Conceptual Model Uncertainty in the Management of the Chi River Basin, Thailand

by

Tussanee Nettasana

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2012

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

With increasing demand and pressures on groundwater resources, accurate and reliable groundwater prediction models are essential for sustainable groundwater management. Groundwater models are merely approximations of reality, and we are unable to either fully characterize or mathematically describe the true complexity of the hydrologic system; therefore, inherent in all models are varying degree of uncertainty. A robust management policy should consider uncertainties in both the imprecise nature of conceptual/numerical models and their parameters. This study addresses the critical question of whether the use of multiple conceptual models to explicitly account for conceptual model uncertainty improves the ability of the models to assist in management decisions.

Twelve unique conceptual models, characterized by three alternative geological interpretations, two recharge estimations, and two boundary condition implementations, were formulated to estimate sustainable extraction rates from Thailand's Thaphra Area, where increasing groundwater withdrawals may result in water level declination and saline water upconing. The models were developed with MODFLOW and calibrated using PEST with the same set of observed hydraulic head data. All of the models were found to reasonably produce predictions of the available heads data. To select the best among the alternative models, multiple criteria have been defined and applied to evaluate the quality of individual models. It was found that models perform differently with respect to different evaluation criteria, and that it is unlikely that a single inter-model comparison criterion will ever be sufficient for general use. The chosen alternative models were applied both individually and jointly to quantify uncertainty in the groundwater management context. Different model-averaging methods were assessed in terms of their ability to assist in quantifying uncertainty in sustainable yield estimation.

The twelve groundwater simulation models were additionally linked with optimization techniques to determine appropriate groundwater abstraction rates in the TPA Phu Thok aquifer. The management models aim to obtain maximal yields while protecting water level decline. Despite similar performances among the calibrated models, total sustainable yield estimates vary substantially depending on the conceptual model used and range widely, by a factor of 0.6 in total, and by as much as a factor of 4 in each management area. The comparison results demonstrate that simple averaging achieves a better performance than formal and sophisticated averaging methods such as Maximum Likelihood Bayesian Model Averaging, and produce a similar performance to GLUE and combined-multiple criteria averaging methods for both validation testing and management applications, but is much simpler to implement and use, and computationally much less demanding.

The joint assessment of parameter and conceptual model uncertainty was performed by generating the multiple realizations of random parameters from the feasible space for each calibrated model using a simple Monte Carlo approach. The multi-model averaging methods produce a higher percentage of predictive coverage than do any individual models. Using model-averaging predictions, lower optimal rates were obtained to minimize head constraint violations, which do not ensue if a single best model is used with parameter uncertainty analysis.

Although accounting for all sources of uncertainty is very important in predicting environmental and management problems, the available techniques used in the literature may be too computationally demanding and, in some cases, unnecessary complex, particularly in data-poor systems. The methods presented here to account for the main sources of uncertainty provide the required practical and comprehensive uncertainty analysis and can be applied to other case studies to provide reliable and accurate predictions for groundwater management applications.

Acknowledgements

I would like to express my cordial gratitude to my co-supervisor Dr.James Craig and Dr. Bryan Tolson for guidance, encouragement, patience, and providing me with facilities to pursue my work, and for ensuring financial support for the last year of my studies. I would like to express my sincere gratitude to my co-supervisor, Dr. Jon.Sykes, for his support and guidance. Without their help, this thesis would not be possible.

I would like to extend thanks to my review committee: Dr. Rene Therrien, Dr. Don Burn, Dr. Sriram Narasimhan, and Dr. Dave Rudolph for their insightful suggestions/comments. I am heartily thankful to my English teacher, Mary Mcpherson whose encouragement and support from the initial to the final level enabled me to develop my English.

It is an honor for me to acknowledge Thai people for my Ph.D. financial support through Royal Thai Government Scholarship from Department of Groundwater Resources, Ministry of Mineral Resource and Environment of Thailand. My academic development has also been backed up by Department of Groundwater Resources and my colleagues at DGR.

I also wish to give many thanks to my friends who make my Ph.D. study time be wonderful, Rungsima, Monrudee, Anuchart, Lucy, Sander, and others. Thank you all for sharing good times and for the thoughtful support through the hard times.

Finally, I would like to express my appreciation to my family whose dedication, love and persistent confidence in me, has provided happiness and helped me cope with difficulties.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

With increasing demand and pressures on groundwater resources, accurate and reliable model predictions of groundwater flow and contaminant transport are essential for sustainable groundwater management. Groundwater models function as predictive tools for exploring complex groundwater systems, predicting impacts of alternative hydrological or developmental scenarios, evaluating recharge and aquifer storage processes, and assessing the sustainable yield of groundwater resources. Therefore, groundwater management models can be effective tools to support important policy decisions concerning allocation policies and remediation of contamination in aquifers. The usefulness of these models for decision making depends upon their ability to provide accurate predictions. However, groundwater models are merely approximations of reality: we are unable to either fully characterize or mathematically describe the true complexity of hydrologic systems. Due to various sources of uncertainties, including conceptual model, parameter, and scenario uncertainty, predictions from these models are therefore also uncertain (US.EPA, 1997; Meyer et al., 2007).

The first and most important step in developing groundwater models is to build a conceptual model of the underlying system. A conceptual model can be defined variously as "the set of assumptions that represent our simplified perception of the real system that is to be mathematically modeled" (Bear and Verruijt, 1987); "a pictorial representation of the groundwater system, frequently, in the form of a block diagram or a cross section whose purpose is to simplify the field problem and organize the associated data so that the system can be analyzed more readily" (Anderson and Woesser, 1992); or as a "framework that represents the qualitatively and quantitatively component of the site and circumstances being analyzed to serve for mathematical evaluation of system behavior and performance" (Neuman and Wierenga, 2003). Therefore, a conceptual model is a representation of how the hydrogeological system or process operates or behaves and contains qualitative and subjective interpretations. As a consequence of this necessary simplification, uncertainty is inherent in any conceptual model.

Conceptual models are the basis of numerical groundwater flow and transport models. It is still common practice to consider only a single conceptual model. In the modelling process, the conceptual model is implicitly assumed to be given, accurate, and unique (Neuman and Wierenga, 2003). In general, once a single conceptual model of flow and transport is successfully calibrated against observed data, the conceptual model uncertainties are neglected and uncertainty analyses are performed considering only parameter uncertainty. Existing approaches for coping with conceptual model uncertainty are not generic and are not widely used in practice.

The analysis of conceptual model uncertainties has been receiving increasing attention in recent groundwater studies (e.g., Neuman, 2003; Neuman and Wierenga, 2003; Bredehoelft, 2003; Bredehoelft, 2005; Poeter and Anderson, 2005; Troldborg et al., 2007; Rojas et al., 2008a; Wöling and Vrugt, 2008). Researchers cite many forms of conceptual error, including incorrect hypotheses and neglected or inappropriate representation of

relevant processes due to a limited capability to exactly describe the complexity of a hydrological system. Many researchers acknowledge that the bias and uncertainty caused by an inadequate conceptualization are typically much larger than those caused by a suboptimal set of model parameter values (Carrera and Neuman, 1986a, 1986b; Neuman, 2003; Neuman and Wierenga, 2003; Bredehoelft, 2003; Bredehoelft, 2005; Poeter and Anderson, 2005). Lack of hydrological and hydrogeological data can also result in conceptual model uncertainty, leading to imprecise model components, including parameterization, boundary conditions, sources, and sinks. These model components in complex aquifer systems are usually conceptualized on the basis of several hypotheses and assumptions. Different, but reasonable, assumptions can lead to several plausible candidate conceptual models. Analyses of hydrologic systems based on a single conceptual model are therefore prone to statistical bias and underestimation of uncertainty. This flaw, in turn, can lead to errors in policy or design based on these models.

To avoid underestimation of prediction errors, analyzing predictions generated by several alternative models has been encouraged by many researchers (Carrera and Neuman, 1986a, 1986b; Bredehoeft, 2003; Neuman, 2003; Neuman and Wierenga, 2003; Ye et al., 2004; Bredehoelft, 2005; Hojberg and Refsgaard, 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006; Meyer et al., 2007; Refsgaard et al., 2007; Seifert et al., 2008; Rojas et al., 2008a, 2008b). There seems to be significant advantage to using multiple different models for predictive analysis, and considering their individual ability to fit the experimental data should provide important information about key hydrogeologic processes affecting flow and transport through aquifer systems of interest.

It is likely that successful decision making using groundwater management models requires more than a single “best” model. In addition, the uncertainties inherent in such tools should ideally be characterized, aggregated, and quantified. Therefore, this research aims to quantify various uncertainties associated with model development and prediction

by aggregating the results from multiple conceptual models. In addition, it hopes to offer insight into the utility of applying a set of alternative models to a real-world management dilemma in the Thraphra aquifer, Thailand. The proper application of model-averaging and uncertainty assessment techniques in groundwater management could help policy makers and stakeholders understand and implement the best among proposed policies to achieve groundwater management goals.

1.2 Overview of Research

The study addresses the critical question of whether the use of multiple conceptual models improves the ability of the models to assist in management decisions. To achieve this goal, a number of alternative groundwater models of the Thaphra area of the Chi River basin in Thailand have been constructed, calibrated, and compared using a proposed multi-criteria approach and published information criteria. They are assessed for predictive capabilities in a groundwater management optimization of sustainable yields. The efficacies of alternative model selection and aggregation methods are assessed with respect to their ability to improve model predictive capacity as compared to the use of individual models or aggregations of subsets of the alternative models.

1.2.1 Research Objectives

The research demonstrates an uncertainty assessment approach to explicitly account for conceptual model uncertainty in groundwater management applications. This study addresses the questions of whether such an approach improves the models' usefulness in management decisions, and how to use them to assist in decision making. To achieve this primary objective, a number of specific secondary objectives were defined:

- To propose and investigate effective methods to quantify conceptual model uncertainty in groundwater model applications.
- To examine the impact of different conceptual models upon prediction uncertainty.
- To quantify the contribution of conceptual model uncertainty and parameter uncertainty in groundwater management decisions using a real-world case study of the Chi River basin, NE Thailand.

In this thesis, assorted deterministic optimization models and model-averaging techniques were combined to generate ensemble predictions of groundwater management scenarios. Sustainable yield estimates for the Thaphra Area, northeast Thailand, were generated using multiple conceptual models, and then evaluated in order to illustrate this methodology. In order to achieve these goals, the following tasks were performed:

1. A set of twelve alternative conceptual/mathematical models for the real-world groundwater management site were developed. These plausible conceptual models were developed based on three different geological interpretations (hydraulic conductivity distributions), two different recharge estimations, and two different boundary conditions. Models were developed with MODFLOW2000 (Harbaugh et al., 2000).

2. All models were calibrated to a subset of measured data. The 12 different groundwater flow models were individually calibrated to an objective function that minimizes the mismatch between observed and predicted hydraulic heads, using PEST (Doherty, 2004) calibration techniques. Convergence criteria and other variables used by MODFLOW were the same to ensure that all models were calibrated to the same level, meaning that the same bases were used to evaluate all.

3. All models were assessed post-calibration in terms of their evaluation performance. In order to evaluate how well the models describe the system in question, the performance of each was tested by comparing their predictions with independent field data using a variety of model quality criteria, including information criteria, which are currently popular metrics for comparing conceptual models. In this step, the 12 alternative models and different model-averaging methods were compared in terms of their ability to improve the prediction error of the Thaphra aquifer system in year-2008 conditions. Forty water level measurements were used to compare the predictive quality of the best model determined using different criteria and the individual models.

4. The calibrated models were ranked and weighted based upon the various model quality criteria. A new multi-criteria approach was also proposed to evaluate the set of alternative models. The multi-criteria are the combination of quantitative and qualitative measures, which include the following performance metrics: travel time error; calibration absolute residual mean (ARM) for the observation wells located in the entire area; calibration ARM for the observation wells located in the critical area; evaluation ARM for the observation wells located in the entire area; and evaluation ARM for the observation wells located in the critical area. Moreover, the appropriateness and utility of different metrics were compared, and the relative usefulness of model-quality metrics in calibration, evaluation, and model selection/aggregation were judged.

5. The models were used to evaluate system-wide and local sustainable yields in the Thaphra aquifer, and thus form a basis for management decisions, by combining a simulation model with an optimization model. In the management step, a genetic algorithm (GA) was used to obtain the optimal extraction rates. Different model-averaging methods were used to establish ‘best estimates’ of sustainable yields from the multiple models. Information obtained from this

step was used to investigate whether using multiple conceptual models would actually aid in management.

6. The relative contribution of conceptual model uncertainty and parameter uncertainty was quantified for the estimation of sustainable yield from the best model and model averages.

1.3 Outline of the thesis

This thesis is organized as follows.

Chapter 2 (background) contains a brief overview of groundwater flow, sources of uncertainty in groundwater prediction, and methodologies for quantifying predictive uncertainty in groundwater modeling applications. This literature review surveys the different methods used for assessment of conceptual model and parameter uncertainty. Emphasis is placed on the multiple model-averaging methods that can account for uncertainties arising from how alternative conceptual models are defined.

Chapter 3 describes the development of the methodology to handle conceptual model uncertainty in groundwater modeling applications. The implementation of this method is illustrated using a three-dimensional groundwater flow model of the real-world aquifer system. The development and calibration of alternative conceptual models for the study area are presented in this chapter. Additionally, the predictive capability of these models is demonstrated, and model evaluation for these alternative conceptual models is presented. Much of this work closely follows that of Nettasana et al. (2012).

Chapter 4 presents the formulation of the optimization problem, including the objective function definition and decision variables with constraints for the real-world application

of groundwater management. This chapter demonstrates and compares different model-averaging methods for quantifying conceptual model uncertainty.

Chapter 5 presents the effects on groundwater management applications in the presence of parameter uncertainty. This chapter demonstrates a joint estimation of the predictive uncertainty, including parameter and conceptual model uncertainty. Additionally, uncertainties obtained from model-averaging approaches are compared with a parameter uncertainty analysis to quantify the contribution of conceptual model and parameter uncertainty parameter as applied to groundwater management models.

Chapter 6 briefly summarizes the main conclusions, recommendations for future research, and contributions to the scientific literature.

Chapter 2

Background

2.1 Groundwater Flow Modeling

Groundwater models are an important tool commonly used in the field of environmental science. Models have been applied to investigate a wide variety of hydrogeologic conditions and activities, such as assessing the water-supply potential of aquifers, and predicting the movement of contaminants through subsurface environments. In general, models are conceptual descriptions or approximations that describe physical systems or processes using mathematical equations. By mathematically presenting a representation of a hydrogeological system, reasonable alternative scenarios can be predicted, tested, and compared. The applicability or usefulness of a model depends upon how closely its mathematical equations and parameters approximate the physical system being modeled and how closely the conceptual model upon which the mathematical model is based reflects reality. In order to evaluate the applicability or usefulness of a model, it is necessary to understand the physical system and the assumptions embedded in the derivation of the mathematical equations. The assumptions typically involve the geometry of the aquifer, heterogeneity or anisotropy of sediments or bedrock within the

aquifer, hydraulic property values, distribution recharge, boundary conditions, contaminant transport mechanisms and any chemical reactions.

The governing equation describing the saturated three-dimensional movement of groundwater assuming constant density can be derived by combining Darcy's law with the mass conservation equation and can be written as

$$\frac{\partial}{\partial x}\left(K_x \frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y \frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_z \frac{\partial h}{\partial z}\right)+q_s = S_s \frac{\partial h}{\partial t} \qquad (2\text{-}1)$$

where h is the hydraulic head [L]; t is the time [T]; K_x, K_y, and K_z are the principal components of hydraulic conductivity along the x, y, and z coordinate axes [LT^{-1}]; q_s is the volumetric flow rate of fluid sinks/sources per unit volume of the aquifer [T^{-1}]; and, S_s is the specific storage [L^{-1}].

Equation (2-1) describes three-dimensional groundwater flow under non-equilibrium conditions in a heterogeneous and anisotropic saturated medium, provided the principal directions of hydraulic conductivity are aligned with the coordinate axes [McDonald and Harbaugh, 1988]. When combined with a particular set of boundary and initial conditions and known parameter values (e.g., $K_x(x,y,z)$), the groundwater flow equation constitutes a mathematical representation of a groundwater flow system. A solution to equation (2-1) provides the distribution of head throughout the flow system as a function of space and time, h (x, y, z, t).

For steady-state conditions, time dependent variations in groundwater flow are neglected ($\partial h/\partial t = 0$), and the right hand side of equation (2-1) cancels out.

Three types of boundary conditions can be specified for a given location and time along the boundary of the model domain: (1) specified-head boundaries (Dirichlet conditions)

for which head is given, (2) specified-flow boundaries (Neuman conditions) for which the derivative of head (flux) across the boundary is given, and (3) head-dependent flow boundaries (Cauchy or mixed boundary conditions) for which flux across the boundary is calculated given a boundary head value.

The most widely used numerical groundwater flow model code is MODFLOW, which is a three-dimensional model, originally developed by the U.S. Geological Survey (McDonald and Harbaugh, 1988). Ground-water flow within the aquifer is simulated in MODFLOW using a block-centered finite-difference in space, with a backward difference in time. Layers can be simulated as confined, unconfined, or a combination of both. Flows from external stresses such as flow to wells, areal recharge, evapotranspiration, flow to drains, and flow through riverbeds can also be simulated. The advantages of MODFLOW include numerous Graphical User Interfaces (GUIs) for data preparation, easy exchange of data in standard form, extended worldwide experience, continuous development, and the availability of its source code.

2.2 Calibration Methods

Most input parameters of groundwater flow and transport models are estimated from incomplete data due to the difficulties and costs of direct measurements. Therefore, the development of groundwater and transport models requires calibration or adjustment so that specific systems can be represented. Calibration is an iterative method that attempts to determine a set of model parameters, boundary conditions, and stresses that produce model-computed observations that match field-measured observations within a pre-established range of error (Anderson and Woesner, 1992; Bredehoeft, 2002; Doherty, 2004). This model should provide sufficient justification for reasonably good predictions. The calibration method is also called parameter estimation or the inverse modeling method.

The inverse methods in groundwater modeling are generally used to estimate parameters from observations of a system's response, such as hydraulic heads, flows, and concentrations. An inverse problem can be characterized by the method it uses to describe spatial variability (a parameterization approach), the forward equation it uses to relate parameters to measurements, the performance criterion it uses to define optimal parameter estimates, and the solution technique it uses to find the optimum value of parameters (McLaughlin and Townley, 1996).

An inverse problem can be solved manually or automatically. The manual trial-and-error calibration process is very time consuming and expensive, and it often produces a nonunique solution (Poeter and Hill, 1997; Anderson and Woesser, 1992; Yeh, 1986). In contrast, automated methods for calibration and model analysis are more effective. Automated methods can be classified as either deterministic or stochastic. Deterministic methods (often called gradient-based methods) are based on the theory of regression and classified into direct and indirect methods (Yeh, 1986). Parameters are treated as dependent variables in the deterministic direct method, but treated as independent variables in the indirect method. There are many advantages of the automated deterministic indirect method. For example, it is not labor-expensive or time consuming, and it provides a substantial number of statistics to qualify the calibration process. These statistics also help users to recognize and reduce different sources of uncertainty in modeling procedures (Poeter and Hill, 1997; Hill, 1998, Hill and Tiedeman, 2007).

In view of the complexity of real systems and shortage of available data sets, inversion is often used to arrive at working results; however, inversions are often troubled by the problems of non-uniqueness, insensitivity, and instability, regardless of how model calibration is achieved (Hill and Tiedement, 2007). Non-uniqueness, also referred to as equifinality in the literature (Beven and Freer, 2001), occurs when different combinations of parameter values equally match the observations. Insensitivity is said to exist when the

observations do not contain enough information to support the estimation parameters. Instability is present when slight changes in parameter values or observations lead to large changes in the simulated results. Measurement errors in the data can also degrade calibration efforts. These problems can be alleviated by adopting calibration methods that recognize these difficulties and addressing the difficulties in a quantitatively meaningful way.

Nonlinear regression is frequently applied in the calibration of groundwater models in order to improve model fit and to quantify the uncertainty inherent in parameter estimates. The optimization criterion in nonlinear regression is specifically to minimize the sum of squared differences between computed and observed values, typically of the hydraulic heads, flow rates, and concentrations (Hill, 1998, Hill and Tiedeman, 2007). Minimization is completed by a modified Gauss-Newton procedure that systematically updates parameter values in an iterative approach to locating minima on the Sum of Squared Error (SSE) surface. In this automated procedure, all parameters are adjusted simultaneously until the termination criteria are satisfied. UCODE_2005 (Poeter et al., 2005), PEST (Doherty, 2004), and MODFLOW-2000 (Hill et al., 2000) are examples of groundwater calibration software packages that implement the above procedure.

One of these codes, PEST (Doherty, 2004), is used in this research. The Gauss-Lavenberg-Marquardt nonlinear scheme is implemented in PEST. The Lavenberg-Marquardt algorithm is a gradient-based optimization strategy that combines the Gauss-Newton algorithm and the method of gradient descent. It provides a numerical solution to the mathematical problem of minimizing any sum of squared deviations between computed and observed values.

Groundwater model calibration problems are nonlinear parameter estimations (optimization problems). Since these problems involve fitting the parameters of a nonlinear function to a set of observed data points, nonlinear regression optimization

algorithms and theory can be utilized to improve model fit and to quantify the uncertainty inherent in parameter estimates. The optimization objective in nonlinear regression is specifically to minimize the sum of squared differences between computed and observed values, typically of the hydraulic heads, flow rates, and concentrations (Hill, 1998, Hill and Tiedeman, 2007). Although this least squares estimation problem can be tackled by any minimization algorithm, local gradient-based search with the Levenberg-Marquardt algorithm (Cooley, 1977; Carrera et al., 2005) has been shown to be a very efficient and effective technique (Cooley, 1985). The Levenberg-Marquardt algorithm is a gradient-based optimization strategy that combines the Gauss-Newton algorithm and the method of gradient descent. UCODE_2005 (Poeter et al., 2005), PEST (Doherty, 2004), and MODFLOW-2000 (Hill et al., 2000) are examples of groundwater calibration software packages that implement some version of the Levenberg-Marquardt algorithm.

The PEST software (Doherty, 2004), is used in this research to solve all model calibration problems. PEST implements a robust variant of the Gauss-Marquardt-Levenberg method of nonlinear parameter estimation. The optimization systematically updates parameter values in an iterative approach to locating minima on the Sum of Squared Error (SSE) surface. In this automated procedure, all parameters are adjusted simultaneously until the termination criteria are satisfied. Every time the optimization algorithm requires a new parameter set to be evaluated, the PEST software changes the model parameters and runs the simulation model. Parameter values are updated based largely on derivative information for the SSE surface.

PEST calculates all partial derivatives of observations with respect to parameters numerically using finite differences. Normally, the technique of derivative calculation is referred to as the method of “forward differences”. If the increment is either too large or too small the approximation will degrade the optimization performance. To battle such inaccuracy, PEST allows derivatives to be calculated using “central differences”. Using this method, two model runs are needed to calculate a set of observation derivatives

corresponding to any parameters. As it calculates derivatives, PEST records the sensitivity of each parameter with respect to the observation. Variables governing the operation of the Gauss-Marquardt-Levenberg method in determining the optimum upgrade vector can be adjusted prior to repeating the calculation. Hence one can interact with PEST, helping it in its determination of optimum parameter values in challenging situations. PEST has become the industry standard in calibration of various environmental problems (Moore and Doherty, 2006). Based on the Gauss-Marquardt-Levenberg method, the nonlinear parameter estimation algorithm used by PEST is uniquely robust and powerful, having been developed specifically for use with complex environmental models. Therefore, PEST is selected here to assist in data interpretation and in model calibration.

2.3 Sources of Uncertainty

Model predictions are uncertain due to incomplete knowledge of the underlying system, natural variability of the subsurface and field conditions. This incompleteness leads to sources of uncertainties, including those due to conceptual model uncertainty, parameter uncertainty, and scenario uncertainty (US.EPA, 1997; Meyer at el, 2007). Conceptual model uncertainty arises due to the necessary simplification of real world processes, the limitations of information needed to characterize the system, inappropriate representation of processes, and use of inappropriate surrogate variables. Examples of such uncertainty may arise from incorrect boundary conditions and zonation geometry, and inappropriate selection of forcing functions (Carrera and Neuman, 1986c). Parameter uncertainty arises from uncertainties and inadequacies in the model parameter values. The true values of model parameters, such as hydraulic conductivity and recharge, are never known exactly because of measurement errors, heterogeneity, and scaling issues (Gaganis and Smith, 2006). Scenario (or future) uncertainty arises from descriptive errors, and incomplete prediction of the future behavior of a system. Scenario uncertainty affects mainly forcing

terms. Scenario elements that may affect the hydrogeologic characteristics of a system include climate events (e.g. floods, changes in rainfall), change in engineered components, and human activities (e.g., changes in groundwater exploitation policies, land use activity).

Conceptual model uncertainty is considered to be the main source of uncertainty in groundwater model applications, and focusing solely on the optimization of model parameters may not compensate for conceptual model uncertainty (Carrera and Neuman, 1986a, 1986b; Bredehoeft, 2003; Neuman, 2003; Neuman and Wierenga, 2003; Ye et al., 2004; Bredehoelft, 2005; Hojberg and Refsgaard, 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006; Meyer et al., 2007; Refsgaard et al., 2007; Seifert et al., 2008; Rojas et al., 2008a, b).

2.4 Quantification of Uncertainty

It is difficult to separate and evaluate conceptual model uncertainty and parameter uncertainty. The most common approach to assessing uncertainty in hydrogeologic modeling is to assume that the conceptual model is appropriate and address parameter uncertainty only (Meyer et al., 2007). Alternatively, when conceptual model uncertainty is assessed, the parameter values are often assumed to be known perfectly (Meyer et al., 2007). In actual complex systems, both the true model and true parameter values are unknown; thus accuracy in quantifying either model error or parameter error is a function of the magnitude of the other.

2.5 Quantification of Parameter Uncertainty

The primary factors that contribute to hydrogeologic uncertainty are the unknown spatial heterogeneity of the subsurface (Meyer et al, 2007) and the unknown spatial and temporal distribution of inputs such as recharge. Values of hydraulic conductivity, K, and transport properties of the subsurface may vary over several orders of magnitude in the same small area due to a large variation in particle and fracture sizes of hydrogeologic units. Detailed sampling to determine the exact nature of the subsurface is impossible. Interpolation between measurements and other indirect methods is required to estimate properties at unmeasured locations.

The analysis of parameter uncertainty has received much attention in the literature. It is common practice, when a model is defined, to quantify the errors associated with estimating its parameters (Carrera and Neuman, 1986 a, b; Rubin and Dagan, 1987; McLaughlin and Townley, 1996; Poeter and Hill, 1997; Hill and Tiedeman, 2007). Techniques for addressing the uncertainty in model parameters include mathematics or sensitivity analysis (Morgan and Henrion, 1990, Dou et al., 1995, and Meyer et al., 2007), fuzzy set theory (Dou et al., 1995), and probabilistic or stochastic approach (USEPA, 1997).

Sensitivity analysis provides a framework for dealing with parameter uncertainties of unknown structure. In this analysis, a base simulation using the best available estimates of model input parameters is first conducted. During each subsequent simulation, one model parameter is varied by a certain interval from the values used in the base case, while other parameters are kept constant. The result of each simulation is compared with that of the base case to calculate a sensitivity coefficient for the varied parameter. In addition to the overall uncertainties in the model outcome, there are also the uncertainties in given parameters whose influence increases as the sensitivity coefficient of these

parameter increases. Although the sensitivity method is simple, flexible, and versatile, its result may not be indicative of the output range, since the model output is not a simple linear function of the parameters. The use of the upper and lower limits for all parameter values does not guarantee that these limits of the solution will be generated, given the non-linear nature of the problem and correlations between uncertain parameters.

Alternative approaches have modified the search method for finding the best-and-worst case scenarios without using interval mathematics (Brooks et al., 1994). Although the best-and-worst case scenarios provide valuable information, the distribution of possible outputs near the parameter limits can have a strong influence on the tails of output distributions, which play an important role in decision-making scenarios. The impreciseness of many inputs to groundwater flow and contaminant transport models have led to formulations of fuzzy groundwater flow models (Dou et al., 1995), which can be useful in decision-making scenarios.

The most commonly used approach to uncertainty analysis in groundwater flow and contaminant transport modeling is probabilistic analysis, which is designed to generate the nature of output uncertainty. Probabilistic uncertainty analysis can take the form of analytic methods based on stochastic partial differential equations (Dou et al., 1995) or sampling-based methods such as Monte Carlo analysis (USEPA, 1997). Analytical methods using linear and non-linear first-order approximations of 95% confidence intervals that can be attained with gradient-based parameter calibration methods are commonly used to evaluate the optimized parameters (Vecchia and Cooley, 1987; Christensen and Cooley, 1999). However, the method's reliability is assured only with well-posed problems with known error distributions. In contrast, probabilistic analysis using sampling-based methods such as Monte Carlo analysis can provide practical output probability distributions with reasonable intervals even for ill-posed problems. Monte Carlo simulation is the most commonly applied approach used in parameter uncertainty assessment (USEPA, 1997; Carrera et al., 2005), and it can be used as a benchmark for

validating other types of uncertainty estimation methods (Christensen and Cooley, 1999; James and Oldenberg, 1997).

The results produced by Monte Carlo analysis are considered to be probabilistic approximations, which strongly depend on input posterior parameter distributions. Stochastic input parameter distributions are generally selected based upon the characteristics of the globally optimum parameter set. According to this method, random values of stochastic parameters are generated following their respective probabilistic characteristics. Each set of the generated input parameters is employed to compute the equivalent model output. The probability distribution of the simulated model output are then determined through a statistical analysis.

One of the greatest concerns when using the Monte Carlo method is in specifying a suitable probability density function (pdf) for each uncertain parameter. The available field data are rarely adequate for this purpose, and in effect, the pdf is usually postulated on the basis of a few scattered data points. Thus, the uncertainty expressed in this pdf may result in large uncertainty in the model output. A second problem in Monte Carlo analysis is the intensive computational demand. A typical Monte Carlo analysis may require several hundred or thousand runs for the output pdf to be statistically meaningful. Nevertheless, the computational effort required is linear to the number of the uncertain parameters, rather than exponential, as it is in the analytical methods. Monte Carlo assessment remains the most attractive approach because of its conceptual simplicity, its general applicability, and its ability to quantify fully the uncertainty in the model output.

James and Oldenburg (1997) used the first-order second-moment (FOSM) approach to investigate the uncertainty of simulated trichloroethylene concentrations due to parameter uncertainty and variation in conceptual models. TCE concentrations at a site of potential human exposure occur and can be attributed to uncertainty in permeability, porosity, diffusivity, chemical solubility, and adsorption within a single conceptual model. For the

examples considered, the linear FOSM analysis generally captures the uncertainty range calculated by the Monte Carlo method. James and Oldenburg (1997) calibrated a number of conceptual mathematical models to available observations and retained only those calibrated models that had historically reproduced observations adequately. The authors produced a prediction using each calibrated model, assessed the corresponding predictive uncertainty due to uncertainty in the model parameters, and averaged the predictions as well as their ranges of uncertainty by assigning an equal weight to the results of each model. Significant uncertainties in simulated calculations also show that significant output uncertainty is introduced by conceptual model variation.

Beven and Binley (1992) proposed a methodology for calibration and uncertainty estimation of distributed hydrologic models based on the method referred to as GLUE (generalized likelihood uncertainty estimation). The strategy picks up the identification of multiple alternative structural models and the assumption of a prior probabilistic model of parameter uncertainty for each. Each structural model, combined with its corresponding parameter uncertainty model, is used to generate Monte Carlo realizations of simulated hydrologic behaviors and to compare the results with observations made during the same period. Likelihood measures are defined to estimate the degree of correspondence between simulated and observed data. If a likelihood measure falls below the acceptance criterion, the model realizations are discarded. The combination of the model structure and parameter set that both pass this test is retained to make predictions of system responses under selected future scenarios. Each prediction is weighted by a corresponding normalized likelihood measure, to produce a likelihood-weighted cumulative distribution of all available predictions.

Kunstmann et al. (2002) used conditional and unconditional first-order second-moment (FOSM) analysis to quantify the uncertainty in groundwater flow and solute transport equations. A second analysis was performed to derive the covariance matrices for heads and concentrations from the relevant governing equations. The analysis was conducted

for steady-state conditions and considered hydraulic conductivity and recharge uncertainty. The objective was to quantify the exploitation potential of an aquifer in terms of its mean annual recharge and its uncertainty bound. Kunstmann et al. (2002) reported that the uncertainty bounds obtained by the FOSM method correspond well with the results obtained by Monte Carlo analysis. However, the FOSM method is much more advantageous because of its greater computational efficiency.

