shows there is an obvious tradeoff between model accuracy (as estimated by the calibration error) and model complexity or heterogeneity (as given by the regularization term). The rectangles, diamonds, circles and triangles indicate the solutions found by the interactive multi-objective GA (IMOGA). The triangles correspond to solutions with the worst ranks (4 or more), the open squares correspond to solutions with 'average' ranks (rank 3), and the diamonds and circles correspond to solutions ranked 'good' (rank 2) and 'very good' (rank 1), respectively. Transmissivity fields corresponding to solutions from the interactive Pareto front (one for each rank) are shown in the lower part of Figure 6.5.

As for the Freyberg case study (Section 4.3.2), Figure 6.5 shows that solutions that would be considered 'dominated' or sub-optimal with respect to only the quantitative objectives are part of the interactive Pareto front. These are solutions that would never be found by a purely quantitative approach, which would find solutions only along the MOGA front. Some clear user preferences are also apparent from Figure 6.5. The solutions deemed the worst by the expert correspond to solutions with the lowest calibration errors. These solutions are essentially 'over calibrated,' with heterogeneities that are not found plausible by the expert. Solutions found to be 'average' (rank 3) have the least regularization error (i.e. they are the smoothest fields, fitting the transmissivity data and base field the best) but have higher calibration errors. This indicates that this expert tends to prefer 'simpler' solutions with higher calibration errors over more accurate but implausible solutions with unacceptable spatial distributions of transmissivities. The IMOGA also enables the expert to find solutions that have the best combination of both calibration accuracy and regularization. These solutions, ranked the highest by the expert (ranks 1 and 2), are spread in the middle range of the head error and regularization tradeoff curve. As can be seen from the lower part of Figure 6.5, these solutions are more heterogeneous than rank 3 fields but less heterogeneous than rank 4 and 5 fields. An important criterion used by the expert to rank the solutions was that the "…*T field be contiguous with like regions within domain (i.e., that depositional/erosional features were continuous)"* (quote from expert-feedback correspondence). The rank 4 and 5 solutions have isolated islands of low and high transmissivities while rank 1 and 2 solutions have more connected heterogeneities. As an example of how significant this

effect can be, two solutions (indicated by a dashed circle in Figure 6.5) from the non-interactive Pareto front are compared in Figure 6.6. Both of these solutions are quite similar both in terms of the head error and the regularization measure (the H_{err}/T_{err} is 3.14/2.66 for one and 3.18/2.01 for the other), but one has been given a rank of 1 and the other a rank of 4 by the expert (Figure 6.5). Figure 6.6 shows the transmissivity field corresponding to these solutions. The rank 4 solution has an isolated island of high transmissivities in the southern edge of the site (shown by the dotted circle), while the rank 1 solution has a more continuous transmissivity field.

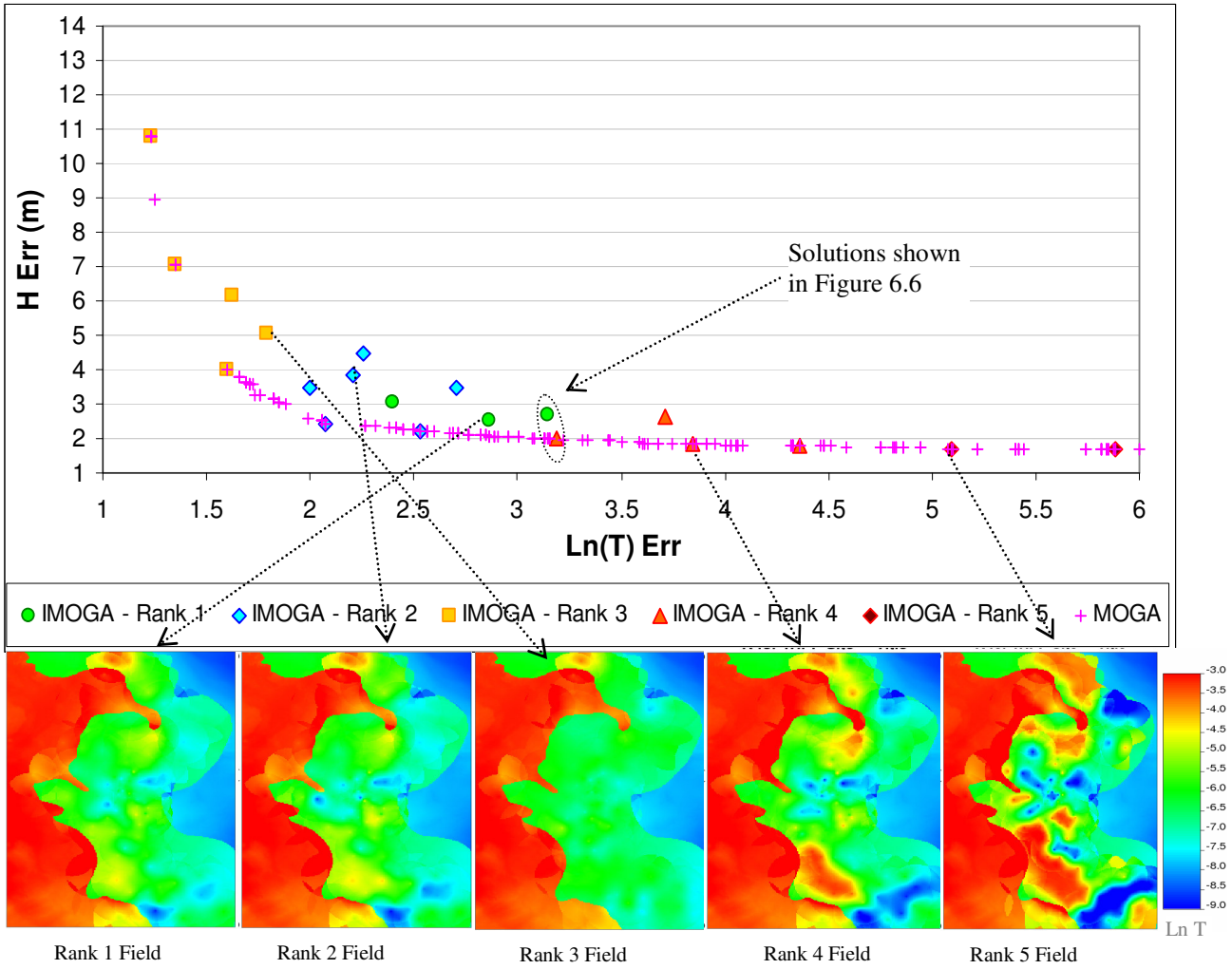


Figure 6.5 IMOGA Pareto Front with the WIPP Site (transmissivity fields from different parts of the Pareto front with different ranks are shown below)

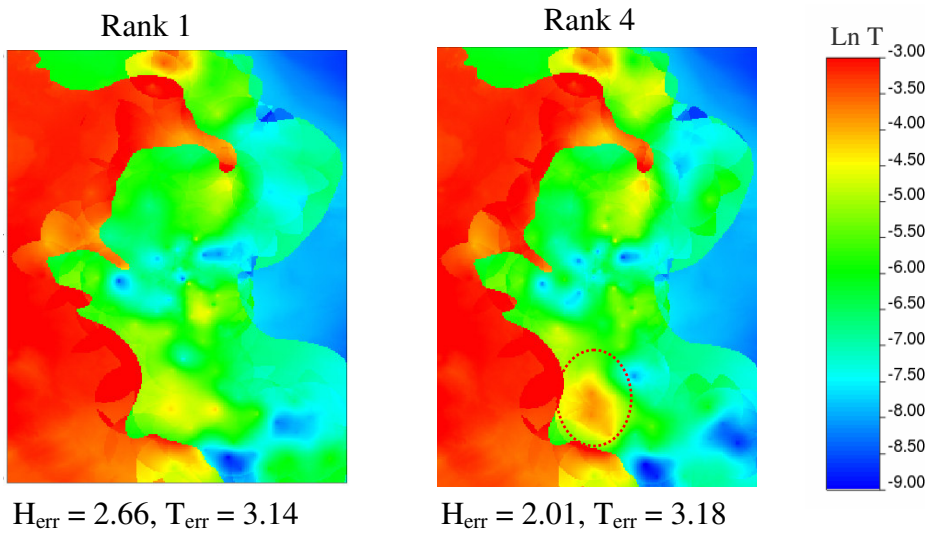


Figure 6.6 Comparison of Solutions from IMOGA Front

6.3.2 Uncertainty Analysis for the WIPP site

Section 6.3.2 discussed the large-scale transmissivity fields identified by the IMOGA. The methodology outlined in Section 6.2.2.2 was then used to generate 50 transmissivity realizations for each IMOGA solution (leading to an ensemble of a total of 1000 transmissivity fields). The head errors were re-calculated for each of these realizations to investigate if the conditioning on the head data is maintained by the realizations. The resulting calibration errors for each set of realizations are shown in Figure 6.7 (note that since the realizations honor the data and pilot point values exactly, the regularization error remains the same for all realizations). Figure 6.7 shows that the small-scale heterogeneity does not disturb the head error much. Overall, the calibrated head response and head response of the realizations differ by less than 10% on average.

Each realization was used with the particle tracking simulation to predict the travel times and travel paths for the conservative particle to the edge of the WIPP boundary. These

travel times are shown in Table 6.1 with and without the effect of the small-scale variability. It can be seen from Table 6.1 that the transmissivity fields without small-scale heterogeneity (as given directly by the IMOGA) lead to much longer travel times (note that shorter travel times are more conservative from a decision-making perspective). The travel times predicted from the fields without small-scale variability are on average 10 times longer (and in some cases more than 30 times longer) than those for the fields with small-scale variability. This shows the importance of including small-scale variability when conducting uncertainty analysis for contaminant travel time predictions.

Figure 6.8 shows the travel times for the realizations of the IMOGA solutions with respect to the two calibration objectives. The dashed yellow line shows the approximate trend of the travel times in terms of the calibration and regularization error. It can be seen from Figure 6.8 the travel times tend to decrease the most close to the 'knee' of the Pareto front, with an increase in the average travel times (as well as the variance in the travel times) towards the extremes of the two objectives. It is interesting to note that neither the most homogenous nor most heterogeneous transmissivity fields have the shortest (most conservative) travel times. It is, in fact, the solutions in the middle-range of heterogeneity that lead to the shortest travel times. Comparing this figure with Figure 6.5 reveals that these are solutions that have been given the highest rank (ranks 1 and 2) by the expert. Recall that the expert tended to prefer heterogeneities that displayed continuity and connectivity in the heterogeneity of the calibrated field. It is these connected transmissive features that lead to the shortest travel times. Figure 6.8 also shows that the largest average and variance in travel times are generally for the most

heterogeneous solutions, which also correspond to the lowest calibration errors and highest regularization error. Comparing this figure with Figure 6.5 reveals that these were also the solutions with the lowest human ranks (ranks 4 and 5) that had isolated islands of large-scale heterogeneity.

The realizations for each IMOGA solution (large-scale transmissivity) can be used to construct a cumulative distribution function (CDF) for the travel time for that particular large-scale structure. Figure 6.9 shows these CDFs for each of the 20 IMOGA solutions. The range within a particular CDF is indicative of the effect small-scale heterogeneity has on the travel times of the particle. Most of the predicted travel times are within 100,000 years, with a few in the 100,000 to 300,000 range.

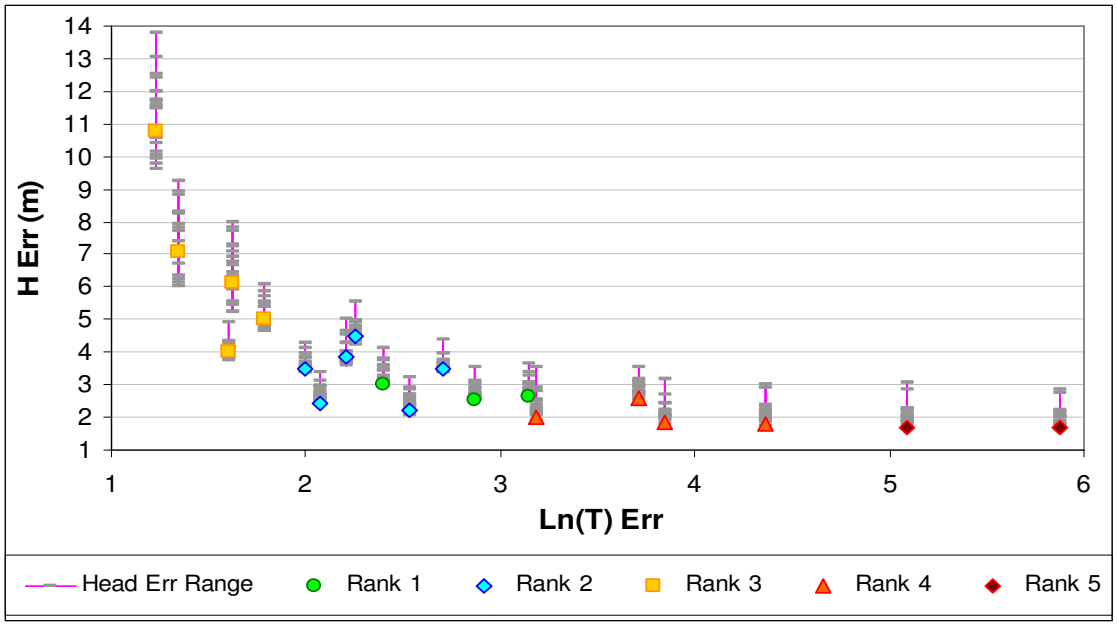


Figure 6.7 Calibration Objectives for Transmissivity Realizations

Table 6.1 Travel Times with and without small-scale variability

Index	Average Travel Time with Small Scale variability - T_s (years)	Travel Time without Small Scale variability - T (years)	Ratio T/T_s
1	60443.04	827748.36	13.69468
2	46175.53	667404.83	14.45365
3	31333.17	522667.62	16.68097
4	30778.46	214926.37	6.983013
5	28118.99	297739.00	10.58854
6	32298.48	117364.31	3.633741
7	50882.01	134420.98	2.641817
8	72247.79	126088.17	1.745218
9	113427.10	139405.57	1.229032
10	193475.10	156521.64	0.809001
11	118431.12	348781.05	2.945012
12	41891.09	1387275.23	33.11624
13	61665.87	254416.67	4.125729
14	76537.88	715369.14	9.346603
15	80479.60	884700.46	10.99285
16	40846.27	364727.70	8.929278
17	34864.08	580513.09	16.65075
18	58301.87	510399.20	8.754423
19	29589.80	328249.70	11.09334
20	35185.53	764766.78	21.73526