2.6 Quantification of Conceptual Model Uncertainty

Conceptual models have many uncertainties due to both the scarcity of data and subjectivity of many modeling decisions. Modelers are forced to make simplistic assumptions of reality. Model errors are introduced in, for example, the parameterization, discretization, parameter zonation and boundary conditions selected. Uncertainties in the conceptual model have been recognized as a main source of uncertainty in model prediction (Usunoff et al., 1992; Neuman and Wierenga, 2003; Hojberg and Refsgaard, 2005). However, the existing approaches for coping with conceptual model error are not generic and are not widely used in practice. In general, the conceptual model uncertainties are neglected, and uncertainty analyses are performed considering only parameter uncertainty and using only a single conceptual model.

A familiar approach to assessing conceptual model uncertainty includes validation or post-audit tests. In the post-audit, the model predictions are compared with existing data that have been omitted from the model calibration or new data. The post-audit tests are most effective when the new data represents stress conditions or aspects of the system that differ from those represented in the model calibration data (Hill and Tiedeman, 2007). The disadvantage of this method is that the model structure quality can only be assessed against the output variables for which the field data are available.

Another approach in dealing with conceptual model error is to select the most plausible models from a set of realistic options, an approach known as Model Selection. The basic idea of this method is to minimize selected model error rather than quantify individual model errors. Different models can be compared in terms of model fit, residual distribution, and confidence intervals for parameters and predictions. The model selection approach is advocated by many researchers (Burnham and Anderson, 2002; Neuman and Wierenga, 2003; Poeter and Anderson, 2005; Ye et al., 2004, 2005, 2008a, 2008b), but has limitations in that the scenarios of alternative conceptual models cannot be fully known and cannot ensure that the multiple models adequately sample the complete range of plausible models.

Refsgaard et al. (2006) proposed a new framework for analyzing the uncertainties due to model structure errors when models are used for making extrapolations beyond their calibration period. In this framework, a suite of conceptual models is independently calibrated and a pedigree analysis is performed to assess the overall reliability of the multiple models. Pedigree conveys an evaluative account of the production process of information, and indicates different aspects of the underpinning and scientific status of the knowledge used. A pedigree is expressed by means of a set of pedigree criteria used to assess these different aspects. The pedigree approach integrates different types of knowledge, such as qualitative and quantitative uncertainty, to estimate the impact of model structure uncertainty on model predictions. The pedigree analysis does not give an indication of the relative quality of the various model structures and, consequently, it cannot be integrated in a quantitative uncertainty analysis in terms of model probabilities. Nevertheless, the statements obtained from pedigree analysis should be available as the best possible scientifically based characterization of uncertainties and, as such, should be made available to those involved in the decision making process. However, the full framework needs to be tested in real water resource management case studies.

When the set of model alternatives is defined, the options for addressing conceptual model uncertainty include: 1) evaluating each alternative and selecting the best model,

based on specific criteria or prediction results, or 2) evaluating each alternative and combining the results using appropriate weighting schemes (Meyer et al., 2007). The model selection methods may be carried out through an informal comparison (James and Oldenburg, 1997; Hojberg and Refsgaard, 2005) or through evaluation of formal model selection criteria such as AIC (Akaike Information Criterion), AIC_c (second-order-bias-corrected AIC), BIC (Bayesian Information Criterion), and KIC (Kashyap Information Criterion) (Burnham and Anderson, 2002; Poeter and Anderson, 2005; Neuman and Wirenga, 2003). Combining the results using weighting schemes may be carried out through the likelihood-based weighting of Beven and Freer (2001), the model likelihood weighting of Burnham and Anderson (2002, 2004), and the model probability weighting of Draper (1995). The details of alternative model selection are described in a later section.

Analyzing conceptual model (model structure) uncertainties has been receiving increasing attention in recent groundwater studies. Incorrect hypotheses, neglect of relevant processes and inappropriate representation of processes introduce errors into a conceptual model. Many researchers acknowledge that the bias and uncertainty caused by an inadequate conceptualization are typically much larger than those caused by an inadequate suboptimal set of model parameter values, and focusing only on the optimization of model parameters does not allow compensation for conceptual model uncertainty (Carrera and Neuman, 1986a, 1986b; Bredehoeft, 2003; Neuman, 2003; Neuman and Wierenga, 2003; Ye et al., 2004; Bredehoelft, 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006; Meyer et al., 2007; Refsgaard et al., 2007; Seifert et al., 2008; Rojas et al., 2008a). Lack of hydrological and hydrogeological data result in conceptual model uncertainty because of imprecise model components, including parameterization, boundary conditions, sources, and sinks. These model components in complex aquifer systems are usually conceptualized on the basis of a number of hypotheses and assumptions. Several assumptions can lead to several plausible candidates for conceptual models. Model predictions based on one selected model are

subject to significant uncertainty. Analyses of hydrologic systems based on a single conceptual model are therefore prone to statistical bias and underestimation of uncertainty. To avoid prediction errors underestimation, jointly analyzing predictions generated by several alternative models has been encouraged by many researchers (Carrera and Neuman, 1986a, 1986b; Bredehoeft, 2003; Neuman, 2003; Neuman and Wierenga, 2003; Ye et al., 2004; Bredehoelft, 2005; Hojberg and Refsgaard, 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006; Meyer et al., 2007; Refsgaard et al., 2007; Seifert et al., 2008; Rojas et al., 2008a) and is the focus of this study.

Approaches for dealing with multiple models to analyze uncertainty in prediction can be divided into two categories. The first approach is to evaluate each alternative and select the best model, based on certain criteria or prediction results. Predictions are made on the assumption that the best model is the true model and uncertainty exists solely in the estimation of parameters. This approach neglects uncertainty in the choice of models, uncertainty that may be important, especially if several models have similar criterion scores but provide significant differences in predictions. A second approach is to evaluate each alternative and combine predictions using model-averaging or weighting schemes (Draper, 1995; Hoeting, 1999). The fundamental goal of model-averaging is to arrive at a set of plausible models that can be weighted according to criteria, usually associated with their likelihood given a set of data. This weighting reflects the degree to which each model is trusted. Predictions are then derived as a weighting average of the predictions from each model in the plausible set, and prediction uncertainty is estimated as a weighted sum of the within- and between-model variance.

2.7 Model averaging Approaches

Model-averaging has gradually received increased attention as an alternative to model selection in quantifying conceptual model uncertainty. When multiple plausible models are implemented, each model has its own strength and weakness in prediction. Thus, model averaging may provide a better predictive capability than relying on a single model.

Model averaging can be defined as a method of combining results from several plausible models into a single set of predictions by a measure of model performance. In groundwater modeling applications, several methods have been proposed for dealing with model averaging, these methods including Generalized Likelihood Uncertainty Estimation (GLUE), Bayesian Model-averaging (BMA), Qualitative approaches, and Multi Criteria Decision Making (MCDM) approaches. The methodologies used to quantify uncertainty associated with model averaging are described as follows.

2.7.1 Generalized Likelihood Uncertainty Estimation (GLUE)

GLUE was developed by Beven and Binley (1992) to deal with model non-uniqueness in environmental modeling. GLUE is based on the concept of "equifinality," that is, many sets of parameters in combination with many conceptual models will produce equally likely modeling results when compared to limited observed system responses (Beven and Freer, 2001; Beven, 2006). This concept rejects the idea of a single correct representation of a system in favor of many acceptable system representations. In the GLUE methodology, the feasible parameter space within a single model structure is first sampled by Monte Carlo simulation to produce many equally likely parameter combinations (realizations). The output corresponding to each realization (or model

alternative) is evaluated with respect to any observed system responses using the chosen likelihood measure or measures to reflect the performance of individual models in reproducing the behavior of the system being studied. Only those realizations (or models) that satisfy the likelihood measure (e.g., a maximum sum-of-squared weighted residuals), also known as the behavioural threshold, are retained for further analysis, and the non-behavioural realizations (models) are rejected. A posterior likelihood for each model is then calculated as a function of the error between the observed outputs and those predicted by the model. The weights (or probabilities) for each model are estimated by normalizing the likelihoods. The GLUE concept can be applied to the evaluation of multiple model structure or alternative conceptual models. Discrete alternatives can be considered as well as alternative parameter sets provided that the likelihood measures are used in the evaluation of each model.

The fundamental features of GLUE is its flexibility with respect to likelihood measure choice. Various likelihood measures based on goodness-of-fit metrics have been proposed. One likelihood measure commonly used to measure the closeness between model predictions and observations in the GLUE literature is given by the inverse weighted variance. The likelihood measure can be defined according to Singh et al. (2010) as

$$L_k = \sum_l \left(\frac{\sigma_l^2}{\sigma_{e,k}^2}\bigg|_l \right)^N \qquad (2\text{-}2)$$

where L_k is the likelihood measure for model k; l is the number of state variables (data types); $\sigma^2_{e,k}$ is the variance of the residuals for model k; σ_i^2 is the variance of the observations of the data type l; and N is a shape factor chosen by the user such that values of $N >> 1$ tend to give higher weights (likelihoods) to models that better agree with the data, Values of $N = 0$ tend to make every model have equal likelihood, and $N=1$ is most often used in GLUE applications. The variance of the errors $\sigma^2_{e,k}$ for data type l is given by

$$\sigma_{e,k}^2\Big|_l = \frac{SSR}{n}\Big|_l \; or = \frac{SSWR}{n}\Big|_l \qquad (2\text{-}3)$$

where *SSR* is the sum-of-squared residuals and *SSWR* is the sum-of-squared weighted residuals for the k^{th} model predictions and observations (of data type *l*), while *n* is the number of observations (for data type *l*).

Another form of the likelihood measure based on the residual variance is the proportion of the observed variance explained by the model called the Nash-Sutcliffe efficiency index (Nash and Sutcliffe, 1970) given by:

$$L_k = \sum_l \left[1 - \frac{\sigma_{e,k}^2}{\sigma_l^2}\Big|_l \right]^N \qquad (2\text{-}4)$$

A further likelihood function of the residual variance is the exponential likelihood function (Beven, 2000):

$$L_k = \sum_l \exp\left[-N \frac{\sigma_{e,k}^2}{\sigma_l^2}\Big|_l \right] \qquad (2\text{-}5)$$

The likelihoods are normalized so that their sum is equal to one, gives the GLUE weight for model *j*:

$$\omega_k(GLUE) = \frac{Pr_k L_k}{\sum_{i=1}^{n} Pr_k L_k} \qquad (2\text{-}6)$$

where L_k is one of the likelihood functions described above, Pr_k is the prior weight given to each model (typically based on the modellers' expert judgment), and *n* is the total number of models being considered.

The GLUE method has been used in a wide variety of applications. However, this approach has been criticized for not being formally Bayesian, resulting in parameter and

predictive distributions that are statistically incoherent, unreliable, and should not be used (Montari, 2005, Mantovan and Todini, 2006; Vogel et al., 2008). The concerning aspects of its methodology involve (1) the lack of theory for deciding an appropriate likelihood function and the rejection criterion used to distinguish between behavioural and non-behavioural models; (2) the lack of a statistical basis for the likelihood and threshold measures used for model selection and weighting; (3) the computational burden required due to the need for extensive Monte Carlo simulations; and (4) the fact that GLUE does not require the model structure and parameters to be optimized, which could lead to overestimation of predictive uncertainty.

Beven (2006) has argued that the formal Bayesian model-averaging (BMA) approach is a special case of GLUE and is applicable under certain assumptions, and model selection can be used within the GLUE methodology to reduce uncertainty. Mugunthan and Shoemaker (2006) have shown that optimization or model selection can be used to generate alternative models for GLUE, leading to efficiency improvements for the GLUE framework by reducing the need for Monte Carlo simulations to generate model alternatives. Regarding the debate between the GLUE and Bayesian methods, Beven (2009) further argues that "the best approach to estimating model uncertainties is a Bayesian statistical approach, but that will only be the case if all the assumptions associated with the error model can be justified" and that "simple assumptions about the error term may be difficult to justify as more than convenient approximations to the real nature of the errors." He ended by cautioning that "making convenient formal Bayesian assumptions may certainly result in over estimating the real information content of the data in conditioning the model space."

2.7.2 Bayesian Model-averaging (BMA)

Bayesian Model-averaging (BMA) techniques were suggested by Draper (1995), Kass and Raftery (1995), and Hoeting et al. (1999) and are based on a formal Bayesian formulation for the posterior probabilities of different conceptual models. BMA is achieved by estimating posterior model probabilities (PMPs) for each model in addition to posterior density of the parameters. PMPs are the non-negative scalar values that sum to one and are the relative probability of the model being true, given the data. The ratio of PMP values for any two models gives the relative support of those models and is known as the Bayes factor. Estimation of PMPs requires specification of prior model probabilities. These priors encapsulate the prior belief that the model is the true model. The prior model probabilities for the models are typically obtained using expert elicitation (Ye et al. 2005, 2008b) or given equal weight based on a non-informative prior. Model averaged or ensemble predictions can be estimated using the PMPs to weight individual model predictions based on their prediction performance. Predictions from better-performing models receive higher weights than those of poorer-performing models. Therefore, BMA avoids the need to choose one model over the others; instead, competing models are assigned different weights based on the observed dataset (Wasserman, 2000).

While the parameter posterior density is estimated using the model likelihood, estimation of PMPs requires estimation of the integrated model likelihood (also called the marginal model likelihood), which is the likelihood multiplied by the prior integrated over all possible parameter values. Model PMPs are then estimated using the individual model marginal likelihood divided by the sum of marginal likelihoods over the set of models under consideration.

In the Bayesian framework, if Δ is the quantity of interest predicted by a set of K alternative models, then its distribution conditioned on dataset D is calculated using Bayes' rule as follows (Hoeting et al., 1999):

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|M_k, D)\, p(M_k|D) \qquad (2\text{-}7)$$

where $p(\Delta|M_k,D)$ is the predictive probability of Δ for model M_k, and $p(M_k,D)$ is posterior model probability of M_k and can be computed using Baye's theorem as

$$p(M_k|D) = \frac{p(D|M_k)\, p(M_k)}{\sum_{l=1}^{K} p(D|M_l)\, p(M_l)} \qquad (2\text{-}8)$$

where $p(M_k)$ is the prior probability of model M_k (similar to Pr_k used in Equation (2-6) for GLUE), and $p(D|M_k)$ is the likelihood of model M_k (a measure of consistency between model predictions and site observations *D*). This model likelihood is given by

$$p(D|M_k) = \int p(D|\theta_k, M_k) p(\theta_k|M_k) d\theta_k \qquad (2\text{-}9)$$

Here, θ_k is the parameter set associated with model *k*; $p(\theta|M_k)$ is the prior probability of the parameters, and $p(D|\theta_k,M_k)$ is the joint probability of model *k* and is a function of the errors with respect to the field data (*D*). The prior probabilities for the parameters, $p(\theta|M_k)$, can either be a subjective value or based on an expert's prior information and philosophical beliefs. The posterior model probabilities are consequently rather sensitive to the specification of the prior probabilities, or that prior probabilities should not dominate the likelihood function as supported by data. If there is no informative support, it is reasonable to assign equal prior probability as a neutral choice (Hoeting et al., 1999; Wasserman, 2000). In BMA, the prior model probabilities also represent a prior model weight equal to one.

In general, the first two moments of Δ are used to quantify the uncertainty. For model M_k, parameter uncertainty is quantified by the mean, $E[\Delta|D,M_k]$, and variance, $Var\ [\Delta|D,M_k]$, which can be obtained using either Monte Carlo simulation or stochastic methods. The posterior mean and variance quantify both the parametric and model uncertainty. The first term and second terms on the right hand side of Equation 2-11 are the within-and between- model variance, respectively.

$$E\left[\Delta\middle|D\right]=\sum_{k=1}^{K}E\left[\Delta\middle|D,M_k\right]p\left(M_k\middle|D\right) \tag{2-10}$$

$$Var\left[\Delta\middle|D\right]=\sum_{k=1}^{K}Var\left[\Delta\middle|D,M_k\right]p\left(M_k\middle|D\right) + \sum_{k=1}^{K}\left(E\left[\Delta\middle|D,M_k\right]-E\left[\Delta\middle|D\right]\right)^2 p\left(M_k\middle|D\right) \tag{2-11}$$

In groundwater modeling, the most commonly used BMA concepts are the Maximum Likelihood Bayesian Model-averaging (MLBMA) method (Neuman 2003) and the Information-theory based method (Poetor and Anderson, 2005). Both methods are based on the use of model selection criteria, which are derived as by-products of the calibration of groundwater models using methods such as Maximum Likelihood (ML) or Weighted Least Squares (WLS). The use of model selection criteria includes ranking alternative conceptual models, eliminating some of them, or weighting and averaging model predictions through the calculation of posterior model weights using a generic formula. The most commonly used model selection criteria are the Akaike Information Criterion (AIC) (Akaike, 1974), modified Akaike Information Criterion (AICc) (Hurvich and Tsai, 1989), Bayesian Information Criterion (BIC) (Schwartz, 1978), and Kashyap Information Criterion (KIC). KIC and BIC are the suggested criteria in MLBMA, while AIC, AICc are the preferred criteria for the information-theory based method of Poetor and Anderson (2005).

2.7.2.1 Maximum Likelihood Bayesian Model-averaging (MLBMA)

Neuman (2003) proposed a variant of the BMA approach called MLBMA. MLBMA approximates the integral in Equation 2-9 by using likelihood measures such as the Kashyap Information Criterion (KIC) or the Bayesian Information Criterion (BIC), which are evaluated for each model calibrated to the maximum likelihood estimator for the parameter set.

The MLBMA assemble models that have been calibrated to observed data using maximum likelihood estimation. The model likelihood is then estimated using

$$p(D|M_k) \; \alpha \; \exp(-\frac{\Delta_k}{2}) \qquad (2\text{-}12)$$

where Δ_k is the difference between the BIC or KIC measure for the k^{th} model and the minimum BIC or KIC value among all competing models. Assuming a multi-Gaussian error distribution with unknown mean and variance for the model likelihood in Equation 2-12, the BIC and KIC terms can be written as:

$$BIC_k = n \ln (\sigma_k^2) + p_k \ln (n) \qquad (2\text{-}13)$$

and

$$KIC_k = (n - p_k) \ln (\hat{\sigma}_{e,k}^2) - 2 \ln p(\hat{\theta}_k) - p_k \ln(2\pi) + \ln \left| X_k^T \omega X_k \right| \qquad (2\text{-}14)$$

where n is the number of observations, p_k is the number of parameters for model k, $\hat{\theta}_k$ is the maximum likelihood estimator for the parameters from model k, $p(\hat{\theta}_k)$ is the prior probability (either assessed from field data or through expert elicitation) for the parameter estimate, and $\hat{\sigma}_{e,k}^2$ is the maximum likelihood estimator for the variance of the error

residuals (e) estimated from the weighted sum-of-squares residuals for model j with the maximum likelihood estimator for the parameters as

$$\hat{\theta}_{e,k}^2 = \left. \frac{e_j^T \omega e_k}{n} \right|_{\theta_k = \hat{\theta}_k} \tag{2-15}$$

where e_k is the calibration error vector, n is the number of samples, $\hat{\theta}_k$ is the maximum likelihood estimator for the parameters, and ω is a weight factor, which theoretically is given by the covariance between the data points. It is common to assume uncorrelated data leading to a diagonal matrix with the variance of the data points along the diagonal. In many cases, the unbiased "least-square" formulation may be used, where, instead of n, (n-p_k) is used in the denominator, with p_k being the number of calibrated parameters in the model k. Also note that for the purpose of simplicity and without loss of generality, we have assumed only a single data type (unlike the GLUE formulations presented in Equations 2-2 to 2-6, which were for multiple data types).

The MLBMA model weights can be calculated by

$$w_k\left(MLBMA\right) = \frac{\exp(-0.5\Delta_k)\, p(M_k)}{\sum_k \exp(-0.5\Delta_k)\, p(M_k)} \tag{2-16}$$

where $p(M_k)$ represents prior probabilities of the models.

Singh et al. (2010) have pointed out the key aspects of the KIC- and BIC-based model weights; (1) the use of the k term, which can vary from 0 (for the model with the minimum KIC or BIC metric) to many orders of magnitude higher (for the models with higher KIC and BIC metrics) and (2) the exponential weighting in Equation 2-16 that tends to assign most of the posterior weights to the relatively few models exhibiting slightly better agreement with the data. The distribution of weights becomes narrower as the number of observations increases, since the value of n linearly affects the minimal

values of BIC and KIC. This statement is reasonable because with more data there needs to be less uncertainty amongst competing models (Poeter and Hill, 2007). However, Beven (2009) has pointed out that this is only desirable if the error structure assumed by the averaging technique is consistent with the "real" error structure. Otherwise, model-averaging techniques such as MLBMA may overestimate the information content of the data while conditioning the model.

Domingos (2000) has compared BMA with other model-averaging techniques and shown that BMA tends to underestimate the predictive uncertainty. However, others such as Minka (2000) have contended that these results are hardly surprising because, by definition, techniques like BMA, and especially MLBMA, are built on the intrinsic assumption that there is *a unique* model of reality (i.e., there is only one mode in the conditional distribution—representing the most likely model). This is confirmed in the original MLBMA paper by Neuman (2003), where he presents the fundamental assumption for this technique so:—"only one of the (alternative) models is correct even in the event that some produce similar predictions for a given set of data." Thus, *MLBMA is more a model selection technique than a model-averaging methodology*. Model selection (or ranking) is simply based on the relative magnitude of the BMA criterion (either BIC or KIC), and thus is not affected by the exponential dependence on *n*.

The formulations shown earlier require the models to be well calibrated (normally distributed errors, etc.) and the residual variance ($\grave{\sigma}^2_{e,k}$) assessed using the calibrated parameters. In fact, the error distribution used is typically unimodal, with the mode approximated by the "calibrated" model. In the case of highly parameterized models, there is bound to be non-uniqueness in the parameter domain (and thus multimodality in the calibration response surface). The applicability of MLBMA and BMA in such cases is not clear. In such cases, it is advisable that the dimensionality of the model parameters be reduced (thereby introducing some level of uniqueness in the calibrated parameter set) before applying this methodology.

2.7.2.2 Information- Theory-based Model-averaging

Information theory assesses the relative model performances as the likelihood of a model and can be assumed to be related to the value of "information" they provide. The familiar information-theory-based model averaging in use is the AIC. Although there are fundamental differences between the Akaike Information Criterion-based model-averaging (AICMA) and the Bayesian approach are fundamental differences, the two frameworks work analogous. The AIC is used to approximate the Kullback-Leibler (K-L) metric, a measure of the loss of information when an imperfect model (M_k) is used to approximate the "real" (and unknown model *f*). The K-L distance (*I*) between model M_k and *f* is defined as (Singh et al., 2010)

$$I\ [f, M_k] = \int f(x) \log\left(\frac{f(x)}{p(M_k|\theta_k)}\right) dx \qquad (2\text{-}17)$$

where $f(x)$ is the real distribution and $p(M_k|\theta_k)$ is the distribution of model M_k given the set of calibrated parameters θ_k. However, for groundwater models, K-L information cannot be computed since the real distribution *f* (e.g., hydraulic conductivity values, boundary conditions, and fluxes) is not known. The relative K-L information can be approximated using the AIC (Akaike, 1973) given by

$$AIC_k = n \ln(\hat{\sigma}^2_{e,k}) + 2p \qquad (2\text{-}18)$$

To further correct for the bias introduced from small sample sizes, a modified AIC equation (Hurvich and Tsai 1989; Poeter and Anderson 2005) has been proposed as follows:

$$AICc_k = n \ln(\hat{\sigma}^2_{e,k}) + 2k + \frac{2p(p+1)}{n-p-1} \qquad (2\text{-}19)$$

where the extra term in Equation 2-19 as compared to Equation 2-18 accounts for second-order bias that may result from a limited number of observations, for example, when *n/p* < 40. This work uses the AICc metric as defined in Equation 2-19 for likelihood estimation.

The AICMA model weights can be written as

$$w_k\left(AICMA\right) = \frac{\exp(-0.5\Delta AIC_k)\,p(M_k)}{\sum_{k=1}^{K}\exp(-0.5\Delta AIC_k)\,p(M_k)} \qquad (2\text{-}20)$$

The AICMA is based on an information theory framework; it assumes that *all* models are approximations and it is impossible to perfectly capture reality. The goal for AICMA therefore is to select models with increasing complexity as the number of observations increases; however, the goal for MLBMA is to seek models with consistent complexity (i.e., constant *k*), regardless of the number of observations (since the penalty term for model complexity is not dependent on the number of observations). Use of the FI matrix in the KIC calculation yields lower probabilities for more complex models, if such complexity is not supported by the data, by which alleviating some of the problems with the consistent complexity assumption.

The AICMA approach shares some of the behavior, in terms of posterior weight distribution, of MLBMA due to the use of the Δ term and exponential weighting in Equation 2.20, use that results in larger weights being given to models that show optimal or near optimal error residuals. The definition of AICc (such as that of KIC and BIC) exhibits a linear dependence on *n*, which implies that the AICc weights are proportional to $(1/\hat{\sigma}^2_{e,k})^n$; whereas the GLUE weights are proportional to $(1/\hat{\sigma}^2_{e,k})$. The weights are the major source of difference in inferring posterior model probabilities among GLUE and MLBMA or AICMA.

2.7.3 Variance Window-Based MLBMA

The previous section highlighted the issue with MLBMA and AIVMA distributing most of the model weights to a few models that showed marginally better calibration performance. Tsai and Li (2008) have proposed an approach to address this issue by using the concept of a "variance window" to modify the MLBMA scheme. The motivation for their work was the realization that BMA tended to assign most of the weights to a few models that exhibit marginally better calibration performance (due to exponential weighting and the *k* term used in Equation 2-15). Tsai and Li (2008) contended that this stringency in the model-averaging criteria is a result of the underlying assumption of "Occam's windows" (Madigan and Raftery, 1994) that only accepts models in a very narrow performance range. Occam's window is defined by Raftery (1995) as the range within which the model performance of two competing models is statistically indistinguishable—that is, if the difference between the calibration metrics of two models (with the same complexity) is less than the Occam's window, then both will be accepted.

Raftery (1995) pointed out that for sample sizes between 30 and 50 data points, an Occam's window of 6 units in the BIC metric (BIC in Equation 2-11) roughly corresponded to a significance level of 5% (in *t* statistics) in conventional hypothesis testing terms. Over the years there has been growing realization that this Occam's window for model acceptance may be too restrictive, leading to biased results (Hoeting et al., 1999; Tsai and Li, 2008). To reduce this overweighting and the resulting bias, Tsai and Li (2008) introduce the concept of a "variance window" as an alternative to an Occam's window for selection with the BMA. The variance window is determined by including a scaling factor α with BIC (and KIC), where α is given by

$$\alpha = \frac{s1}{s2\sigma_D} \qquad (2\text{-}21)$$

where σ_D is the standard deviation of the error chi-square distribution for the "goodness-of-fit" criterion used in formulating KIC or BIC (see Tsai and Li (2008) for details). The variance of the chi-square distribution is given by $2n$ (i.e., $\sigma_D = \sqrt{2n}$), where n is the number of observations, $s1$ is the size of the Occam's window corresponding to the given significance level, and $s2$ is the width of the variance window in the unit of σ_D. As the width of the variance window becomes larger, α becomes progressively smaller than 1. Note that since the minimum size of the variance window is the Occam's window, the value of α is never larger than 1. When the concept of this variance window is incorporated into the model-averaging process, the posterior model probabilities (also the model-averaging weights) are given by

$$w_k\left(MLBMA\right) = \frac{\exp(-0.5\alpha\Delta_k)}{\sum_{k=1}^{K}\exp(-0.5\Delta_k)} \qquad (2\text{-}22)$$

where $\Delta_k = (BIC_k - BIC_{\min})$ or $\Delta_k = (KIC_k - KIC_{\min})$. It can be seen that α is a multiplicative factor that when multiplied with ΔBIC or ΔKIC (as the case may be) reduces the impact the exponential term has on the weighting. For $\alpha = 1$, the weighting is identical to the BIC or KIC based weights, and for $\alpha = 0$ all models are equally weighted irrespective of their calibration performance. Tsai and Li (2008) also provide a table for recommended values of α corresponding to different significance levels and variance window sizes, which are shown in Table 2-1.

Table 2-1 Scaling Factors for Different Sizes of Variance Window and Significance Levels using *n* Observation Data (From Tsai and Li, 2008)

	σ_D	$2\sigma_D$	$4\sigma_D$
Significance level 5%	$4.24/\sqrt{n}$	$2.12/\sqrt{n}$	$1.06/\sqrt{n}$
Significance level 10%	$6.51/\sqrt{n}$	$3.26/\sqrt{n}$	$1.63/\sqrt{n}$

Tsai and Li (2008) originally derived the variance window concept only for Bayesian model-averaging. It is not entirely clear if a similar α factor can be applied to AIC-based likelihoods, and if so then what significance level and variance size would such factors correspond to. Singh et al. (2010) used the variance window concept with the KIC-based cumulative distribution function (CDF) to compare the performance of different groundwater model-averaging techniques for quantifying the impacts of model uncertainty on groundwater model predictions.

2.7.4 Qualitative Approaches

Refsgaard et al. (2006) proposed a new framework to address conceptual model uncertainty when models are used for making predictions beyond their calibration period. In this framework, a suite of conceptual models is independently calibrated and a pedigree analysis is performed to assess the overall reliability of the multiple models. Pedigree is expressed by means of a set of pedigree criteria to assess these different aspects. The pedigree approach integrates different types of knowledge, such as qualitative and quantitative uncertainty, to estimate the impact of model structure uncertainty on model predictions. Regardless of the framework's flexibility, the pedigree analysis does not provide an indication of the relative quality of the various model structures in quantitative terms. As a result, it cannot be integrated in quantitative uncertainty analysis, a significant drawback of this method since the inclusion of informative and proper prior knowledge about the alternative conceptual models, in terms of consistent prior model probabilities, will provide useful information to further reduce predictive uncertainty (Ye et al., 2005; Rojas et al., 2009). Nevertheless, the statements obtained from pedigree analysis should be available as the best possible scientifically based characterization of uncertainties and, as such, be made available to those involved in the decision-making process. However, the full framework needs to be tested in real water resource management case studies.

2.7.5 Multi-Criteria-Decision-Making (MCDM) Approaches

Multi-Criteria-Decision-Making (MCDM) approaches have been identified as a useful strategy to evaluate various model calibrations (Zeleny, 1982; Sziadarovszky eta al., 1986; Swaify and Yakowitz, 1998; Kuchanur, 2006). Multi Objectives help decision-makers to compare, rank, and organize the available choices (alternative simulation models) based on the identified evaluation criteria. These approaches are currently being widely used in public policies formulation. However, traditional multi-objective programming cannot quantify and incorporate the subjective preferences of stakeholders and decision-makers. Thus, Fuzzy Multi-Criteria-Decision-Making (FMCDM) was developed by Kuchaner (2006) to rank the alternative models not just based on their performances on goodness-of fit but also on a comprehensive set of subjective and objective measures. This approach can help to evaluate the models quantitatively, although the method for integrating the combined qualitative and subjective preferences is not yet clearly apparent.

Although, no one disagrees that predictions need to be evaluated using alternative models, diverse opinions exist on what alternative models should be included. Burnham and Anderson (2004) and Poeter and Anderson (2005) have proposed that all models in a candidate set be incorporated in a predictive uncertainty analysis, whereas others have recommended a more selective approach. The argument for including all models is that the models that do not provide an acceptable match to the observations are automatically assigned smaller model probabilities in the analysis, and that leaving them in simply allows all underlying conceptual models to be represented. The argument for a more selective approach is that results from clearly unreasonable models can be confusing to decision-makers and the public (Hill and Tiedeman, 2007).

2.8 Application of the Multiple Conceptual Model Method

Neuman and Wierenga (2003) provided a comprehensive strategy for constructing alternative conceptual models of subsurface flow and transport, selecting the best models, and using them jointly to render optimum predictions under uncertainty.

Harrar et al. (2003) presented an example of a multiple conceptual model method based on six alternative geological interpretations for a glacial-till aquifer system in western Denmark. Each model was calibrated against hydraulic heads and flows. Six flow models were used to predict the steady-state impact of a proposed well field, and to simulate particle tracking and solute transport. The predictive simulations show that simple models of heterogeneity produce capture zones similar to those of more complex models, but with different travel times and solute breakthroughs.

Ye et al. (2004) applied the Maximum Likelihood Bayesian Model-averaging (MLBMA) method to seven geostatistical models of air permeability variations at the Apache Leap Research Site in central Arizona. They used adjoint state maximum likelihood cross validation to compare its predictive capabilities with those of each individual model. The authors found the averaged model contained more information (had a smaller log score) and showed better predictive performance (showed wider predictive coverage) than any individual model considered.

Poeter and Anderson (2005) presented methods to rank the alternative models and discussed the use of AIC, AICc, BIC, and KIC criteria in groundwater applications. The prediction analysis was performed using all models and model-averaging for AICc and KIC. The results showed that the best fit to calibration data does not guarantee the most

accurate predictions at all locations in the model. They reported that model-averaging increases the width of model prediction confidence intervals in the best model.

Hojberg and Refsgaard (2005) investigated to which extent parameter uncertainty analysis may encompass model structure errors in a groundwater model. They used three different conceptual models, based on three different hydrogeological interpretations, for a multi-aquifer system in Denmark. Each model was calibrated inversely against piezometric heads and streamflows. The three models provided equally good and very similar predictions of groundwater heads, including well field capture zones. A parameter uncertainty analysis was carried out by Monte Carlo simulations. A comparison of the predictive uncertainties in the conceptual models shows large differences between the uncertainties in the conceptual models because their importance increase when predictive simulations consider data types that are extrapolated from the data types used for calibration.

Ye et al. (2006) assessed the conceptual model uncertainty of five recharge models within the modeling framework of the Death Valley Regional Flow System (DVRFS) model. Maximum Likelihood Bayesian Model-averaging (MLBMA) was used for conceptual model uncertainty assessment. In a Bayesian framework, the recharge model uncertainty is assessed, a priori, using expert judgments gathered from expert elicitation in the form of prior probabilities of the models. The prior probabilities indicate that no single model can be selected and other models cannot be discarded. The uncertainty is then evaluated by updating the prior probabilities to estimate posterior model probabilities. Based on calibration results, three information criteria (AIC, BIC, and KIC) are evaluated to rank the models. The posterior probabilities of the five recharge models, evaluated using KIC, are used as weights to average head predictions, which gives posterior mean and variance. The posterior quantities incorporate both parametric and conceptual model uncertainties. Posterior variance of MLBMA is larger than the variance of any single model, since conceptual model uncertainty is also addressed.

Kuchanur (2006) developed steady-state groundwater flow models to evaluate alternative models that reconcile simulation model uncertainties and applied interval optimization approaches to estimate groundwater availability and to prioritize management efforts in semiarid regions such as South Texas.

Troldborg et al. (2007) investigated uncertainty related to model conceptualizations and evaluated the effects on predictions of groundwater age and environmental tracer concentrations. Four different conceptual models based on alternative geological interpretations of a complex Quaternary aquifer complex were constructed and auto-calibrated to observations of head and discharge data. The four models showed major differences in predictions of age and concentrations, and none of the four models performed particularly well in the extrapolation step. The authors concluded that a single conceptualization may be adequate in characterizing natural behaviour of a field system after calibration against observations, because the calibration process is able to compensate for conceptual model errors through biased parameter values. However, it is critical to take model conceptual uncertainty into account when using the models to make predictions beyond the calibration period.

Rojas et al. (2008a) combined Generalized Likelihood Uncertainty Estimation (GLUE) and Bayesian Model-averaging (BMA) to assess model predictive uncertainty that arises from errors in model structure, input, and parameters. The likelihood measures of acceptable simulators assigned to the plausible models based on their ability to reproduce observed system behaviour were integrated over the joint input and parameter space to obtain the integrated model likelihood. The model probability was used to weight the model predictions in the BMA ensemble predictions. A three-dimensional hypothetical setup was used to demonstrate the methodology. Results showed that the predicted groundwater budget varied considerably among competing models; and a set of 16 head observations used for calibration did not allow differentiating between the models. BMA

provided average predictions that were more conservative than individual predictions obtained for individual models.

Diks and Vrugt (2010) compared model-averaging methods that weigh models using different techniques, without always requiring that the weights sum up to one. The methods were applied to two sites and compared in term of their predictive performance measured by out-of-sample root mean squared prediction error.

Ajami and Gu (2010) use the Bayesian Model-averaging (BMA) approach of Raftery et al. (2005) to assess uncertainty in a suite of biogeochemical models of various levels of complexity to simulate the fate and transport of nitrate at a field site in California. Their results demonstrate that whereas single models, regardless of their complexity levels, are incapable of representing all active processes at the site, the 95% uncertainty bounds of BMA bracket 90% to 100% of the observations.

Tsai used a variance-window (Tsai and Li 2008) version of Maximum Likelihood (ML) BMA (MLBMA; Neuman 2003; Ye et al. 2004) to quantify model uncertainty in managing groundwater within a thick sandy aquifer in Louisiana where saltwater intrusion is of concern. Alternative models are postulated to reflect uncertainty in conceptualizing hydraulic head boundaries and geostatistical parameterization through variogram models. The results show that using model-average predictions in the management problem requires relatively high injection and pumping rates are required to avoid violating constraints associated with multiple models.

Morales-Casique et al. (2010) also use variance-window with MLBMA to quantify uncertainty associated with five variogram and gas flow models for unsaturated fractured tuff in Arizona. Cross-validation indicates that, whereas MLBMA is far superior to individual models in one validation test (as measured by predictive coverage and log score), it is second to last in another such test, the estimated weights depending on calibration data.

Singh et al. (2010) develop an Interactive Multi-Objective Genetic Algorithm (IMOGA) to assess model uncertainty at the Waste Isolation Pilot Plant (WIPP) site in New Mexico. IMOGA can use either MLBMA or the Generalized Likelihood Uncertainty Estimation method (GLUE; Beven and Binley 1992) to estimate model weights. Their approach incorporates subjective expert knowledge in the weight estimation process. The authors find GLUE to yield more uniform weights than MLBMA, which tends to favor one model that fits observed data best.

Reeves et al. (2010) applied GLUE to assess model uncertainty in simulating radionuclide flux at the Climax Mine area of the Nevada Test Site. Model uncertainty arises from 25 possible hydrostratigraphy recharge combinations at the regional scale, as described by Ye et al. (2010a). The breakthrough of a conservative radionuclide is used to evaluate the influence of conceptual model and parameter uncertainty on radionuclide mass flux predictions.

2.9 Short Comments on Contribution in Relation to Previous Work

Formal uncertainty analyses have focused only on model parameters and input data as the principal sources of uncertainty in management model predictions. These analyses may underestimate management model uncertainties. During the past decade there has been a surge in the development of techniques for quantifying conceptual model uncertainty associated with groundwater flow and mass transport (Neuman and Wierenga (2003); Harrar et al. (2003); Ye et al. (2004); Poeter and Anderson (2005); Hojberg and Refsgaard (2005); Kuchanur (2006) ; Troldborg et al. (2007) Rojas et al. (2008a, b); Ajami and Gu (2010); Singh et al. (2010); Reeves et al. (2010). Despite this progress, the

quantification of conceptual models' uncertainties using multi-model ensembles as applied to groundwater management have not been explored.

Simulation-optimization models have been widely used to estimate the optimal and sustainable yield of groundwater (Gharbi and Peralta, 1994; Barlow et al., 2003; Czaenecki et al., 2003; McPhee and Yeh, 2004; Das and Detta., 1999, 2001; Qahman et al., 2005). However, no research has been found that accounts for conceptual model uncertainty in the estimation of sustainable groundwater yield.

Quantification of prediction uncertainties is important, especially when combined simulation-optimization models are used for decision making in groundwater management. Uncertainties must be determined and quantified since they can help policy-makers to prioritize their efforts to reduce these uncertainties when making risk-informed policy decisions. The conclusions made from these models and the formulation of relevant policies should be conducted with the objective of providing robust management plans.

The traditional approach for management modeling is a deterministic model. This model is intended to achieve optimal operation policies using a single model with single parameter estimation without quantification of conceptual model uncertainty. In contrast, stochastic simulation-optimization models are often used to quantify prediction uncertainty arising from imprecise model parameters, such as hydraulic conductivity values (Tung, 1986; Wagner and Gorelick, 1987; Wagner and Gorelick, 1989; Morgan et al., 1993; Feyen and Gorelick, 2004; Singh and Minsker, 2008), and the type of boundary conditions (Feyen and Gorelick, 2004; Feyen and Gorelick, 2005). Although these models can account for any source of uncertainty or model complexity, their drawback is computational burden. Thus, these models are usually employed for relatively simple models, including ones for which greater complexity in hydrogeologic setting requires greater intensive computational effort.

Tsai (2010) first introduced a variance-window-based BMA method to deal with model structure uncertainty in groundwater optimization models. The methodology was implemented to manage saltwater intrusion into the Baton Rouge area, Louisiana. His study focuses on model structure uncertainty in the boundary condition values of the groundwater model and in the semivariograms of hydraulic conductivity. However, his study did not explore the contribution of both conceptual and model parameter uncertainty.

Therefore; this study intends to demonstrate how multiple models for the same groundwater system can be used to quantify uncertainty and project its impact on groundwater management decisions. The efficacy of model selection and averaging methods are assessed with respect to their ability to improve model predictive capacity as compared to the use of individual models or ensembles of alternative models. Most importantly, uncertainties obtained using model-averaging approaches are integrated with a parameter uncertainty to quantify the contribution of both conceptual model and parameter uncertainty as applied to groundwater management models.

Chapter 3

Multi-model Development, Calibration, and Assessment

This chapter focuses on the construction and calibration of multiple groundwater flow models in order to explicitly account for conceptual model uncertainty in management decisions. Quantification of uncertainty has not been widely conducted for simulation-optimization models; thus, it is important not only to develop the framework but also to demonstrate the application of this framework in a real-world management context, thereby illustrating the validity of this approach.

The methodology presented in Chapter 3 is tested using data collected from the Thaphra Area, part of the Chi River Basin, located in northeastern Thailand, where groundwater is the main supply source. Pumping from closely spaced wells in two high volume pumping areas in this site has resulted in the development of deep cones of water-level depression in the vicinity. Therefore, efficient sharing out of the groundwater resources in this area is needed.

In this chapter, a description of the study area and its hydrogeological processes are presented and used to construct multiple conceptual models. Subsequently, groundwater flow simulation analyses using MODFLOW are described, the objective being formulation of flow models for the management problem.

3.1 Study Area Description

3.1.1 Background

Groundwater resources in the Thaphra Area (TPA) of the Phu Thok aquifer have been exploited for domestic, industrial, and agricultural purposes for over three decades. Fresh groundwater in this aquifer is underlain by brackish and saline groundwater, which is attributed to underlying rock salt in the underlying Maha Sarakham Formation. In the TPA, many villages and most of the local beverage industry depend entirely upon groundwater resources. As a result, groundwater withdrawal rates in the TPA have considerably increased since the late 1980s, leading to sustained water level decline. Records from monitoring wells by Department of Groundwater Resources (DGR) indicate that water levels in some wells (such as wells located at the center of TPA) have declined up to 8 m since monitoring commenced in 1962. Continued pumping without proper management of groundwater abstraction may lead to a regional groundwater depression and consequently result in a decline of fresh water head and the upconing of saline water into the aquifer.

The assessment of declining groundwater resources in the TPA has received substantial attention since the late 1980s, prompting further studies of the regional groundwater system. General hydrogeologic and groundwater quality data in the Khon Kaen Province, where the TPA is located, were initially compiled by Wongsawat et al. (1989a) to assess the potential of groundwater resources. Consequently, a 1:100,000 scale groundwater availability map for Khon Kaen Province was constructed to support a provincial water

management plan. Pumping tests, as well as monthly water level and water quality monitoring, were performed during the period 1989-1990 (Wongsawat et al. (1989b).

A numerical groundwater flow model was developed by Srisuk (1994) to analyze the regional groundwater flow regime and determine the groundwater balance for the entire Khon Kaen Drainage Basin, with a secondary focus on the local hydrogeology of the Ban Nong Khrai Nun Area, which is a part of the study area. The steady-state groundwater flow model for the year 1989 was shown capable of reproducing the general flow pattern of the aquifer. At that time, it was assumed that groundwater abstraction did not alter the overall flow pattern within the region. In the year 1998-2001, the Department of Mineral Resources (DMR) developed a medium-scale study (DMR et al., 2001) which included hydrogeological mapping, geophysical surveys, hydraulic testing, hydrochemistry, and water level and water quality monitoring. Moreover, a transient-state groundwater flow and mass transport model was developed, from the 1998 to 2000 data, to predict the impact of groundwater extraction (under different scenarios) for the years 2005 to 2010. They discovered that there were three areas potentially at high risk of saline water intrusion. Upon their recommendation, long term monitoring of water levels and water quality within the heavily exploited areas commenced.

Later, a preliminary study on aquifer storage recovery and contamination vulnerability in this area was carried out by the Department of Groundwater Resources (DGR, 2007 and DGR, 2008). The transient flow and transport models were also developed in the period from 2007-2008 as part of these studies. However, these models were calibrated over a short period of time with sparse data to support any hypotheses regarding salt transport from the Maha Sarakham formation. Therefore, these models do not satisfactorily reproduce the observed total dissolved solids (TDS) within the model domain. It is likely that the available data and incomplete knowledge about the primary salt transport mechanism are not adequate for simulating the effect of saline intrusion and upconing

within the aquifer. However, these studies have provided useful data and improved our understanding about the Thaphra aquifer system.

In the current study, a new model is developed as a tool for assessing the impact of alternative groundwater management plans and is designed to integrate the most current knowledge and hydrogeological information available for this area (building upon the successes of earlier modeling efforts). Unlike the aforementioned models, the model developed here is calibrated using PEST (Doherty 2004). In addition, alternative conceptual models are developed to quantify predictive capability of the model in the context of making groundwater management decisions.

3.1.2 General Description

The study area covers an area of 150 km^2 and is located in the Muang and Ban Haet districts of Khon Kaen Province and the Kosum Phisai district of Maha Sarakham Province, in the Northeastern part of Thailand (Figure 3-1). It is a part of the Chi River Basin, which consists of gently sloping and undulating hilly terrain. The elevation of this area is in the range of 150 to 220 m above mean sea level (m amsl), with the highest altitude located in the south (Figure 3-2). The TPA has a tropical monsoon climate, with an average annual rainfall and potential evaporation of 1,210 and 1,802 mm, respectively.

The primary river within the study area, the Chi River, flows from southwest to north. The flow of the Chi River varies extensively; about 90% occurs during the wet season of May through October, and about 10% occurs during the dry season of November through April. Due to the high intensity of rainfall, flooding of large areas drained by contributing streams often occurs towards the end of the rainy season. At the Thaphra station, the average runoff is 64 m^3/s (RID, 2009). Wetlands, swamps, and saline soils are commonly found in the lower elevations along the Chi floodplain.

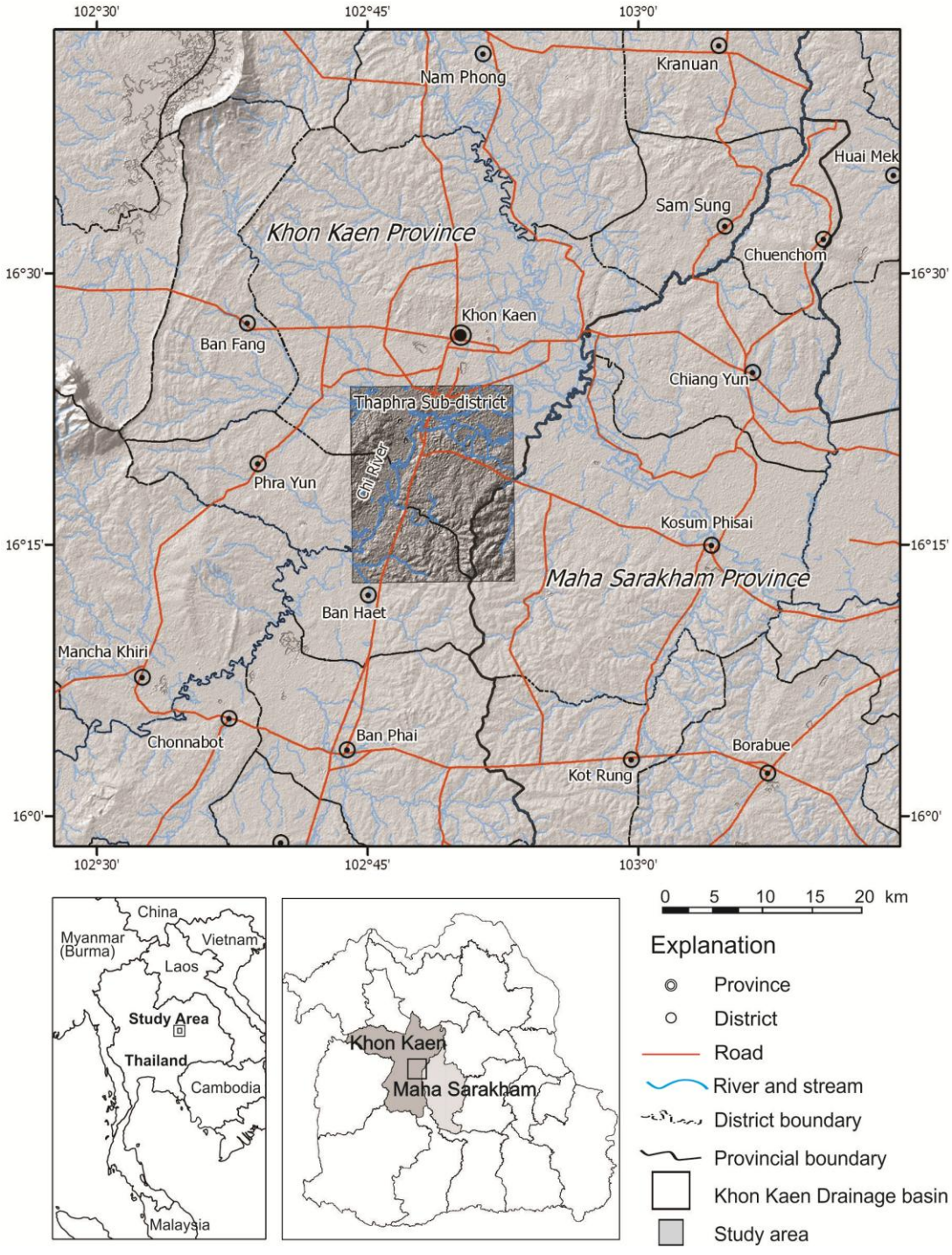


Figure 3-1 Location of the study area

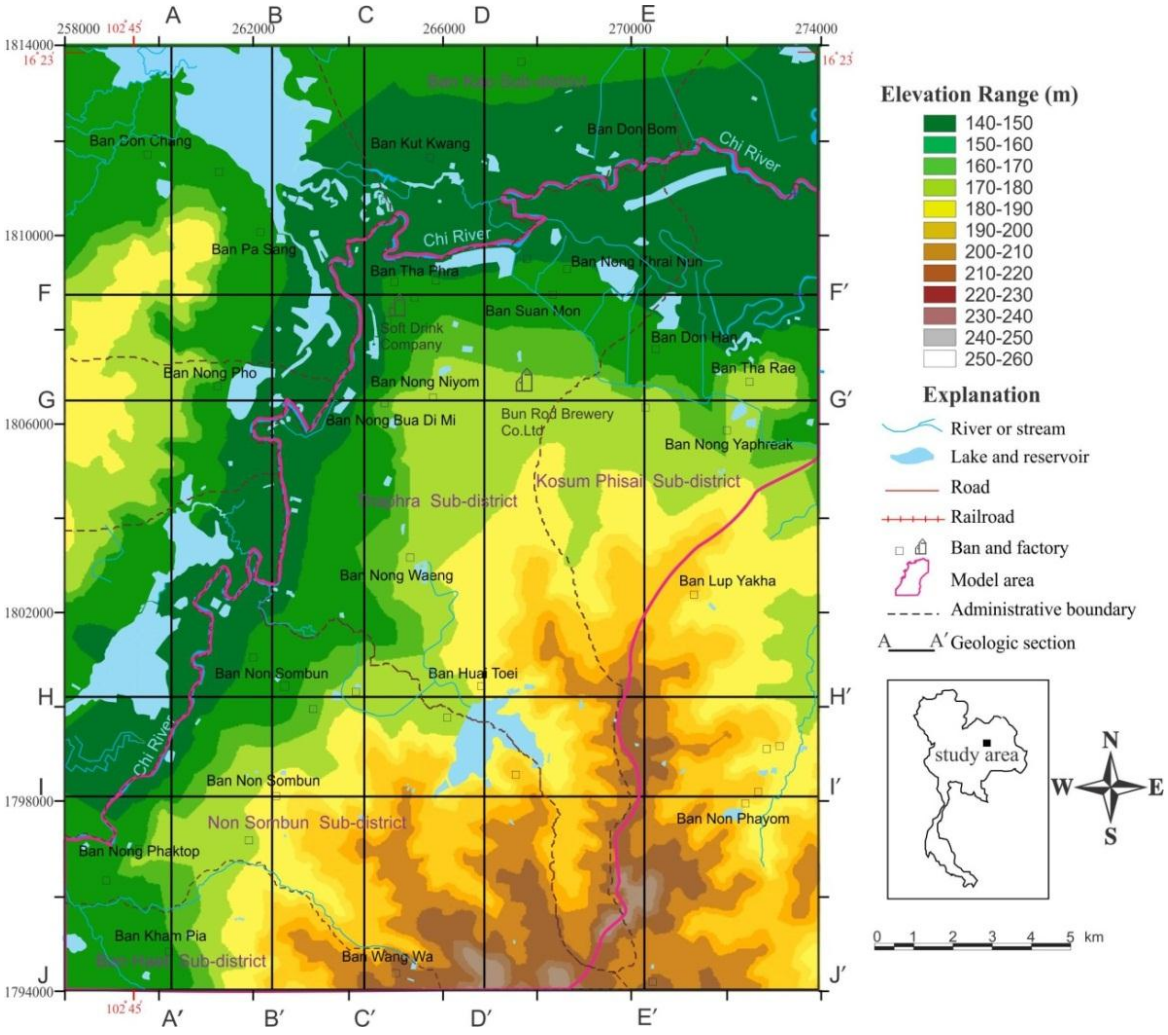


Figure 3-2 Topography of the study area, Thaphra area, part of the Chi River Basin, Northeast Thailand

3.1.3 Geology

The Thaphra area is underlain by three primary formations. From oldest to youngest, the strata are the MahaSarakham (KTms), Phu Thok (Tpt), and Alluvium (Qa) Formations, as depicted in Table 3-1. The Alluvium Formation, which is composed of 5 to 20 m of a combination of sand, gravel, silt, and clay, is the dominant surface formation along the banks of the Chi River. This alluvium is underlain by the Phu Thok Formation, which is composed of brick-red to reddish-brown sandstone interbedded with siltstone and mudstone (Suttayarak, 1985).

Table 3-1 Stratigraphic classification and lithology of the Thaphra Aquifer

Geologic Era, Period	**Formation**	**Sub-Formation**	**Lithology**
Quaternary	Alluvium (Qa)		Alluvium: sand, silt, clay and laterite (5-20 m)
Tertiary	Phu Thok (Tpt)	Middle Phu Thok	Sandstone, siltstone and shale: reddish brown to orange brown (50-250 m)
		Lower Phu Thok	Claystone and siltstone (100-400 m)
Cretaceous	Maha Sarakham (KTms)	Maha Sarakham	Rock salt interbedded with clay or claystone, potash, anhydrite, and gypsum

The Phu Thok formation can be divided into three members, which range in age from Upper Cretaceous to Lower Tertiary (DMR, 2002). The upper unit of the Phu Thok Formation consists of brick-red, coarse-grained, thick-bedded, and large cross-bedding sandstone. The middle unit consists of cross-bedded, fine-grained sandstone interbedded with siltstone. The lower unit is composed of reddish brown siltstone interbedded with gypsum-disseminated claystone. Within the study area, only the middle and the lower units of the Phu Thok Formation are found. According to the interpretation of seismic

reflection imagery, the total thickness of the Formation in the study area varies from 250 to 600 m (DMR et al., 2001). The thinner portions of the Phu Thok Formation are located at the base of the Chi River in the north and north-east, whereas the thickest parts are found in the south of the study area.

Underlying the Phu Thok Formation, the Maha Sarakham Formation is comprised of interbedded rock salts, clay/claystone, potash and anhydrite. According to Japakasetre (1985), the units making up the Maha Sarakham Formation are, from bottom to top, anhydrite, lower rock salt, lower clay/claystone, middle rock salts, middle clay/claystone, upper rock salt, and disseminated gypsum/anhydrite, respectively. This formation does not occur as outcrops in the study area because it is covered by the Phu Thok Formation and floodplain deposits. However, its presence is critically important with regard to water quality in the region due to the existence of rock salt layers. From seismic reflection interpretation (DMR et al., 2001), the rock salt layer found in exploration well (well number TP1) has been interpreted to be the upper rock salts of the Maha Sarakham formation.

3.1.4 Hydrogeology

The main aquifer of the study area involves the sandstone and siltstone of the Phu Thok Formation. In the north, fresh groundwater in the sandstone and siltstone are directed underlain by brackish water caused by the rock salt of the Maha Sarakham Formation at depths of 100 to 350 m below the ground surface.

The Phu Thok aquifer in the study area is presented in only two units: the Middle Phu Thok (MPT) and Lower Phu Thok (LPT) Aquifers (Table 3-1). The MPT aquifer, with a thickness of 50-250 m, consists of fine to medium sandstone and siltstone. The formation is easily fractured and is usually a productive aquifer, with well yields ranging from 15 to 50 l/s. Highly fractured zones are commonly identified through mud loss while rotary drilling at the rim of syncline structure parallel to the Chi River.

The transmissivity, horizontal hydraulic conductivity, and storage coefficient for Phu Thok aquifer have been estimated through the interpretation of pumping test data obtained from the DGR database (DGR, 2005; DGR 2007; DGR 2008). Results from 25 pumping tests with observation wells and 70 single well tests were used to estimate the transmissivity and horizontal hydraulic conductivities of the sandstone and siltstone of the Phu Thok Formation. The well test-derived hydraulic conductivity of the MPT aquifer varies from 0.0004 to 90 m/d. Storativity values range from 1E-07 to 1.5E-03. The LPT aquitard consists of clay and claystone, with an average thickness of about 150 m. The unit is underlain by one to three layers of rock salt, depending upon location. The hydraulic conductivity of the LPT Aquifer is two to four orders of magnitude lower than that of the MPT Aquifer, ranging from to 8E-5 to 1E-2 m/d.

Figure 3-3 shows hydraulic head contours and flow directions generated from the average water level measurements in the monitoring wells over the 2007 period. Regional hydraulic heads in the sandstone and siltstone of the Phu Thok Formation range from 140 to 200 m amsl. In the north and northwest region, along the Chi floodplain, flowing artesian wells can be found. Recharge occurs primarily in the south where the surface elevation is relatively high and the water table is relatively deep. In the recharge area, the average depth of the water table is 10-30 m below the ground surface. For the most part, groundwater flows towards the north and northwest discharge to the Chi River.

Due to generally increased withdrawals since the start of heavier industrial development in the early 1990s, water levels have declined from the predevelopment highs in some areas such as at the soft drink company site and Nong Khrai Nun Village (depicted in Figure 3-3). As recently as 1990, groundwater in Nong Khrai Num village was obtained from flowing artesian wells. The hydraulic heads in these wells were 3 to 6 m above the ground surface in 1989 but declined to approximately 5 to 8 m below the ground surface by 1992 and flowing artesian wells are no longer presented in the Nong Khrai Num village.

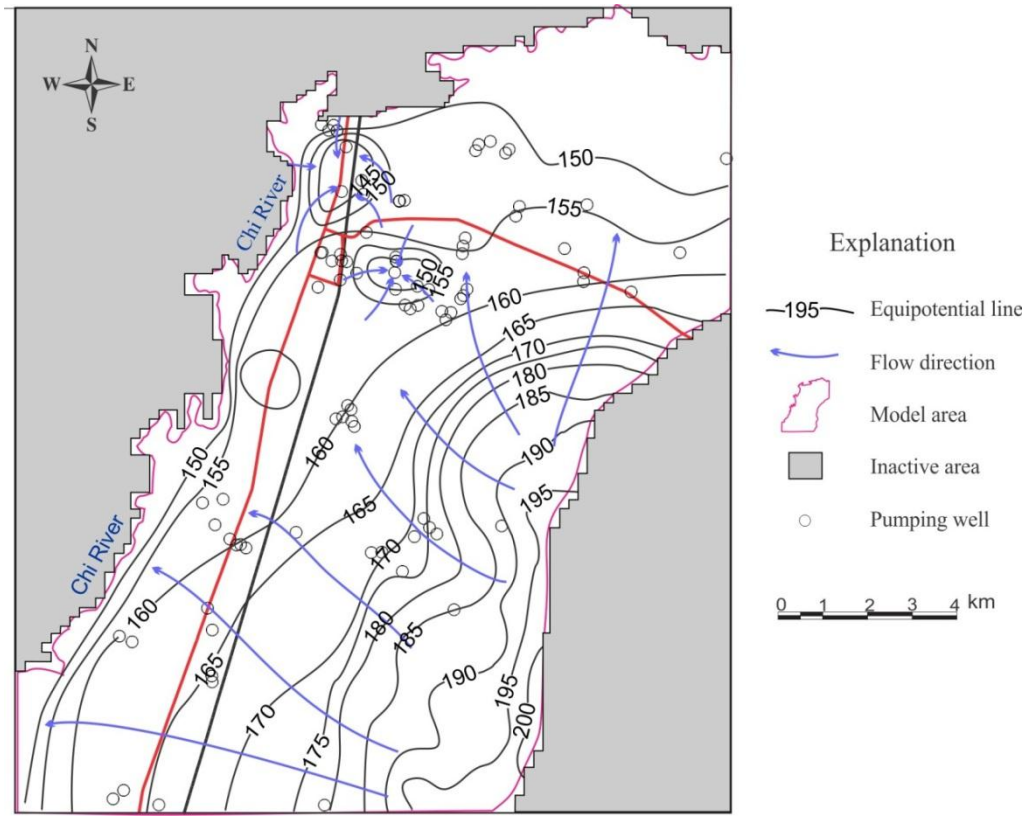


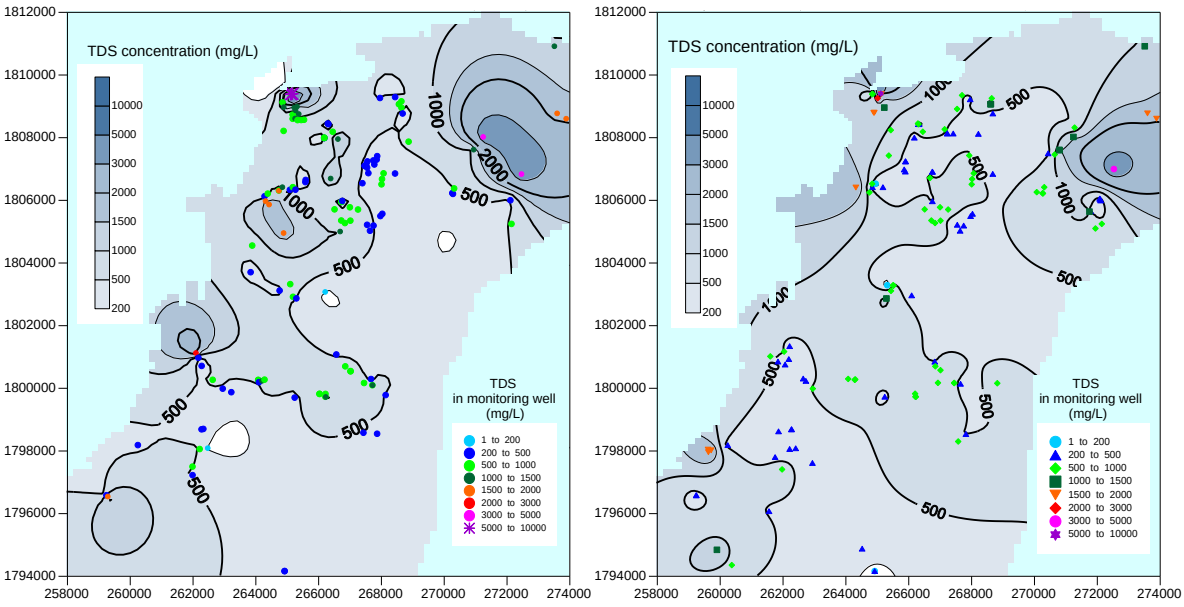
Figure 3-3 Groundwater contour map for the Phu Thok aquifer, Thaphra area, in the year 2007

3.1.5 Water Quality

The principal chemical constituents of concern in the Thaphra area groundwater are total dissolved solids (TDS) and chloride, both signatures of brackish water derived from the MahaSarakham rock salts. The total dissolved solids (TDS) of groundwater in the Thaphra area range from 100 to 8,000 mg/L, but are generally less than 1,000 mg/L. TDS concentrations are lowest in the recharge area located in the south and southeast of the study area. Approximately 40% of samples taken from 120 wells obtained from DGR database (located primarily in the north) from January to February 1999 contained TDS concentrations of less than 500 mg/L, which is the secondary Drinking Water Regulation limit for TDS (US.EPA, 2009). Less than 10% of the samples exceed 1,500 mg/L. The highest sampled concentrations were observed in the Thaphra Nao wells, located in the

northern part of the study area where the rock salt of the Maha Sara Formation is less than 250 m in depth. Chloride concentrations range from 5 to 5,000 mg/L. During the same monitoring period, approximately 35% of groundwater wells exceeded the guideline limits of 250 mg/L, 12% of wells exceeded 500 mg/L, and 3% exceeded 1,000 mg/L. The highest concentrations were found in the same locations, where the highest concentrations of TDS were observed.