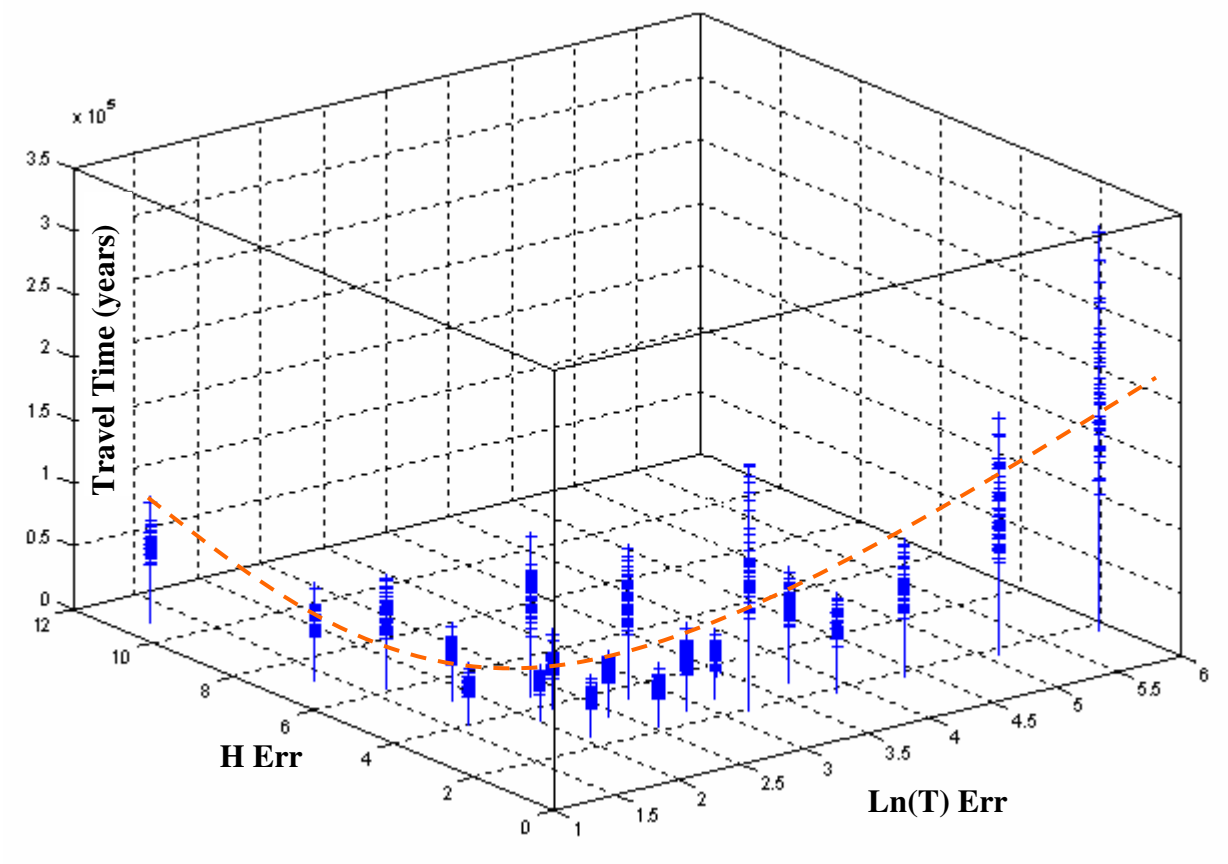


Figure 6.8 Travel Times for the IMOGA solutions

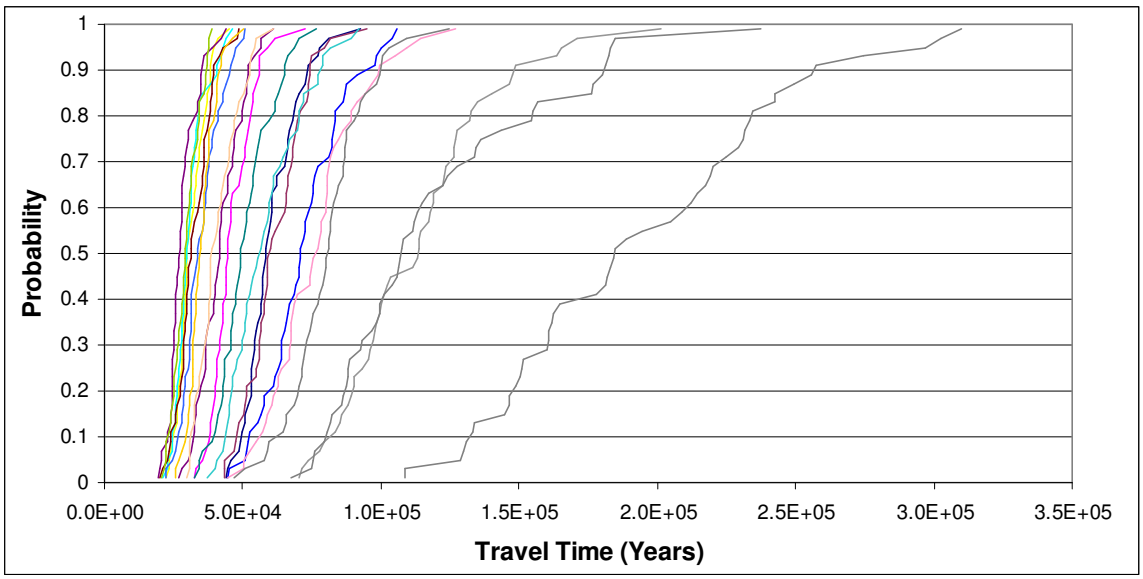


Figure 6.9 CDFs of travel times for different IMOGA solutions (each color correspond to a particular large-scale transmissivity field)

The model averaging scheme presented in Section 6.2.2.1 is used to combine the probability distributions given in Figure 6.9. The 20 solutions found by the IMOGA are given subjective probabilities by the expert. The MLBMA and GLUE weights with and without the subjective probabilities are shown in Table 6.2. Table 6.2 shows that MLBMA gives a weight of almost 1 to just one model and essentially rejects all of the others (with near zero weights). Consequently, there is no effect from including the subjective probabilities between different models, because only one model remains after the MLBMA weighting (the subjective probabilities are in effect multiplied by zero for all but one of the models). The GLUE approach, on the other hand, yields more well-distributed weights and here, in fact, subjective weighting *does* make a difference in the relative weighting.

Table 6.2 Weighting for IMOGA Solutions

Index	$Ln(T)_{err}$	H_{err}	Subj. Probs	MLBMA Weights without Subj. Probs	MLBMA Weights with Subj. Probs	GLUE Weights without Subj. Probs	GLUE Weights with Subj. Probs
1	1.2324	10.7813	0.25	2.23E-162	2.20E-162	0.0167	0.0174
2	1.3456	7.0512	0.25	6.97E-44	6.97E-44	0.0240	0.0342
3	1.6006	3.9956	0.25	1.00E+00	1.00E+00	0.0529	0.0752
4	2.0733	2.4247	0.3	1.06E-33	1.27E-33	0.0856	0.1460
5	2.5335	2.2061	0.3	1.06E-88	1.27E-88	0.0693	0.1181
6	3.1441	2.0029	0.35	1.17E-179	1.60E-179	0.0546	0.1086
7	3.8448	1.8347	0.09	0.00E+00	0.00E+00	0.0435	0.0223
8	4.3641	1.7905	0.09	0.00E+00	0.00E+00	0.0354	0.0181
9	5.0925	1.7043	0.01	0.00E+00	0.00E+00	0.0287	0.0016
10	5.8795	1.6606	0.01	0.00E+00	0.00E+00	0.0227	0.0013
11	3.1855	2.6629	0.09	2.96E-187	1.10E-187	0.0301	0.0154
12	2.8676	2.5500	0.35	2.61E-138	3.70E-138	0.0405	0.0805
13	3.7103	2.6098	0.09	0.00E+00	0.00E+00	0.0231	0.0118
14	2.7042	3.4775	0.3	3.33E-121	4.00E-121	0.0245	0.0417
15	2.2588	4.4939	0.25	1.56E-78	1.56E-78	0.0210	0.0298
16	1.7895	5.0306	0.25	1.86E-35	1.86E-35	0.0267	0.0379
17	2.2090	3.8326	0.3	3.67E-61	4.41E-61	0.0302	0.0515
18	1.6266	6.1389	0.25	3.67E-44	3.67E-44	0.0217	0.0308
19	1.9973	3.4808	0.3	1.91E-32	2.29E-32	0.0448	0.0764
20	2.4007	3.0327	0.35	7.17E-75	1.00E-74	0.0408	0.0812

The weights in Table 6.2 are then used to combine the CDFs shown in Figure 6.9 to get one combined CDF for the predicted travel times (note that since the MLBMA gives all weights to one model, only the GLUE weights actually 'combine' the different CDFs). In addition, the same analysis is conducted for the MOGA solutions with no user interaction. Comparing the CDFs with and without user interaction can show the effect interaction has on the distribution of the travel times for the WIPP site. These CDFs, along with the ensemble weighted average for the solution set with and without interaction, are shown in Figure 6.10. Figure 6.10 also shows the CDF and average for the *DOE/WIPP* [2004] study. Some critical issues are brought forth in Figure 6.10. Comparing the MLBMA CDF (shown by the dashed line) with the GLUE (combined) CDF shows that the latter has a larger range of travel times (as seen by the long tail in the GLUE CDF). This is to be expected since the MLBMA CDF corresponds to only one of the CDFs shown in Figure 6.9, while the GLUE CDF is a combination of the different CDFs and thus the uncertainty across the different large-scale models is reflected in this CDF. This is similar to the work done by *Domingos* [2000] who compared BMA with other model averaging techniques and showed that BMA tends to underestimate the predictive uncertainty. More importantly for this work, since MLBMA chooses a single model, both the interactive and non-interactive CDFs are essentially the same (see discussion in previous paragraph). For this reason, the GLUE CDF is used for further analysis.

Figure 6.10 also shows that the ensemble without interaction predicts a much longer average travel time. The average travel time for the ensemble with user interaction (40,690 years) is approximately 10,000 years less than the average travel time for the ensemble without user interaction (51,538 years). Since shorter travel times are more

conservative from a decision making point of view, interaction leads to a more conservative estimate of this prediction. Another feature that is clear in Figure 6.10 is that both the interactive and non-interactive ensembles have travel times that are, on average, larger than the *DOE/WIPP* [2004] CDF. This is primarily because *DOE/WIPP* [2004] worked with multiple base fields, each with different large-scale heterogeneities to represent the fractures in the Culebra media (see Section 6.2.2.2 for a discussion on the difference between this work and the *DOE/WIPP* [2004] geostatistical framework). Since this work uses a single average base field, these large-scale heterogeneities were averaged out, thus leading to longer travel times. However, Figure 6.11 shows that the predictions from the interactive ensemble are closer to the *DOE/WIPP* [2004] than the non-interactive predictions. The interactive ensemble average represents an improvement of more than 40% over the non-interactive ensemble. In essence, this means that the expert interaction is able to compensate for some of the structural simplifications that are part of the conceptual model. By allowing connected heterogeneity patterns (especially with large transmissivity values) the expert accounted for some of the effects that fractured media would have in this aquifer.

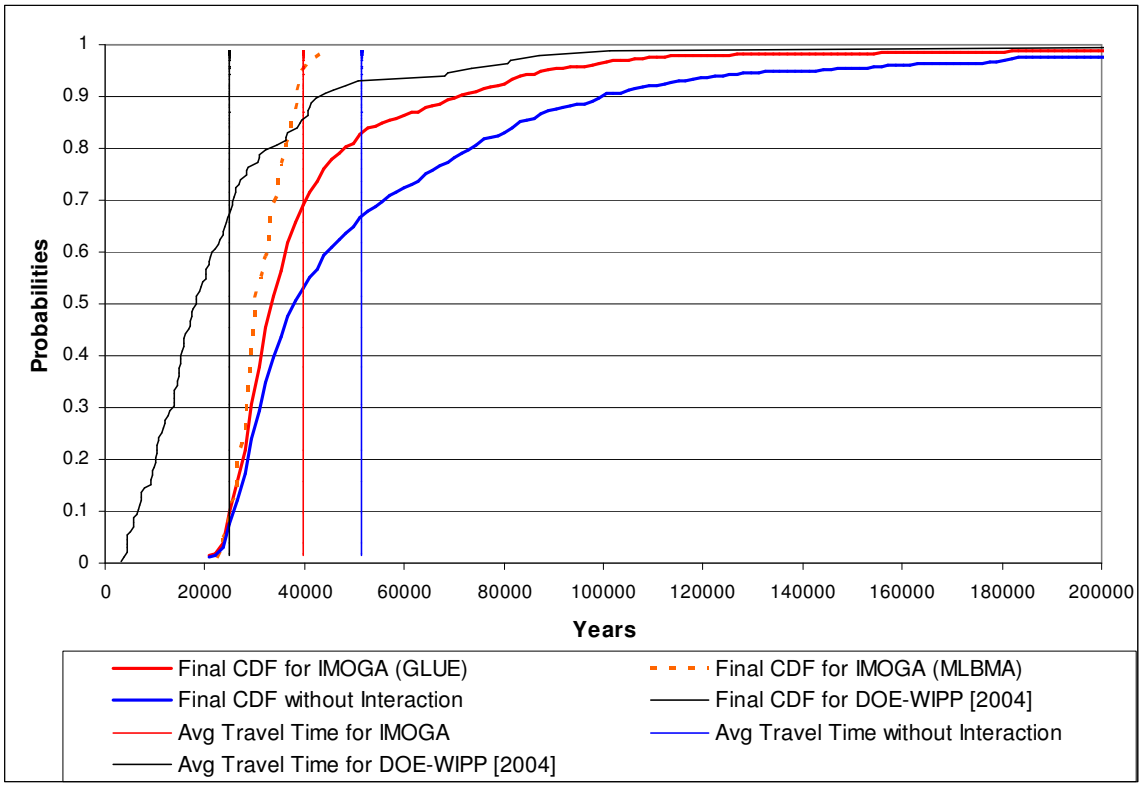


Figure 6.10 CDFs and average travel times for the WIPP site with and without interaction compared to the CDF for *DOE/WIPP* [2004]

Finally, the particle travel paths for representative solutions from the IMOGA front are shown in Figure 6.11. These paths correspond to the same transmissivity fields shown in Figure 6.5 and are shown in order of decreasing calibration error (and increasing regularization error). This figure shows the effect the calibration objectives (both quantitative and qualitative) have on the particle travel paths. It can be seen from Figure 6.11 that as the fields become more heterogeneous and fit the head data better (i.e. the regularization error increases and head error decreases) the overall path of the particle within the WIPP site seems to shift from a northwest-southeast direction to a more curved path towards the southwest side of the WIPP boundary. The longer, curved paths

correspond to the longer travel times that were seen for such solutions in Figure 6.8. While the overall direction of flow for Rank 1, 2, and 3 solutions remain the same, rank 2 and 3 solutions show less variance in the travel paths compared to the rank 4 and 5 solutions. These trends were found to be consistent with the other solutions belonging to each rank. In general, as the transmissivity becomes more heterogeneous, the particles take paths that are more convoluted, responding to local groundwater gradients. Moreover, the small-scale heterogeneity can establish new preferential pathways by connecting (or disconnecting) the large-scale heterogeneities, thus leading to more variance in travel times as seen in Figure 6.8.

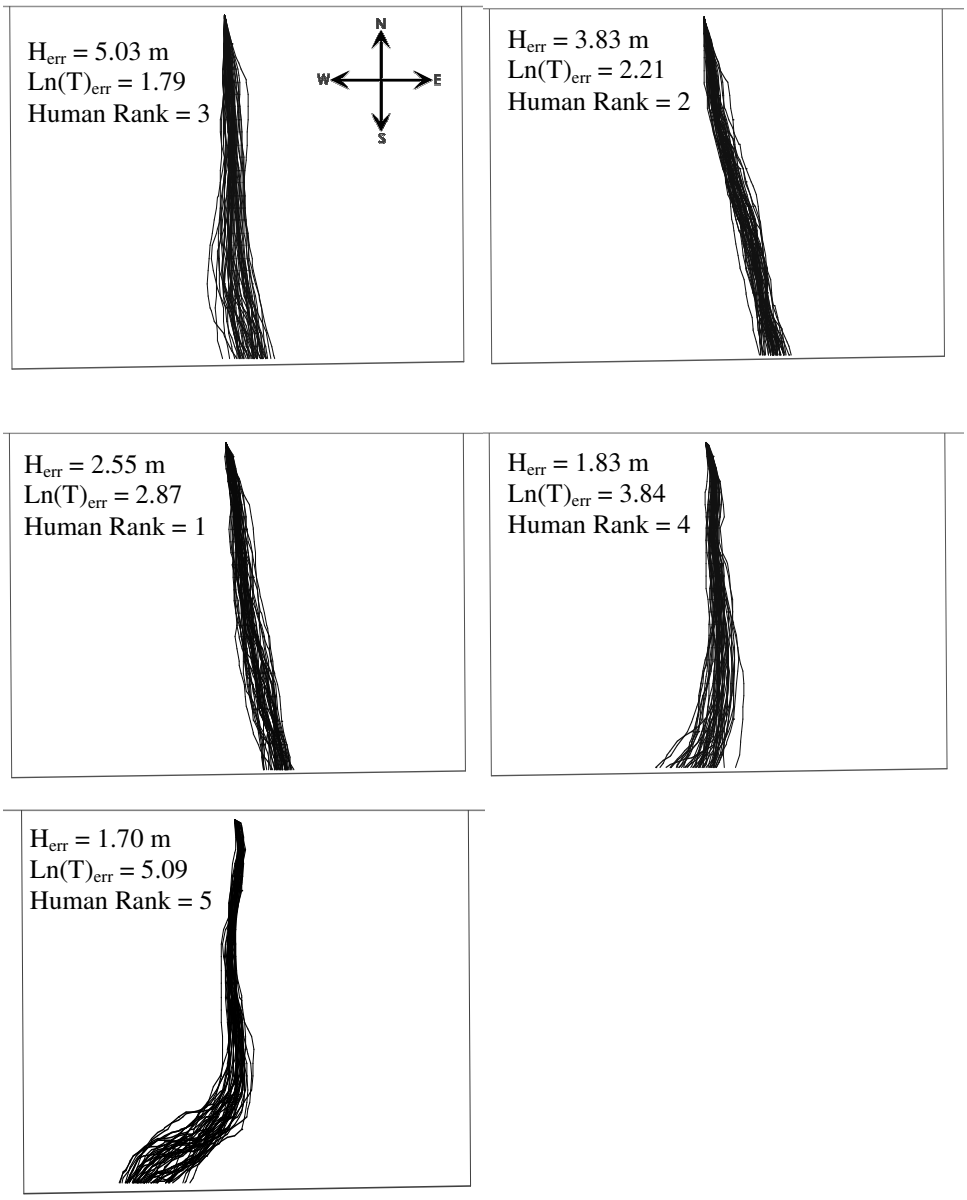


Figure 6.11 Particle travel paths for representative IMOGA solutions (the WIPP site boundary is shown by grey rectangles)

6.3.3 Effect of Different Users on IMOGA

This section presents the IMOGA results for the interactive sessions conducted by the scientists from Sandia National Labs – Dr. S. McKenna and Dr. R. Beauheim. Figures 6.12 and 6.13 show the IMOGA Pareto fronts after two independent, short (5

generations) interactive sessions conducted with each of these experts (both runs were seeded with the same population as that for Dr. Walker). As can be seen from these figures, the IMOGA adapts to different user preferences. In general, the rankings given by Dr. McKenna (Figure 6.12) are similar to that of Dr. Walker. Highly heterogeneous solutions with low calibration errors and highly smooth solutions with high calibration errors are both ranked low (ranks 3 and 4). The solutions with a mid-range of heterogeneity with acceptable calibration errors are ranked high (rank 2). No rank 1 solution was found, either due to the fact that the IMOGA was run for very few generations or because the expert was a conservative ranker. In contrast to Dr. Walker's interactive runs (who showed a preference for smoother solutions over highly heterogeneous ones), this expert seemed to split his votes between the overly smooth and the overly heterogeneous solutions (ranks 3 and 4 exist on both sides of rank 2).