Maps depicting spatial distributions of TDS and chloride concentrations at well depths range from 30 to 150 m are shown in Figure 3-4. These maps also compare the chloride concentrations measured in January and February 1999 with the concentration measured in June through August 2007 (data obtained DGR database). As shown in Figure 3-4, the elevated concentrations of TDS and chloride were observed only in a small area in the northern part of the study area. The slight increases in average TDS and chloride concentrations between the year 1999 and 2007 suggest that saline water intrusion and upconing is not currently a widespread problem in the TPA, and is not significantly worsening, despite increased pumping.



(a) (b)

Figure 3-4 TDS concentration contour map for the Phu Thok aquifer: a) January-February 1999, b) June-August 2007

3.1.6 Groundwater Withdrawals

Groundwater in the Thaphra area has been extracted from the Phu Thok aquifer for agricultural, industrial, and domestic uses for over four decades, with most withdrawals drained from the Middle Phu Thok aquifer. Estimated current withdrawals are based on a combination of metered data (for larger industrial users such as brewery, bottled water and soft drink companies) and residential estimates derived from typical pump type and pump capacity. Groundwater withdrawals from the Upper Phu Thok aquifer have increased from approximately 0.5 Mm^3/y (million cubic meters per year) in 1984 to 3.4 Mm^3/y in 2007 as shown in Figure 3-5. Approximately 73 % of the total withdrawal in the region is used for industry, 24 % for domestic use, and 3 % for agricultural purposes. Moreover, approximately 80% of the total withdrawal occurred only in a small area of 20 km^2 in the north of the study area, causing the development of deep cones of depression in the vicinity of two pumping centers, seen in Figure 3-3. Thus, this small area is considered to be the "critical area".

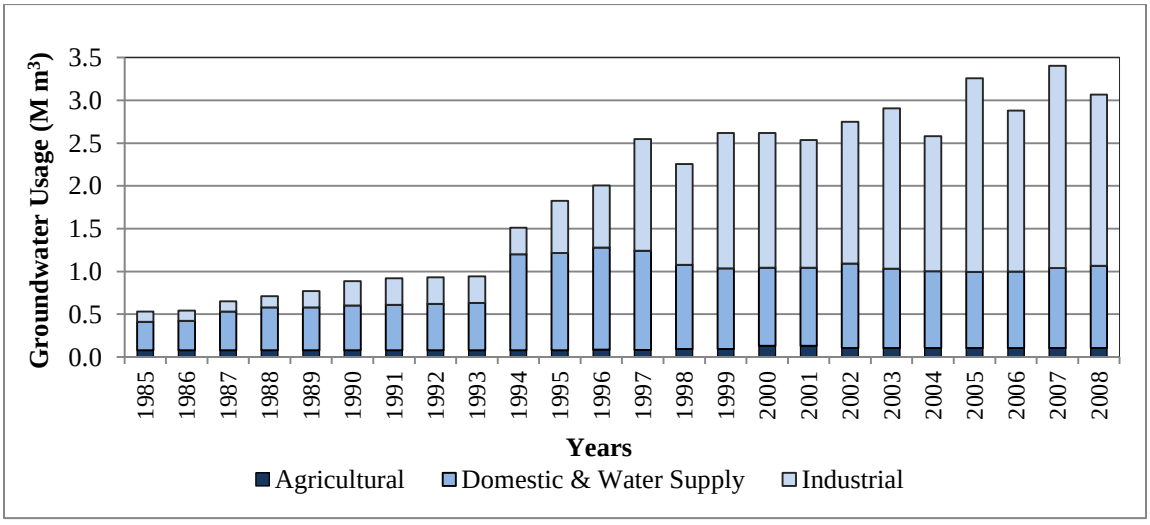


Figure 3-5 Estimated groundwater usage in the study area

3.2 Alternative Conceptual Model Development

The development of a groundwater model involves various subjective judgments, simplifications, and assumptions that can be predominantly attributed to the incomplete knowledge and data about the system being modeled. Model uncertainties can originate from errors, associated with parameterization, hydrogeologic processes (such as recharge, evapotranspiration), boundary conditions, and initial conditions. Identifying the full range of plausible conceptual models and sources of uncertainty is desirable but not feasible in practice. Based on the characteristics of the Thaphra study site and the available data, a number of possible conceptual models are developed considering the uncertainties in the simulation modeling process. Here, a set of unique conceptual models that potentially contribute significant uncertainty are generated for testing a number of hypotheses regarding the impact of conceptual uncertainty on groundwater management. The twelve plausible conceptual models, based on three different hydrogeologic unit interpretations, two different recharge estimations, and two different boundary conditions, are developed and presented in the following section.

3.2.1 Hydraulic Property Distributions

Three conceptual models of site hydrostratigraphy are developed that represent a degree of uncertainty in the hydraulic property values. The hydraulic property distributions have been interpreted based on lithological data from 120 boreholes, 65 pumping tests, and the interpreted data obtained from geophysical surveys (200 vertical electrical soundings) and 6 lines of seismic reflection surveys. These data were obtained from DGR's database. Location of borehole data, pumping tests, and geophysical surveys are shown in Figure 3-6 to 3-7, and 3-8, respectively. The lithological data and the pumping test locations are clustered in the northern part (Thaphra Sub-district) of the study area. There is relatively less information available in the south and south-eastern part of the study area.

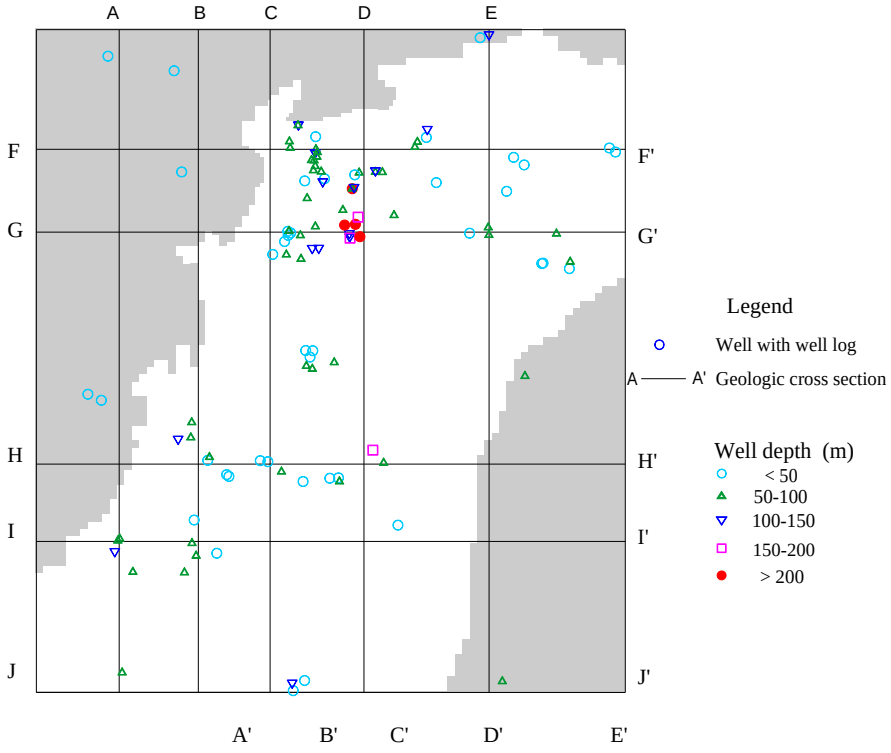


Figure 3-6 Location of borehole with lithological data

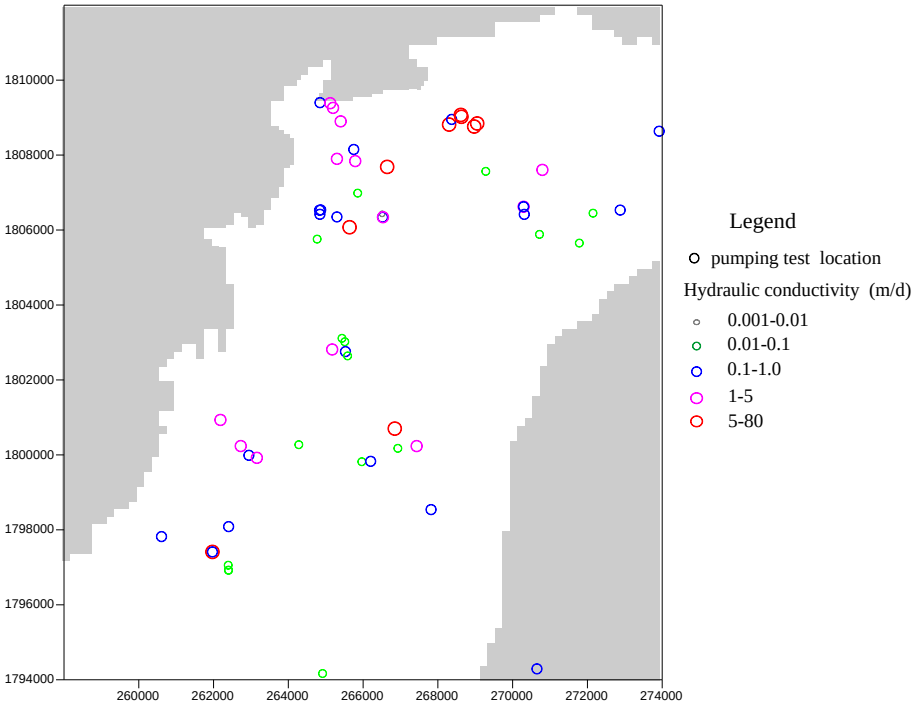


Figure 3-7 Location of wells with hydraulic conductivity measurements in the study area

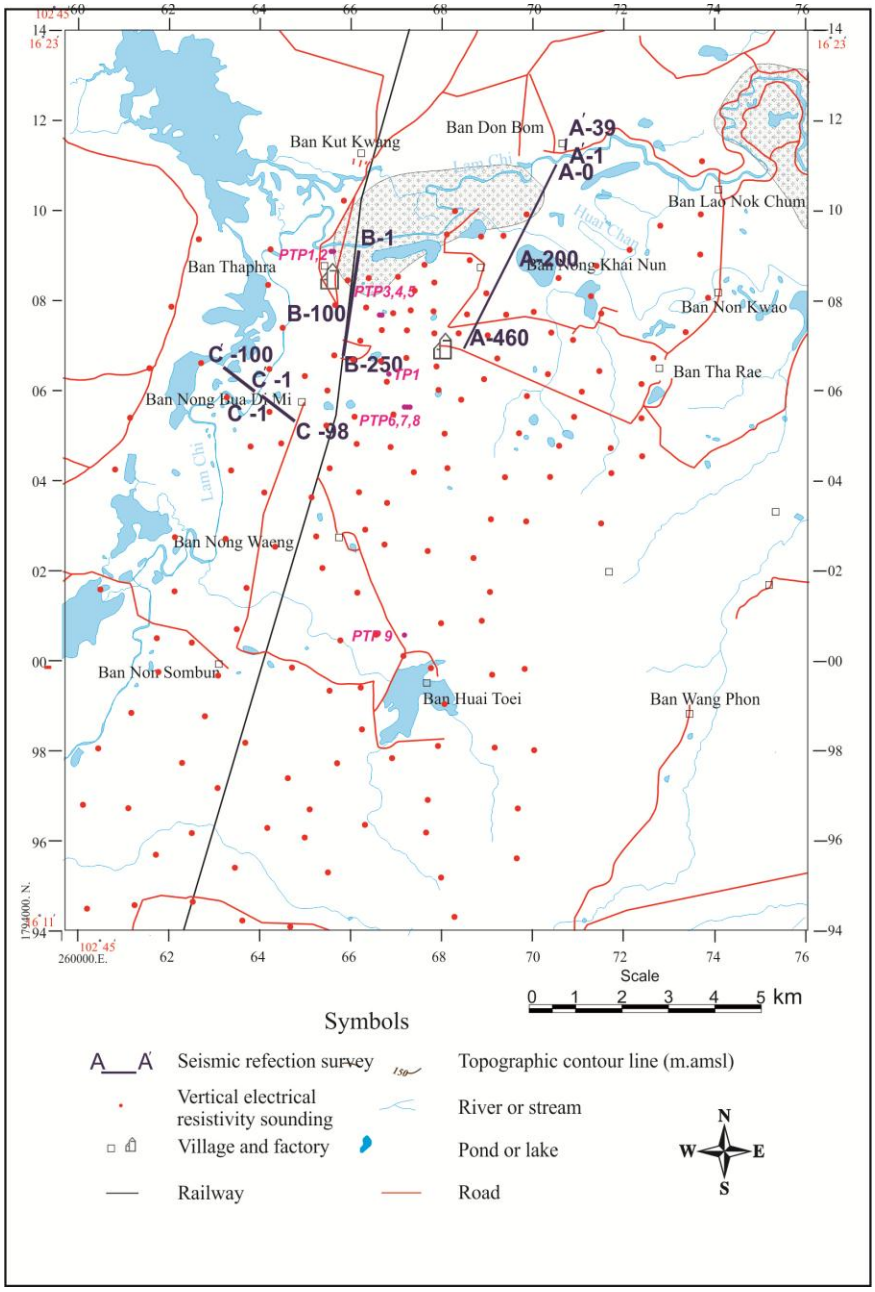


Figure 3-8 Location of seismic reflection survey and vertical electrical resistivity sounding (data obtained from DGR's database)

Most of the lithological data (94 of 120 wells) are available at depths less than 100 meters and only 6 wells provide information at depths greater than 200 meters. Values of hydraulic conductivity based on 65 pumping test analyses range from 0.001-80 m/d (Figure 3-7). No

measurements for vertical hydraulic conductivity are available in the study area. Based on the limited data to characterize the distribution of hydraulic properties, the geophysical survey data were used to help define and classify the hydrogeologic units. The combined interpretation of resistivity, seismic reflection data, lithological data, and hydraulic property values from pumping test analysis were used to construct geological sections containing thickness and boundary between different hydrogeologic units. The uncertainties in any hydrogeologic unit distribution arise since hydrogeologists cope with incomplete, inferred or interpretive data. These data must be interpolated between limited borehole data, a task that is subjective in nature. While it is impossible to generate a perfect sampling of plausible parameter distributions, the intent here was to generate hydrogeological models similar to those that would be built by three different hydrogeologist teams, and thus be independent of one another. That is, the differences in the geological interpretation of the three models (H1, H2, and H3) are due in part to data interpolation differences to represent subjective interpretation and actual knowledge of the system by the individual geologists.

In Model H1, seven hydrogeological units were identified. The Middle Phu Thok formation has been divided into four sub-units according to lithology and hydraulic properties as shown in Table 3-2. Model H2 was conceptualized by seven hydrogeological units. Fractures were assumed to be less frequent at the deeper levels of the Middle Phu Thok aquifer. Model H3 was constructed as the most detailed model. Eight hydrogeological units were conceptualized from interpreting vertical profiles based on borehole information, resistivity, and seismic reflection data. In model H3, the Lower Phu Thok aquifer was assumed to be a more permeable layer than it is in models H1 and H2. The thickness and geometry of each model are shown in Figure 3-9.

Table 3-2 Hydrogeologic unit classification and lithology of the Thaphra Aquifer

Sub-Formation	Lithology	Hydrogeologic unit	Hydraulic conductivity parameter		
			Model H1	Model H2	Model H3
	Alluvium: sand, silt, clay and laterite	Qa	Qa	Qa	Qa
Middle Phu Thok	Sandstone, siltstone and shale: reddish brown to orange brown	MPT1	MPT1	MPT1	MPT1
			MPT2	MPT2	MPT2
			MPT3	MPT3	MPT3
			MPT4	MPT4	MPT4
				MPT5	MPT5
Lower Phu Thok	Claystone, mudstone, interbedded with siltstone	LPT	LPT	LPT	LPT1
					LPT2

3.2.2 Recharge Models

The recharging of aquifers is a complex process and depends upon several factors, including precipitation, evapotranspiration, runoff, land use, land cover characteristics, soil moisture holding capacity, and water level elevations. Multiple methods can be used to measure recharge, for example gravity lysimeters, chloride mass-balance models, or by interpreting the water-table fluctuation in wells. Unfortunately, such studies have not been carried out in the Thaphra area. Therefore, the recharge rates must be obtained through calibration; however, information regarding the likely spatial distribution of recharge may be discerned from information about land use and soil type, as described below..

Two common recharge zone delineation approaches were chosen to represent one source of uncertainty associated with the conceptualization of the aquifer system. For the first recharge model, the net recharge was distributed spatially into three zones according to only soil type. Surface soils with moderate-to-well, moderate, and poor drainage characteristics were mapped into the three recharge zones used in the model (Figure 3-10).

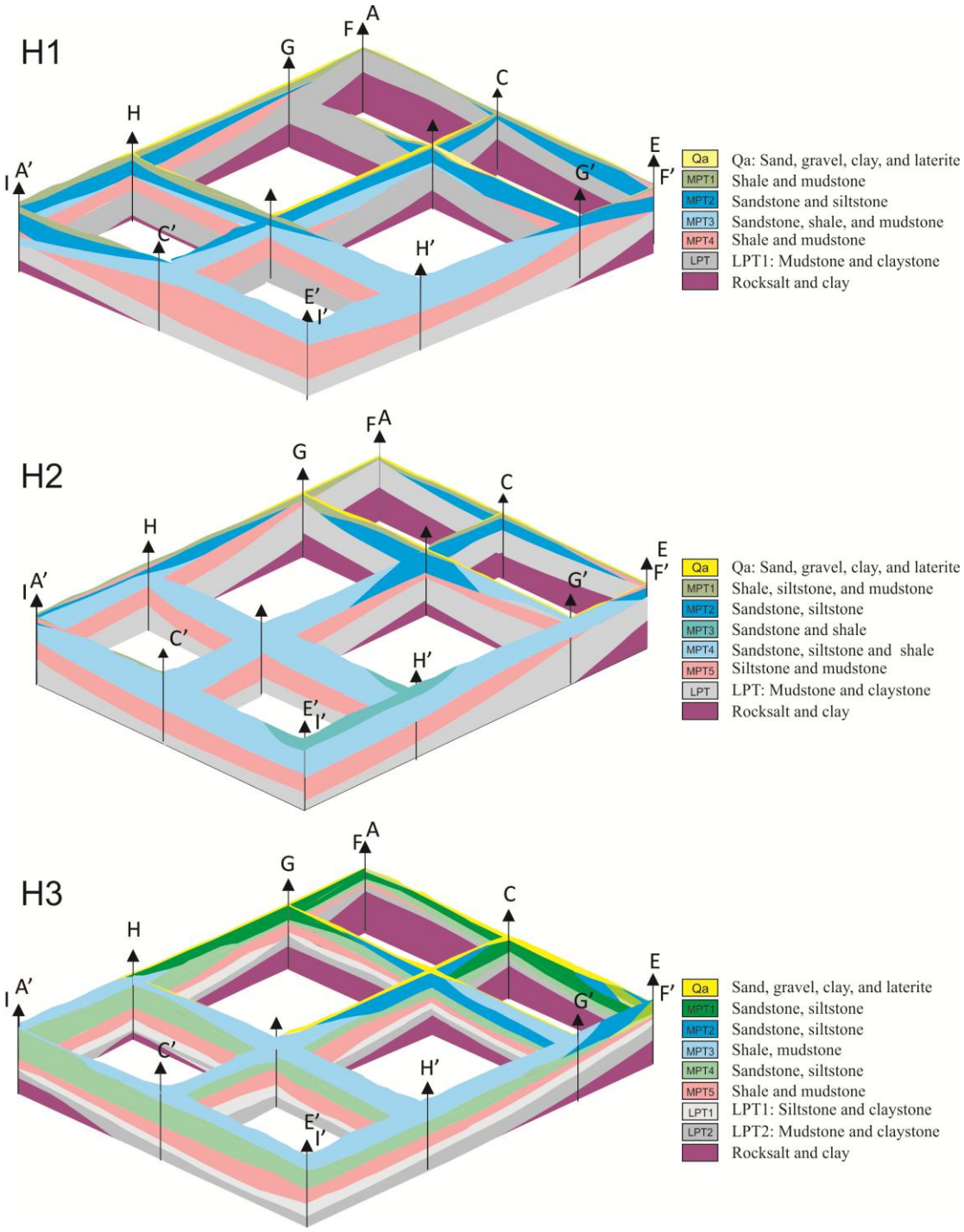


Figure 3-9 The geometry and distribution of model H1 (a), Model H2 (b), and Model H3. The cross-sections are defined in Figure 3- 2

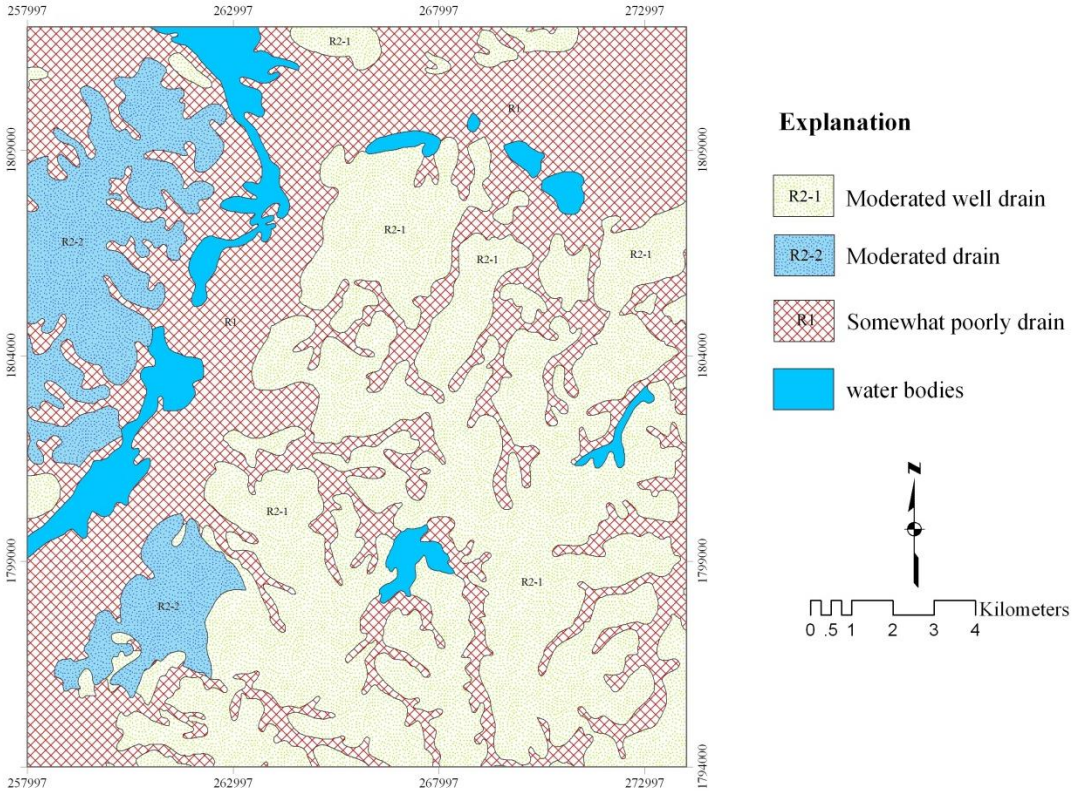


Figure 3-10 Recharge Model 1 for the TPA model domain

In the second, more refined, recharge model, the land surface was discretized into nine zones, each zone delineated based on a combination of soil type and land use/land cover characteristics (Figure 3-11). The land use was delineated into five categories rice paddy, other agriculture (crop), grass, forest (wood), or residential. This was overlain with a map of low, moderate, and high potential soil drainage to generate a map of eleven different zones. Note that only the zone boundaries were determined using these methods. In both cases, the magnitude of recharge assigned to each of these zones was estimated by an inverse simulation procedure that produced the best fit of water level observations, with zones of unique property combinations assumed to have identical recharge signatures.

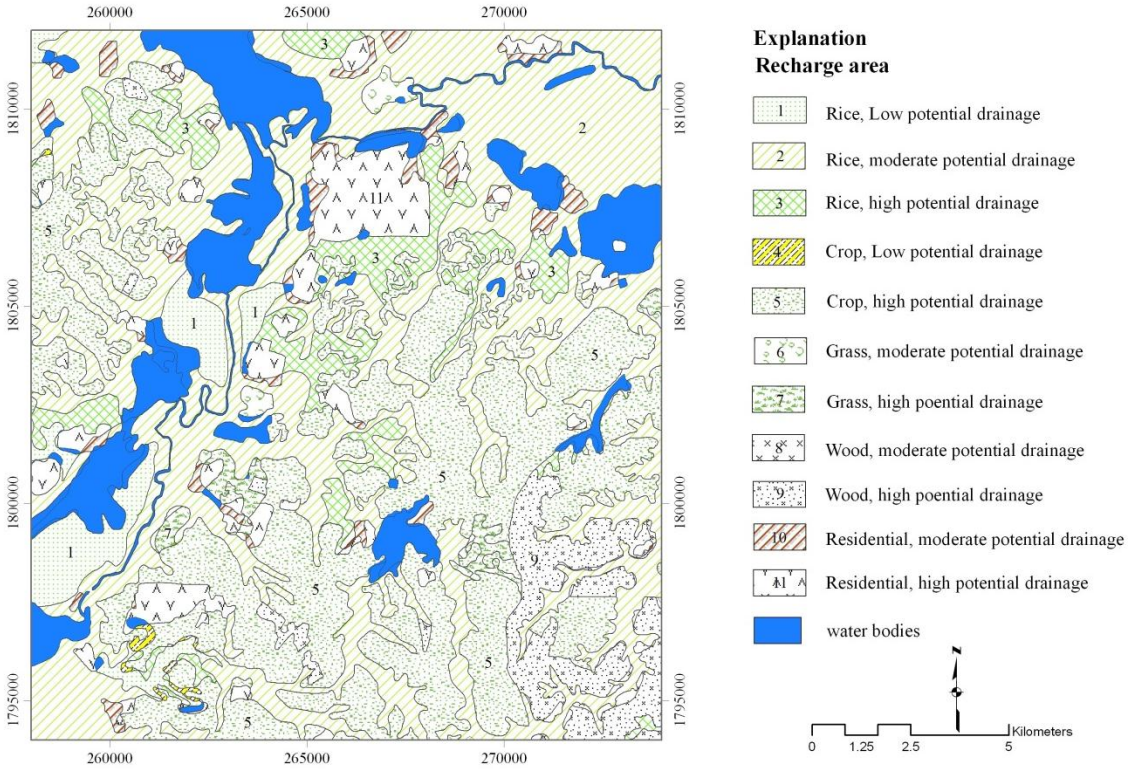


Figure 3-11 Recharge Model 2 for the TPA model domain

3.2.3 Boundary Conditions

Boundary conditions assigned in the alternative conceptual models are divided into two sets (Boundary-1(B-1) and Boundary-2 (B-1)), as shown in Figure 3-12. For both models, the underlying rock salt of the Maha Sarakham Formation forms the impervious bottom boundary, reaching depths of approximately 100 to 250 m below ground surface in the north to 550 m below ground surface in the south. The topmost layer is treated as an unconfined aquifer, with the others treated as confined aquifer. Head dependent flow conditions (MODFLOW's River Package) are used to characterize groundwater interaction between aquifers and the Chi River and Huai Toei Reservoir in the top layer. In B-1, no-flow boundary conditions are applied in all layers at the lateral boundaries to coincide with presumed groundwater divides in the west and south. In B-2, general head boundary

conditions (GHBs) are used to characterize groundwater inflows and outflows between the lateral boundaries located in the west and south. The GHBs are also used for the lateral boundary under the Chi River in the deeper layers. The boundary conditions assigned in the simulation models are depicted in Figure 3-12.

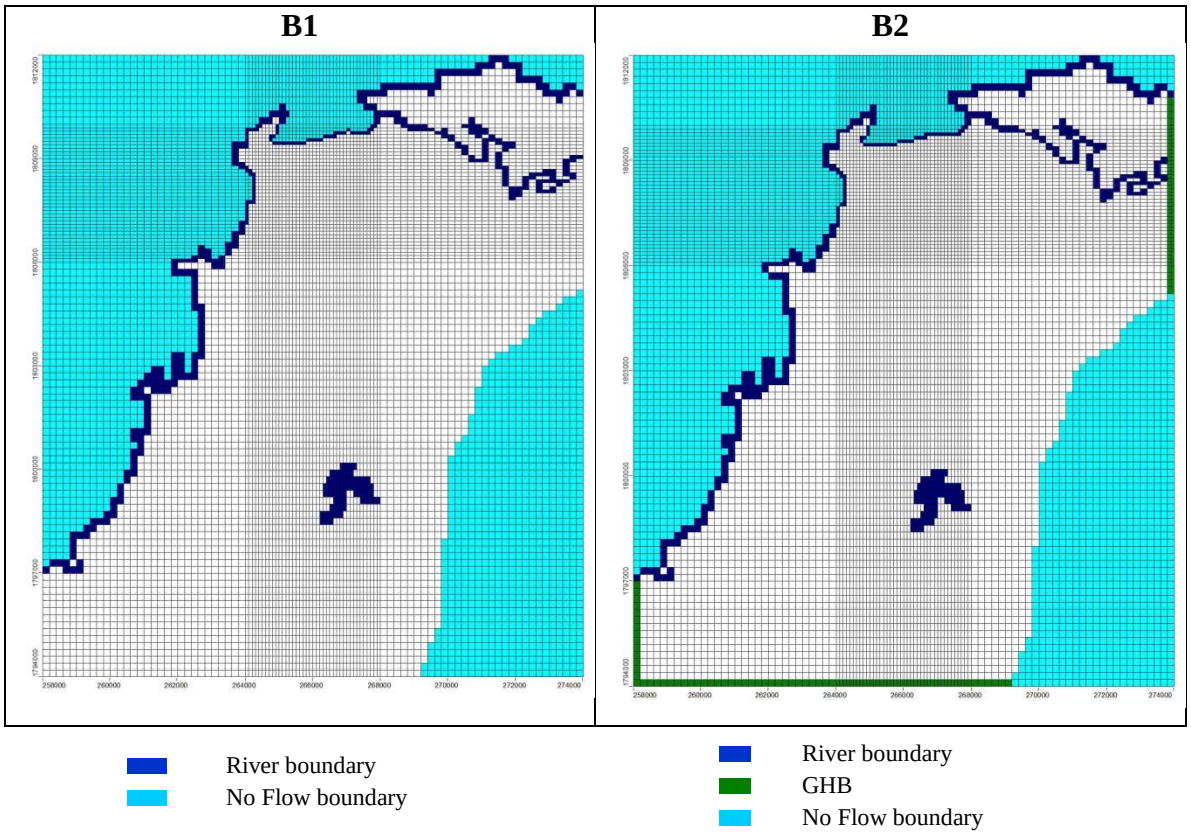


Figure 3-12 Boundary conditions of the two boundary models

3.3 Numerical Modeling

The twelve three-dimensional finite-difference numerical models are developed using MODFLOW (Harbaugh et al. 2000) and MT3D (Zheng and Wang, 1999), with the Groundwater Vista graphical user interface (ESI, 2007). Steady-state simulations were executed using MODFLOW-2000 (Harbaugh et al., 2000). Identical discretization was assigned in all models to minimize biases caused by different numerical errors in each model.

3.3.1 Discretization

The modeled domain is 16 by 18 km and 520 m in depth. The domain is oriented north-south and discretized into 110 rows, 100 columns, and 15 layers, with 89,951 active cells. The grid cell size is 200 m in both the x and y-directions and is refined to 100 by 100 m in the area of interest where the groundwater wells are dense and extraction rates are high (Figure 3-13). A digital elevation model is used to define the upper surface of the aquifer system. The thicknesses of model layers 1 to 7 vary with changes in topography, while the thicknesses of model layers 8 to 12 are kept constant with a vertical discretization of 40 m. The base model was tested at two different levels of discretization, with the finest model simulated using 50m x 50m grid cells; there was minimal difference in model results (i.e., <0.1m head difference at observation points), and the level of discretization was therefore deemed appropriate. Figure 3-13 shows a three-dimensional view of the model grid geometry and boundary conditions.

3.3.2 Recharge

The net recharge is applied to the top of the active portion of the top model layer by zone. The evapotranspiration (ET) package was not used here since there was insufficient information to parameterize the ET model. The net recharge is identified through calibration by dividing the land surface into different zones as shown in Figure 3-10 and 3-11. In model Recharge l (R1), one zone was insensitive and in Recharge 2 (R2) five zones were insensitive to model observations, so these zones were combined with recharge zones with having similar recharge values. The distributions of recharge specified in the numerical models are shown in Figure 3-14. The magnitudes of recharge assigned to each of these zones are adjusted during steady-state model calibration.

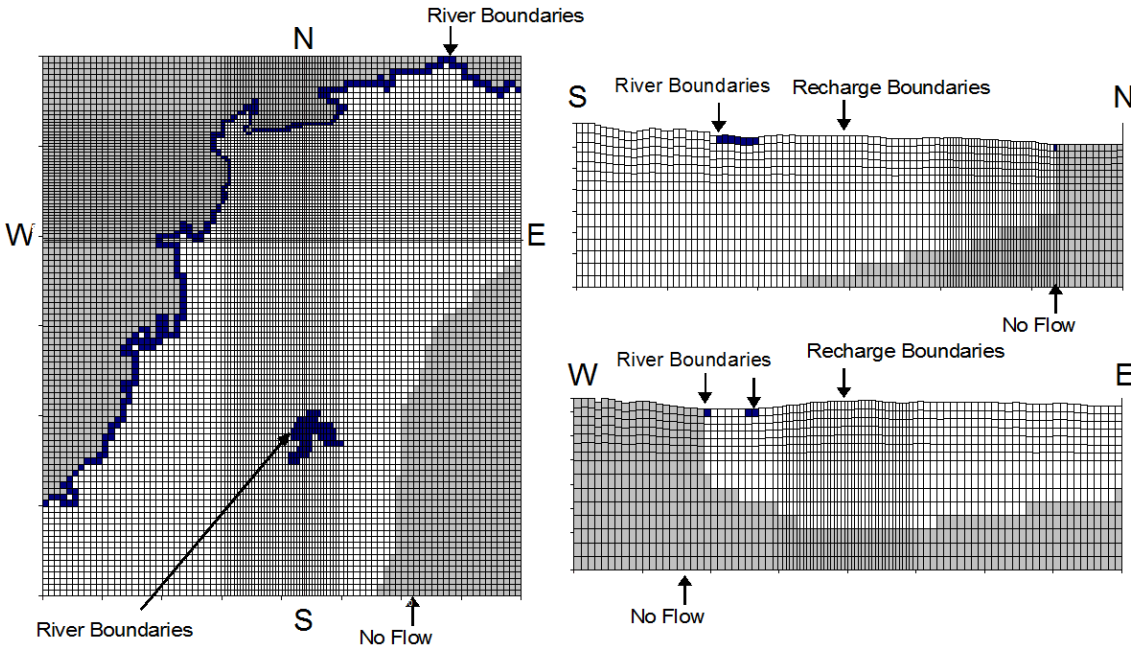


Figure 3-13 Illustration of model geometry, grid, and boundary conditions

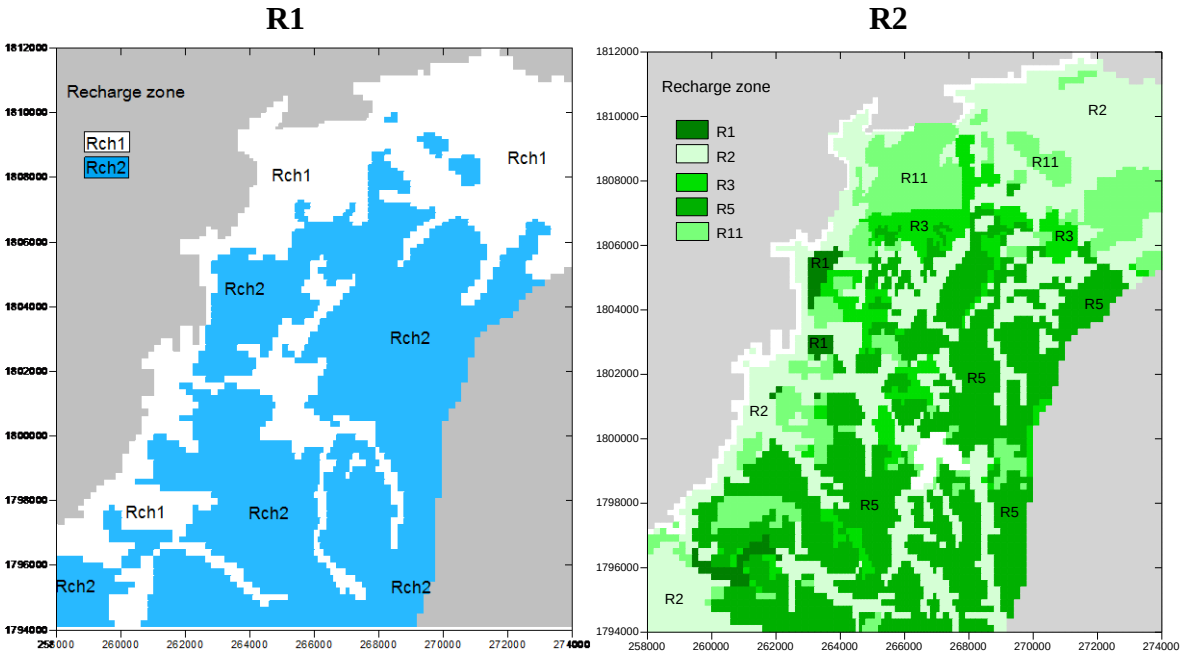


Figure 3-14 Recharge zone for the TPA Numerical model domain

3.3.3 Pumping

Pumping from wells is simulated using the well package. Wells in the Thaphra area are typically completed with perforations that extend to multiple intervals and thus to multiple layers in the models. Substantial volumes of groundwater discharge from the aquifer system through pumped wells are shown by the model layers in Figure 3-15. More than 70 percent of the groundwater withdrawal is completed in model layers 3 to 5 and mostly withdraw from hydrogeologic unit MPT4 and MPT2 in all Hydrogeologic models.

Pumping rates during 1998-1999 are used for the steady-state model since the pumping in this period was quite steady. In this year, the daily total withdrawal from 38 industrial wells and 60 public-supply and domestic wells was 6,600 m^3/d. Return flows of pumping through subsequent infiltration of excess water likely occurred, mainly with domestic and agricultural wells, whereas those of industrial wells are scant and not accounted for in the

modeling. A total of 165 wells are incorporated into the model; the locations of pumping wells are shown in Figure 3-16.

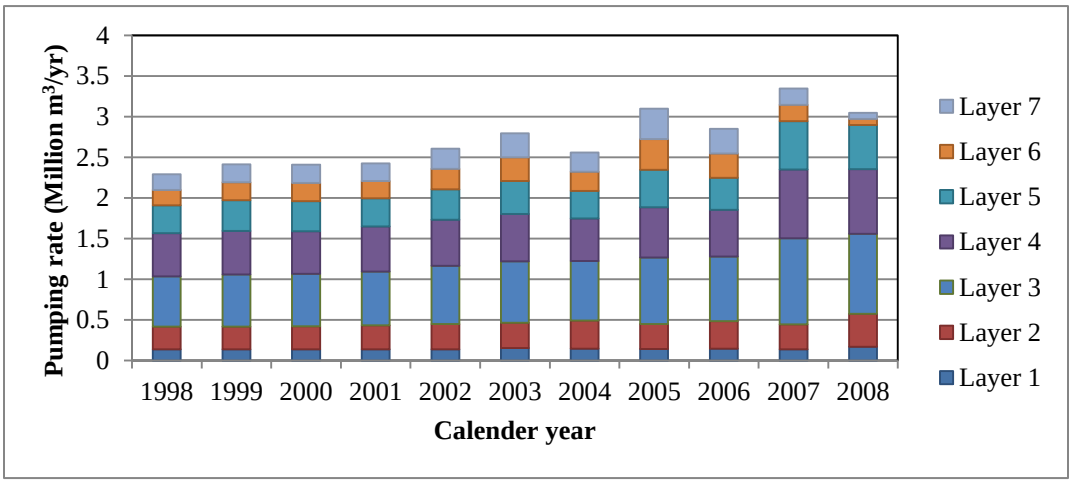


Figure 3-15 Pumping by model layers in the Thaphra model

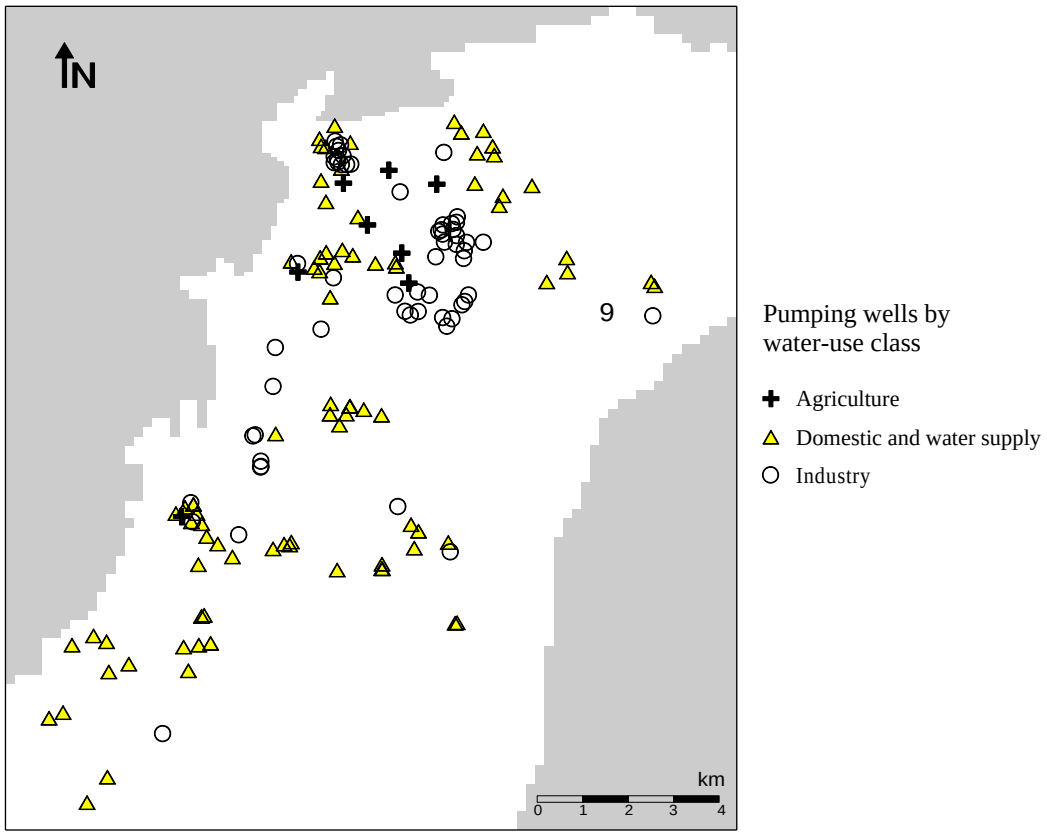


Figure 3-16 Location of groundwater withdrawal wells in the Thaphra area

3.4 Model Calibration

The alternative groundwater flow models were calibrated for steady-state conditions using the same set of head observations. Both manual and automatic calibrations were used to achieve optimum parameter sets for each of the 12 alternative models. Model calibration was designed to avoid overfitting to the data while simultaneously maximizing our ability to represent the physical system. The automatic calibration was executed using the PEST software program (Doherty, 2004). PEST was used to calibrate horizontal and vertical hydraulic conductivity, recharge rates, and boundary-flux model parameters. PEST uses non-linear regression to determine an optimal parameter vector, *b*, that minimizes an objective function, *S*(*b*), of the form

$$S(b) = \sum_{i=1}^{75} \omega_i \left[h_i - h'_i \right]^2 \qquad (3\text{-}1)$$

where h_i is an observed head; h'_i is the simulated head; and ω_i is the weight for the observation *i*. For the head measurement data taken from wells, well elevations determined by surveying in the field were assigned weight =1, and weight = 0.5 was assigned for those from the digital elevation model.

The steady-state calibration attempts to reproduce water level conditions for the year 1998, a year when the groundwater system was assumed to be at or near quasi-steady-state conditions because water levels, recharge, and pumping conditions that occurred during 1998 were similar to the long term average conditions. A review of long-term rainfall records from the two stations located in or near the model domain indicated small departures from the long- term average condition (i.e., for the years between 1987 and 1998) of only 37 or 30 mm for the two stations in 1998. In addition, 1998 contains the most reliable and numerous data for calibration. The average annual data for water levels, well extractions, recharge, and river stage conditions for this period are used in this

calibration. During steady-state calibration, several manual trials were first undertaken for each conceptual model. Each of the twelve models was then individually auto calibrated with the PEST software.

3.5 Calibration Data

The observations in the Thaphra area for the period of May 1998 to December 2008 were obtained from DGR's database. The available water level measurements were initially reviewed for use in the calibration process. It was found there is an acute lack of continuous records of long-term measured hydraulic heads. Most hydraulic head observations are manual measurements obtained from several DGR projects during different periods and for different purposes. For the measurements at some locations (e.g., the Ban Thaphra Nao School or the Livestock Development and Research Center Office at Ban Thaphra), abrupt decreases in ground water levels over time have been observed, which might be caused by incorrect measurements or clogged well screens. These measurements were removed from the observation data considered in the calibration exercise.

Steady-state head observations are specified in the database as an elevation, meaning they implicitly possess the uncertainty in the land-surface elevation of the well cap. Although land-surface elevations have been surveyed at many wells, it was found that the land elevations for a large portion of the monitoring wells vary with the different source documents. Discrepancies are likely a by-product of different elevation benchmarks being used by different investigators. Further, some altitudes are derived from 1:50,000-scale topographic maps with contour intervals of 10 m. As a result, many water-level observations have an associated error of roughly one half of the value of the contour interval (5 m). Since the observations are based on measurements of unknown accuracy, it is assumed that the accuracy level is the same for all observations. Therefore, the

weights assigned for these observations during calibration are only assigned considering the well elevation uncertainty. In this study, the higher weight of 1.0 was assigned only for the head measurements taken from wells for which the well elevations were determined by surveying with consistent elevations, whereas the lower weight of 0.5 was assigned for observations with inconsistent reporting and those derived from the digital elevation model.

Average hydraulic heads of years 1998 to 1999 for 75 observation wells within the model domain were used as calibration targets for the steady-state simulation. These wells were screened at varying intervals, ranging from model layers 1 through 11, although most head observations (80%) are from wells completed in the top three layers. Within the model domain, the observed hydraulic heads range from 138.5 to 186.4 m amsl, and the range of observation is approximately 48 m.

Compiling and analyzing information relevant to the development of groundwater simulation and optimization models indicates significant data gaps, in particular, on recharging, storage coefficients, and limited long-term time-series data on hydraulic heads. Therefore, one of the important aspects of this thesis is to illustrate these limitations and develop methodologies to overcome the challenges posed by limited data.

3.6 Calibration Parameters

Parameters considered for estimation included hydraulic conductivity values for all of hydrogeologic units defined in three hydrogeological models and all recharge zones in recharge models.

Only the sensitive components of hydraulic conductivity are estimated. These are the horizontal hydraulic conductivity of the MPT units. The horizontal and vertical hydraulic

conductivity are subsequently linked through the anisotropy ratio, a=Kv/Kh. A value of a=0.1 is used for all hydrogeologic units. The anisotropy ratios were assigned based upon the results of trial-and error calibration.

Tables 3-3 to 3-5 list initial values of the calibrated parameters and the ranges used to constrain parameter variation within PEST for all alternative models. The initial parameter values and upper and lower bounds for hydraulic conductivity were derived from pumping test data obtained from DGR's database and the initial parameter for recharge were based on previous calibration results (DMR et al. 2001; DGR 2008). Log-transformed parameters were calibrated for all parameters.

Table 3-3 Initial values and upper and lower limits for parameter estimated in the calibration for Hydrogeologic model H1.

Parameter (m/d)	Rock description	Lower bound	Upper bound	H1R1B1	H1R1B2	H1R2B1	H1R2B2
K_Qa	sand, gravel	0.001	20	5	5	5	5
Kx_MPT1	shale, mudstone	$1.0x10^{-3}$	10	0.05	0.05	0.05	0.05
Kx_MPT2	sandstone, siltstone	$1.0x10^{-3}$	20	2	2	2	2
Kx_MPT3	sandstone, shale	$1.0x10^{-4}$	10	0.02	0.02	0.02	0.02
Kx_MPT4	shale, mudstone	$1.0x10^{-3}$	10	1	1	1	1
Kx_LPT	mudstone, claystone	$1.0x10^{-5}$	$1.0x10^{-1}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$
Rch1		$4.2x10^{-5}$	$7.0x10^{-5}$	$5.6x10^{-5}$	$5.6x10^{-5}$		
Rch2		$2.1x10^{-4}$	$3.5x10^{-5}$	$2.8x10^{-4}$	$2.8x10^{-4}$		
R1		$2.8x10^{-4}$	$4.6x10^{-4}$			$3.7x10^{-4}$	$3.7x10^{-4}$
R2		$2.2x10^{-4}$	$3.6x10^{-4}$			$2.9x10^{-5}$	$2.9x10^{-5}$
R3		$3.4x10^{-4}$	$5.8x10^{-4}$			$4.6x10^{-4}$	$4.6x10^{-4}$
R5		$1.8x10^{-4}$	$3.0x10^{-4}$			$2.4x10^{-4}$	$2.4x10^{-4}$
R11		$7.5x10^{-6}$	$1.2x10^{-5}$			$1.0x10^{-5}$	$1.0x10^{-5}$

"K" represents hydraulic conductivity values, and "Rch" and "R" represents recharge rate values. Empty space indicates that the parameter was not present in a model.

Table 3-4 Initial values and upper and lower limits for parameter estimated in the calibration for Hydrogeologic model H2

Parameter (m/d)	Rock description	Lower bound	Upper bound	H2R1B1	H2R1B2	H2R2B1	H2R2B2
Kx_Qa	sand, gravel, and clay	$1.0x10^{-3}$	10	0.1	0.1	0.1	0.1
Kx_MPT1	shale, mudstone	$1.0x10^{-4}$	10	$5.0x10^{-3}$	$5.0x10^{-3}$	$5.0x10^{-3}$	$5.0x10^{-3}$
Kx_MPT2	sandstone, siltstone	$1.0x10^{-2}$	20	2	2	2	2
Kx_MPT3	Sandstone	$1.0x10^{-4}$	1	0.1	0.1	0.1	0.1
Kx_MPT4	Mudstone, claystone	$1.0x10^{-3}$	20	1	1	1	1
Kx_MPT5	sandstone and shale	$1.0x10^{-4}$	10	0.5	0.5	0.5	0.5
Kx_LPT	shale and mudstone	$1.0x10^{-5}$	$1.0x10^{-1}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$
Rch1		$3.8x10^{-5}$	$6.3x10^{-5}$	$5.0x10^{-5}$	$5.0x10^{-5}$		
Rch2		$1.5x10^{-4}$	$2.5x10^{-5}$	$2.0x10^{-4}$	$2.0x10^{-4}$		
R1		$3.0x10^{-4}$	$5.0x10^{-4}$			$4.0x10^{-4}$	$4.0x10^{-4}$
R2		$7.5x10^{-7}$	$1.3x10^{-6}$			$1.0x10^{-6}$	$1.0x10^{-6}$
R3		$1.9x10^{-4}$	$3.1x10^{-4}$			$2.5x10^{-4}$	$2.5x10^{-4}$
R5		$3.4x10^{-4}$	$5.6x10^{-4}$			$4.5x10^{-4}$	$4.5x10^{-4}$
R11		$7.5x10^{-5}$	$1.3x10^{-4}$			$1.0x10^{-4}$	$1.0x10^{-4}$

"K" represents hydraulic conductivity values, and "Rch" represents recharge rate values. Empty space indicates that the parameter was not present in a model.

Table 3-5 Initial values and upper and lower limits for parameter estimated in the calibration for Hydrogeologic model H3

Parameter (m/d)	Description	Lower bound	Upper bound	H2R1B1	H2R1B2	H2R2B1	H2R2B2
Kx1_Qa	Sand, gravel, and clay	$1.0x10^{-3}$	10	0.1	0.1	0.1	0.1
Kx_MPT1	sandstone, siltstone	$1.0x10^{-3}$	10	0.3	0.3	0.3	0.3
Kx_MPT2	sandstone, siltstone	$1.0x10^{-3}$	80	4.5	4.5	4.5	4.5
Kx_MPT3	sandstone, shale	$1.0x10^{-3}$	5	$5x10^{-3}$	$5x10^{-3}$	$5x10^{-3}$	$5x10^{-3}$
Kx_MPT4	sandstone, siltstone	$1.0x10^{-3}$	10	2	2	2	2
Kx_MPT5	siltstone and mudstone	$1.0x10^{-3}$	10	1	1	1	1
Kx_LPT1	shale, mudstone	$1.0x10^{-6}$	1	$6.5x10^{-3}$	$6.5x10^{-3}$	$6.5x10^{-3}$	$6.5x10^{-3}$
Kx_LPT2	Mudstone, claystone	$1.0x10^{-6}$	1	$3.0x10^{-3}$	$3.0x10^{-3}$	$3.0x10^{-3}$	$3.0x10^{-3}$
Rch1		$1.4x10^{-4}$	$2.4x10^{-4}$	$1.9x10^{-4}$	$1.9x10^{-4}$		
Rch2		$4.4x10^{-6}$	$7.4x10^{-5}$	$5.9x10^{-5}$	$5.9x10^{-5}$		
R1		$3.0x10^{-4}$	$5.0x10^{-4}$			$4.0x10^{-4}$	$4.0x10^{-4}$
R2		$1.9x10^{-5}$	$3.3x10^{-5}$			$2.5x10^{-5}$	$2.5x10^{-5}$
R3		$1.9x10^{-4}$	$3.1x10^{-4}$			$1.0x10^{-4}$	$1.0x10^{-4}$
R5		$3.4x10^{-4}$	$5.6x10^{-4}$			$2.7x10^{-4}$	$2.7x10^{-4}$
R11		$7.5x10^{-7}$	$1.3x10^{-6}$			$1.0x10^{-6}$	$1.0x10^{-6}$

"K" represents hydraulic conductivity values, and "Rch" and "R" represents recharge rate values. Empty space indicates that the parameter was not present in a model

3.7 Calibration Results

Results of model calibration were evaluated both quantitatively and qualitatively. Quantitative analysis is conducted to examine the quality of the calibration in term of goodness-of-fit between observations and simulations. Qualitative analysis is conducted to help clarify how the flow system behaves for different conceptualizations. Typically, groundwater models are considered calibrated if they meet the following conditions: (1) the non-linear regression converges with minimized sum of squares weighted residuals (SSWR), (2) the regression-determined parameter values are within the range of pre-defined reasonable values, (3) parameters are not excessively correlated, (4) the simulated hydraulic-property distribution within hydrogeologic units is reasonable for the model scale, and (5) positive and negative model residuals are randomly distributed (Hill, 1998).

The performance of the calibrated models were evaluated using the match between observed and simulated hydraulic heads in terms of goodness-of-fit to the goals as follows (proposed by ESI, 2007).

- The absolute residual mean (ARM) divided by range in head for all targets should be less than 0.1 (10%)
- The residual standard deviation divided by range in head for all targets (also called normalized root mean squared error (NRMS)) should be less than 0.1 (10%)

The residual is calculated as the observed head minus the model-computed head. In this study, the range in heads is approximately 50 meters. Therefore, the ARM and residual standard deviation should be less than 5 meters. The calibration statistics from of the 12 models are presented in Table 3-6. The table shows the sum of squared weighted

residuals (SSWR), residual mean, absolute residual mean, and residual standard deviation (Res.Std. Dev).

The statistics in term of goodness-of-fit for the 12 TPA calibrations meet and greatly exceed the calibration goals described previously. The absolute residual means divided by the total ranges in heads are only 2.1 to 2.6%; the standard deviations divided by the total ranges are only 2.9 to 3.4%. The residual means and absolute residual mean in each model layer for 12 TPA models are listed in Table 3-7 and 3-8, respectively. The residuals show a small negative bias in layer 1 for all models. Model H2R2B1 has the smallest bias in layer 1, 5, 6, and 11. Model H1R2B1 has the smallest absolute residual mean in layer 3, 4, and 6. The distribution of the smallest residual mean and absolute residual mean are found only in hydrogeologic models H1 and H2. It is shown that hydrogeological models H1 and H2 perform better in simulated hydraulic heads in each model layer than those of hydrogeologic models H3.

Table 3-6 Calibration statistics for 12 alternative models in the TPA

Model	SSWR (m^2)	Residual Mean (m)	Abs. Res Mean (m)	Res. Std. Dev (m)	Abs. Res Mean Divided by Range	Std.Dev Divided by Range
H1R1B1	144.5	0.02	1.04	1.39	0.022	0.029
H1R1B2	150.7	0	1.08	1.42	0.022	0.030
H1R2B1	145.4	0.01	1.06	1.39	0.022	0.029
H1R2B2	151.6	0	1.07	1.42	0.022	0.030
H2R1B1	189.8	0.03	1.13	1.59	0.025	0.033
H2R1B2	190.0	-0.03	1.15	1.59	0.024	0.033
H2R2B1	142.1	0.05	1.03	1.41	0.021	0.029
H2R2B2	178.1	0.03	1.23	1.13	0.024	0.033
H3R1B1	198.1	0.08	1.22	1.62	0.025	0.034
H3R1B2	196.6	-0.09	1.21	1.62	0.025	0.034
H3R2B1	181.3	-0.34	1.23	1.52	0.026	0.032
H3R2B2	187.0	-0.034	1.23	1.56	0.026	0.032

Table 3-7 Residuals mean (m) by model layer for the 12 alternative models in the TPA

Model	Model layer						
	1 (5*)	2 (35*)	3 (18*)	4 (10*)	5 (3*)	6 (3*)	11 (1*)
H1R1B1	-0.43	0.19	0.27	**0.02**	0.30	-1.29	-2.58
H1R1B2	-0.48	0.13	0.23	0.06	0.25	-1.18	-2.37
H1R2B1	-1.30	-0.10	0.50	0.23	-0.57	-0.38	1.87
H1R2B2	-1.31	0.18	0.33	0.15	0.44	-1.24	-2.52
H2R1B1	-0.30	0.08	0.23	0.43	0.83	-1.25	-2.88
H2R1B2	-0.83	**0.03**	0.51	0.14	0.95	-1.69	-3.71
H2R2B1	**-0.12**	-0.15	0.60	0.70	**-0.22**	**-0.36**	**0.68**
H2R2B2	-1.05	-0.22	**-0.17**	-0.09	0.79	-1.27	-4.16
H3R1B1	-0.91	-0.50	-0.04	0.03	0.83	-1.28	-3.35
H3R1B2	-1.80	-0.58	0.38	0.16	0.67	-1.52	-3.16
H3R2B1	-1.38	-0.56	0.20	0.34	0.62	-1.05	-2.48
H3R2B2	-1.19	-0.48	0.20	0.25	0.59	-1.24	-2.28

* represents number of head observation in each model layer

Table 3-8 Absolute residuals mean (m) by model layer for the 12 models in the TPA

Model	Model layer						
	1 (5*)	2 (35*)	3 (18*)	4 (10*)	5 (3*)	6 (3*)	11 (1*)
H1R1B1	1.24	1.02	0.90	1.40	1.00	1.74	2.58
H1R1B2	1.31	1.02	0.93	1.36	**0.74**	1.74	2.37
H1R2B1	1.60	1.14	**0.84**	**1.03**	1.70	**0.76**	1.87
H1R2B2	1.48	0.98	0.98	1.27	1.57	1.53	2.52
H2R1B1	2.07	0.96	1.28	1.39	2.49	1.69	2.88
H2R1B2	1.08	0.93	1.15	1.48	3.39	2.23	3.71
H2R2B1	**1.08**	1.06	1.40	1.58	1.50	1.41	**0.68**
H2R2B2	1.15	**0.85**	1.02	1.46	3.46	2.78	4.16
H3R1B1	2.10	1.22	1.65	1.11	2.50	1.69	3.35
H3R1B2	2.12	1.14	1.36	1.21	2.00	1.52	3.16
H3R2B1	1.87	1.16	1.31	1.05	2.11	1.05	2.48
H3R2B2	1.63	1.17	1.27	1.19	1.78	1.28	2.28

* represents number of head observation in each model layer

A more detailed assessment of calibrated model error can be done by examining the spatial distribution of head residuals. Figures 3-17 shows contours of hydraulic heads simulated in the TPA, and the spatial distributions of head residuals. The simulated heads obtained from the twelve models are consistent with observed groundwater levels and have a similar pattern of spatial distribution in the northern area where the density of observation wells are high. However, there are some differences in head contours in the southeast area where the elevations are high and the data is too limited to replicate the actual hydraulic heads. The simulated hydraulic head contours from model layer 3 illustrate the general direction of horizontal ground-water movement in the Phu Thok aquifers, which mimic the topography. Figure 3-18 is a scatter plot of head residuals and contours of hydraulic heads in model layer 3 simulated by the four models that have smallest SSWR values (shown in Table 3-6). Similar scatter plots and hydraulic head contours are shown in Appendix A. These plots show that positive and negative residuals in all models are quite randomly distributed, indicating no significant model bias.

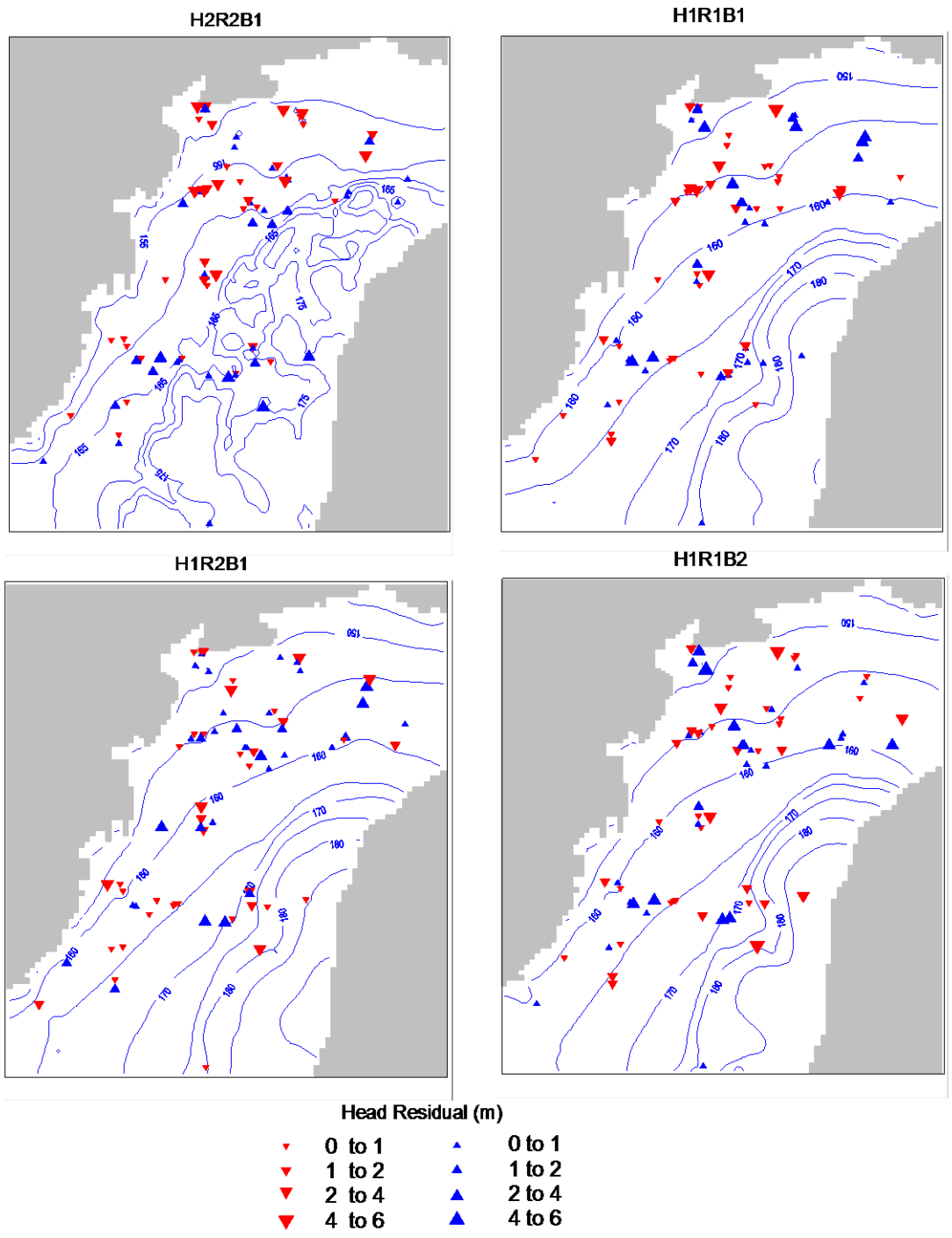


Figure 3-17 Scatter plots of heads residual and contours of hydraulic heads at model layer 4 in TPA simulated by the smallest SSWR models (H2R2B1, H1R1B1, H1R2B1, and H1R1B2)

Figure 3-18 shows the calibration plots of 75 observed heads versus simulated heads obtained from the four models that have smallest SSWR values. For other models, the scatter plots are shown in Appendix B. These plots illustrate that calibrations of the TPA are reasonable for the model scale, although there are a few outliers. Of the 75 targets, only a few in err by more than 5 meters. The calibration results demonstrate that it is possible to obtain an almost identical calibration performance for all 12, and thus these models can be accepted and used for predictive simulations.