Figure 6.13 shows the IMOGA results for a different type of expert. This expert did not find *any* solution that could be given a rank higher than 3. Feedback from the expert indicated that this expert was looking for a very specific feature in the transmissivity fields - connectivity between the two high transmissivities entrants (the two red 'strips' of large transmissivity that are projecting into the green region from the west in Figure 3.5) at the edges of the Salado dissolution zone - that was not found in any of the solutions. In the absence of such a feature, the expert's preferences were exactly correlated with the calibration errors and thus he ranked solutions with low calibration errors the highest. The effect this behavior has on the IMOGA front is that the final Pareto front is essentially the same as the original non-interactive Pareto front. Thus, in the absence of

informative interaction from the expert, the IMOGA remains converged on the non-interactive front.

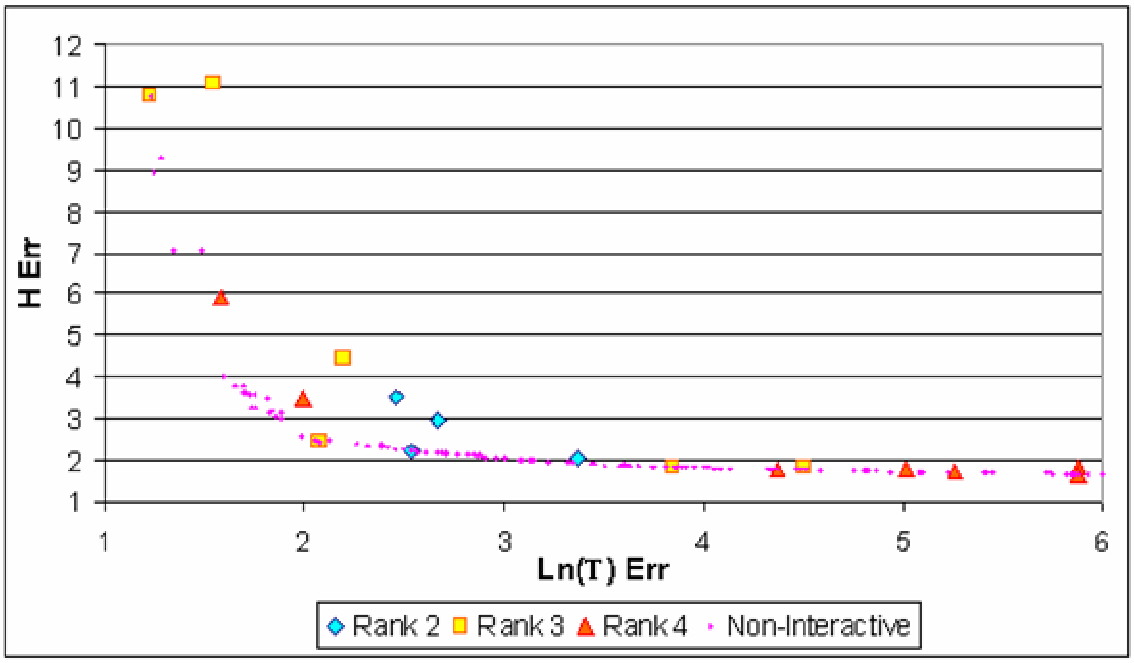


Figure 6.12 IMOGA Pareto front for interactive session with Dr. McKenna

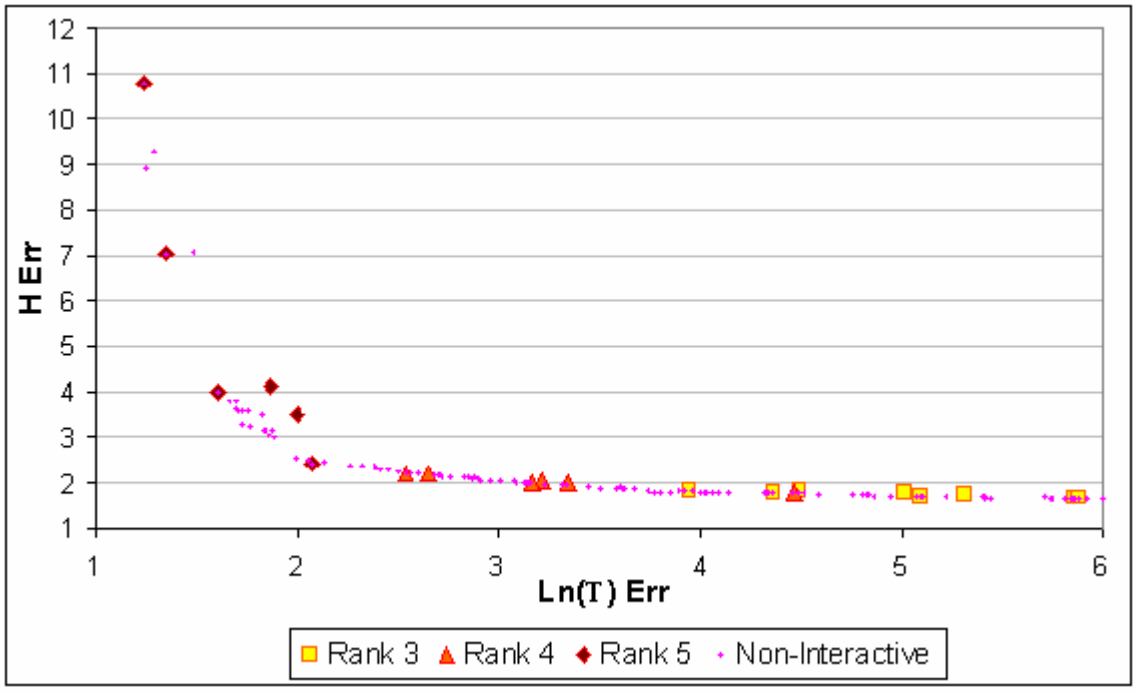


Figure 6.13 IMOGA Pareto front for interactive session with Dr. Beauheim

6.4 Summary and Conclusions

The third phase of this research addressed the issue of conceptual and stochastic uncertainty for a real-world application based on the WIPP site. A predictive uncertainty framework, which could be combined with the efficient interactive optimization framework proposed in chapters 4 and 5, was developed to assess these types of uncertainties and incorporate them into model predictions. The large-scale transmissivity fields identified through the interactive process were considered alternative models for the average transmissivity field for the model. Each transmissivity field was then given a weight that included the effects of calibration errors, conditioning on field data, model complexity (as measured by the regularization term), and the subjective preferences of the expert. Both the MLBMA and GLUE approaches were tested for these weightings. It was seen that GLUE led to more uniformly distributed weights, while MLBMA tended to overweight one solution with the best weighted sum of squared residuals for the calibration and regularization terms.

To develop stochastic simulations for the field-scale application, a geostatistical approach called 'direct sequential simulation with locally varying means' was implemented. In order to maintain the head calibration errors, it was necessary to not only maintain the large-scale average and covariance of the transmissivity field but also constrain the variability in regions with high calibration error sensitivity. For this problem it was seen that the calibration errors were most sensitive to the transmissivities along the model boundaries and thus the uncertainty in these regions was constrained when creating stochastic simulations. In future work, these types of sensitivities could be identified

through a post-optimization sensitivity analysis to identify spatial regions that had the strongest impact on the calibration error. In essence, this would be similar to the adjoint-sensitivity approach that *Rama Rao et al*. [1995] used in their pilot point methodology.

Direct sequential simulation with locally varying means (constrained along the model boundaries) was seen to preserve the optimal large-scale structure and (hence) the optimal calibration error found by the IMOGA. As an aside, this methodology can prove to be particularly useful for many practical geostatistical applications because it is valid without the stringent multi-Gaussian assumption (unlike the sequential Gaussian simulation approach), which is difficult to verify for most real-world applications [*Caers*, 2000].

Results for the WIPP site with the expert (Dr. D. Walker) showed that the expert tended to prefer transmissivity fields that had connected heterogeneity patterns. Overly smooth and highly heterogeneous fields with isolated hot spots were not preferred by this expert. The results showed that though the final IMOGA solution set was close to the non-interactive Pareto front, it contained solutions with preferential spatial characteristics that could not have been found by the non-interactive run. Predictive uncertainty analysis for the WIPP site revealed that expert preferences had a significant impact on the distribution of the model predictions. In this case the prediction consisted of calculating travel times of a conservative solute released from the middle of the WIPP site. Due to the simplifications that were used in this modeling framework the predicted travel times were in general higher than those found by *DOE/WIPP* [2004]. This was due to the fact that

the *DOE–WIPP* [2004] also included an additional source of regional heterogeneity based on stochastic realizations of fracture interconnectivity. While the proposed stochastic simulation approach can include small-scale variability to the IMOGA results without disturbing the calibration errors it is very difficult to include the regional scale of heterogeneity in the post-calibration process, without destroying the calibration errors of the optimal spatial structures found by the IMOGA (since the head errors are sensitive to these large-scale transmissivity trends). To incorporate stochastic simulations at the regional scale within the IMOGA framework future work will investigate extending the IMOGA to consider more than one base field during the optimization process.

Despite these approximations, results showed that the expert's selection of transmissivity fields with connected heterogeneity patterns could reduce these modeling errors. In fact, the IMOGA ensemble of transmissivity fields predicted an average travel time that was over 40% closer to the *DOE/WIPP* [2004] prediction, and more than 10,000 years less (more conservative) than the prediction from the non-interactive run. These results indicate that using expert knowledge about the plausibility of the transmissivity field not only leads to more plausible model calibration but it can also compensate for certain modeling simplifications leading to more reliable and realistic results.

The caveat, of course, is that for this to be true the expert *should* have some knowledge about the site and should be contributing in a constructive fashion during interactive optimization. The issue of reliability of the expert is an important factor in such an approach. Two questions arise at this stage – first, how does one assess the reliability

and/or uncertainty in the expert information provided to the IMOGA; and second, how would the IMOGA respond to multiplicity and conflict in information from different experts. These remain potential future research topics. One way that the reliability of expert information has been addressed is by using 'confidence ratings' provided by the expert him/herself [*Babbar*, 2006]. This allows the expert to express his or her own conception of uncertainty in subjective information. While such an approach is useful in distinguishing between 'experts' and 'novices' (see *Babbar* [2006] for details) it is not useful in assessing the uncertainty or bias in a confident user's preferences. Such uncertainty and bias can only be evaluated 'externally,' for example by allowing multiple experts to interact with the IMOGA. Extending the IMOGA to multiple experts remains a future area for research that is discussed in more detail in Section 7.2. At this stage, some of these issues were addressed by evaluating the effect different experts have on the IMOGA results for the same problem. Preliminary interactive sessions by two scientists from Sandia National Laboratory revealed some key aspects about the IMOGA. One of the Sandia experts (Dr. S. McKenna) had preferences very similar to the primary expert for this study (Dr. D. Walker) and identified transmissivity fields with mid-range transmissivities as the most preferable. The other expert had remarkably different preferences and was ranking transmissivity fields based on the presence (or absence) of a specific geological feature (connectivity of the two large-transmissivity entrants). Further discussions revealed that the basis of the preferences of this expert was new data *that were not included* in this formulation of the WIPP case study. The current data set, in fact, has measurements of low transmissivity in the very regions that this expert expects high transmissivities and thus without including additional data it was impossible for the

IMOGA to find solutions that met the expert's preferences (since all of the transmissivity fields used within the IMOGA honor the field measurements exactly). This finding motivates the need for more 'active' user participation in the calibration process, wherein the expert could be allowed to manually change the transmissivity fields such that features that were not altogether consistent with the field data could also be explored during the interactive process.

However, these findings also reveal a strength of the IMOGA framework. In this case, since none of the IMOGA solutions satisfied the expert, the rankings given by the expert were in essence non-informative and IMOGA was seen to converge to a set of solutions that was Pareto optimal with respect to only the quantitative objectives. In other words, if the expert has additional information that can be used to find new solutions, the IMOGA uses this information to explore the space surrounding the numerically optimal solution space. On the other hand, if there is no additional information provided by the expert, the IMOGA gives the same result as non-interactive optimization. Thus, the IMOGA is seen to display robustness to non-informative interaction by maintaining the quantitatively optimal solution space.

In conclusion, the IMOGA is a demonstrably adaptive framework that allows the expert to incorporate his or her knowledge in the model calibration process, allowing him or her to explore the solution space within reasonable bounds of the numerically optimal solution space. The issue of resolving input from multiple experts is one that remains to be solved. This works lays the foundation for future research in this area.

7 CONCLUDING REMARKS

Theory and calculation are not substitutes for judgment, but are the bases for sounder judgment

~Dr. Ralph B. Peck, Professor Emeritus, UIUC

The last and concluding chapter of this dissertation is divided into two sections. The first section presents the important findings and conclusions for each phase of this research. The second section discusses future areas for extending the research presented in this thesis.

7.1 Summary of Research Findings

This research has developed an integrated framework where multiple sources of qualitative and quantitative information about real-world sites can be used to design and parameterize better hydrologic models. The novel aspect of this methodology, called interactive multi-objective genetic algorithm (IMOGA), is the interactive component in which qualitative expert knowledge and judgment can be expressed within a multi-objective optimization setting. For such a methodology to be applicable in the real world, it is important that the amount of interaction required from the user be minimized. Thus, this work also addressed the issue of user fatigue in such interactive systems, developing a two-step methodology to systematically reduce the amount of user interaction while still maintaining high solution quality and reliability.

The advantage of such an interactive strategy for building better groundwater models was first demonstrated for a hypothetical aquifer in Chapter 4. Results indicated that using

purely quantitative criteria can lead to implausible model parameters that may lead to biased predictions. Using the IMOGA allowed the expert (the author) to include his understanding of the site's spatial characteristics, leading to more plausible and robust parameters. It is noteworthy that the IMOGA finds a gamut of solutions that represent the best tradeoff between quantitative and subjective criteria. Thus solutions that are quantitatively optimal are not over-ridden by the expert's preferences (or biases). On the other hand, if any conflict exists between the expert's knowledge about the site and the data, this is identified in the IMOGA's final Pareto front. The predictive performance of parameter fields found from strictly quantitative optimization was compared with those found from the interactive approach, and it was shown that the latter had better predictive power both for the calibration scenario (with the interactive results approximately 8% better than the non-interactive results) and for validation scenario (with the interactive results approximately 13% better than the non-interactive results) that was not used for calibration. More importantly, the value of such human interaction was shown to increase as 'hard' data available from the field site became sparser (with the interactive results performing up to 30% better for the calibration), indicating that the expert can partially compensate for the lack of field data. This demonstration served as a proof-of-concept to show that the IMOGA *can* incorporate subjective judgment in the search for optimal parameters and that such information can lead to important gains in the model quality for the hypothetical aquifer.