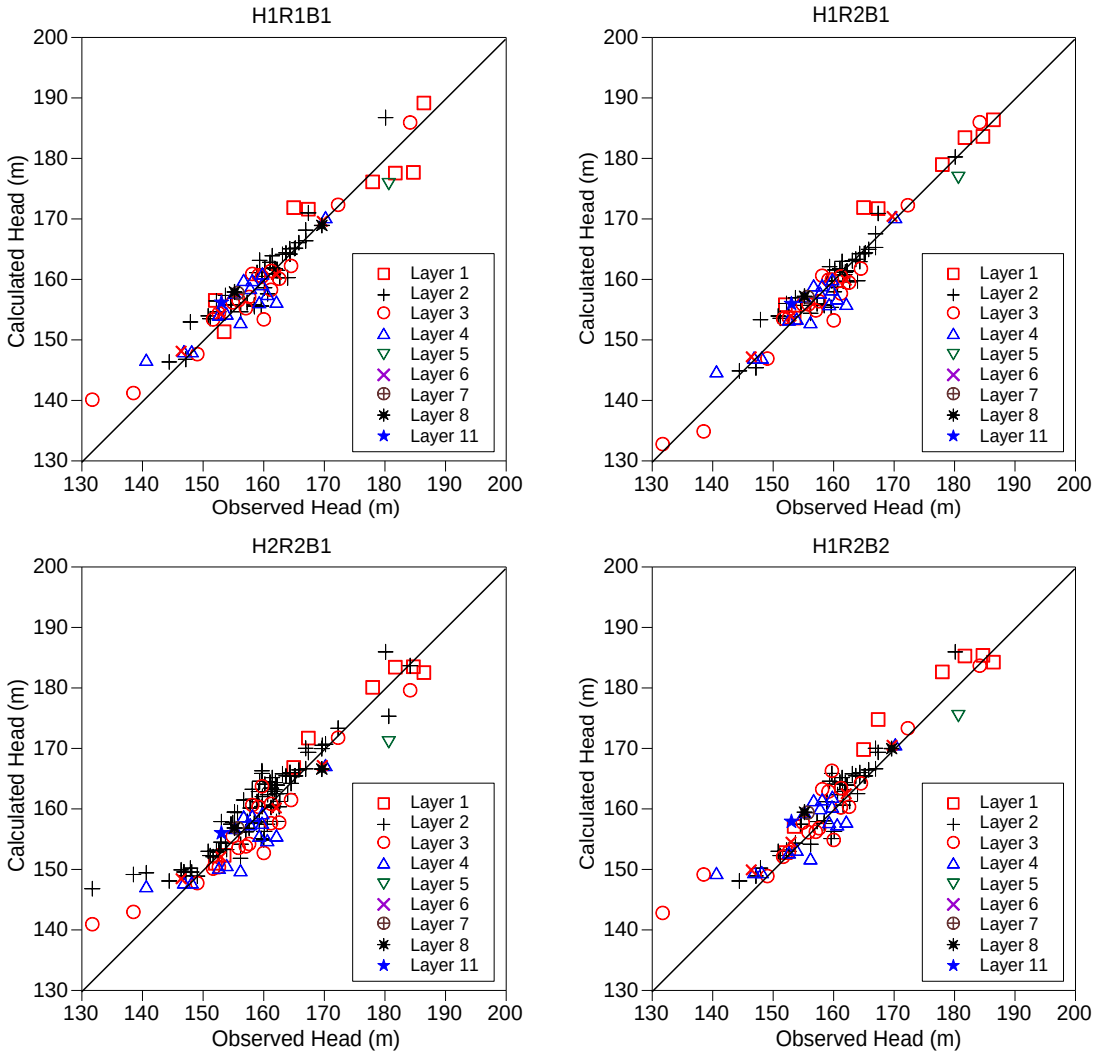


Figure 3-18 Observed vs. simulated heads for the smallest SSWR models (H1R1B1, H1R2B1, H2R2B1, and H1R1B2)

The simulated mass balances for the steady-state conditions for all models are presented in Table 3-9. The percentage discrepancy in mass balance is very small (less than 0.01 % for all models), indicating that the simulation results are accurate in term of mass balance. Most of the inflow to the model (72-99 %) comes from precipitation recharge. The remaining inflow is from the Chi River and Huai Toei reservoir losses. About three quarters of the total discharge in the model is baseflow to the Chi River. About 15-30 % of the discharge is pumping from wells. Overall, the withdrawal rate is still much lower than the rate of replenishment. The recharges in hydrogeologic model H2, especially for model H2R2B1 and H2R2B2, are higher than those in hydrogeologic model H1 and H3. However, there is no clear evidence which is the most impacted by uncertainties among these hydrostratigraphic, recharge, and boundary condition models.

Table 3-9 Simulated mass balance for the steady-state condition

Model	Rate into the model (m^3/d)				Rate out of the model (m^3/d)			
	Recharge	River Leakage	GHB	Total In	Well	River Leakage	GHB	Total Out
H1R1B1	24,809	98		24,907	6,625	18,281		24,906
H1R1B2	25,813	167	3,021	29,001	6,604	15,224	7,218	29,046
H1R2B1	24,893	188		25,081	6,604	18,452		25,056
H1R2B2	27,762	102	5,402	33,266	6,572	22,870	3,823	33,265
H2R1B1	26,058	44		26,102	6,625	19,474		26,099
H2R1B2	27,036	37	8,248	35,321	6,625	19,546	9,150	35,321
H2R2B1	39,193	159		39,352	6,625	32,727		39,352
H2R2B2	35,754	47	13,450	49,251	6,617	24,484	18,151	49,252
H3R1B1	22,220	15		22,235	6,582	15,653		22,235
H3R1B2	20,696	33	17,496	38,225	6,625	28,278	4,158	39,061
H3R2B1	21,930	34		21,964	6,625	16,140		22,765
H3R2B2	22,914	67	18,253	41,234	6,604	31,479	3,163	41,246

The optimized parameters estimates for the 12 models are listed in Tables 3-10 to 3-12. The final values of hydraulic conductivity parameters generally are within the range of

available field measurements in all models. The estimated parameter values and their 95% confidence intervals are presented graphically in Figures 3-19 to 3-21.

Table 3-10 Calibrated parameter values for hydrogeologic model H1 (unit: m/d)

Parameter	Field Estimate	H1R1B1	H1R1B2	H1R2B1	H1R2B2
Kx_Qa	0.01-5.5	0.69	1.46	0.3	1.41
Kx_MPT1	0.001-5	0.04	0.04	0.04	0.04
Kx_MPT2	0.1-80	1.67	1.54	1.32	2.7
Kx_MPT3	0.01-3.5	0.051	0.03	0.13	0.06
Kx_MPT4	0.35-13.5	1.26	1.73	8.2	1.97
Kx_LPT	$4.0x10^{-5}$-0.04	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$
Rch1		$3.6x10^{-6}$	$6.12x10^{-5}$		
Rch2		$1.81x10^{-4}$	$1.70x10^{-4}$		
R2				$6.11x10^{-6}$	$1.66x10^{-5}$
R11				$2.25x10^{-5}$	$1.94x10^{-5}$
R3				$1.00x10^{-4}$	$4.50x10^{-4}$
R5				$3.23x10^{-4}$	$4.93x10^{-4}$
R1				$5.0x10^{-4}$	$5.00x10^{-4}$

Table 3-11 Calibrated parameter values for hydrogeologic model H2 (unit: m/d)

Parameter	Field Estimate	H2R1B1	H2R1B2	H2R2B1	H2R2B2
Kx_Qa	0.01-5.5	0.044	0.38	0.05	1.02
Kx_MPT1	0.001-5	0.74	0.006	$1.00x10^{-2}$	$2.22x10^{-2}$
Kx_MPT2	0.1-80	1.755	1.836	1.833	1.946
Kx_MPT3	0.01-5	0.022	0.027	0.012	0.013
Kx_MPT4	0.001-3.5	0.008	$8.25x10^{-3}$	0.008	0.01
Kx_MPT5	0.35-13.5	3.531	3.83	8.02	8.95
Kx_LPT	$4.0x10^{-5}$-0.04	$1.00x10^{-4}$	$1.00x10^{-4}$	$1.0x10^{-4}$	$1.00x10^{-4}$
Rch1		$4.34x10^{-5}$	$1.21x10^{-5}$		
Rch2		$2.09x10^{-4}$	$2.30x10^{-4}$		
R1				$6.00x10^{-4}$	$3.95x10^{-4}$
R2				$2.17x10^{-5}$	$1.50x10^{-6}$
R3				$2.01x10^{-4}$	$1.92x10^{-4}$
R5				$5.00x10^{-4}$	$4.75x10^{-4}$
R11				$1.00x10^{-4}$	$1.40x10^{-4}$

Table 3-12 Calibrated parameter for hydrogeologic model H3 (unit: m/d)

Parameter	Field Estimate	H3R1B1	H3R1B2	H3R2B1	H3R2B2
K_Qa	0.01-5.5	0.10	0.12	0.11	0.13
K_MPT1	0.001-5	0.41	0.27	0.09	0.04
K_MPT2	0.1-80	5.66	3.04	5.09	5.00
K_MPT3	0.001-0.5	$2.94x10^{-3}$	$6.13x10^{-3}$	$3.91x10^{-3}$	$3.14x10^{-3}$
K_MPT4	0.5-10	3.56	4.43	3.84	3.79
K_MPT5	0.35-13.5	1.79	0.67	1.74	1.93
K_LPT1	$4.0x10^{-5}$-0.4	$6.47x10^{-3}$	$1.10x10^{-3}$	0.15	$2.97x10^{-3}$
K_LPT2		$3.28x10^{-3}$	$4.0x10^{-3}$	$6.04x10^{-4}$	$6.53x10^{-6}$
Rch1		$5.90x10^{-5}$	$7.70x10^{-5}$		
Rch2		$1.97x10^{-4}$	$1.23x10^{-4}$		
R1				$5.00x10^{-4}$	$4.97x10^{-4}$
R2				$2.52x10^{-5}$	$2.52x10^{-4}$
R3				$8.44x10^{-5}$	$7.68x10^{-5}$
R5				$2.83x10^{-4}$	$2.92x10^{-4}$
R11				$1.00x10^{-6}$	$1.00x10^{-6}$

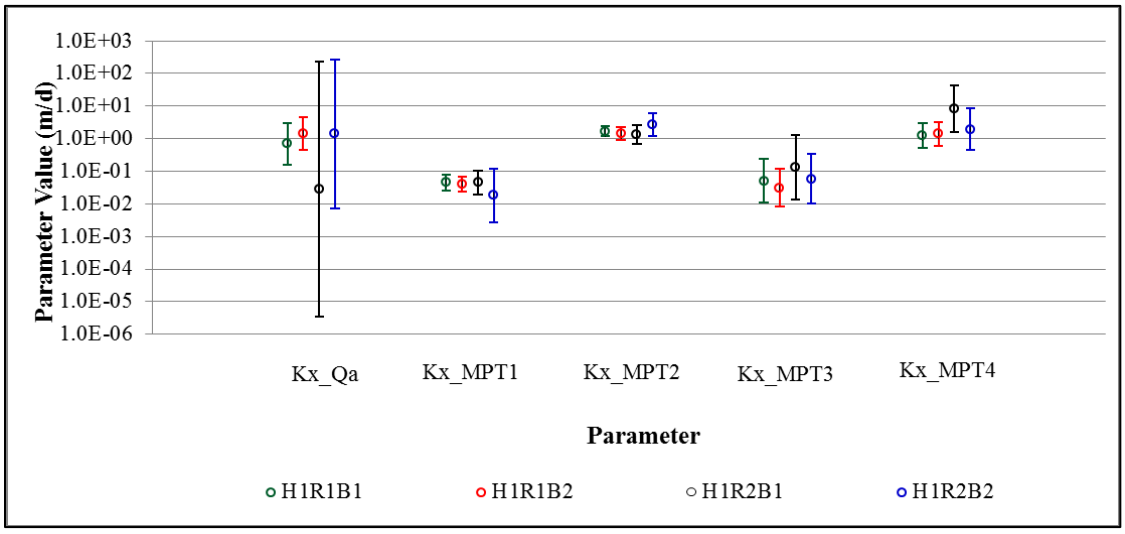


Figure 3-19 Calibrated parameter values and 95% upper and lower linear confidence limits for hydrogeologic models H1

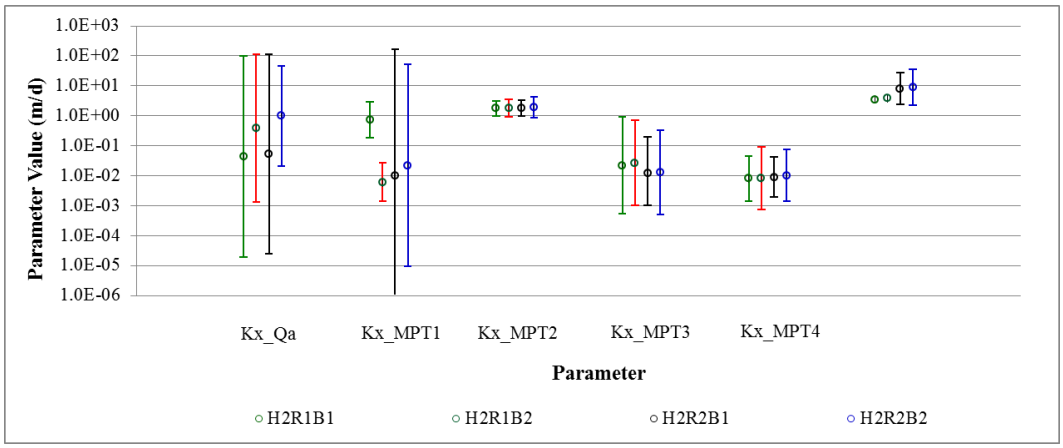


Figure 3.20 Calibrated parameter values and 95% upper and lower linear confidence limits for hydrogeologic models H2

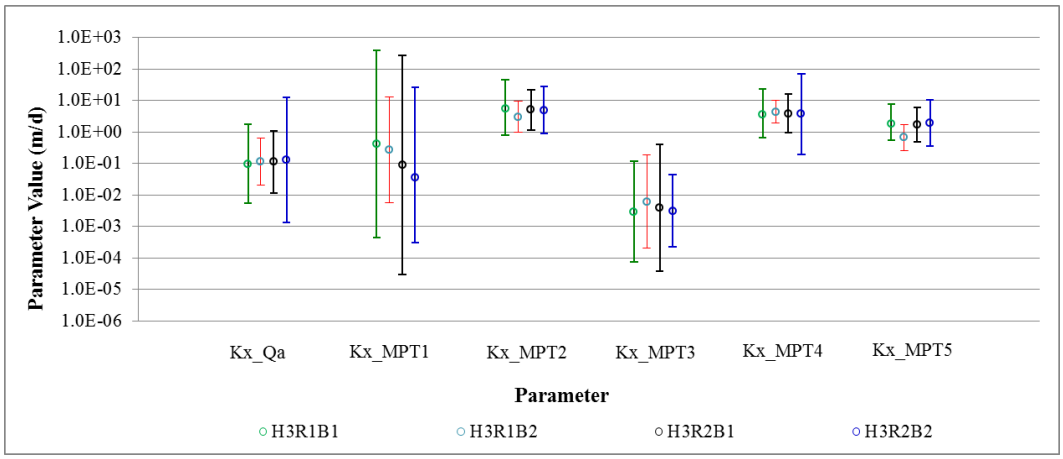


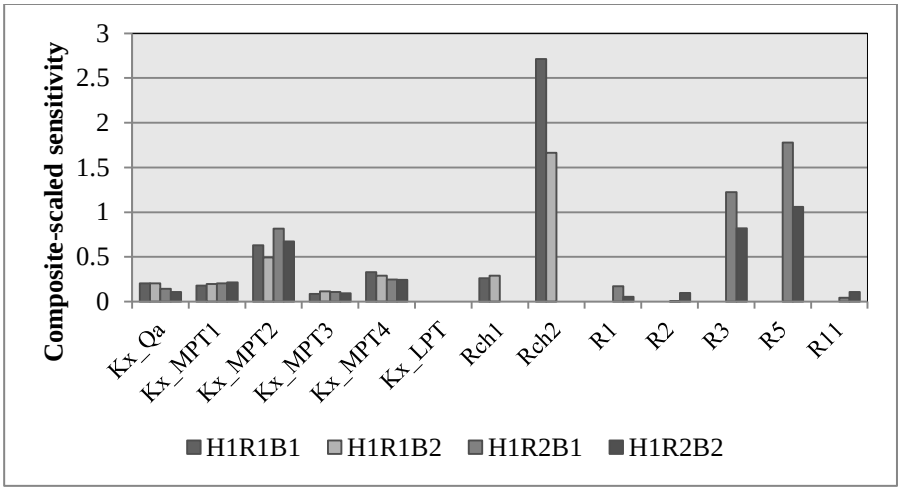
Figure 3-21 Calibrated parameter values and 95% upper and lower linear confidence limits for hydrogeologic models H3

The hydraulic conductivity values of Qa and MPT1 have the widest confidence limits (models H1 and H2 for Qa, and models H2 and H3 for MPT1). . The hydraulic conductivity of MPT2, MPT4, and MPT5 are well determined and slightly different for all model and their 95 % confidence intervals are relatively small. The estimated parameters for hydraulic conductivity values are in good agreement of field measurements. The calibrated recharge rates are all physically plausible and the recharge rates in each zone are vary from 0.5 mm/yr. to 182.5 mm/yr.

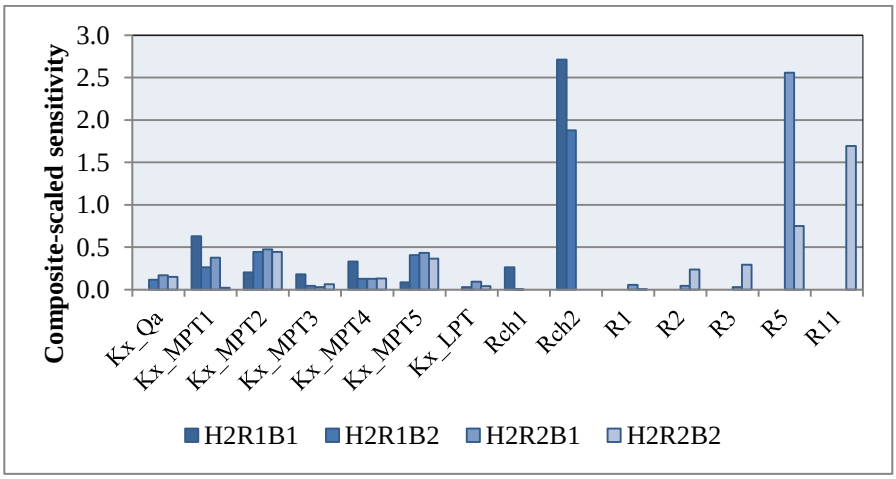
3.8 Sensitivity Analysis

The sensitivity of model-simulated heads and flows to parameters was used to assist model calibration. Composite-scaled sensitivities (CSS) can be used to assess the importance of each parameter to the calculation of model-simulated heads and flows (Hill and Tiedeman, 2007). Parameters with larger CSS have greater importance and greater influence on the model simulation than parameters with smaller CSS. Generally, model observations provide enough information to estimate parameters that have CSS greater than 1 (Hill and Tiedeman, 2007). The CSS values can be obtained as part of the automatic calibration process in PEST and are shown in Figure 3-22.

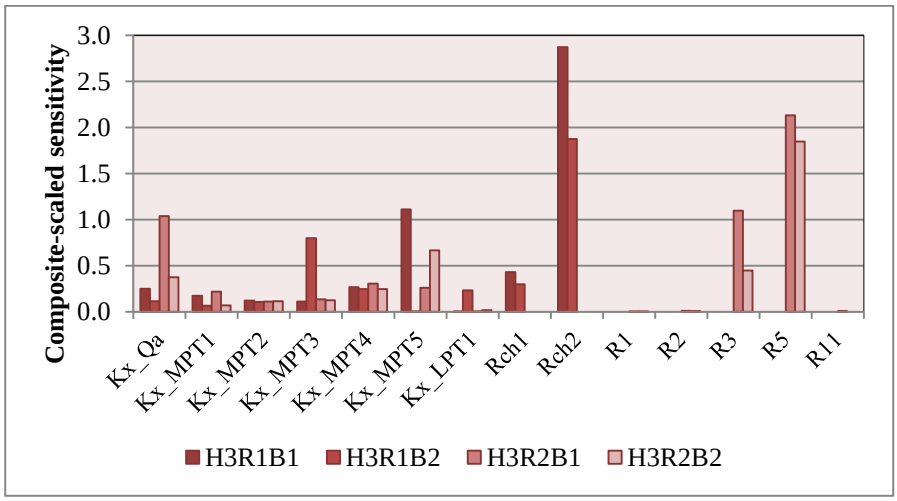
The highest CSS values were observed for the recharge parameters (Rech2, R3, and R5) and thus indicate that simulated heads at observation locations are more sensitive to recharge than to hydraulic conductivity parameters (Figure 3-19). Only parameters Kx_Qa, Kx_MPT3, and Kx_MPT5 for Hydrogeologic model H3 have CSS values greater than 1. The small CSS values for many hydraulic conductivity parameters may be due to the distribution of head observations located mainly in the top three model layers and thus provide insufficient information to estimate more than half of the parameters used. In addition, head observations alone are not enough information for the estimation of each parameter (commonly the only calibration data available in data-poor aquifer systems).



(a)



(b)



(c)

Figure 3-22 Composite-scaled sensitivity of model parameters to head observations for the steady state of the TPA for (a) hydrogeologic models H1 (b) hydrogeologic models H2, and (c) hydrogeologic models H3

Parameter correlation coefficients (PCC) can be used with composite scaled sensitivities to yield a useful sensitivity analysis. The PCC values are calculated for each pair of model parameters to indicate whether parameter values can be estimated uniquely, given the constructed model and the observation provided. The PCC values can varies from -1.00 to 1.00. If the PCC for a pair of parameters is equal to or very close to -1.00 or 1.00, the two parameters cannot be determined uniquely. If the absolute values of PCC are less than approximately 0.95, then it is likely that all parameters can be estimated uniquely (Hill, 1998; Hill and Tiedeman, 2007). Maximum (PCC calculated by PEST using the final parameter values for all models are shown in Table 3-13. Based on the above 0.95 guideline, PCC values are all less than this indicating that all model parameters can be uniquely identified.

Table 3-13 Maximum absolute parameter correlation coefficient for final parameter values calculated by PEST

Parameter	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2	H3R1B1	H3R1B2	H3R2B1	H3R2B2
Kx_Qa	0.68	0.55	0.72	0.76	0.74	0.69	0.61	**0.82**	0.38	0.89	0.69	0.77
Kx_MPT1	0.30	0.23	0.33	0.47	0.53	0.73	0.73	0.39	0.70	0.51	0.57	0.32
Kx_MPT2	0.37	0.32	0.64	0.64	**0.84**	**0.85**	**0.83**	0.76	0.33	0.24	0.45	0.29
Kx_MPT3	0.47	0.76	0.49	0.47	0.58	0.57	0.34	0.47	**0.89**	**0.94**	**0.87**	0.64
Kx_MPT4	0.59	0.65	**0.78**	**0.83**	0.30	0.41	0.46	0.55	0.70	0.40	0.46	0.42
Kx_MPT5					0.74	0.23	0.34	0.27	0.83	0.42	0.69	**0.80**
Kx_LPT					0.84	0.16	0.30	0.33	**0.89**	**0.94**	**0.87**	0.54
Kx_LP2												
Rch1	**0.89**	**0.84**			0.74	0.69			0.70	0.54		
Rch2	**0.89**	**0.84**			**0.84**	**0.85**			0.83	0.89		
R1			0.72	0.76			0.73	0.47			0.43	0.64
R2			0.72	0.77			**0.83**	**0.82**			0.84	0.46
R3			0.62	0.58			**0.83**	0.71			0.48	0.71
R5			**0.78**	**0.83**			**0.83**	0.76			0.57	**0.80**
R11			0.64	0.77			0.63	0.71			0.54	0.64

3.9 Model Averaging

Several methods for assigning model weights have been developed in the hydrologic and statistics literature. The fundamental goal of model-averaging is to get, by weighing predictions from a set of plausible models based on their relative performance, an averaged prediction that is "most representative" of actual system response. These weighting approaches typically require model calibration results and prior model probabilities for computing posterior model probabilities or model weights. In this study, two approaches are implemented based on 1) the GLUE framework (Beven and Binley, 1992) and 2) information criteria approaches (Neuman, 2003; Poetor and Anderson, 2005). In addition to the formal criteria approaches, a combined multi-criteria approach is proposed in Section 3.9 to evaluate the set of alternative models and thus use the available information efficiently.

Following the GLUE method (Beven and Binley, 1992), the likelihood function of GLUE does not consider the principle of parsimony; only model fit is used to evaluate model probability. In this study, the likelihood metric is based on the variance of the residual between the measured and computed hydraulic head (SSWR) and can be calculated as the inverse of SSWR. Model weights are normalized so that their sum is equal to one; therefore, the GLUE weight for model *k* can be calculated as

$$\omega_k(GLUE) = \frac{p(M_k)SSWR_k^{-N}}{\sum_{k=1}^{K} p(M_k)SSWR_k^{-N}} \qquad (3\text{-}1)$$

where $p(M_k)$ is the prior probability for model *k*; $SSWR_k$ is the sum of squared weighted residuals for model *k* with $k = (1,\ldots,K)$, *K* is the total number of models being considered, and *N* is a shape factor. N=1 is normally used in GLUE applications and is used here.

For the information criteria, posterior model probabilities or model weights of the twelve models can be calculated as (in Neuman et al., 2003)

$$w_k = \frac{\exp(-0.5\Delta IC_k)\, p(M_k)}{\sum_k \exp(-0.5\Delta IC_k)\, p(M_k)} \qquad (3\text{-}2)$$

where $\Delta IC_k = IC_k - IC_{min}$ with IC_k being any of the information criteria described in section 2.7 for a given model M_k; IC_{min} is the minimum value obtained from model M_k with $k = (1,\ldots,K)$; and $p(M_k)$ is the prior probability for model k. IC-based criteria are calculated using Equations 2-13, 2-14, 2-18, and 2-19, respectively.

Both the GLUE and information criteria approaches contain the prior model probability term ($p(M_k)$). Various methods for assigning prior model probability can be found in the Bayesian model literature. Prior model probability distribution can be calculated based on the modelers' expert judgment, which reflects their beliefs regarding the relative plausibility of each model and considers their consistency with available data and their understanding of the aquifer system. Alternatively, a uniform prior model probability distribution can also be assigned, reflecting no prior preference on the plausibility of the multiple conceptual models (Meyer et al, 2007: Rojas et al., 2008a).

In this study, uniform prior model probability was employed. For the 12 alternative models, an equal prior probability of 1/12 or 8.3% was assigned to each model. Table 3-14 shows model weights calculated from different criteria for the 12 alternative models. The model weights for SSWR using the GLUE approach are more uniformly distributed than with other approaches and receiving weights within a range of 0.07-0.1. When using GLUE or the SSWR metric, very little preference is shown for one model over another. The information criteria, on the other hand, show strong preference for a single model. In the case of AIC, AICc, and BIC, most of the model weight is assigned to model H1R1B1 with the weights of 0.87, 0.94, and 0.85, respectively, indicating that other models will make only insignificant contributions in the ensemble model-averaging. In the case of

KIC, the model weight is assigned only to model H1R1B2 (0.99), indicating that only model H1R1B2 will contribute to the predictive distribution estimation.

Table 3-14 Model weights and ranks for different model criteria

Model	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2	H3R1B1	H3R1B2	H3R2B1	H3R2B2
$p(M_k)$[a]	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12
SSWR[b]	144.5	150.7	145.4	151.6	189.8	190.0	142.1	178.1	198.1	196.6	181.3	187.0
Nk[c]	11	12	14	15	12	13	15	16	12	13	15	16
AIC	75.4	81.4	84.7	90.9	98.7	101.7	86.1	106.2	101.9	104.2	104.3	109.9
AICc	71.2	76.3	77.7	82.8	93.6	95.7	77.9	96.9	96.8	98.3	96.2	100.5
BIC	69.8	74.8	75.9	80.9	92.1	94.1	76.1	94.9	95.3	96.7	94.3	98.5
KIC	68.7	59.3	104.1	118.3	79.4	79.9	123.2	152.6	60.6	87.7	86.2	95.4
GLUE wts	0.097	0.093	0.097	0.093	0.074	0.074	0.099	0.079	0.071	0.071	0.078	0.075
AIC wts	0.867	0.066	0.034	0.003	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000
AICcwts	0.939	0.047	0.009	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000
BIC wts	0.850	0.069	0.040	0.003	0.000	0.000	0.038	0.000	0.000	0.000	0.000	0.000
KIC wts	0.009	0.991	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SSWR Rank	2	4	3	5	9	10	1	6	12	11	7	8
AIC Rank	1	2	3	5	6	7	4	10	9	11	8	12
AICc Rank	1	2	3	5	6	7	4	10	9	11	8	12
BIC Rank	1	2	3	5	6	7	4	9	10	11	8	12
KIC Rank	2	1	9	10	3	4	11	12	7	6	5	8

a) $p(M_k)$ = prior model probability b) SSWR=sum of squared weighted residual
c) Nk= number of parameters)

3.10 Model validation

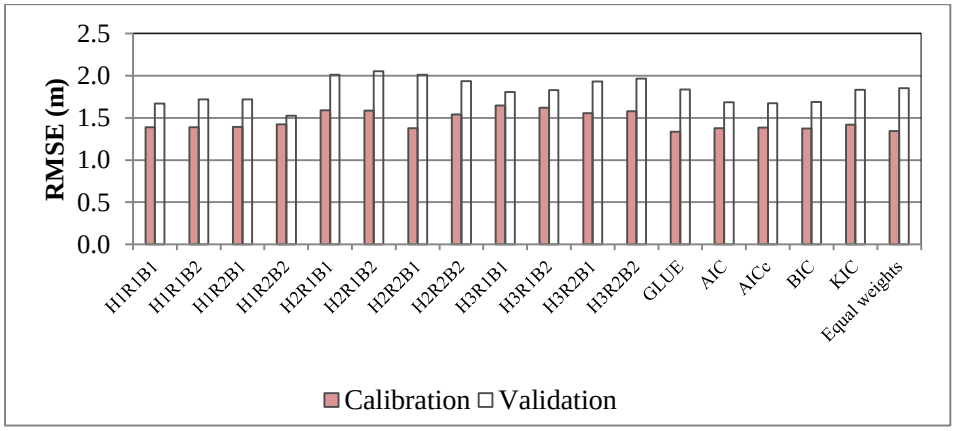
Model performance during calibration is normally used as a measure of the predictive capability of each model. However, calibration is not capable of testing the performance of the model in predictive mode. Incorrect simulation models can still be adequately calibrated. It is clear that multiple models can yield comparable results. The capability of these models to make predictions should additionally be evaluated with independent data

not used for calibration. The performance of each of our calibrated models and different model-averaging methods were compared using model predictions against a validation data set to determine their ability to accurately predict system response.

In this study, two years (1998-1999) of data were used for model calibration, and two other years (2007–2008) were subsequently used to evaluate the prediction performance. The groundwater pumping, recharge rates and river stage conditions for 2007-2008 were averaged to represent the steady state conditions. Based on available data for the 2007-2008 period, average head measurements from 40 observation wells were used to compare the predictive performance of these models. The root mean squared errors (RMSE) and absolute residual mean (ARM), commonly used for evaluating the accuracy of deterministic predictions (Duan et al., 2007; Diks and Vrugt, 2010), were used to evaluate the accuracy of head predictions and associated prediction uncertainty bounds using the individual models and different model-averaging methods

Generally, it is desirable to achieve calibrated models that are consistent and perform well in validation. Figure 3-23 shows the RMSE and ARM for each model and model averaging predictions obtained from the calibration and validation periods. There is some degradation between the validation and calibration period for this case study in terms of performance measures of RMSE and ARM statistics. The correlation between calibration and validation periods show no consistent relationship between calibration and validation prediction; the best model in the calibration period in terms of RMSE and ARM (H2R2B1) does not produce the most accurate prediction in the validation period; in turn, it produces the second worst result in validation period. Model H1R2B2 produces the fifth rank in calibration, but it is the first during validation testing. Considering model performance in both periods, the H1 hydrogeologic models are likely the most reliable predictors and the H3 hydrogeologic models are the worst.

(a)



(b)

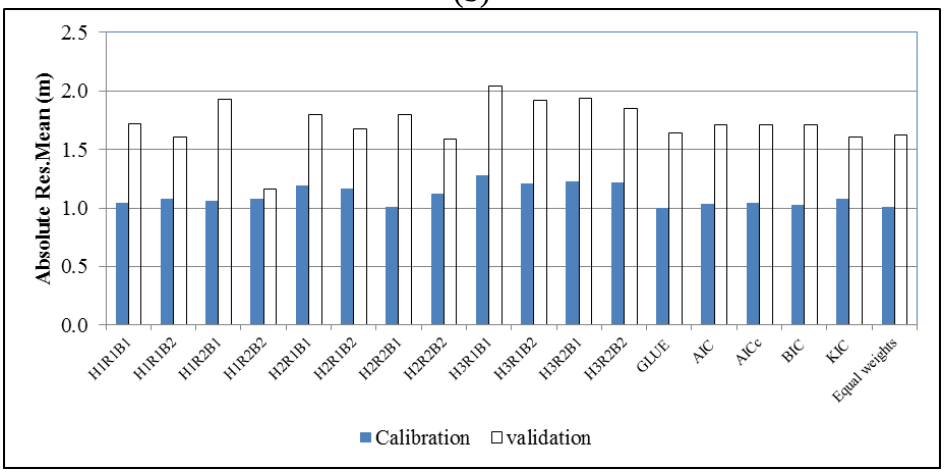


Figure 3-23 Comparisons of predictive performance between calibration and validation periods for alternative models and model-averaging methods using (a) root mean squared error and (b) absolute residual mean measure

Most model averaging methods produce closer predictions to the calibration data than do the individual models. None of the model averaging methods produce predictions that are closer to the validation head data than model H1R2B2; however, their prediction quality is significantly increased as compared to the majority of the individual models. GLUE, KIC, and equally weighted model-averaging produce results close to that of the most accurate individual model and produce predictions that are better than the other model-averaging methods.

The calibration and validation results show that the different models have different strengths, so it is difficult to select a single best model that matches all observations well. Using model-averaging methods, the contribution of each model differs from one set of observations to another depending on their performance in term of the measure employed. Ensemble models contribute to model predictions in each part of a domain based on their weights; therefore, their jointly made predictions do not underestimate the predictive uncertainty as the single best model does. They are likely to make better predictions than most single models, or at least obtain prediction performance levels that are close to the best performing model. An interesting observation that may be made is that simple model averaging methods, such as equal weight or GLUE averaging methods, perform just as well as a sophisticated MLBMA method such as KIC.

Apart from measures that indicate the predictive capabilities of a model such as RMSE and ARM, an alternative performance metric is its predictive coverage, that is, the percentage of observed data that fall within a given prediction interval around predicted values (Wintle et al., 2003; Vrugt et al., 2009, Morales-Casique et. al., 2010). Prediction intervals for the predictions of any single model and model averaging methods are computed with the point-wise standard deviation obtained from PEST software on the assumption that the head residual is normally distributed and set for a confidence level of 90%. Ninety percent prediction intervals for each model and model-averaging method can be calculated in the following way:

$$\text{The 90\% prediction intervals on } X = X \pm z\sigma \qquad (3\text{-}3)$$

where X is the calculated head for each calibrated model; σ is the standard deviation of the model prediction; and z is the standard normal random variable associated with the prediction interval selected (for a 90% prediction interval z=1.645).

Figure 3-24 shows the predictive coverage (percentage of head observations contained in the 90% prediction intervals) and the width of the prediction intervals for the alternative

models and different model-averaging methods in the validation period. Typically, the wider the prediction interval, the greater the uncertainty, and the less we can infer about the true value. Therefore, the desirable approach for predictions is to capture most observations with the narrowest prediction intervals. In the validation period, head predictions at most locations are fairly accurate and were captured by the 90% prediction intervals. As seen in Figure 3-24, the average 90% predictive width for individual models range from 5 to 7.1 m and for different model-averaging methods it ranges between 5.5 and 6.0 m. The percentage of observations contained in 90% prediction intervals for individual models vary from 87.5 to 90 and for model-averaging methods vary from 87.5 to 90. Lower or higher percent coverage is an indication of too narrow or too wide uncertainty intervals.

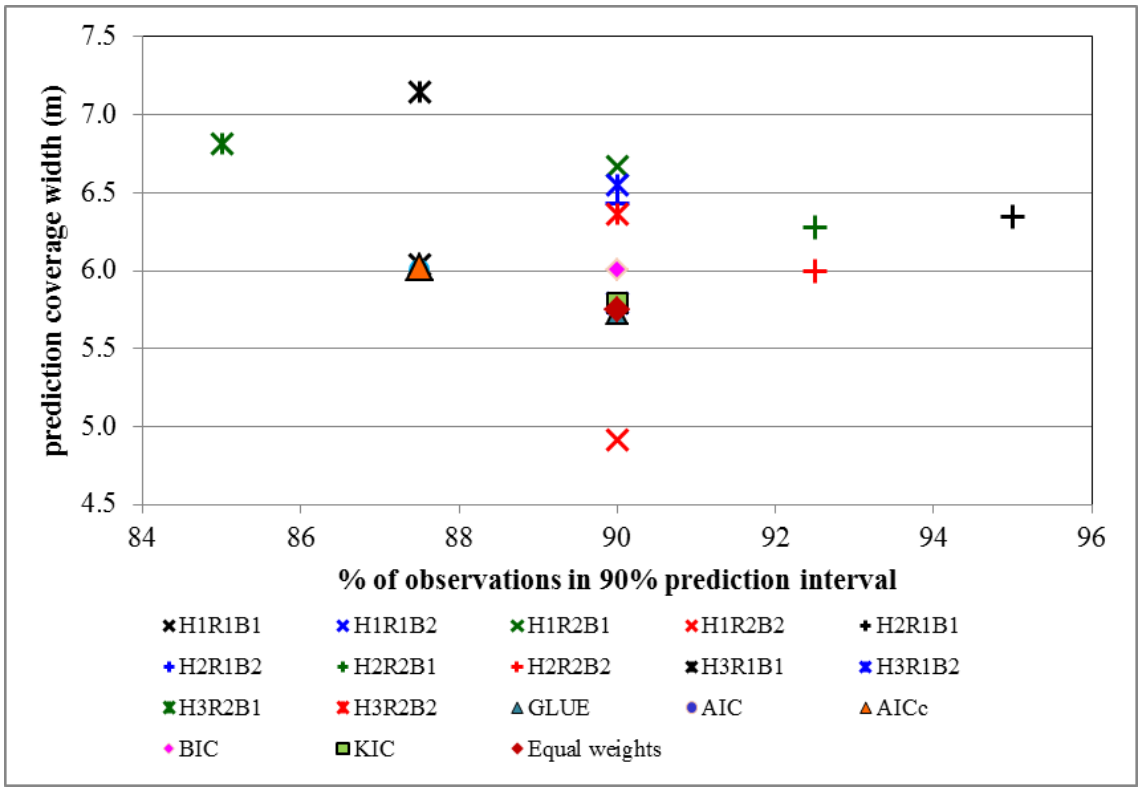


Figure 3-24 Percentages of head predictions captured by 90% prediction intervals and average 90% prediction intervals for alternative models and different model averaged during validation period.

However, due to the small number of data points in this study, every point that falls outside the prediction intervals has a significant impact on the percent coverage (±2.5%). Among individual models, model H1R2R2 has the best predictive coverage contains in 90% prediction bounds (a desirable characteristic for head forecasting) and the narrowest predictive coverage width and Model H3R1B1 has the worst. When averaged, all model averaging methods yielded slightly more appropriate predictions than the individual models, and the AIC and AICc models yielded the highest predictive coverage of all the model averaging techniques, with coverage of 91%. Among model averaging methods, KIC and equal weight are the best because they produce the narrowest predictive interval and the highest percent coverage.

Model H1R2R2 clearly outperforms the model averaging methods in the validation as the prediction interval width is smaller and achieves the same predictive coverage; however, it does not predict better than the model-averaging methods do in the calibration period. Even if model H1R2B2 achieves results closest to the head observations in both calibration and validation, it does not guarantee that it can perform well in other predictions; it is neither a correct nor true model since the specification of a correct model is an unachievable task (George, 1999). Although, model results can achieve consistency with observations both in the present and past, it does not guarantee that the model will perform at the same level when used to predict the future (Oreskes et al., 1994). Therefore, using a single best model is likely not adequate to quantify conceptual model uncertainty.

Comparing the predictive capabilities of a model such as RMSE and ARM values (seen in Figure 3-20), it makes clear that all model-averaging methods except KIC perform better than any individual model in the calibration period. Although no model-averaging method performs better than the best individual predictors in validation period, they all perform better than most individual models. These results likely indicate the strength of

implementing model-averaging in prediction, as this method produces results that are better or at least close to the predictions made using the best model.

Due to the limited data available predictions obtained from both periods are inevitably uncertain. Therefore, while it does not appear to have been done in the literature, it is reasonable and likely warranted to combine the weights from both prediction periods. Moreover, different model performance metrics rank all models differently. It is suggested that multiple model performance metrics, including those generated during validation, should be used in model ranking.

Typically, model averaging approaches have the prior model probability term ($p(M_k)$) shown in Equation (3-1) and (3-2), allowing for the subjective preference likelihood of experts to be included with quantifiable probability measures when computing posterior model probabilities or model weights. Prior model probabilities can be assigned using the subjectivity in the expert's knowledge of the site or expert elucidation (Pohlmann, et al., 2007; Ye, 2010; Singh et al., 2010). In this study, validation results were used instead of expert judgment to evaluate the alternative model uncertainty. The inclusion of validation results can give informative prior probabilities of the models. It is expected that using informative probabilities can obtain more accurate prediction results.

The validation results are aggregated into prior model probabilities using the GLUE method as follows:

$$p(M_k) = \frac{SSWR^*_k{}^{-N}}{\sum_{k=1}^{K} SSWR^*_k{}^{-N}} \qquad (3\text{-}4)$$

where $p(M_k)$ is the prior model probability; $SSWR^*$ is the sum of squared weighted residuals in the validation period, and K is the number of alternative models.

Aggregated prior model probabilities are plotted in Figure 3-25. Model H1R2B2 has the highest prior weight, and model H3R1B1 has the lowest weight, indicating that they receive the highest and lowest confidence, respectively.

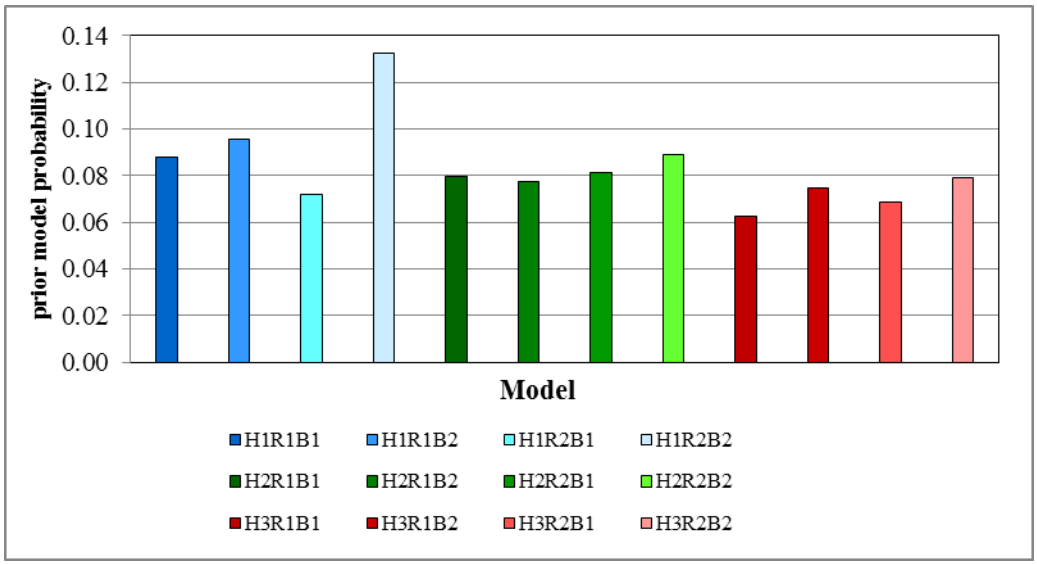


Figure 3-25 Prior model probabilities using validation results of the 12 alternative models

Table 3-15 lists posterior model probabilities for two sets of prior model probabilities. In the first set, all models are treated as equally likely, and each model has a prior probability of 8.3 percent. In this case, posterior probability is solely determined by the quality of model fit measured by the SSWR. In the second set, the prior model probabilities obtained from validation results are combined with the calibration results to calculate the posterior model probabilities using Equation (3-1) and (3-2). It is shown that KIC gives a weight of 0.99 (shown in bold in the table) and gives near zero weights to the others. Thus, there is no effect from incorporating the validation results between different alternative models because only one model includes after the model-averaged weighting (the validation weights are in effect multiplied by zero for all except one of the models). For AIC, AICc and BIC, there is also very slight impact from including validation results

between different models. The influence of including subjective probabilities is shown when using GLUE weighting. Model H1R2B2 with received fifth rank weight (0.93) in calibration results is now received the highest weight (1.43) when including subjective probability. The weights shown in Table 3-15 are then used to evaluate the uncertainty in the sustainable yield estimates in the next chapter.

Table 3-15 Weighting for 12 alternative models

Model	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2	H3R1B1	H3R1B2	H3R2B1	H3R2B2
GLUE[1]	0.097	0.093	0.097	**0.093**	0.074	0.074	0.099	0.079	0.071	0.071	0.077	0.075
GLUE[2]	0.101	0.105	0.082	**0.146**	0.070	0.068	0.095	0.083	0.053	0.063	0.063	0.070
AIC[1]	0.867	0.066	0.034	0.003	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000
AIC[2]	0.868	0.072	0.028	0.004	0.000	0.000	0.028	0.000	0.000	0.000	0.000	0.000
AICc[1]	0.939	0.047	0.009	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000
AICc[2]	0.937	0.051	0.007	0.001	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000
BIC[1]	0.850	0.069	0.040	0.003	0.000	0.000	0.037	0.000	0.000	0.000	0.000	0.000
BIC[2]	0.852	0.075	0.033	0.005	0.000	0.000	0.035	0.000	0.000	0.000	0.000	0.000
KIC[1]	0.009	**0.991**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KIC[2]	0.008	**0.992**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

1) weights with equal priors

2) weights with unequal priors obtained from validation test

3.11 Multiple Criteria Decision Making Approach for Model-averaging

Traditionally MCDM approaches are used to find the best alternative in a set of decisions. Here, the basic approach of MCDM is instead used to generate a model quality metric that can be used either for selection or for model weighting. The primary objective of this thesis is to determine how best to use multiple conceptual models to quantify

model prediction uncertainty. We are forced to either select the best model or somehow aggregate models to inform policy. Even using the information criteria and GLUE (using SSWR likelihood), it is still unclear which are the best model and/or metric to use and how to screen for the best model. One approach we can use is MCDM. MCDM approaches have been used to help select the best strategy from a number of potential alternatives in water resources planning and management (Shafike et al., 1992; Bose and Bose, 1995; Tkach and Simonovic, 1997). An MCDM approach can help decision-makers to compare, rank, and organize the available choices (in this case, alternative simulation models) based on identified evaluation criteria. This methodology can evaluate and rank alternative models based on either their model fit performances or a comprehensive set of subjective and objective measures.

As shown in previous sections, evaluating multiple models using different measures such as SSWR, and ARM can yield different ranking and weights. Given incomplete knowledge of a system and limited available data for the modeler, the reasonable conclusion is that the debates about conceptual models exist in response to the judgments and conceptualizations made as part of a simulation model's development. Therefore, it is reasonable to develop a set of performance measures, instead of relying on a single criterion. Ultimately, having a set of evaluation criteria will help user select and weigh multiple models.

The set of performance metrics for comparing and ranking multiple models are based two groups of data: those generated through calibration and validation. The measures used here are as follows

Calibration

1. SSWR for the entire area
2. SSWR for the critical area
3. ARM for the entire area
4. ARM for the critical area

Validation

5. SSWR for the entire area
6. SSWR for the critical area
7. ARM for the entire area
8. ARM for the critical area

Two weighting methods are used to calculated model weights: GLUE and the fuzzy set method. The concepts from fuzzy set theory (Zedeh, 1965) were used to assess qualitatively each one of the performance measures, based on the level of satisfaction gained from different criteria. A fuzzy MCDM methodology has been used by Kuchenur (2006) to assist the selecting the best among identified alternative models and weighing model performances.

The specific quantification of model performance criteria is not always possible, nor is it always necessary. When the values of variables cannot be precisely specified and are better qualified as high or low, sufficient or insufficient, or good or fair, such variables are assumed to be uncertain or fuzzy. If the values are uncertain, probability distributions using fuzzy membership functions can be used to quantify them. Both probability distributions and fuzzy membership functions of these uncertain or qualitative variables can be included in quantitative evaluation. The form or shape of a membership function depends upon individual subjective judgments; in this study a linear membership is assumed.

As done in Kuchenur (2006), the fuzzy membership function for each performance metric is here calculated using the formula below.

$$Fuzzified\ PM_i = \frac{Max(PM)}{Max(PM) - Min(PM)} - \frac{PM_i}{Max(PM) - Min(PM)} \qquad (3\text{-}5)$$

where *Max(PM)* is the maximum value of the performance metric in all models
Min(PM) is the minimum value of the performance metric in all models

PM_i is the value of the performance metric i

Table 3-16 summarizes all 12 models' preference with respect to the eight identified metrics. The smallest number represents the greatest likelihood model. As shown in Table 3-16, no one model performs best for all evaluation criteria. For example, model H2R2B1 has the best calibration statistics but not the best validation statistics. The best model identified during the calibration process is not the best in the model validation. Moreover, the best model identified for the entire area may not produce the best prediction in the area of interest.

Table 3-16 Summarizes all 12 models' preference with respect to the identified metrics

Model	Performance metric							
	1 (m^2)	2 (m^2)	3 (m)	4 (m)	5 (m^2)	6 (m^2)	7 (m)	8 (m)
H1R1B1	144.5	1.04	78.2	1.21	82.3	1.77	134.7	1.59
H1R1B2	150.7	1.08	80.8	**1.24**	76.8	1.66	124.0	1.49
H1R2B1	145.4	1.06	79.3	1.18	93.2	**1.91**	164.3	1.75
H1R2B2	151.6	1.07	80.6	1.21	**57.2**	**1.28**	**89.2**	**1.12**
H2R1B1	189.8	1.19	89.4	1.08	82.7	1.82	148.9	1.65
H2R1B2	190.0	1.17	86.5	1.09	75.7	1.68	152.7	1.60
H2R2B1	**142.1**	**1.01**	**75.9**	**1.06**	73.2	1.68	145.7	1.63
H2R2B2	178.1	1.13	84.4	1.23	74.9	1.50	132.9	1.47
H3R1B1	**198.1**	**1.28**	91.4	1.22	**95.0**	1.89	**188.6**	**1.85**
H3R1B2	196.6	1.21	**94.4**	1.19	77.8	1.73	158.6	1.70
H3R2B1	181.3	1.22	91.9	1.22	94.3	1.84	171.6	1.74
H3R2B2	187.0	1.22	91.6	1.20	78.8	1.69	149.5	1.62
Max	198.1	1.28	94.4	1.24	95.0	1.91	188.6	1.85
Min	142.1	1.01	75.9	1.06	57.2	1.28	89.2	1.12

Remarks: 1=SSWR for the entire area; 2= SSWR for the critical area; 3= ARM for the entire area; 4=ARM for the critical area; 5=SSWR for the entire area; 6= SSWR for the critical area; 7= ARM for the entire area; 8 =ARM for the critical area
1-4 metrics for calibration period; and 5-8 for validation period

A significant advantage of this methodology over the formal criteria is that it can combine both quantitative and qualitative criteria to evaluate and rank a model. Moreover, it is flexible and can incorporate stakeholder preference into the analysis. Each metric can be ranked in different scales. A high score can be assigned to a high preference metric and a low score can be assigned to a low preference. In this study, an equal score is assigned for all metrics. The models were ranked based on calculations conducted using Equation 3-5; the results are shown in Table 3-17.

Table 3-17 Summary of all 12 model weights with respect to the identified metrics

Model	Performance metrics								Total weight
	1	2	3	4	5	6	7	8	
H1R1B1	0.17	0.14	0.15	0.04	0.07	0.06	0.11	0.09	**0.103**
H1R1B2	0.15	0.12	0.13	0.00	0.10	0.10	0.13	0.12	**0.106**
H1R2B1	0.16	0.13	0.14	0.08	0.01	0.00	0.05	0.03	**0.076**
H1R2B2	0.14	0.12	0.13	0.04	0.21	0.25	0.20	0.25	**0.167**
H2R1B1	0.03	0.05	0.05	0.21	0.07	0.04	0.08	0.07	0.074
H2R1B2	0.03	0.07	0.07	0.20	0.11	0.09	0.07	0.08	0.090
H2R2B1	0.17	0.16	0.17	0.24	0.12	0.09	0.09	0.07	**0.140**
H2R2B2	0.06	0.09	0.09	0.02	0.11	0.16	0.11	0.13	**0.098**
H3R1B1	0.00	0.00	0.03	0.03	0.00	0.01	0.00	0.00	0.009
H3R1B2	0.00	0.04	0.00	0.06	0.10	0.07	0.06	0.05	0.048
H3R2B1	0.05	0.03	0.02	0.02	0.00	0.03	0.03	0.04	0.030
H3R2B2	0.03	0.04	0.03	0.06	0.09	0.09	0.08	0.08	0.061
Sum	1	1	1	1	1	1	1	1	1

Using fuzzy MCDM methodology, model H1R2B2 performed best with respect to the identified performance criteria (but only nominally) and received a weight of 0.16. Overall, H1 hydrogeologic models always perform the best, whereas the H3 ones always perform worse than the others. MCDM model weights calculated using fuzzy set and GLUE methods are compared in Table 3-18. They are some differences in model weights between the two methods. Model H1R2B2 receives weight of 0.167 and 0.103 when

calculated using Fuzzy Set and GLUE weights, respectively. And H3 hydrogeologic models get much smaller weights of Fuzzy than those of GLUE weights.

Table 3-18 MCDM model weights using fuzzy set and GLUE methods

Model	Fuzzy weights	GLUE weights
H1R1B1	0.103	0.086
H1R1B2	0.106	0.088
H1R2B1	0.076	0.081
H1R2B2	0.167	0.103
H2R1B1	0.074	0.080
H2R1B2	0.090	0.082
H2R2B1	0.140	0.089
H2R2B2	0.098	0.086
H3R1B1	0.009	0.072
H3R1B2	0.048	0.078
H3R2B1	0.030	0.075
H3R2B2	0.061	0.080
Sum	1	1

3.12 Impact of Ensemble Size on Performance of Model Averaging

In this section, SSWR performance or SSWR weights were used to test the impact of ensemble size on performance of model averaging prediction. The two, four, and six models which received the smallest SSWR weights have sequentially been removed from the twelve model ensemble (ensemble A). Therefore, there are 10, 8, and 6 models remaining in the ensembles A1 to A3, respectively. The members of these ensembles are shown in Table 3-19.

In contrast, the three ensembles B1 to B3 were created by sequentially removing two models which received the largest and second largest SSWR weights from the ensemble model A. Consequently, ensemble B1, B2, and B3 consist of 10, 8, and 6 models, respectively. Table 3-19 lists new model weights of the different ensemble sizes. These model weights for different ensemble were used to predict the hydraulic heads in calibration and validation periods as described previously in section 3.4 and section 3.9, respectively.

Table 3-19 Summary model weights for different ensemble size

	Model weights						
	GLUE	A1	A2	A3	B1	B2	B3
H1R1B1	0.097	0.113	0.137	0.174			
H1R1B2	0.093	0.109	0.131	0.167	0.116		
H1R2B1	0.097	0.113	0.136	0.173	0.120		
H1R2B2	0.093	0.108	0.131	0.166	0.115	0.151	
H2R1B1	0.074	0.086			0.092	0.120	0.167
H2R1B2	0.074	0.086			0.092	0.120	0.167
H2R2B1	0.099	0.115	0.139	0.177			
H2R2B2	0.079	0.092	0.111	0.141	0.098	0.128	
H3R1B1	0.071				0.088	0.115	0.160
H3R1B2	0.071				0.089	0.116	0.161
H3R2B1	0.077	0.090	0.109		0.096	0.126	0.175
H3R2B2	0.075	0.088	0.106		0.093	0.122	0.170

Table 3-20 summarizes statistics of the ensemble prediction performance generated using information contained in ensembles A1 to A3 and B1 to B3. The quality of the fit between observed heads and ensemble predicted weighted average heads generally decrease with decreasing size of the ensemble when the best model with respect to the

SSWR were excluded from the ensemble (see model ARM, RMSE and Coverage in validation all degrade moving from ensemble B1 to B3). In contrast, the performance measures for ensembles A1 to A3 are more stable regardless of whether the lowest weighted 2, 4 or 6 models were not included in the ensemble. For example, the least stable performance metric considering ensembles A1 to A3 is the coverage in validation (decreasing from 87.5% to 82.5%).

Table 3-20 Summary statistics for predictive performance in different ensemble models.

Model Performance Metric	Ensemble number						
	A	A1	A2	A3	B1	B2	B3
ARM, calibration (m)	1.08	1.06	1.05	1.08	1.08	1.12	1.17
ARM, validation (m)	1.47	1.43	1.44	1.4	1.46	1.43	1.57
RMSE, calibration (m)	2.23	2.19	2.15	2.23	2.23	2.44	2.64
RMSE, validation (m)	3.03	2.95	2.95	2.83	3.00	2.94	3.46
Coverage, calibration (%)	97.3	96.0	96.0	96.0	96.0	93.3	93.3
Coverage, validation (%)	90.0	87.5	85.0	82.5	77.5	67.5	65.0
Average width, calibration (m)	4.87	4.83	4.91	4.91	5.14	5.34	4.87
Average width, validation (m)	5.72	5.65	5.65	5.54	5.70	5.64	6.12

Remark: coverage and average width are given for the 90% prediction bounds.

In comparison to ensemble A (all twelve models in the ensemble), the results above demonstrate that model performance metrics for this case study are insensitive to small reductions (removing two models) in ensemble size – even if the models not considered were the two most highly weighted models. In addition, comparing ensemble A3 to ensemble A shows that as long as the highest weighted models are included in the ensemble, not including low weighted models (six in this case) in the ensemble has little impact on model performance metrics. Unfortunately it is not clear how this result can be generalized to other case studies.

3.13 Summary

This chapter has evaluated three sources of model uncertainty for groundwater flow modeling in the TPA. Uncertainty in the hydrogeological model, boundary conditions and recharge components results from different interpretations of geological and geophysical data, different postulation of boundary conditions, and different techniques for recharge estimation, respectively. These models were calibrated against the head observations. The calibration results demonstrate that these 12 models can attain an almost identical calibration performance, perhaps implying that head observations cannot clearly discriminate among different competing models.

Different model weighting approaches were used to rank the multiple conceptual models. Model weights obtained from information-based model averaging methods (AIC, AICc, BIC, and KIC) dominated in only one or two models, meaning that only one or two models are needed for making predictions. The GLUE or SSWR weights were more evenly distributed than those of IC-based model selection criteria; all models received weights within a range of 7 to 10%. The performance of each of the calibrated models and different model-averaging methods were compared using model predictions against different data sets to determine their performances in describing the system. Using model-averaging methods significantly increased the prediction performance, which were much better than those of the individual models. This finding indicates the strength of using model-averaging methods, which are able to make better predictions than is a single model, or at least to make predictions that are close to those of the best models. Importantly, the model averaging procedure in this case study is not sensitive to small changes in the ensemble size and the procedure is also not sensitive to larger changes in

the ensemble size (removing six of twelve models) so long as these larger changes involved ignoring the poorest quality models.

Based on the calibration and validation results, evaluating multiple models using different measures such as SSWR, and ARM can yield different ranking and weights. Specifically, IC-based metrics tend towards model selection rather than averaging, which may be misleading in the context of management. Given incomplete knowledge of a system and data-poor aquifer systems, there are bound to be discussions about the conceptual models that exist in response to the judgments and conceptualizations made as part of a simulation model's development. Thus, it is reasonable to develop a set of performance measures, instead of relying on a single criterion. Here, the fuzzy MCDM approach was used to rank and weigh the multiple conceptual models. This approach combined eight performance measures from both the calibration and validation processes. The model weights from various criteria are used in the next chapter in a model averaging method to quantify model prediction in groundwater management. In addition, the impact of ensemble size was evaluated with the SSWR weights. The results show that removal of the worst two models has a minimal effect but removal of the best two models result in model performance degradations, but it is unclear how much of this degrades is a function of the specific original size. It is demonstrated that there is little sensitivity of validation results to ensemble size. These imply that the impact of ensemble size on decision making would be the same as validation results.

Chapter 4
Conceptual Model Uncertainty in a Groundwater Management Context

4.1 Introduction

This study addresses the question of whether the use of multiple conceptual models improves the ability of numerical models to assist in management decisions. The models developed in Chapter 3 are used here in a management context, with the goal of using the collection of models to identify a sustainable pumping strategy for the TPA. The efficacy of model selection and averaging methods are assessed for whether they improve model predictive capacity over the use of either individual models or ensembles of alternative models.

The aim of this chapter is to assess various means by which multiple models can be used to quantify uncertainty and provide utility for groundwater management decisions. Different model-averaging methods (GLUE, MLBMA (BIC and KIC), AICMA (AIC, AICc), and the Multi-Criteria Decision Making approach (MCDM)) are assessed in terms of their ability to assist in quantifying uncertainty in groundwater management problems.

In this study, groundwater simulation models are linked with optimization techniques to estimate sustainable yields from the aquifer and identify the best management strategy, which in this case is appropriate allocation of pumping rights in areas at high risk from water resource over-exploitation.

In Chapter 3, a number of conceptual numerical models of the Phu Thok aquifer were developed and assessed purely in terms of calibration/validation performance. Here, these calibrated models are applied in a management context in an attempt to maximize total groundwater extraction rates while adhering to water level constraints, i.e., to determine the "sustainable yield" of the aquifer. Substantial uncertainty in the "best" management option exists because drawdowns depend on the unknown spatial distribution of hydrogeologic properties, recharge rates, and boundary conditions. Different conceptual models will therefore lead to different optimal pumping strategies.

Optimization models are implemented to determine appropriate groundwater abstraction rates in the TPA Phu Thok aquifer. The management models aim to obtain maximal yields while protecting against water level decline and prevent subsequent saline water upconing. The optimization models were solved using the MGO (Modular Groundwater Optimizer) developed by Zheng and Wang (2003). Here, a genetic algorithm (GA) is used to obtain the optimal extraction rates.

In section 4.2, the groundwater management model used for evaluating sustainable yield is introduced. In section 4.3, the sustainable yield results from the individual conceptual models are reported and contrasted. An attempt is made to holistically assess the range of results and determine how individual model quality metrics correlate to sustainable yield estimates. Various schemes for averaging and reporting management model results are evaluated, with particular attention paid to the utility of various model averaging approaches in the context of management.

4.2 Groundwater Management Problem

4.2.1 Formulation of the Groundwater Management Model

Groundwater sustainability can be defined as the development and use of groundwater to meet current and future proposes at rates that can be maintained indefinitely without causing unacceptable environmental, economic or social consequences (Alley et al., 1999). The definition of "unacceptable" is essentially subjective, depending upon the specific situations and most decision-makers tend to work within a finite period or project lifetime, as opposed to an indefinite timeframe, for which they wish to avoid unacceptable consequences. In this study, groundwater quality is not considered in the assessment of sustainability due to lack of data and little evidence of increasing brine upconing in the study area. Rather, "sustainable" groundwater development is here defined as a withdrawal rate from the aquifer that can be maintained indefinitely without causing depletion of groundwater levels below an acceptable level.