Chapter 5 addressed the challenging problem of user fatigue in such interactive systems. The two-step methodology to combat user fatigue consisted of: (1) clustering similar

parameter fields to optimally select a subset of the solutions to be shown to the expert and (2) training a machine learning model to learn from an archive of previously ranked solutions to provide the ranks of the remaining unevaluated solutions in every generation. These approaches for clustering and machine learning were tested on a rigorous test bed comprising of multiple off-line IMOGA sessions. The approaches included hierarchical and spectral clustering (implemented as the N-cuts algorithm), and decision trees and naïve Bayes learning models. N-cuts clustering was found to outperform ordinary hierarchical clustering, consistently leading to well defined and perceptually similar clusters (N-cuts clustering led to an improvement of as much as 60% in cluster accuracy and 40% in cluster separability). Including information about the spatial characteristics of the parameter fields in the form of nested spatial moments was found to improve results for both unsupervised classification (with up to 30% improvement in clustering accuracy) and supervised clustering (with up to 12% improvement in prediction accuracy). The naïve Bayes learning model, though simple, was found to perform better for small training datasets, consistently giving better results with lower prediction variance. However, the performance of both algorithms was comparable for larger training datasets (which would be the case towards the end of an IMOGA run when many user evaluated solutions are available). Including clustering with machine learning led to the most significant improvements in the machine learning performance because clustering ensured that a wide range of solution types were sampled. Overall, N-cuts clustering in conjunction with naïve Bayes with nested spatial moments as input led to an improvement of more than 20% compared to learning with non-spatial information and randomized training. Moreover, clustering also improved the reliability of the IMOGA by

ensuring that a larger range of potentially good (or bad) solutions were evaluated directly by the expert (overall the variance of the training data set increased by an average of 16% with N-cuts clustering) so that any prediction errors from the learning algorithm would have minimal effect on the IMOGA's solution quality. This framework was tested with 100% and 50% user interaction (again with the author as expert) for both hypothetical and field scale case studies, and it was demonstrated that the solutions from both trials were very similar. This indicated that user fatigue could be reduced by up to 50% without compromising the solution quality of the IMOGA.

Chapter 6 extended the interactive calibration framework by including predictive uncertainty analysis of the results obtained from the IMOGA. As such, the IMOGA proposed in this study is a deterministic system. However, a framework was developed to assess both the conceptual and stochastic uncertainty in the solutions found by the IMOGA for the WIPP site. Key to this development was creating conditional simulations to compensate for the smoothing imposed by the parameterization used to create the parameter field. A simulation algorithm called 'direct sequential simulation' was shown to work the best for the IMOGA, preserving both the covariance and the local mean structure of the IMOGA solutions, ensuring that the optimal calibration errors are not overly disturbed. To include the conceptual uncertainty in the large-scale solutions found by the IMOGA, a statistical framework was developed to incorporate all of the calibration criteria (both quantitative and qualitative) within a sampling scheme.

Chapter 6 also applied the IMOGA framework (with some modifications in the calibration formulation) to a field scale application (based on the WIPP site) employing an actual site expert. Results for the WIPP site with a real site expert identified some important aspects about the expert's judgment. Over-calibrated solutions with significant (implausible) heterogeneity were deemed the worst by the expert. Smooth solutions with continuous spatial features were found to be 'acceptable,' indicating clearly that this expert tended to prefer simpler solutions (with higher calibration errors) over more complex and implausible solutions with unacceptable spatial distributions of transmissivities. The expert also identified above average and good solutions that had the best combination of both calibration accuracy and fit with transmissivity data.

Given the set of alternative transmissivity fields, predictive uncertainty analysis was undertaken for the WIPP site. Ensemble averaging was performed to identify the probability distribution of the prediction of interest (particle travel times) and it was seen that including interaction again led to significant gains (the average predicted travel time was up to 40% closer to the more sophisticated *DOE/WIPP* [2004] model, and up to 10,000 years more conservative compared to the non-interactive results). This indicated that the IMOGA allowed the expert to compensate for some of the conceptual model's approximations and deficiencies.

7.2 Future Research

The research presented in this dissertation can provide a starting point for future research in many different areas. This section discusses some possible ramifications and enhancements for this interactive multi-objective approach:

- The interactive system can be enhanced by allowing the expert more active participation. The results with different experts for the WIPP case indicated that at least one expert felt the need to explore transmissivity fields that were different from those conditioned on the transmissivity data. In the current IMOGA framework, the modeler interacts 'passively' with the IMOGA, only providing subjective fitness for potential solutions identified by the algorithm. In a more participatory environment, he or she could manually change certain features in the IMOGA solutions to better represent their knowledge of the site. The expert could also be allowed to adjust the crossover and mutation operations of the GA to better search the solution space in areas that he or she finds promising. In essence, these changes would represent a shift from an 'interactive' approach to a more 'human-based' approach that has been proposed by *Kosorukoff* [2001].
- The interactive system needs to be extended to consider uncertainty during the optimization process. This research proposed a methodology to include stochastic uncertainty during post-calibration uncertainty analysis. However, the scale of heterogeneity that can be included in this framework is restricted by the sensitivity of the calibration objective. Heterogeneities at larger scales would need to be included as multiple stochastic prior fields that the IMOGA would then calibrate within a probabilistic optimization framework. In effect, this would require a *probabilistic multi-objective interactive* optimization framework (work on probabilistic multi-objective optimization such as *Singh et al* [2003] could prove useful in this respect).

- The current research was applied to a steady-state groundwater problem for estimating two-dimensional hydraulic conductivity fields. However, many groundwater models are three dimensional and may have other uncertain parameters that also need to be estimated (e.g., recharge, layer thickness, and boundary conditions). While the overall framework for the IMOGA for such multi-dimensional multi-parameter problems would remain the same, the fundamental challenge remains in visualizing such high-dimensional data for the expert. An important factor that should always be considered when building such interactive systems is to only show the expert information about which he or she has some knowledge (or is interested in). Thus, for a multi-layer groundwater model, if the expert only has knowledge about the first few aquifer layers, only these layers need be shown during the interactive phase. If the entire three-dimensional field *is* visualized for the expert, then 'interactive cuboids' that can be manipulated by the user (allowing him or her to change perspective, viewing angle, and zoom level) could be a useful visualization tool. GIS overlays can also be used to show the expert information about multiple spatial characteristics (such as recharge and aquifer thickness) simultaneously. Overall, extending the IMOGA for multi-dimensional and multi-parameter problems is intrinsically linked with research on developing efficient tools for visualization and data exploration.
- The IMOGA also provides a promising calibration approach for watershed models where there are numerous (correlated) inputs and parameters (such as evapo-transpiration, infiltration, soil texture, soil curve numbers, land use, and water demands), many of which are purely parametric and difficult to measure

directly from the modeled system. The interactive approach can prove effective in the better estimation of such parameters by allowing the expert to include his or her judgment for preferred values of such parameters based on his or her modeling expertise. Finally, the interactive approach can be particularly useful for integrating environmental, social, and political factors (such as sustainability, equity, water rights, social ethics/norms, and regulatory policies) in water resources applications such as water supply or watershed planning and management.

- The research presented in this dissertation addressed the problem of calibrating a single model. Often, especially in hydrological or meteorological applications, multiple conceptual models need to be assessed and calibrated. The results from these multiple models are then combined using model-averaging paradigms similar to those discussed in Section 6.2.2. The non-uniqueness and equifinality inherent in the groundwater calibration problem also exists (and is in fact more pronounced) when considering multiple models. Thus, an important future direction for research on the IMOGA would be to use it to identify multiple alternative conceptual models, each calibrated with respect to multiple calibration criteria. Such an interactive framework would not only allow the expert to assess different calibrations for a given model but also allow him or her to compare the performance and plausibility of alternative models for the environmental system being modeled. Some work on multiple model selection and combination has been undertaken by *Neuman* [2003] and *Beven and Freer* [2001], however a comprehensive (interactive) optimization framework such as the IMOGA has not

been applied to this problem. One of the challenges in simultaneously assessing and calibrating multiple conceptual models using the IMOGA would be the varying number of parameters (or decision variables) that need to be calibrated for each model. To adapt the IMOGA to this domain, research on variable length genetic algorithms such as the 'messy GA' (*Goldberg et al.*, 1991) and the 'variable GA' (*Bandyopadhyay et al.*, 2001) can prove to be particularly useful.

- Finally, the proposed framework needs to be extended to consider multiple users and stakeholders. In practice, many environmental and water resources problems are beset with conflicting requirements expressed by different parties. The interactive system could, in such cases, be used as a conflict resolution framework by allowing all users to assess the tradeoffs and consequences of their individual and collective preferences. Initial experiments with multiple users have indicated that there can be both shared and conflicting opinions between different experts. Such multi-user sessions can in fact be used to address the issue of assessing the reliability of the human expert - a possible strategy could be to have multiple users vote on different solutions and use majority preferences (weighted by some metric of user 'expertise') to drive the IMOGA. Such multi-user sessions can also provide a powerful discovery and innovation tool by revealing predominant conflicts (and agreements) within a group of stakeholders as well as identifying promising solutions that may be more acceptable to the decision-making group. In conclusion, the issue of extending the IMOGA for multiple users remains one of the most exciting (and challenging) future directions for this research.

REFERENCES

Adams, E.E., and Gelhar, L.W. (1992), Field study of dispersion in a heterogeneous aquifer, 2 spatial moment analysis, *Water Resources Research*, Vol. 28, No.12, pp.3293-3307.

Aksoy, A. and Teresa B. (2000), Culver, Effects of sorption assumptions on the aquifer remediation designs, Ground Water 38(2): 200-208.

Anderson, M. and Woessner, W.W. (1992), *Applied Groundwater Modeling*, Academic Press, San Diego.

Alcolea, A., Carrera, J., Medina, A. (2006), Pilot points method incorporating prior information for solving the groundwater flow inverse problem, *Advances in Water Resources* (29), 1678-1689.

Aly, A.H., and R.C. Peralta (1999), Optimal Design of Aquifer Cleanup Systems under Uncertainty Using A Neural Network and A Genetic Algorithm, *Water Resources Research*, 35(8), pp. 2523-2532

Babbar, M., Minsker, B. S., and Takagi, H. (2004), Interactive Genetic Algorithm Framework for Long Term Groundwater Monitoring Design, American Society of Civil Engineers (ASCE) Environmental & Water Resources Institute (EWRI) World Water & Environmental Resources Congress 2004 & Related Symposia, Salt Lake City, UT.

Babbar, M., Minsker, B. S. and Takagi, H. (2005) Expert Knowledge in Long-Term Groundwater Monitoring Optimization Process: The Interactive Genetic Algorithm Perspective, American Society of Civil Engineers (ASCE) Environmental & Water Resources Institute (EWRI) World Water & Environmental Resources Congress 2005 & Related Symposia, Anchorage, AK.

Babbar, M. (2006), Interactive Genetic Algorithm Framework for Long Term Groundwater Monitoring Design, PhD thesis, University of Illinois.

Bailey, K. R. and Fitzpatrick, B. G. (1996), Estimation of groundwater flow parameters using least squares, *Tech. Rep. CRSC-TR96-13*, North Carolina State University, Center for Research in Scientific Computation.

Ball, G.H., and Hall, D. J. (1965), ISODATA, a novel method of data analysis and classification, *Tech. Rep.*, Stanford University, Stanford, CA.

Banzhaf, W. (1997), *Interactive Evolution*, IOP Publishing Ltd and Oxford University Press.

Barry, D. A., and Sposito, G. (1990), Three-dimensional statistical moment analysis of the Stanford/Waterloo Borden tracer Data, Water Resources Research, 26:1735.

Bengio, Y., Vincent, P., and Paiement, J.-F. (2003), Learning Eigenfunctions of Similarity: Linking Spectral Clustering and Kernel PCA, Technical Report 1232, Departement d'Informatique et Recherche Operationnelle, February 28th, 2003.

Beven, K.J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour., 16, 41-51.

Beven, K.J. (2000), Uniqueness of place and the presentation of hydrological processes. Hydrol. Earth Syst. Sci., 4, 203-213.

Beven, K.J. (2006), On undermining the science?, *Hydrol. Processes* 20, 3141–3146.

Beven, K.J. and Binley, A. (1992), The Future of Distributed Models: Model Calibration and Uncertainty Prediction, *Hydrological Processes*, 6: 279-298.

Beven, K.J and Freer, J (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrology*, 249, 11-29.

Biles, J.A. (1994), GenJam: A Genetic Algorithm for Generating Jazz Solos. In Proceedings of the 1994 International Computer Music Conference, ICMA, San Francisco.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK.

Boggs, J.M., Young, S.C, Beard, L.M., Gelhar, L.W. Rehfeldt, K.R. and Adams, E.E. (1992), Field study of dispersion in a heterogeneous aquifer, 1 Overview and site description, *Water Resources Research*, Vol. 28, No. 12 pp. 3281-3291.

Brezillon, P., and Pomerol, J.-Ch. (1997), Joint cognitive systems, cooperative systems and decision support systems: A cooperation in context, *Proc. of the European Conference on Cognitive Science*, Manchester, UK, April, pp. 129-139

Brill, E. D., Jr. (1979), "The use of optimization models in public sector planning," *Management Sci.,* vol. 25, no. 5, pp. 423-432

Brill, E. D., Jr., Chaw, S. and Hopkins, L. D. (1982), "Modeling to generate alternatives: The HSJ approach and an illustration using a problem in land use planning, *Management Sci.*, vol. 28, no. 3, pp. 221-225

Brill, E.D., Jr., Flach, J.M., Hopkins, L.D., and Ranjithan, S. (1990), "MGA: A Decision Support System for Complex, Incompletely Defined Problems," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 20, no. 4, pp. 745-757

Caers, J. (2000), Adding Local Accuracy to Direct Sequential Simulation, *Mathematical Geology*, Vol. 32, No. 7, 2000

Caers, J. (2000), Direct sequential indicator simulation, In: Proceedings of the 6th International Geostatistics Congress, Kleingeld, W. and Krige, D (eds.), Cape Town, South Africa, April 10-14, 2000. 12p.

Capilla, J.E., Gomez-Hernandez, J.J., and Sahuquillo, A. (1998), Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric head data--3. Application to the Culebra Formation at the Waste Isolation Pilot Plan (WIPP), New Mexico, USA, *Journal of Hydrology*, Volume 207, Number 3, pp. 254-269(16).

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L. J. (2005), Inverse problem in hydrogeology, *Hydrogeology Journal*, 13(1): 206-222. ISSN: 1431-2174

Carrera, J., and Neuman, S.P. (1986), Estimation of aquifer parameters under transient and steady state conditions, 1, Maximum likelihood method incorporating prior information, *Water Resources Research*, 22(2), 199–210.

Chiles, J. P. and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*. John Wiley, New York.

Cho, S.-B., and Lee, J.-Y. (2002), A human-oriented image retrieval system using interactive genetic algorithm, *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 32, no. 3, pp. 452-458.

Christensen, S, and Cooley, R.L. (1999), Evaluation of confidence intervals for a steady state leaky aquifer model, *Adv. in Water Resour.*, 22(8):807 –817.

Cieniawski, S. E. (1993). An Investigation of the Ability of Genetic Algorithms to Generate the Tradeoff Curve of a Multi-Objective Groundwater Monitoring Problem, Masters Thesis University of Illinois: Urbana, IL.

Clifton, P.M., Neuman, S.P. (1982), Effects of kriging and inverse modeling on conditional simulation of the Avra valley aquifer in southern Arizona, *Water Resources Research*, 18(4):1215 –1234.

Coello, C.A.C. (1999), An Empirical Study of Evolutionary Techniques for Multiobjective Optimization in Engineering Design, Doctoral dissertation, Tulane University: New Orleans, L.A.

Cole, C.R., Bergeron, MP, Wurstner, S.K., Thorne, P.D. Orr, S. and McKinley, M.I. (2001), *Transient Inverse Calibration of Hanford Site-Wide Groundwater Model to Hanford Operational Impacts—1943 to 1996*, PNNL-13447, Pacific Northwest National Laboratory, Richland, WA.

Cooley, R. L. (1977), A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 1, Theory and numerical properties, *Water Resources Research*, 13(2), 318–324.

Cressie, N.A. (1993), *Statistics for Spatial Data*, Probability and Mathematical Statistics. John Wiley & Sons, New York.

Corney, D.P.A. (2002), Intelligent analysis of small data sets for food design, PhD thesis, University College, London.

Cristianini, N. and Shawe-Taylor, J. and Kandola, J. (2002), Spectral kernel methods for clustering, In *NIPS*, 14.

Dagan, G. (1985), Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resources Research*, *2*1(1), 65–72.

Deb, K., Agrawal, S., Pratap, A. and Meyarivan, T. (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.

Deb, K., and Chaudhuri, S. (2005), I-EMO: An Interactive Evolutionary Multi-objective Optimization Tool, KanGAL Report Number 2005003, PReMI 2005, 690-695

Deutsch, C.V., and Journel, A.G. (1998), *GSLIB: Geostatistical Software Library*, Oxford University Press, New York.

Delhomme, J.P. (1979), Spatial variability and uncertainty in groundwater flow parameters : a geostatistical approach, *Water Resources Research*, v. 15, 2, p. 269-280.

Deutsch, C.V. and Journel, A.G. (1992), *GSLIB Geostatistical Software Library and User's Guide*, Oxford University Press, New York.

Diday, E. and Simon, J. C. (1976), Clustering analysis, In *Digital Pattern Recognition*, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, 47–94.

Dhillon, I. S., Guan, Y., and Kulis B. (2004), A Unified View of Kernel k-means, Spectral Clustering and Graph Partitioning, UTCS Technical Report #TR-04-25, June 30. http://www.cs.utexas.edu/users/UTCS/techreports/index/html/Abstracts.2004.html#TR-04-25

Diday, E. and Simon, J.C. (1976), *Cluster analysis - Digital Pattern Recognition*, Springer-Verlag.

DOE/WIPP (2004), WIPP Compliance Recertification Application, DRAFT-3231, Available at http://www.wipp.energy.gov/library/CRA/CRA_Index.htm

Doherty, J. (2003), Groundwater model calibration using pilot points and regularization, *Ground Water*; 41(2): 170-177.

Doherty, J. (2004), PEST: Model-independent parameter estimation, user manual, version 5, Watermark Numer. Comput., Brisbane, Australia.

Domingos, P. (1997). Why does bagging work? A Bayesian account and its implications. KDD'97. http://www.cs.washington.edu/homese/pedrod/kdd97.ps.gz

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. ICML'00. http://www.cs.washington.edu/homese/pedrod/mlc00b.ps.gz

Domingos, P. and Pazzani, M. (1997), On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, 29:103-130.

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. Roy. Statist. Soc. Ser.*, B 57: 45–97

Duda, R. O., Hart, P.E., and Stork, D.G. (2001), *Pattern Classification*, (2nd Edition), Wiley-Interscience

Eggleston, J., and Rojstaczer, S. (1998), Identification of Large-Scale Hydraulic Conductivity Trends and the influence of trends on Contaminant Transport, *Water Resources Research*, vol. 34, no. 9, pp. 2155-2168.

Elfeki, A.M.M., and Rajabiani, H.R. (2002), Simulation of plume behavior at the Macrodispersion Experiment (MADE1) site by applying the coupled Markov chain model for site characterization. In: Computational Methods in Water Resources (CMWRXIV) (S.M. Hassanizadeh, R.J. Schotting, W.G. Gray, & G.F. Pinder (Eds.)), p. 55-662, Amsterdam: Elsevier Science B.V.

Elkan, C. (1997), *Boosting and Naive Bayesian learning*, Technical report, Department of Computer Science and Engineering, University of California

Fonseca, C. M. and Fleming, P. J. (1993), Genetic Algorithms for Multi-objective Optimization: Formulation, Discussion and Generalization, In Stephanie Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, 416-423.

Freyberg, D. L. (1988), An exercise in Ground-Water Model Calibration and Prediction, *Ground Water*, V. 26, No. 3, May-June.

Gelhar, L. W. (1993), *Stochastic Subsurface Hydrology*, Prentice-Hall, Engle-wood Cliffs, N. J.

Ghahramani, Z. (2004) Unsupervised Learning, In *Advanced Lectures on Machine Learning,* Bousquet, O., Raetsch, G. and von Luxburg, U. (eds), LNAI 3176. Springer-Verlag.

Giacobbo, F., Marseguerra, M., Zio, E. (2002), Solving the inverse problem of parameter estimation bygenetic algorithms: the case of a groundwater contaminant transport model, *Annals of Nuclear Energy*, 29 (8), 967-981.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, Reading, MA

Goldberg, D. E. (2002), *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*, Boston, MA: Kluwer Academic Publishers.

Gómez-Hernández, J.J., Sahuquillo, A., Capilla, J.E. (1997), Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data: 1-Theory., *Journal of Hydrology*, 204 (1-4): 162-174.

Gómez-Hernández JJ, Hendricks Fransen HJ, and Sahuquillo A. (2003), Stochastic conditional inverse modeling of subsurface mass transport: A brief review of the self-calibrating method. *Stoch Env Res Risk*, A 2003;17:319-328.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford Univ. Press, New-York, 483.

Gupta, H.V., L. Bastidas, S. Sorooshian, W.J. Shuttleworth, and Z.L. Yang (1999), Parameter Estimation of a Land Surface Scheme Using Multi-Criteria Methods, GCIP II Special Issue of the Journal of Geophysical Research Atmospheres, Vol. 104, No. D16, p. 19491-19503.

Harsanyi, J. C. and Chang, C.-I. (1994), Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach, *IEEE Trans. Geosci. Remote Sensing*, vol. 32, no. 4, pp. 779—785.

Hansen, T. M., and Mosegaard, K. (2007), VISIM : Sequential simulation for linear inverse problems. *Computers and Geosciences*, (to be published 2007). Available at http://imgp.gfy.ku.dk/visim/

Helton, J.C., Anderson, D.R., Baker, B.L., Bean, J.E., Berglund, J.W., Beyeler, W., Garner, J.W., Iuzzolino, H.J., Marietta, M.G., Rechard, R.P., Roache, P.J., Rudeen, D.K., Schreiber, J.D., Swift, P.N., Tierney, M.S., and Vaughn P. (1995), Effect of Alternative Conceptual Models in a Preliminary Performance Assessment for the Waste Isolation Pilot Plant, *Nuclear Engineering and Design*, Vol. 154, 1995, pp. 251-344.

Hendry, C. (1996), Understanding and Creating Whole Organizational Change through Learning Theory, *Human Relations*, 49(5), 621-641. Available: OVID: Accession Number: 01221235. Kerby, Joe Kent (1975), Consumer Behavior: Conceptual Foundations, New York: Dun-Donnelley Publishing Corp.

Hernandez, A.F., Neuman, S.P., Guadagnini, A., Carrera, J. (2003), Conditioning mean steady state flow on hydraulic head and conductivity through geostatistical inversion, Stochas. Env. Res. Risk Assess., 17(5):329 –338.

Hill, M.C. (1992), A computer program (MODFLOWP) for estimating parameters of a transient, three-dimensional, ground-water flow model using nonlinear regression, U.S. Geological Survey.

Hill, M. C. (1998), Methods and Guidelines for Effective Model Calibration, *U.S. Geological Survey Water-Resources Investigations Report 98-4005*, Denver, Colorado.

Hill, M.C., Cooley, R.L., Pollock, D.W. (1998), A controlled experiment in ground water flow model calibration, *Ground Water*, 36(3):520 –535.

Chan Hilton, A.B., and Culver, T.B. (2000), Optimizing groundwater remediation design under uncertainty, *Proceedings of the Joint Conference on Water Resource Engineering and Water Resources Planning and Management*, Minneapolis, MN.

Hoeksema, R.J., and Kitanidis, P.K. (1984), An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, *Water Resources Research*, v. 20, 7, p. 1003-1020.

Hoeksema, R.J., Kitanidis, P.K. (1989), Predictions of transmissivities, heads, and seepage velocities using mathematical models and geostatistics, *Advances in Water Resources*, v. 12, 2, p. 90-102.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999), Bayesian model averaging: A tutorial, *Statist. Sci.*, 14(4): 382–417.

Holland, J. H., (1975), *Adaptation in natural and artificial systems*, Ann Arbor: The University of Michigan Press.

Holt, R.M., Beauheim, R.L., and Powers D.W. (2005), Predicting Fractured Zones in the Culebra Dolomite, Dynamics of Fluids and Transport in Fractured Rock, Geophysical Monograph Series 162, AGU, 10.1029/162GM11.

Iqbal, Q., and Aggarwal, J. K. (2002), CIRES: A system for content-based retrieval in digital image libraries, in Invited Session on Content-based Image Retrieval: Techniques and Applications, in *Proc. 7th International Conference on Control Automation, Robotics and Vision (ICARCV)*, pp. 205--210, December 2002

Jain, A., Murty, M. and Flynn, P. (1999), Data Clustering: A Review, *ACM Computing Surveys*, 31(3), September.

Jain, A.K., and Dubes, R. C. (1988) *Algorithms for Clustering Data*, Prentice-Hall advanced reference series, Prentice-Hall, Inc., Upper Saddle River, NJ.

Johnson, S.C. (1967), Hierarchical Clustering Schemes, *Psychometrika*, 2:241-254.

Jörgensen, C. (2003), *Image Retrieval: Theory and Practice*, Scarecrow Press (Rowman Littlefield).

Journel A.G. (1994), Modeling Uncertainty: some conceptual thoughts, *Geostatistics for the next century*, R. Dimitrakopoulos (ed), Kluwer, Dordrecht, Holland, 30-43.

Kamalian, R., Takagi, H., and Agogino, A.M. (2004), Optimized Design of MEMS by Evolutionary Multi-objective Optimization with Interactive Evolutionary Computation, *Proc. of Genetic and Evolutionary Computation Conf.* (GECCO-2004) pp. 1030-1041.

Kao, J-J, and Liebman, N. (1991), Computer-Aided System for Ground-Water Resources Management, Journal of Computing in Civil Engineering, Vol. 5, No. 3, pp. 251-266.

Karpouzos, D. K., Delay, F., Katsifarakis, K. L., de Marsily, G. (2001) A multipopulation genetic algorithm to solve the inverse problem in hydrogeology *Water Resources Research*, Vol. 37 , No. 9 , p. 2291.

Kashyap, R.L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Anal. Mach. Intel.*, PAMI 4(2): 99–104.

Kass, R.E. and Raftery, A.E. (1995), Bayes factors, *Journal of the American Statistical Association*, 90, 773-795.

Kaufman, L. and Rousseeuw, P. J. (1990), *Finding groups in data*. Wiley, New York.

Keen, P. G. W., and Scott Morton, M. S., (1978) *Decision Support Systems: An Organizational Perspective*. Reading, MA: Addison-Wesley, Inc.

Kirby, M. Sirovich, L. (1990), Application of the Karhunen-Loeve procedure for the characterization of human faces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 12, Issue 1, 103-108.

Kitanidis, P.K. (1997), The minimum-structure solution to the inverse problem, *Water Resources Research*, v. 33, 10, p. 2263-2272.

Kitanidis, P.K., and Vomvoris, E.G. (1983), A geostatistical approach to the inverse problem in groundwater modeling, steady state, and one-dimensional simulations. *Water Resources Research*, v. 19, 13, p. 677-690.

Knopman, D.S., and Voss, C.I. (1989), Multiobjective sampling design for parameter estimation and model discrimination in groundwater solute transport, *Water Resources Research*, v. 25, no. 10, pp 2245-2258.

Kosorukoff, A. (2001), Human-based Genetic Algorithm, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2001, 3464-3469.

Kowalsky, M.B., Finsterle, S., Rubin, Y. (2004), Estimating flow parameter distributions using ground-penetrating radar and hydrological measurements during transient flow in the vadose zone, *Adv. Water Resour.*, 27:583 –599.

Lau, A. and Leong, T. (1999), PROBES: A framework for probability elicitation from experts. Available at http://citeseer.ist.psu.edu/462815.html

LaVenue, A. M., and Pickens, J. F. (1992), Application of a coupled adjoint sensitivity and kriging approach to calibrate a groundwater flow model, *Water Resources Research*, *2*8(6), 1543–1570.

LaVenue, A.M., RamaRao, B.S., de Marsily, G., Marietta, M.G. (1995), Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: Part 2-Application, *Water Resources Research*, v. 31, 3, p. 495-516.

Lee, J.-Y. and Cho, S.-B. (1999), Sparse fitness evaluation for reducing user burden in interactive genetic algorithm, *Proc. Of FUZZ-IEEE'99*, pp. II998-II1003, Aug. 1999.

Llora, X., Sastry, K., Goldberg, D.E., Gupta, A., and Lakshmi, L. (2005), Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness, IlliGAL Report No. 2005009.

Louis, S.J., and Tang R. (1999), Interactive Genetic Algorithms for the Traveling Salesman Problem, *Proceedings of the Genetic and Evolutionary Computation Conference*, Orlando, Florida, vol. 1, pp. 1043-048.

Luxburg, U.V. (2006), A Tutorial on Spectral Clustering, to appear in Statistics and Computing, see also Technical Report 149, Max Planck Institute for Biological Cybernetics. Available at http://www.kyb.tuebingen.mpg.de/bs/people/ule/publications/publication_downloads/luxburg06_TR_v2.pdf

MacQueen, J. B. (1967), Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297

Mandal, M.K., Aboulnasr, T. and Panchanathan, S. (1996), Image Indexing Using Moments and Wavelets, *IEEE Transactions on Consumers Electronics*, Vol. 42, No. 3, pp. 557-565, August 1996.

Madsen, H. (2003), Parameter estimation in distributed hydrological catchment modeling using automatic calibration with multiple objectives, *Advances in Water Resources*, 26, 205-216.

Mantoglou, A. (2003), Estimation of Heterogeneous Aquifer Parameters from Piezometric Head Data using Ridge Functions and Neural Networks, *Stochas. Environmen. Risk. Assess.*, 17:339 –352.

de Marsily, G.H. , Lavedan, G., Boucher, M., Fasanino, G. (1984), Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, *Proc Geostatistics for natural resources characterization*, Part 2.D, Verly et al (ed), Reidel Pub.Co., pp 831 –849.

de Marsily, G., Delhomme, J.P., Coudrain-Ribstein, A., LaVenue, M. A. (2000), Four decades of inverse problems in hydrogeology, In Theory, Modeling, and Field Investigation in Hydrogeology: A Special Volume in Honor of Shlomo P. Neuman's 60th Birthday, Zhang, D., and Winter, C.L., eds., Boulder, Colorado, Geological Society of America Special Paper 348, p. 1-17.

McDonald, M.G., and Harbaugh, A. W. (1988) *A modular three-dimensional finite-difference ground-water flow model*, Techniques of Water Resources Investigations 06 A1, United States Geological Survey.

McKenna, S.A., Doherty, J., Hart, D.B. (2003), Non-uniqueness of inverse transmissivity field calibration and predictive transport modeling, *Journal of Hydrology*, Volume 281, Number 4, October 2003, pp. 265-280(16).

McKenna, S.A., and Hart, D.B. (2003), Conditioning of Base T Fields to Steady-State Heads, Sandia National Labs Analysis Report, Task 3 of AP-088.

McLaughlin, D., and L. R. Townley, (1996) A reassessment of the groundwater inverse problem, *Water Resources Research*, 32(5), 1131–1161.

Medina, A., and Carrera, J. (1996), Coupled estimation of flow and solute transport parameters, *Water Resources Research*, Volume 32, Issue 10, p. 3063-3076.

Michie, D., Spiegelhalter, D.J., and Taylor, C.C. (eds.) (1994), *Machine learning, neural and statistical classification*, Ellis Horwood.

Milligan, G.W., and Cooper, M.C. (1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, 50, 159-179.

Minka, T.P. (2000), Bayesian Model Averaging is not Model Combination, MIT Media Lab note (7/6/00), Available at http://research.microsoft.com/~minka/papers/minka-bma-isnt-mc.pdf

Mitchell, M. T. (1996), *Machine Learning*, McGraw Hill, New York.

Mantovan P and Todini E. (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, 330: 368–381.

Mugunthan, P. and Shoemaker, C.A. (2006) Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, *Water Resources Research*, Vol. 42, W10428, doi:10.1029/2005WR004640.

Moore, C., and Doherty, J. (2005), The role of the calibration process in reducing model predictive error, *Water Resources Research*, 41(5), W05020.

Moore, C., and Doherty, J. (2006), The cost of uniqueness in groundwater model calibration, *Advances in Water Resources*, 29 (4), 605–623.

Nagao, N., Yamamoto, M., Suzuki, K., Ohuchi, A. (1998), Evaluation of the image retrieval system using interactive genetic algorithm, *J. of Japanese Society for Artificial Intelligence*, vol 13, no. 5, pp. 720-727.

National Research Council (NRC) (1994), Alternatives for Ground Water Cleanup, National Research Council, Washington, DC.

National Research Council (NRC) (2001), Conceptual Models of Flow and Transport in the Fractured Vadose Zone, Natl. Acad. Press, Washington, D. C.

Nelson, R.W. (1960), In place measurement of permeability in heterogeneous media, 1. Theory of a proposed method, *Journal of Geophysical Research*, v. 65, 6, p. 1753-1760.

Neuman SP (1973). Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty. *Water Resources Research*, 9 (4): 1006-1021.

Neuman, S.P. (1982), Statistical characterization of aquifer heterogeneities: an overview, In *Recent trends in Hydrogeology*, Geological Society of America Special Paper, 189, p. 81-102, Boulder, Colorado.

Neuman, S. P. (2002), Accounting for conceptual model uncertainty via maximum likelihood model averaging, *in Proceedings of 4th International Conference on Calibration and Reliability in Groundwater Modelling (ModelCARE 2002)*, edited by K. Kovar and Z. Hrkal, pp. 529– 534, Charles Univ., Prague, Czech Republic.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, 17(5), 291– 305, doi:10.1007/s00477-003-0151-7.

Neuman, S.P., and Wierenga, P.J. (2003), A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, NUREG/CR-6805. U.S. Nuclear Regulatory Commission, Washington, DC.

Ng, A.Y., Jordan, M.I., and Weiss, Y. (2002), On spectral clustering: Analysis and an algorithm, In *Advances in Neural Information Processing Systems* (NIPS), 14, MIT Press.

Ng, R.T., and Hand, J. (1994), Efficient and Effective Clustering Methods for Spatial Data Mining, In Proceedings for the 20th International Conference on Very Large Data Bases. http://www.softlab.ntua.gr/facilities/public/AD/DM/clarans.pdf

Nishio, K., Murakami, M., Mizutani, E. and Honda, N. (1997), Fuzzy Fitness Assignment in an Interactive Genetic Algorithm for a Cartoon Face Search, in *Genetic Algorithms and Fuzzy Logic Systems: Soft Computing Perspectives*, E. Sanchez, T. Shibata, and L. A. Zadeh, eds., World Scientific Publishing, pp.175-192.

Ohlander, R., Price, K., and Reddy, D.R. (1978) Picture Segmentation Using a Recursive Region Splitting Method, *Computer Graphics and Image Processing,* 8:313-333.

Ohsaki, M. Takagi, H., and Ohya, K. (1998) An input method using discrete fitness values for interactive GA. Journal of Intelligent and Fuzzy Systems 6(1): 131-145.

Ortiz, J. and Deutsch, C.V. (2002), Calculation of Uncertainty in the Variogram, *Mathematical Geology*, 34(2), p. 169-183.

Oz B., Deutsch C.V., Tran T.T., and Xie, Y. (2003), DSSIM-HR: A FORTRAN 90 program for direct sequential simulation with histogram reproduction, *Computers and Geosciences*, Elsevier, Volume 29, Number 1, February 2003 , pp. 39-51(13).

Pardo-Igúzquiza, E. and Dowd, P.A. (2001), The variance-covariance matrix of the experimental variogram: assessing variogram uncertainty, *Mathematical Geology*, 33, (4), 397-419.

Parmee, I.C., Cvetkovic, C., Watson, A.H., and Bonham, C.R. (2000), Multi-objective satisfaction within an interactive evolutionary design environment, *Journal of Evolutionary Computation*, 8, 197–222.

Poeter, E.P., and Hill, M.C. (1998), Documentation of UCODE: a computer code for universal inverse modeling, U.S. Geological Survey Water-Resources Investigations Report 98 –4080:116 pp.

Pollock, D.W., 1994, *User's Guide for MODPATH/MODPATH-PLOT, Version 3: A particle tracking post-processing package for MODFLOW*, the U.S. Geological Survey finite-difference ground-water flow model: U.S. Geological Survey Open - File Report.

Press, S. J. (1989), *Bayesian Statistics: Principles, Models, and Applications*, John Wiley and Sons, New York.

Prickett, T.A., Naymik, T.G., and Lonnquist, C.G. (1981), *A 'Random-Walk' Solute Transport Model for Selected Groundwater Quality Evaluations*, ISWS/BUL 65/81, Bulletin 65, State of Illinois, Illinois Department of Energy and Natural Resources.

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

Quinlan, J. R. (1994), *C4.5: Programs for Machine Learning,* Morgan Kaufman.

Raftery, A. E., Balabdaoui, F., Gneiting, T., and Polakowski, M. (2003) *Using Bayesian Model Averaging to Calibrate Forecast Ensembles*, Technical Report no. 440, Department of Statistics, University of Washington, December 15, 2003.

Raftery, A. E., Madigan, D. and Volinsky, C. T. (1996), Accounting for model uncertainty in survival analysis improves predictive performance, *Bayesian Statistics*, edited by J. Bernardo et al., pp. 323– 349, Oxford Univ. Press, New York.

Ramarao, B.S., LaVenue, A.M., de Marsily, G.H., Marietta, M.G. (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.

Rao S.V.N., Thandaveswara B.S., Bhallamudi S.M. (2003), Optimal groundwater management in deltaic regions using simulated annealing and neural networks, *Water Resour. Manag.*, 17(6):409 –428.

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

Reed, P., and Minsker, B. (2004), Striking the Balance: Long-Term Groundwater Monitoring Design for Conflicting Objectives, *ASCE Journal of Water Resources Planning and Management*, 130 (2), 140-149.

Rehfeldt, K.R., Boggs, J. M., and Gelhar, L. W. (1992), Field study of dispersion in a heterogeneous aquifer, 3 Geostatistical analysis of hydraulic conductivity, *Water Resources Research*, Vol. 28, No. 12 (1992) 3309.

Ritzel, B.J., Eheart, J.W. and Ranjithan, S. (1994), Using Genetic Algorithms to Solve a Multiple Objective Groundwater Remediation Problem, *Water Resources Research*, 30, (5), 1589-1603.

Robertson, Thomas S. (1970), Consumer Behavior, Glenview, Illinois: Scott, Foresman and Company. Runyon, Kenneth E. (1977), Consumer Behavior and the Practice of Marketing, Columbus, Ohio: Charles E. Merrill Publishing Company.

Roth, C. (1998), Is Lognormal Kriging Suitable for Local Estimation?, *Mathematical Geology*, Vol 30, No. 8, 1998.

Ruan, J. and Zhang, W. (2006), Identification and evaluation of weak community structures in networks, *Proc. of National Conference on Artificial Intelligence*, (AAAI-06), Boston, MA, July 2006.

Rubin, Y. (1991), Transport in heterogeneous porous media: Prediction and uncertainty, *Water Resources Research*, 27(7), 1723–1738.

Rubin, Y., and Dagan, G. (1987), Stochastic identification of transmissivity and effective recharge in steady groundwater flow. 1. Theory., *Water Resources Research*, v. 23, 7, p. 1185-1192.

Rudeen, D.K. (2003) User's Manual for DTRKMF Version 1.00. ERMS# 523246. Carlsbad, NM: Sandia National Laboratories, WIPP Records Center.

Ruspini, E. H. (1969), A new approach to clustering, *Inf. Control*, 15, 22–32.

Russell, P. and Norvig, S. J. (1995), *Artificial Intelligence: A Modern Approach*, Prentice-Hall.

Samper, F. J., and Neuman, S. P. (1989), Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 2. Synthetic experiments, *Water Resour. Res.*, 25, 363– 371.

Rutledge, G. W. (1995), *Dynamic Selection of Models*, PhD Thesis, Stanford University.

Schaffer, J. D. (1984), Some experiments in machine learning using vector evaluated genetic algorithms, Doctoral dissertation, Vanderbilt University: Nashville, TN.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, 6(2), 461– 464.

Sclove, S.L. (1994) Small sample and large sample statistical model selection criteria, *Selecting Models from Data: AI and Statistics IV*, Cheeseman, P. and Oldford, R.W. (ed), New York, NY: Springer-Verlag.

Shi, J. and Malik, J. (2000) Normalized Cuts and Image Segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Singh, A., Minsker, B. S., and Goldberg, D. (2003), Combining Reliability and Pareto Optimality - An Approach Using Stochastic Multi-Objective Genetic Algorithms, American Society of Civil Engineers (ASCE) Environmental & Water Resources Institute (EWRI) World Water & Environmental Resources Congress 2003 & Related Symposia, Philadelphia, PA, 2003.

Soares, A. (2001), Direct Sequential Simulation and Cosimulation, *Mathematical Geology*, Vol 33, No. 8, November.

Solomatine D.P., Dibiki, Y.B., Kukuric, N. (1999), Automatic calibration of groundwater models using global optimization techniques, *Hydrol. Sci. J.*, 44(6):879 –94.

Srinivas, N. and Deb, K. (1994). Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms, *Evolutionary Computation*, 2(3), 221-248.

Stone, H., and Sidel, J. L. (1993), Sensory Evaluation Practices, 2nd ed. Academic Press.

Sun, N.-Z., and Yeh, W. W.-G. (1990), Coupled Inverse Problems in Groundwater Modeling, 1. Sensitivity Analysis and Parameter Identification, *Water Resources Research*, 26(10): 2507-2525.

Sun, N.-Z., and Yeh, W. W-G. (1992), A Stochastic Inverse Solution for Transient Groundwater Flow: Parameter Identification and Reliability Analysis, *Water Resources Research*, 28(12): 3269-3280.

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.

Tikhonov, A. N. (1963) *Solution of incorrectly formulated problems and the regularization method*, Soviet Math Dokl 4, 1035-1038 English translation of Dokl Akad Nauk SSSR 151, 1963, 501-504.

Tokui, N. and Iba, H. (2000), Music Composition with Interactive Evolutionary Computation, in *Proceedings of 3rd International Conference on Generative Art (GA2000)*, Milan, Italy.

Tsai, F. T-C., Sun, N.Z., Yeh, W.G. (2003), Global-local optimization methods for the identification of three dimensional parameter structure in groundwater modeling, *Water Resources Research,* 39(2) 1043.

Turk M., and Pentland, A. (1991) Eigenfaces for recognition, *J. of Cognitive Neuroscience*, v3 nr1, p71-86.

US DOE (1990), Final supplement environmental impact statement: waste isolation pilot plant, Rep. DOE/EIS-0026-FS, US Department of Energy, Washington, DC.

US DOE (1991), Strategy for the waste isolation pilot plant test phase, Rep. DOE/EM/48063-2, US Department of Energy, Washington, DC.

Vailaya, A. Figueiredo, M. A. T., Jain, A. K., Zhang, H. J. (2001), Image classification for content-based indexing, *IEEE Transactions on Image Processing*, 10(1): 117-130.

van der Gaag, L. C., Renooij, S. Witteman, C. L. M., Aleman, B. M. P., and Taal, B. G., How to Elicit Many Probabilities, *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence 1999*. Available at http://citeseer.ist.psu.edu/vandergaag99how.html

Vecchia, A.V. and Cooley, R.L. (1987), Simultaneous Confidence and Prediction Intervals for Nonlinear Regression Models with Application to a Groundwater Flow Model, *Water Resources Research*, vol23, no. 7, pp1237-1250.

Verma, D. and Meila, M. (2003) *A comparison of spectral clustering algorithms*, Technical report uw-cse-03-05-01, University of Washington. Available at http://citeseer.ist.psu.edu/article/verma03comparison.html

Vermeul, V.R., Bergeron, M.P. Cole, C.R., Murray, C.J., Nichols, W.E., Scheibe, T.D., Thorne, P.D. Waichler, S.R., Xie, Y. (2003), *Transient Inverse Calibration of the Site-Wide Groundwater Flow Model (ACM-2): FY 2003 Progress Report*, PNNL-14398, Pacific Northwest National Laboratory, Richland, WA.

Vesselinov, V.V., Neuman, S.P., Illman, W.A. (2001), Three-dimensional numerical inversion of pneumatic cross-hole tests in unsaturated fractured tuff 2. Equivalent parameters, high-resolution stochastic imaging and scale effects, *Water Resources Research*, 37 (12): 019-41.

Vogel, R.M., Batchelder, R., and Stedinger, J.R. (2007) Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) Method, *Water Resources Research*, submitted, January 11, 2007.

Wagner, B.J., and Gorelick, S.M. (1989), Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity; from data to design, *Water Resources Research*, v. 25, 10, p. 2211-2225.

Wang, Q.J., (1991) The genetic algorithm and its application to calibrating conceptual runoff models, *Water Resource Research*, 27(9), 2467-2471.

Wang, J., Wiederhold, G., Firschein, O., Wei, S. (1997), Wavelet-Based Image Indexing Techniques with Partial Sketch Retrieval Capability, International Conference on the Advances in Digital Libraries, Library of Congress, Washington, D. C., May 7-9.

Wang, M. and Zheng, C. (1997), Optimal remediation policy selection under general conditions, *Groundwater*, 35(5), 757-764.

Ward, J. H. (1963), Hierarchical Grouping to optimize an objective function, *Journal of American Statistical Association*, 58(301), 236-244.

Welge, M., Auvil, L., Shirk, A., Bushell, C., Bajcsy, P., Cai, D., Redman, T., Clutter, D., Aydt, R., and Tcheng, D. (2003), *Data to Knowledge (D2K)* - An automated learning group report, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. Available at http://alg.ncsa.uiuc.edu

Wijns, C., Boschetti, F. , Moresi, L. and Takagi, H. (2001), Inversion in geology by interactive evolutionary computation, IEEE Systems, Man, and Cybernetics Conference, 7-10 October 2001, Tucson, USA, *Proc. of IEEE Int'l Conf. on Systems, Man, and Cybernetics*, pp. 1053-1057.

Winograd, T., and Flores, F. (1986) *Understanding computers and cognition: a new foundation for design*, Ablex, Norwood, N.J.

Woodbury, A.D., and Ulrych, T.J. (2000), A full-Baysian approach to the groundwater inverse problem for steady state flow. *Water Resources Research*, 36(8), 2081-2093.

Woods, D.D., Roth, E.M., Benett, K. (1990), Explorations in joint human-machine cognitive systems. In Robertson S. Zachary W. & Black J.B. (Eds.) *Cognition, Computing and Cooperatio*n. Ablex (pp. 123-158).

Xiao, N., Bennett, A.B., and Armstrong, M.P. (2002), Using evolutionary algorithms to generate alternatives for multiobjective site-search problems, *Environment and Planning A*, 2002, vol. 34, pages 639 – 656.

Yakowitz, S., and Duckstein, L. (1980), Instability in aquifer identification: Theory and case study, *Water Resources Research*, 16(6), 1045–1064.

Yan, S., and Minsker, B. S. (2005), Optimal Groundwater Remediation Design Using Trust Region Based Meta-models within a Genetic Algorithm", American Society of Civil Engineers (ASCE) Environmental & Water Resources Institute (EWRI) World Water & Environmental Resources Congress 2005 & Related Symposia, Anchorage, AK.

Yang, J., Zhang, D., Frangi, A. F., Yang, J. Y. (2004), Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131-137.

Yapo, P.O., Gupta, H.V. and Sorooshian, S. (1998), Multi-objective global optimization for hydrologic models, *Journal of Hydrology*, 204, 83-97.

Ye, M., Neuman, S. P., and Meyer, P. D., Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resources Research*, Vol. 40, W05113.

Yeh, W. W-G. (1986), Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resources Research*, 22(1), 95–108.

Yeh, W. W-G., and Yoon. Y.S. (1981), Aquifer parameter identification with optimum dimension in parameterization, *Water Resources Research*, 17(3):664 –672.

Zimmerman, D.A., de Marsily, G., Gotway, C.A., Marietta, M.G., Axness, C.L., Beauheim, R.L., Bras, R.L., Carrera, J., Dagan, G., Davies, P.B., Gallegos, D.P., Galli, A., Gómez-Hernández, J.J., Grindrod, P., Gutjahr, A.L., Kitanidis, P.K., LaVenue, A.M., McLaughlin, D., Neuman, S.P., RamaRao, B.S., Ravenne, C., Rubin, Y. (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.

Zheng, C., and Wang, P.P. (1996), Parameter structure identification using tabu search and simulated annealing, *Adv. Water Resour.*, 19(4): 215-224.

Zheng, X., Cai, D. He, X., Ma, W.-Y. and Lin, X. (2004), Locality Preserving Clustering for Image Database, *ACM Multimedia,* Oct, 2004.

APPENDIX A - DATA FOR THE WIPP SITE

This appendix provides additional data for the WIPP case study discussed in Section 3.2. Table A.1 gives the field measurements for the transmissivities and hydraulic heads from the WIPP site. Figure A.1 shows the histogram for the transmissivity residuals (as well as the average and standard deviation for these residuals). Figure A.2 shows the normal Q-Q plot of the transmissivity residuals with respect to a Gaussian distribution to show that the distribution of the residuals is close to normal. Finally, Figure A.3 shows the model variogram used for kriging and conditional simulations (for Sections 6.2.2.1 and 6.2.2).

Table A.1 Transmissivity and Head Data for the WIPP Site

Index	Well ID	Easting	Northing	Transmissivity	T Residual	Head
1	WIPP-29	596981	3578694	-3	-0.124965978	
2	WIPP-26	604014	3581162	-2.9	0.215976988	921.06
3	WIPP-27	604426	3593079	-3.3	-0.032094958	
4	WIPP-25	606385	3584028	-3.5	-0.013776615	932.7
5	USGS-1	606462	3569459	-3.3	0.289981391	
6	H-7c	608095	3574640	-2.8	0.397944886	
7	H-7b1	608124	3574648			913.86
8	D-268	608702	3578877	-5.7	0.279142755	
9	P-14	609084	3581976	-3.5	0.162120559	
10	H-6b	610594	3585008			934.2
11	H-6c	610610	3584983	-4.4	-0.01524374	
12	P-15	610624	3578747	-7	-0.959381322	
13	WIPP-28	611266	3594680	-3.6	-0.151237057	
14	H-18	612264	3583166	-5.7	0.73158886	937.22
15	H-14	612341	3580354	-6.5	-0.269343758	920.24
16	H-4b	612380	3578483			915.55
17	H-4c	612406	3578499	-6.1	0.052211851	
18	WQSP-1	612561	3583427	-4.5	0.015404188	935.64
19	WQSP-6	612605	3580736	-6.6	-0.32261093	920.02
20	WIPP-13	612644	3584247	-4.1	0.421802134	935.17
21	H-2b2	612661	3581649			926.62
22	H-2c	612666	3581668	-6.2	0.135944192	
23	CB-1	613191	3578049	-6.5	-0.329428823	
24	H-16	613369	3582212	-6.1	0.349623652	
25	H-1	613423	3581684	-6	0.412951363	927.19
26	WQSP-5	613668	3580353	-5.9	0.471781966	917.22

27	DOE-2	613683	3585294	-4	0.694917592	940.03
28	ERDA-9	613696	3581958	-6.3	0.152498092	921.59
29	H-3b2	613701	3580906			917.16
30	WIPP-12	613710	3583524	-7	-0.396269283	935.3
31	WIPP-30	613721	3589701	-6.7	-0.351306277	936.88
32	H-3b1	613729	3580895	-4.7	-0.221314463	
33	WIPP-18	613735	3583179	-6.5	0.068402924	936.08
34	WIPP-22	613739	3582653	-6.4	0.105489781	930.96
35	WIPP-19	613739	3582782	-6.2	0.325981755	932.66
36	WIPP-21	613743	3582319	-6.6	-0.111478687	927
37	WQSP-2	613776	3583973	-4.7	-0.027289774	938.82
38	P-17	613926	3577466	-6	0.247624709	915.2
39	H-9c	613974	3568234	-4	-0.227630764	
40	H-9b	613989	3568261			911.57
41	H-19b0	614514	3580716	-5.2	-0.62242446	917.13
42	WQSP-3	614686	3583518	-6.8	-0.151390932	935.89
43	WQSP-4	614728	3580766	-4.9	-0.288951145	917.49
44	Engle	614953	3567454	-4.3	-0.516318452	
45	DOE-1	615203	3580333	-4.9	-0.210043134	916.55
46	H-11b4	615301	3579131	-4.3	0.253142801	915.47
47	H-15	615315	3581859	-6.8	-0.126309126	919.87
48	H-17	615718	3577513	-6.6	-0.143097525	915.37
49	H-5b	616872	3584801			936.26
50	H-5c	616903	3584802	-6.7	0.02946428	
51	H-12	617023	3575452	-6.7	-0.076472738	914.66
52	AEC-7	621126	3589381	-6.8	-0.110777878	933.19
53	H-10b	622975	3572473	-7.4	-0.014838901	

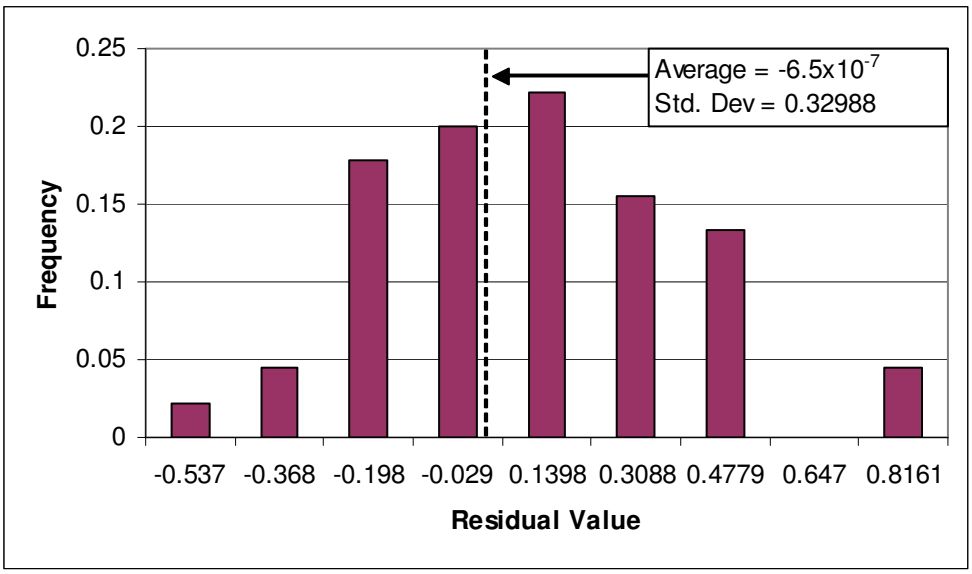


Figure A.1 Histogram and summary statistics of the T residuals

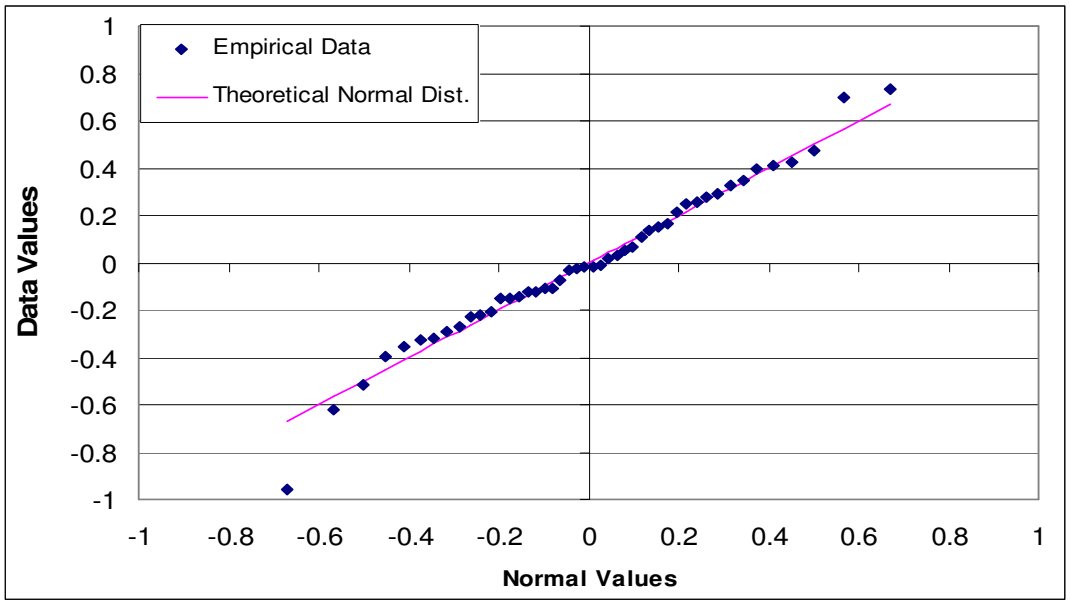


Figure A.2 Normal Q-Q plot to assess normality of T residuals – empirical data on the straight line indicates perfect normality of data

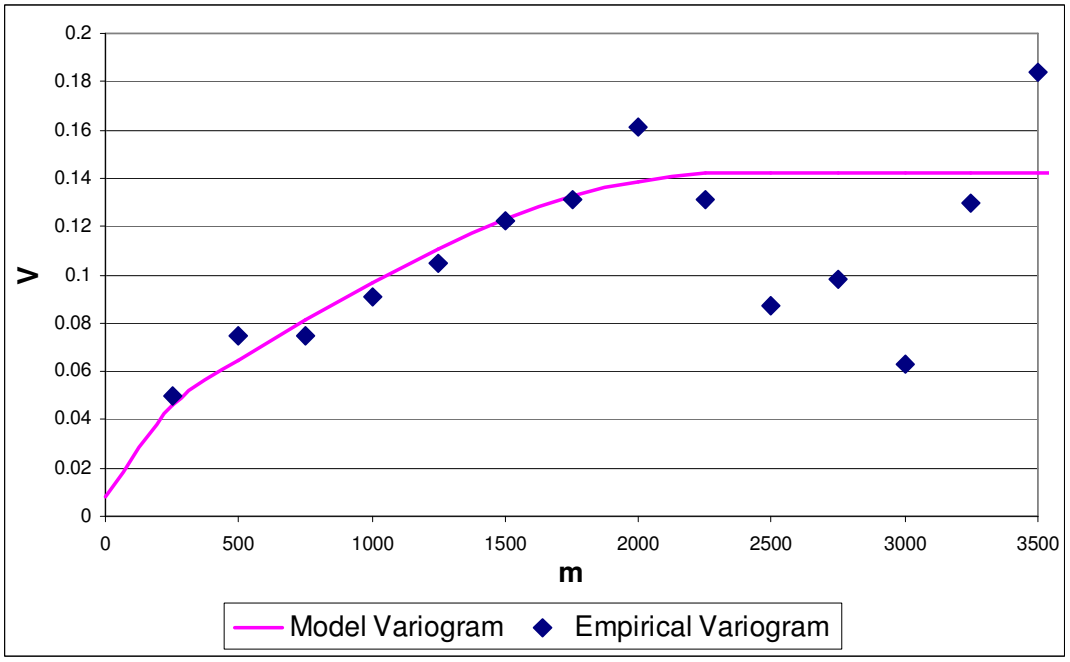


Figure A.3 Empirical and model variogram for the T residuals. The model variogram consists of a nugget of 0.008 and two nested spherical variograms with sills and ranges of 0.02/220 m and 0.114/2330 m, respectively

APPENDIX B – KRIGING THEORY AND ESTIMATION COVARIANCE

This appendix provides the underlying theory behind kriging that is used for much of this research (especially Sections 4.2.1, 4.2.2, 6.2.2.1). The first part of this appendix discusses the mathematical formulation of ordinary kriging. The second part shows the derivation of the estimation covariance that is used in sections 4.2.2 and 6.2.2.1.

Ordinary Kriging

Given *n* data points and *m* pilot points, the log-conductivity field can be generated with ordinary kriging using the following equation [*Deutsch and Journel,* 1998] (here as elsewhere we simplify the notation by using K to mean the log-conductivity, since the values are kriged in log-space):

$$K(u_x)' = \sum_{\alpha=1}^{n_x} \lambda_\alpha \left(K\ (u_o^\alpha) - \overline{K}_x \right) + \sum_{i=1}^{p_x} \lambda_i \left(K(u_{pp}^i) - \overline{K}_x \right) + \overline{K}_x \quad n_x, p_x \in W_x \quad \forall x \qquad \text{(B.1)}$$

where $K(u_x)'$ is the kriging estimate for location u_x, $K(u_o^\alpha)$ is the value of the log conductivity and λ_α the corresponding kriging weight for the α^{th} observation point located at u_o^α, $K(u_{pp}^i)$ is the value and λ_i the kriging weight for the i^{th} pilot point located at u_{pp}^i, $\overline{K}_x$ is the expected value for the log conductivity for location u_x (unknown but assumed to be constant within the kriging window), n_x is the number of data points and p_x the number of pilot points contained within the kriging window W_x for location x. The kriging constraints are given below:

$$\begin{cases} \sum_{\beta=1}^{n_x} \lambda_\beta Cov(u_o^\alpha, u_o^\beta) + \sum_{\gamma=1}^{p_x} \lambda_\gamma Cov(u_o^\alpha, u_{pp}^\gamma) + \mu = Cov\left(u_o^\alpha, u_x\right) \quad \alpha = 1,..., n_x \\ \sum_{\delta=1}^{p_x} \lambda_\delta Cov(u_{pp}^\gamma, u_{pp}^\delta) + \sum_{\alpha=1}^{n_x} \lambda_\alpha Cov(u_{pp}^\gamma, u_o^\alpha) + \mu = Cov\left(u_{pp}^\gamma, u_x\right) \quad \gamma = 1,..., p_x \\ \sum_{\alpha=1}^{n_x} \lambda_\alpha + \sum_{i=1}^{p_x} \lambda_i = 1 \end{cases} \tag{B.2}$$

where $Cov(u^x,u^y)$ is the model covariance between locations u^x and u^y (these could be locations for the data points – subscript o – or pilot points – subscript pp), and μ is the Lagrange multiplier used to impose the non-bias condition given by the third part of Equation B.2.

Once the log-conductivity field has been generated using Equations B.1 and B.2, it is then back-transformed using the following relation from *Roth* [1998], to give the conductivity field to be used for model prediction:

$$K_{BT}(u_x') = \exp\left\langle K(u_x') + \sigma_x^2 / 2 - \mu_x \right\rangle \tag{B.3}$$

where K_{BT} is the back-transformed kriged estimate for location x, K_x is the kriging estimate in the log-normal space, σ_x is the kriging estimation variance for location x (also calculated by the kriging algorithm), and μ_x is the Lagrange multiplier (from Equation B.2) used for ordinary kriging at location x.

Estimation Covariance

The estimation covariance (in other words, the correlation between the expected kriging errors) between kriged locations i and j is given by:

$$C_{pp}^{ij} = E\left\langle \left(K(u_{pp}^i)' - K(u_{pp}^i)\right)\left(K(u_{pp}^j)' - K(u_{pp}^j)\right)\right\rangle \tag{B.4}$$

where C_{pp}^{ij} is the covariance between estimation errors for two pilot points located at u_{pp}^{i} and u_{pp}^{j}, $K(u_{pp}^{i})'$ is the kriging estimate for the i^{th} pilot point (given by equation B.1), $K(u_{pp}^{i})$ is the true value for the i^{th} pilot point, and *E()* is the expected value for the given variable. We can write B.4 in terms of residual from the mean value.

$$C_{pp}^{ij} = E\left\langle \left[K(u_{pp}^{i})' - \overline{K}_i - \left(K(u_{pp}^{i}) - \overline{K}_i\right)\right]\left[K(u_{pp}^{j})' - \overline{K}_j - \left(K(u_{pp}^{j}) - \overline{K}_j\right)\right]\right\rangle \tag{B.5}$$

Expanding B.5 we get:

$$\begin{aligned} C_{pp}^{ij} &= E\left\langle \left(K(u_{pp}^{i})' - \overline{K}_i\right)\left(K(u_{pp}^{j})' - \overline{K}_j\right)\right\rangle + E\left\langle \left(K(u_{pp}^{i}) - \overline{K}_i\right)\left(K(u_{pp}^{j}) - \overline{K}_j\right)\right\rangle \\ &- E\left\langle \left(K(u_{pp}^{i})' - \overline{K}_i\right)\left(K(u_{pp}^{j}) - \overline{K}_j\right)\right\rangle - E\left\langle \left(K(u_{pp}^{j})' - \overline{K}_j\right)\left(K(u_{pp}^{i}) - \overline{K}_i\right)\right\rangle \end{aligned} \tag{B.6}$$

Substituting equations B.1 for $K(u_{pp}^{i})'$ and $K(u_{pp}^{i})'$ into B.6:

$$\begin{aligned} &= E\left(\sum_{\alpha=1}^{n_i}\lambda_\alpha\left(K(u_o^\alpha) - \overline{K}_i\right)\sum_{\beta=1}^{n_j}\lambda_\beta\left(K(u_o^\beta) - \overline{K}_j\right)\right) \\ &+ E\left(K(u_{pp}^{i}) - \overline{K}_i\right)\left(K(u_{pp}^{j}) - \overline{K}_j\right) \\ &- E\left(\sum_{\alpha=1}^{n_i}\lambda_\alpha\left(K(u_o^\alpha) - \overline{K}_i\right)\left(K(u_{pp}^{j}) - \overline{K}_j\right)\right) \\ &- E\left(\sum_{\alpha=1}^{n_i}\lambda_\alpha\left(K(u_o^\alpha) - \overline{K}_i\right)\left(K(u_{pp}^{j}) - \overline{K}_j\right)\right) \end{aligned} \tag{B.7}$$

Simplifying B.7 we get equation B.8, which is the same as Equation 4.3 in Section 4.2.2.

$$\begin{aligned} C_{pp}^{ij} &= \sum_{\alpha=1}^{n_i}\sum_{\beta=1}^{n_j}\lambda_\alpha\lambda_\beta Cov\left(u_o^\alpha - u_o^\beta\right) + Cov\left(u_{pp}^{i} - u_{pp}^{j}\right) \\ &- \sum_{\alpha=1}^{n_i}\lambda_\alpha Cov\left(u_o^\alpha - u_{pp}^{j}\right) - \sum_{\alpha=1}^{n_i}\lambda_\alpha Cov\left(u_o^\alpha - u_{pp}^{j}\right) \end{aligned} \tag{B.8}$$

APPENDIX C – EIGENIMAGE ANALYSIS

This appendix provides details about the calculation of eigenimages and eigenscores that are used to extract spatial information from the conductivity/transmissivity fields. These eigenimages and eigenscores are used in the clustering and machine learning algorithms discussed in Sections 5.2.1.1.

Calculating Eigenimage and Eigenscores

Given *n* images (in this case 2-D conductivity fields) with *r* rows and *c* columns of cells, the algorithm to compute the eigenimages and the corresponding projections is as follows:

1. For each image, align all cells length-wise to form a vector of length *rc*.
2. Stack all *n* vectors to form the super-matrix A of dimensions *rc* by *n* such that each column of A corresponds to an image vector.
3. Calculate the overall mean $\overline{A}$ across all columns of *A*:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} A_{[i]} \tag{C.1}$$

   where $A_{[i]}$ represents the i^{th} column in A.

4. Subtract $\overline{A}$ from *A* to get the centered matrix *A'*.
5. Calculate the *rc*,*rc* covariance matrix *C* for *A'*:

$$C = \frac{1}{rc}(A'^{T}\ A') \tag{C.2}$$

6. Find the eigen-decomposition of the covariance matrix:

$$\Lambda = E^{T} C E \tag{C.3}$$

where Λ is a *rc,rc* diagonal matrix with the eigenvalues (λ_1, $\lambda_{2...}\lambda_n$) of *C*, and *E* is a *rc*,*rc* matrix with columns corresponding to the eigenvectors of C.

7. Sort the Λ and *E* matrix according to λ. Keep only the top *p* eigenvectors that capture a required amount of the variability in the dataset. We use eigenvectors with 95% or more of the total spectral energy. Mathematically this is equivalent to selecting the maximum *p* such that:

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \leq 0.95 \qquad \text{(C.4)}$$

where λ_i is the eigenvalue corresponding to the i^{th} eigenvector. This yields an *rc*,*p* truncated *E'* matrix with the columns representing the top *p* eigenvectors.

8. Realigning the i^{th} columns of *E'* back to *r* rows and *c* columns and multiplying the resulting matrix by the root of the corresponding eigenvalue $\lambda_i^{1/2}$ gives the i^{th} eigenimage for the *A'* dataset (before visualizing these eigenimages it is necessary to add back the average matrix $\overline{A}$).

9. Calculate the *p,n* score matrix Y for *A'*:

$$Y = E'^{T} A' \qquad \text{(C.5)}$$

The i^{th} column in Y represents the projection (or eigenscores) of the i^{th} image over the top *p* eigenimages. In other words, these *p* eigenscores represent the 'presence' or contribution that the *p* eigenimages have in that particular image. High values of the score imply a greater contribution of that eigenimage (or a close match between the image and the eigenimage), low scores represent less

contribution, and negative scores imply an anti-correlation between the patterns in the eigenimage and the patterns seen in the actual image.

APPENDIX D – CLUSTERING THEORY

This appendix provides details about the two kinds of clustering algorithms used in this research (in Sections 5.2.1.2) – hierarchical clustering and spectral N-cuts clustering. In addition, the calculation for the 'Wallace Index' used in Section 5.3.1 is also discussed in this appendix.

Agglomerative Hierarchical Clustering

Given *n* data points to be split into *k* clusters, the algorithm for hierarchical agglomerative clustering is:

1. Start with each data point assigned to a unique cluster – leading to *n* total clusters.
2. Calculate 'linkage' based on a 'similarity metric' (a measure of how close two clusters are – examples are nearest-neighbor, farthest-neighbor, average linkge, etc; see *Duda et al* [2001] for a discussion on different similarity metrics) between all clusters.
3. Combine the clusters with the greatest similarity metric to form *n-1* clusters.
4. Repeat 2 and 3 until *k* clusters are formed.

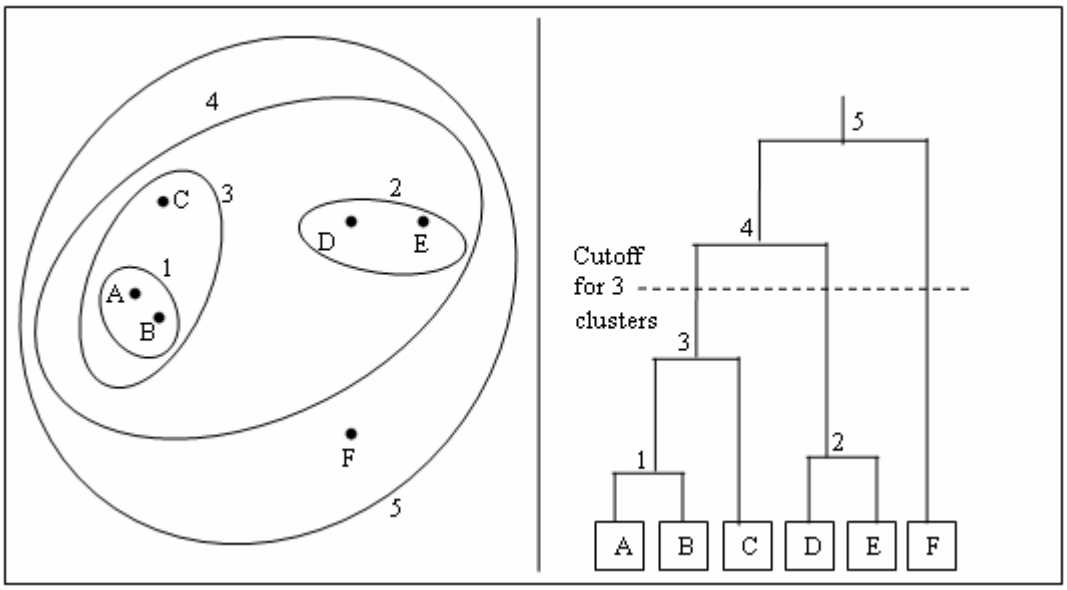


Figure D.1 Example of Agglomerative Hierarchical Clustering for 2-D Data

Figure D.1 shows an example of agglomerative clustering with five two-dimensional data points. The left side of the figure shows the actual data configuration and the clusters formed, while the right side shows a 'dendrogram' [*Duda et al.*, 2001] showing the linkage relationships between the different data points and clusters. Each data point is labeled from A to F, and each cluster is labeled from 1 to 5. The number of the cluster also corresponds to the sequence in which it is created (i.e. cluster 1 with A and B is the first cluster created, followed by 2 with D and E, and so on). As one traverses up the tree the data are clustered into higher 'abstractions' (fewer clusters). To identify 3 unique clusters, the linkage tree would be cut off at the level shown in Figure 5.3.

Spectral N-Cuts Clustering

For *n* data points to be clustered in *k* groups, the N-cuts algorithm is as follows:

1. Construct the *n, n* affinity matrix (also referred to as the heat kernel) for the differences between all the data points. Element i, j in the affinity matrix denotes the 'affinity measure' between data points i and j:

$$A_{i,j} = \exp\left(\frac{-\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \qquad \text{if } i \neq j \quad \text{otherwise } A_{i,i} = 0 \qquad i,j = 1\ldots n \qquad \text{(D.1)}$$

where σ is a scaling parameter (known as the decay coefficient of the heat kernel) that controls how rapidly the affinity measure falls off as the difference in *i* and *j* increases.

2. Define *D* as a diagonal matrix with $D_{i,i} = \Sigma_j A_{i,j}$, i.e. the i^{th} diagonal element in *D* is the sum of *A*'s i^{th} row (or the i^{th} column, as *A* is symmetric). Construct the *n*, *n* Laplacian matrix $L = D^{-1/2} A\, D^{-1/2}$ (theoretically the Laplacian is given by *I* - *L*, but *Ng et al* [2002] use this form instead, to simplify the computation).

3. Find $E = [e_1, e_2, \ldots, e_k]$ largest eigenvectors of L (where *k* is the number of clusters).

4. Normalize E such that each row has unit length

$$E_{i,j} = \frac{E_{i,j}}{\left(\sum_{j=1}^{k} E_{i,j}^2\right)^{1/2}} \qquad i = 1\ldots n;\ j = 1\ldots k \qquad \text{(D.3)}$$

5. E is a *n, k* matrix where each row is a vector in R^k. Each data point is thus associated with a *k* dimensional vector. The data can then be clustered based on

these k dimensions using a standard clustering algorithm (this work uses hierarchical clustering).

Wallace Index for Clusters

The Wallace index between the true (labeled) clustering Γ and the predicted clustering Γ' is defined as:

$$W(\Gamma,\Gamma') = \min\left(\frac{N_{\Gamma,\Gamma'}}{S(\Gamma)}, \frac{N_{\Gamma,\Gamma'}}{S(\Gamma')}\right) \quad \text{(D.4)}$$

where $N_{\Gamma,\Gamma'}$ is the number of pairs of members that appear in the same cluster in both clustering schemes. $S(\Gamma)$ is the total number of linkages of intra-cluster members in the labeled clustering scheme, and $S(\Gamma')$ is the total number of linkages of intra-cluster members in the predicted scheme.

For the ith cluster, λ_i, with n members, the total number of linkages is given by:

$$s(\lambda_i) = 1 + 2 + \cdots (n-1) = \frac{1}{2} n(n-1) \quad \text{(D.5)}$$

$S(\Gamma)$ is simply the sum of all $s(\lambda_i)$ calculated for all the clusters in the clustering scheme Γ. It is noteworthy that the Wallace index is a non-linear function of the number of correct linkages. As the predicted clusters become more and more non-informative (by putting all or most of the members in the same cluster) $\frac{N_{\Gamma,\Gamma'}}{S(\Gamma)}$ approaches a value of 1, however $S(\Gamma')$ starts to get large in such cases leading to a lower $\frac{N_{\Gamma,\Gamma'}}{S(\Gamma')}$, and subsequently a lower Wallace index. The other extreme of most of the members

belonging to individual clusters leads to a low value for $N_{\Gamma,\Gamma'}$ and a subsequent low Wallace index.

APPENDIX E - PARAMETERIZATION OF THE LEARNING MODELS

This appendix presents the results from optimizing the parameters of the machine learning models (decision trees and naïve Bayes) used in Section 5.3.2. The two major parameters for the decision tree and naïve Bayes are the minimum leaf error ratio and the number of bins, respectively. The minimum leaf ratio gives the minimum allowable classification error that defines a leaf node. High values of minimum leaf ratio lead to shorter, more compact decision trees that may not have as much predictive power, while lower values lead to long, deeply branched tree-structures that run the risk of over-fitting the training data. The number of bins is an important parameter for naïve Bayes that splits continuous attributes into categorical values. A large number of bins lead to more complete representation of the true distribution of that particular attribute. However, the frequency of occurrence of very narrow bins within the data set is much lower (especially for sparse datasets) and this can lead to unreliable prediction. Thus trade-offs exist for both the minimum leaf ratio and the number of bins. Offline testing was used to test the decision tree and naïve Bayes to determine the best parameters for each of the prediction models. Figures E.1 and E.2 show the average prediction accuracy (in terms of the root mean square error in ranks) for different minimum leaf error ratios for the decision tree, and different bin sizes for the naïve Bayes model. As can be seen from the figures, both testing and training errors for the decision tree initially decrease with the decrease in the leaf ratio, however leaf ratios lower than 0.01 do not lead to any subsequent improvement in either the training or testing dataset (the testing error in fact slightly increases after LR = 0.01). Thus, the minimum leaf ratio for the decision tree was set at 0.01. The graph for

the prediction error for the naïve Bayes with different bin sizes indicates that a bin size of 10 would be a good alternative for this model (the prediction accuracy is slightly lower for a bin size of 15, but this was seen to give unstable results for the different trials and thus a bin size of 10 was chosen instead).

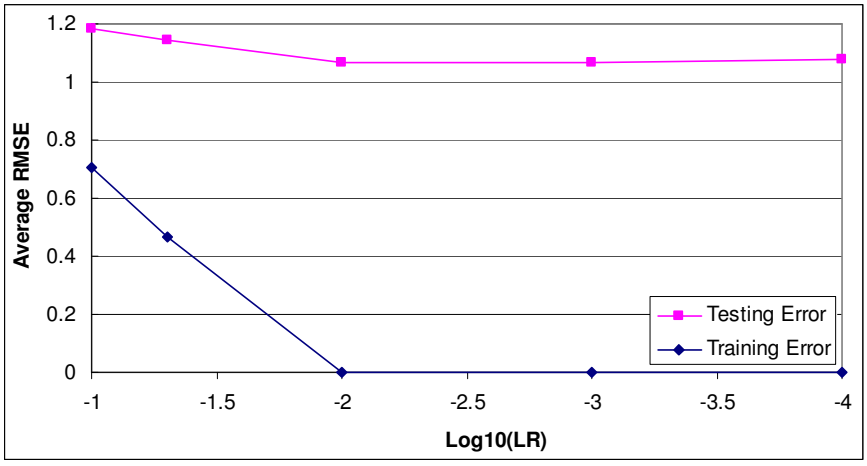


Figure E.1 Prediction accuracy for decision tree for different leaf error ratios

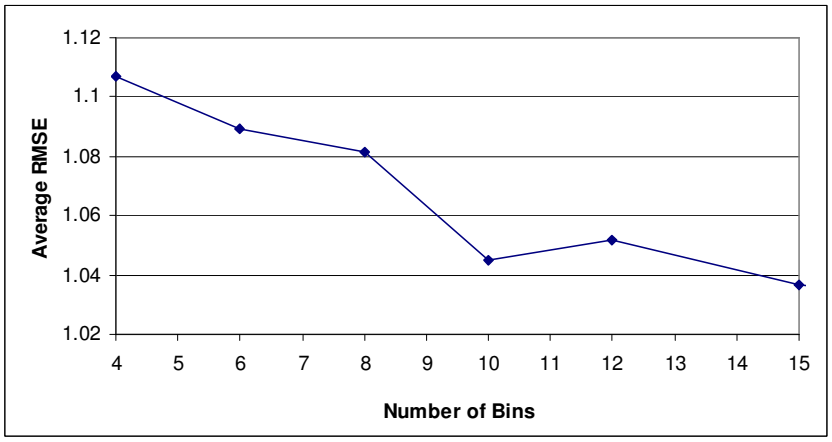


Figure E.2 Prediction accuracy for naïve Bayes for different bin sizes

APPENDIX F - LIST OF ACRONYMS

AI – Artificial Intelligence

AIC – Akaike Information Criterion

BIC – Bayesian Information Criterion

BMA – Bayesian Model Averaging

D2K – Data to Knowledge

DOE – Department of Energy

DSS – Decision Support Systems

DT – Decision Tree

EMO - Evolutionary multi-objective

GA – Genetic Algorithm

GLUE - Generalized Likelihood Uncertainty Estimation

GUI – Graphical User Interface

GW - Groundwater

HBGA – Human-Based Genetic Algorithms

IGA – Interactive Genetic Algorithm

IMOGA – Interactive Multi-Objective Genetic Algorithm

JCS – Joint Cognitive Systems

KIC – Kashyap Information Criterion

MGA – Modeling to Generate Alternatives

MLBMA – Maximum Likelihood Bayesian Model Averaging

MOGA - Multi-Objective Genetic Algorithm

NB – Naïve Bayes

NCSA – National Center for Supercomputing Applications

N-Cuts – Normalized Cuts

NLL – Negative Log Likelihood

NSGA – Non-Dominated Sorting Genetic Algorithm

PDF – Probability Distribution Function

PP – Pilot Points

RMS – Root Mean Square

RMSE – Root Mean Square Error

SGS – Sequential Gaussian Simulation

SNL – Sandia National Laboratory

T - Transmissivity

WIPP – Waste Isolation Pilot Plant

AUTHOR'S BIOGRAPHY

Abhishek Singh received his Bachelor of Engineering (Honors) in Civil Engineering at the Birla Institute of Technology and Science, Pilani, India (2001). In August 2001, he joined the Master of Science program in Environmental Engineering, in the Civil Engineering department at University of Illinois at Urbana Champaign. Soon, after the successful completion of the Masters program in October 2003, he continued his studies towards the PhD degree. In September 2007, he will start work as an Environmental Scientist in the Water Resources Division of INTERA Inc., Austin, TX. His research interests lie in the analysis, modeling and optimization of complex large-scale environmental systems, particularly water-resources problems. He has a keen interest in stochastic analysis, data assimilation and application of data-mining and machine-learning techniques for innovative and efficient solutions to real world problems.