The Modular Groundwater Optimizer (MGO) code developed by Zhang and Wang (2003) and embedded in the Groundwater Vistas software was used to address the optimization problem. The MGO code was selected for its ability to solve a variety of optimization problems with highly non-linear and complex objective functions and to optimize continuous decision variables (pumping/injection rates). It is compatible with MODFLOW, as is required for this study. The MGO code uses a genetic algorithm (GA) in order to find the optimal or near-optimal pumping rates for selected wells in a MODFLOW model subjected to user-specified constraints. A number of researchers have demonstrated that the GA can yield satisfactory results when used to seek optimal solutions to complex problems (McKinney and Lin, 1994; Huang and Mayer, 1997; Aly and Peralta, 1999; Zheng and Wang, 1999; McPhee, 2004; Qahman et al., 2005).

The optimization models were formulated with the objective of maximizing water pumping from wells subject to (1) ensuring that the drawdowns at control locations do not exceed specified values and (2) limiting groundwater withdrawals to lower and upper bounds (well capacity constraints). Steady-state conditions were used in the analysis (rather than transient-conditions) since the maximized pumping rates are intended to represent the sustainable yield of the system, which is the rate that can be maintained indefinitely. Furthermore, groundwater in TPA is mainly exploited to meet industrial and municipal demands; therefore, future groundwater withdrawals are likely to be non-seasonal in nature. The sustainable yields are optimized based on the assumption that the future number and distribution of wells will be the same as in 2008, and that no new wells will be added to the system.

In these models, the decision variables are the withdrawal rates at 155 model cells corresponding to existing well locations. Figure 4-1 shows the location of pumping decision variables and constraints. There are a total of 155 wells and a total of 55 head control locations. These head control locations are selected based on current monitoring locations for future verification, and some clusters of wells were removed to minimize computation. The drawdown constraint is controlled via the lower bounds on hydraulic heads at these control locations. The regulations (or constraints) are that heads at control locations cannot decline more than 10 meters in pumping cells within model layers 2-4. In addition to the head constraints, the models were subjected to the upper and lower bounds for pumping wells. Here, the minimum pumping rates (Q_{min}) are set to 20 m^3/d to ensure that the pumping is occurring at each well for the entire period, and the values representing the upper bounds (Q_{max}) are set to 5 to 10 times the pumping rates from the year 2008.

The objective function of the optimization can be expressed as:

$$\text{maximize } Z = \sum_{i=1}^{N} Q_i \qquad (4\text{-}1)$$

subject to: well constraints

$$Q_{min} \le Q_i \le Q_{max} \qquad i=1,..,N \qquad (4\text{-}2)$$

head constraints

$$d_{j,k} \le D_j^* \qquad j=1,..,55 \qquad (4\text{-}3)$$

where

N	is the total number of pumping wells (N=155)
i	the pumping well index number
j	is the constraint location index number
k	is the model index number
$d_{j,k}$	is the drawdown at the control location j for model k, in meters;
D^*_j	is the maximum drawdown level at the control location j in meters;
Q_i	is the pumping rate from zone i (155 wells), in cubic meters per day;
Q_{min}, Q_{max}	are the lower and upper pumping rate bounds at the pumping location i, respectively, in cubic meters per day.

In the GA, these pumping rates can be satisfied by restricting the population space of the decision variable (pumping rates) to be within the above limits. The GA does not directly deal with head constraints; a penalty function is used here to handle these constraints. In the penalty method, a constrained problem is converted to an unconstrained one by attaching a penalty to constraint violations (Goldberg, 1989). The number of discretization intervals for each pumping rate and the number of simulations per optimization per optimization iteration or population size are set to the default values of 32 and 100, respectively. The uniform crossover method is used with a crossover probability set at 0.5, and the jump mutation probability is set to equal 0.01.

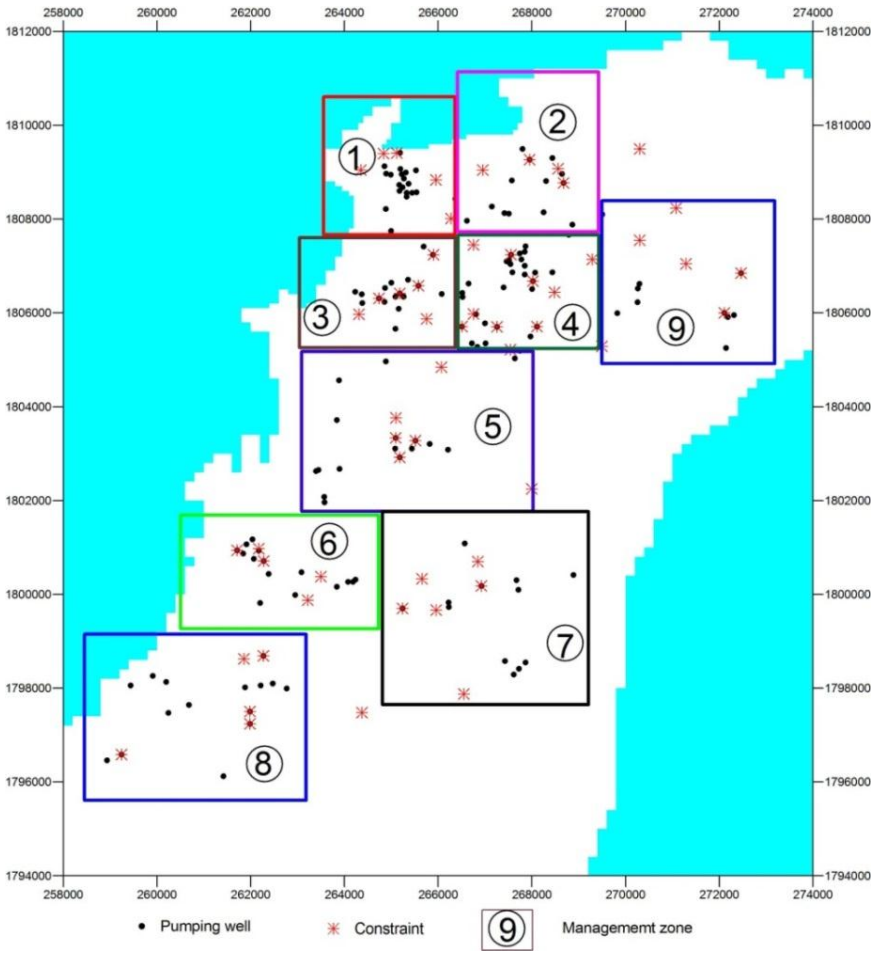


Figure 4-1 Location of optimized pumping wells and management areas

4.2.2 Optimization Model Results

The sustainable yields calculated by MGO from 12 alternative models are summarized by management area, as shown in Table 4-1 and Figure 4-2. Detailed sustainable yields for all pumping wells are shown in Appendix C. No constraint violations occur at the constrained wells for any of these models. For most conceptual models, the lower bound constraints on hydraulic heads are binding at some control locations. The small drawdowns at many head control locations may imply that additional pumping wells can likely be placed in those locations without violating the specified constraints.

Table 4-1 Sustainable groundwater yield (m^3/d) in the 9 management areas for 12 alternative optimization models

Area	H2R2B2	H1R2B1	H1R2B2	H2R1B1	H1R1B2	H1R1B1	H3R1B1	H2R2B1	H3R1B2	H3R2B2	H2R1B2	H3R2B1
1	3,588	4,197	3,949	6,316	5,858	3,345	3,865	3,804	3,358	4,203	4,995	3,245
2	3,661	3,419	3,561	2,554	3,419	2,868	2,265	2,432	3,422	2,048	2,633	2,413
3	3,669	3,197	2,310	3,916	2,923	4,151	3,273	2,817	3,132	3,088	2,297	2,733
4	3,278	3,484	3,086	2,064	3,584	2,503	1,491	2,000	1,542	2,126	1,587	2,123
5	2,419	3,164	2,864	2,587	1,880	1,309	1,768	1,774	1,384	1,736	1,593	2,313
6	4,545	2,694	3,084	2,910	1,235	3,320	2,338	2,587	1,477	1,626	1,272	974
7	3,548	2,758	3,080	1,555	3,230	3,358	3,665	1,652	2,590	2,207	2,394	1,871
8	3,626	3,828	3,690	3,503	3,058	3,219	3,642	2,742	2,519	2,990	2,142	2,539
9	3,211	1,626	2,659	1,845	1,907	1,732	1,139	2,259	2,558	1,771	1,084	1,087
Total	**31,545**	**28,366**	**28,283**	**27,251**	**27,093**	**25,807**	**23,445**	**22,068**	**21,983**	**21,795**	**19,997**	**19,297**

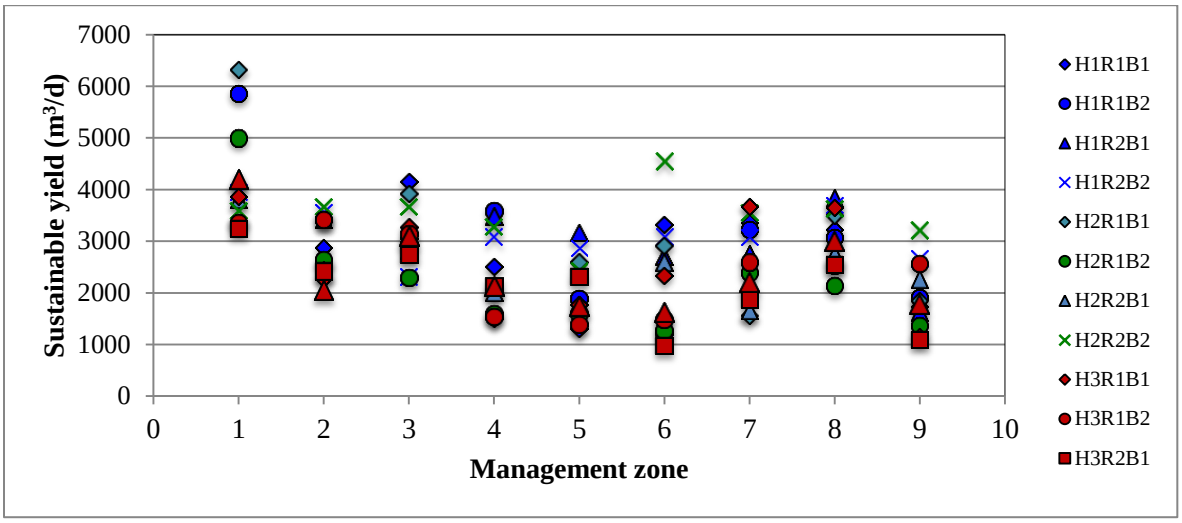


Figure 4-2 Sustainable yields by groundwater management area obtained from 12 alternative models

Despite similar performance among the models in calibration, total sustainable yield estimates vary substantially depending on the conceptual model used, and range from 19,300 to 31,500 m^3/d, i.e., by a factor of 60 %. Generally, the sustainable yield estimates from all hydrogeologic models H3 are less than those of hydrogeologic models H1 and H2. The estimated sustainable yields in each management area in Table 4-1 vary more widely than the total sustainable yields among different alternative conceptual models, in one case by as much as 350 % between models.

The more variability in the sustainable yield estimation, the wider the resultant prediction interval is likely to be for the ensemble and the more uncertainty is reflected. Assessing the reliability of an estimate depends not only on the size of its estimated prediction interval, but also on how large the prediction interval is relative to the estimate (mean value) itself. The relative prediction interval is the width of the prediction interval divided by the mean (Equation 4-4):

$$\text{Relative 90\% prediction interval} = \left(\frac{95^{th}\ Percentile - 5^{th}\ Percentile}{Mean} \right) \qquad (4\text{-}4)$$

where 95^{th} and 5^{th} percentile represent the upper and lower limits for the 95 % prediction interval, respectively.

Moreover, the coefficient of variation (CV), the ratio of the standard deviation to the mean, can be used to compare the variability of prediction intervals. Table 4-2 shows various statistics relates to sustainable yields.

Table 4-2 Statistics of sustainable yield estimates (m^3/d) from 12 optimization models in each management area

Area	Mean	Standard Deviation	Coefficient of variation (%)	Quartile1	Quartile3	5th Percentile	95th Percentile	Relative 90 % PI
1	4227	996	23.6	3530	4401	3300	6064	0.65
2	2891	572	19.8	2427	3420	2167	3606	0.50
3	3125	576	18.4	2796	3372	2304	4022	0.55
4	2406	767	31.9	1897	3134	1519	3529	0.84
5	2066	594	28.8	1700	2461	1350	2999	0.80
6	2338	1060	45.3	1426	2953	1118	3871	1.18
7	2659	735	27.6	2123	3262	1608	3600	0.75
8	3125	550	17.6	2691	3630	2349	3752	0.45
9	1930	640	33.2	1504	2334	1116	2907	0.93
Total	24,744	3840	15.5	21936	27509	19835	29797	0.40

As shown in Figure 4-2 and Table 4-2, the sustainable yields results estimated by alternative models vary in each groundwater management area, especially in the critical areas (management areas 2-4) where the relative prediction ranges from 50 to 84%, indicating relatively high uncertainty. The results from Table 4-2 show high degrees of uncertainty in the sustainable yield estimates, indicating that the contribution of conceptual models is relatively important. However, the variation of sustainable yields does not clearly show which uncertainties have the greatest impact on the results among different hydrogeologic, recharge, and boundary condition interpretations.

The wide range of sustainable yield estimates from the 12 alternative models strongly supports the need to address conceptual model uncertainty in the practice of groundwater management modeling, Further supporting evidence is supplied by (Neuman, 2003; Refgaard et al., 2006; Rojas, 2008a, b; Ye et al., 2010b). Despite similar calibration/validation performance, individual models provide total sustainable yields that differ by a factor of two and regional yields differing by as much as a factor of six.

4.3 Assessing Uncertainty in Sustainable Yield Estimates

In general, the difference in estimated sustainable yields seen in the 12 conceptual models indicates an important contribution of conceptual model uncertainty to predictive uncertainty in a management context. The estimated sustainable yields depend on each groundwater model's accuracy in simulating hydraulic head changes in response to pumping and on the suitability of the objectives and constraints assigned in the optimization exercise. The impacts of the form of objective or constraints in the optimization model are not investigated here. Rather, the focus is on how best to encapsulate the knowledge gained from multiple-model simulation/optimization in a single management decision.

There are now 12 models with 12 different performances in calibration/validation and 12 different sustainable yields. Ideally, one would like to be able to use this information to help us more rigorously determine both the most likely range of sustainable yields and to give the 'best' information to policymakers. Various researchers have proposed the use of information criteria and/or various model weighting schemes, and one would like to determine whether or not these are useful or warranted, and clarify what the strengths and weaknesses of some of these methods may be. The following sections describe a number of experiments used to test hypotheses regarding which averaging methods are most appropriate for assessing conceptual model uncertainty in management problems.

4.3.1 Model Weights for Alternative Models

The ranges of estimated yields from the 12 conceptual models are a useful, but imperfect, measure of conceptual model uncertainty in decision-making models. Because some of these models are bound to be better surrogates for the real world system, they should not be treated as equally valid. In many cases (including this case study), no single model is clearly superior to other competing models. Thus, some researchers have suggested that model-averaging may provide a better predictive capability than relying on a single model (e.g., Rojas et al; 2008a; Singh, 2010; Tsai, 2010). The fundamental goal of model-averaging is to average predictions by weighting predictions from a set of plausible models based on their relative performance.

In this study, a set of different model-averaging approaches including AICMA, MLBMA, GLUE, equal weight averaging, and a combined-Multi-Criteria approach developed in section 3.9 were compared in terms of their prediction coverage in sustainable yield estimates. Because GLUE and MLBMA approaches have prior probability term, as seen in Equations 4-5 and 4-6, that allow expert preferences to be integrated into the weighted scheme, these approaches allow for subjective likelihood or probabilities to be included as prior probabilities when computing posterior probabilities or model weights. In this work, validation results are used instead of expert judgment to obtain prior probabilities, and the use of uniform prior probabilities is compared to the use of subjective prior probabilities.

The model-averaged predictions for a given quantity are calculated as:

$$\bar{Z} = \sum_{i=1}^{K} w_i Z_i \qquad 4\text{-}5)$$

where Z_i is the predicted output value for model i, $\bar{Z}$ is the model-averaged prediction, K is the number of plausible models, and w_i is the weighted or posterior model probability of M_k where $\sum w_i = 1$. The GLUE weight for model k can be calculated as

$$\omega_k(GLUE) = \frac{p(M_k)SSWR_k^{-N}}{\sum_{k=1}^{K} p(M_k)SSWR_k^{-N}} \qquad (4\text{-}6)$$

where $p(M_k)$ is the prior probability for model k; $SSWR_k$ is the sum of squared weighted residuals for model k with $k = (1,\ldots,K)$, K is the total number of models being considered, and N is a shape factor. N=1 is normally used in GLUE applications and is used here.

For the information criteria, the posterior model probabilities or model weights of the twelve models can be calculated as (Neuman et al., 2003)

$$w_k = \frac{\exp(-0.5\Delta IC_k)\,p(M_k)}{\sum_k \exp(-0.5\Delta IC_k)\,p(M_k)} \qquad (4.7)$$

where $\Delta IC_i = IC_k - IC_{min}$ with IC_k corresponds to any of the information criteria described in section 2.7 for a given model M_k; IC_{min} is the minimum value obtained from model M_k with $k = (1,\ldots,K)$; and $p(M_k)$ is the prior probability for model k. IC-based criteria are calculated using Equations 2-13, 2-14, 2-18, and 2-19, respectively.

In both weight calculation approaches, models deemed to be better based on a certain metric receive higher weights than do worse performing models.

Table 4-3 presents model weights generated by the different model averaging methods. These model weights are then used to average predictions of the maximum yield that can be safely pumped from the aquifer while preventing water level decline lower than its corresponding limit. This table shows that the AICMA and MLBMA methods give a weight of almost one to just one or two models and essentially reject all of the others. Subsequently, including the subjective probabilities from different models, make no

difference because only one model remains after the AICMA and MLBMA weighting (the subjective probabilities are multiplied by zero for all but one of the models). In this study, GLUE weight derived from SWWR likelihood and MCDM methods are favored over the others. Considering model performances in calibration and validation processes based only on head residuals, all twelve models have quite similar predictive capability, with no model being clearly better than the others. In addition, the knowledge about the system being modeled and hydrogeologic data is too limited to justify the exclusion of the other 10 models. Eliminating these models will likely lead to the underestimation of conceptual model uncertainty, suggesting that AICMA and MLBMA weighting may be overly subjective. Therefore, all models should be applied in evaluating the uncertainty in the sustainable yield estimates to better demonstrate the application of conceptual model uncertainties.

Table 4-3 Model weights for different model-averaging methods

Model	GLUE	GLUE*	AIC	AIC*	AICc	AICc*	BIC	BIC*	KIC	KIC*	MC_F	MC_G
H1R1B1	0.097	0.101	0.867	0.868	0.939	0.937	0.850	0.852	0.009	0.008	0.103	0.086
H1R1B2	0.093	0.105	0.066	0.072	0.047	0.051	0.069	0.075	0.991	0.992	0.106	0.088
H1R2B1	0.097	0.082	0.034	0.028	0.009	0.007	0.040	0.033	0.000	0.000	0.076	0.081
H1R2B2	0.093	0.146	0.003	0.004	0.000	0.001	0.003	0.005	0.000	0.000	0.167	0.103
H2R1B1	0.074	0.070	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.074	0.080
H2R1B2	0.074	0.068	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.090	0.082
H2R2B1	0.099	0.095	0.030	0.028	0.004	0.004	0.037	0.035	0.000	0.000	0.140	0.089
H2R2B2	0.079	0.083	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.098	0.086
H3R1B1	0.071	0.053	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.072
H3R1B2	0.071	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.048	0.078
H3R2B1	0.077	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.075
H3R2B2	0.075	0.070	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.061	0.080

* with subjective probability obtained from validation test

4.3.2 Assessment of Sustainable Yield Model Averaging Techniques

Different model-averaging methods, including AICMA, MLBMA, GLUE, equal weight averaging, and a combined-Multi-Criteria approach developed in section 3.9, were used to assess the predictive uncertainty in sustainable yield estimates.

Using GLUE and MCDM model-averaged methods, the total sustainable yields were effectively identical at 24,900 m^3/d, while using AIC and KIC, the total optimal yields were 25,900 and 27,100 m^3/d, respectively (Figure 4-3). Comparisons of estimated sustainable yields at management zones using different model-averaging methods are shown in Figures 4-4 and 4-5. It is shown that the total sustainable yield estimates from IC-based criteria seem coincidentally higher than those from MC and GLUE criteria but also are only valid for the total sustainable yields. Moreover, GLUE, MCDM criteria, and the equal weights method estimate similar sustainable yields at every management zone, whereas AIC and KIC criteria estimate differently, especially in the critical area (management areas 2 to 4). It is noted here that one of these classes of averaging methods may be less useful for practical application; it remains to be seen whether these observed differences can be compared in a meaningful way.

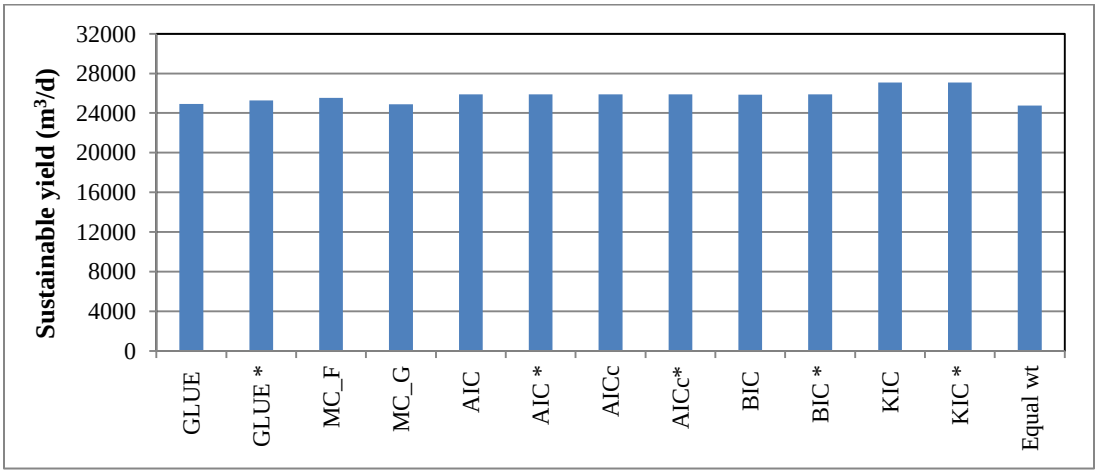


Figure 4-3 Total sustainable yield (m^3/d) for different model-averaging methods

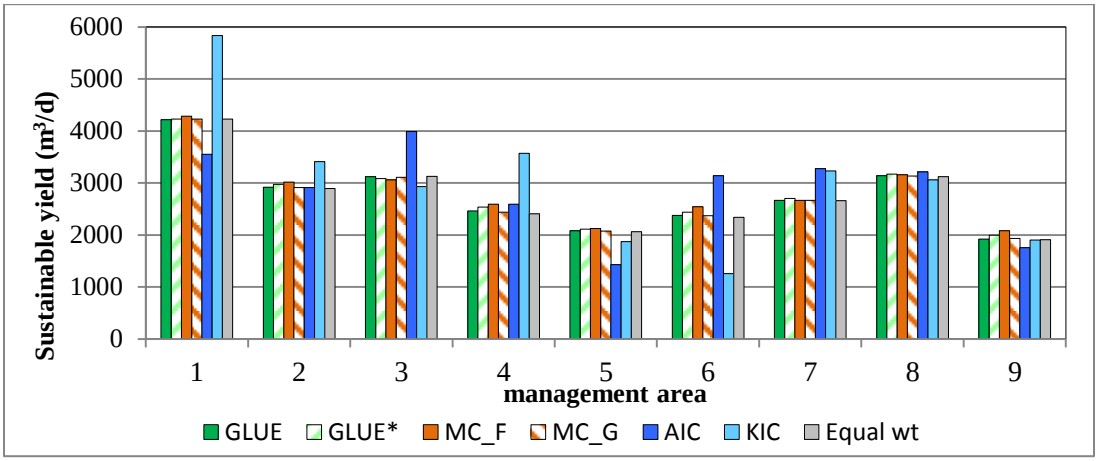


Figure 4-4 Optimized groundwater yield (m^3/d) in the nine management areas

Here, two methods are used to heuristically assess the utility of model-averaged sustainable yields generated using various weighting schemes. First, the performances of different multi-model ensemble methods were examined in terms of predictive intervals, using 90% prediction bounds and quartile 1-3 bounds, estimated from the weighted model-specific prediction intervals. Note the weighting changes in this comparison. Figure 4-6 shows the percentage of predictions or estimations which are bracketed by 90% uncertainty bounds. One hundred percent of pumping rates (155 pumping wells) estimated using GLUE, MCDM, or equal weights fall within the 90% prediction limits, and nine of the ten are covered by inter-quartile ranges of the ensemble estimations. Approximately nine-tenths and seven-tenths of the sustainable yield values from the IC-based model averaging method (AIC, AICc, BIC, and KIC) are contained by the 90% prediction bounds and inter-quartile ranges of the ensemble estimation, respectively.

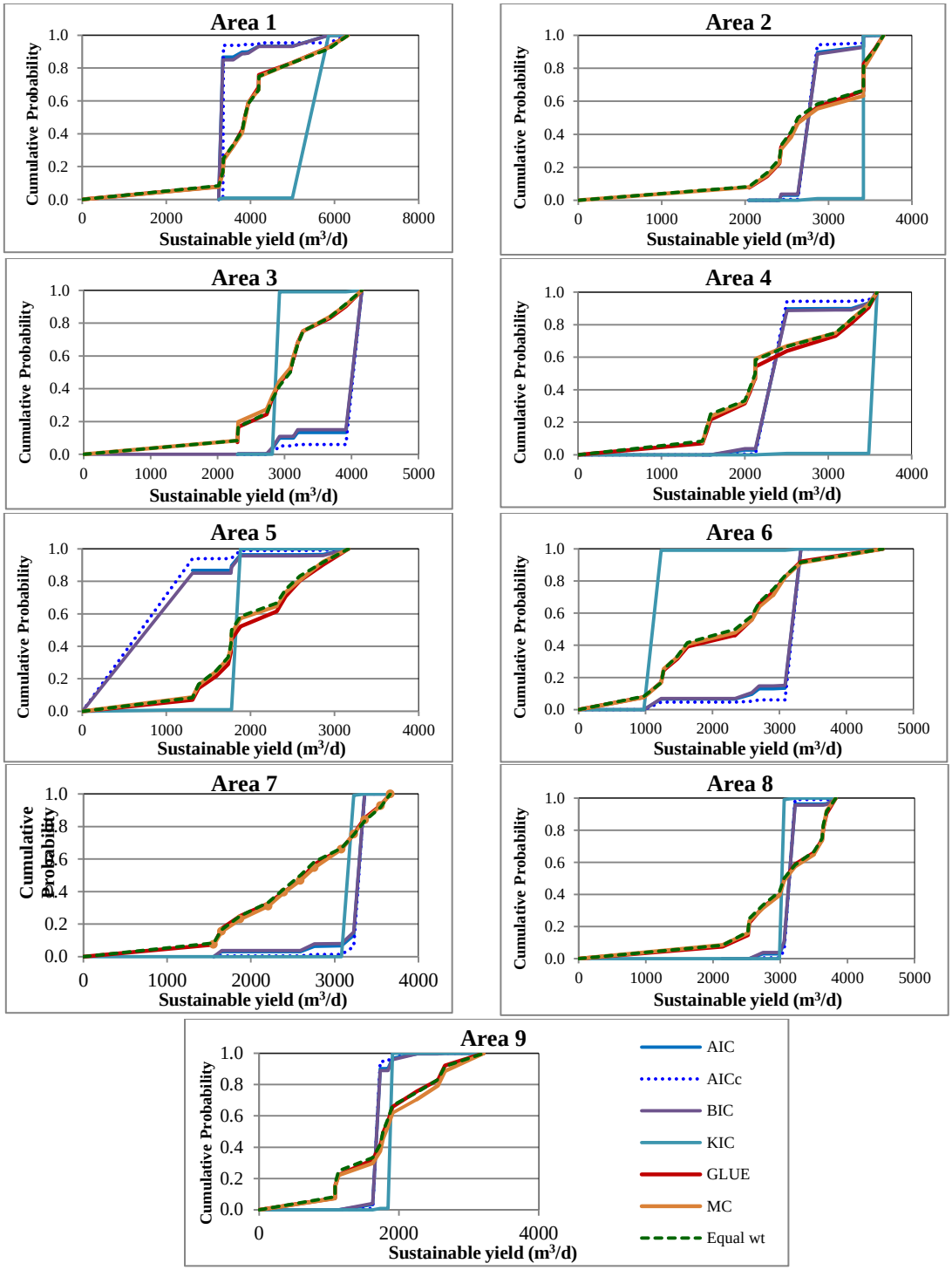


Figure 4-5 Comparison of cumulative probability for sustainable yields among different model-averaging methods

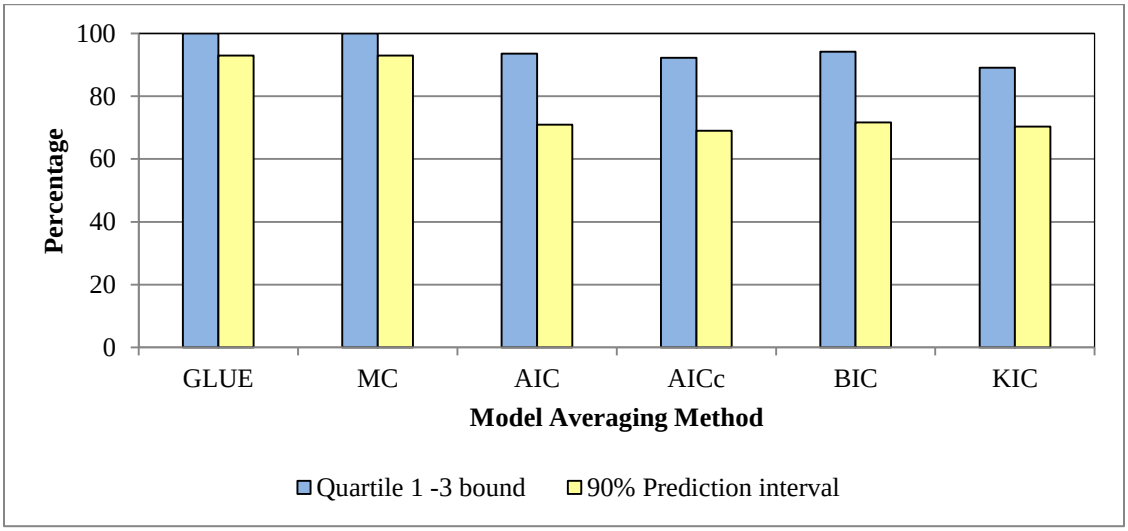


Figure 4-6 Percentage of pumping rates in 90% prediction intervals and inter-quartile of different model averaging methods

Figure 4-7 shows the ranges of sustainable yield estimates from 12 alternative models and the values obtained from different model-averaged methods for each management area. Sustainable yield values from both GLUE and MCDM are captured by the inter-quartile ranges of the ensemble estimations in all management areas. The sustainable yield values from AIC and KIC, however, are not contained by the inter-quartile ranges of the ensemble estimation in multiple management areas. For example, the estimate values are not captured by AIC in management areas 3 and 6, and management areas 1 and 4 for KIC methods.

The results illustrated in Figures 4-6 and 4-7 confirm that relying on a best single model, as is done implicitly with AIC, AIC_C, BIC, and KIC, is prone to produce an over-confident estimation of prediction uncertainty. On the other hand, GLUE and MCDM model-averaged methods produce more generalized estimations of the predictive uncertainty compared to the estimations of AIC and KIC, which are effectively the same as a single model selection approach.

In addition to the test above, the sum of head violations caused by model-averaged pumping rates were compared to address the question of which model-averaging method is the most appropriate for assessing conceptual model uncertainty for sustainable yield estimation. The risk of using the optimal pumping rates from the best model can be tested against model averaging by re-evaluating the sum of head violations caused by those model averaged rates when input into the original 12 models. Figure 4-8 shows the number of head constraint violations when sustainable yield rates obtained using the single best models (model H1R1B1 and model H1R2B2) and different model-averaging methods (such as the AIC, KIC, GLUE, GLUE*, MCDM, and equal weighted) are substituted into the 12 alternative models. The highest number of head constraint violations can be found when using KIC averaged rates, and the lowest number is found in equal weights and GLUE weights, respectively.

Figure 4-9 shows the total count of head constraint violations using the same models and the same model averaging pumping rates as in Figure 4-8.

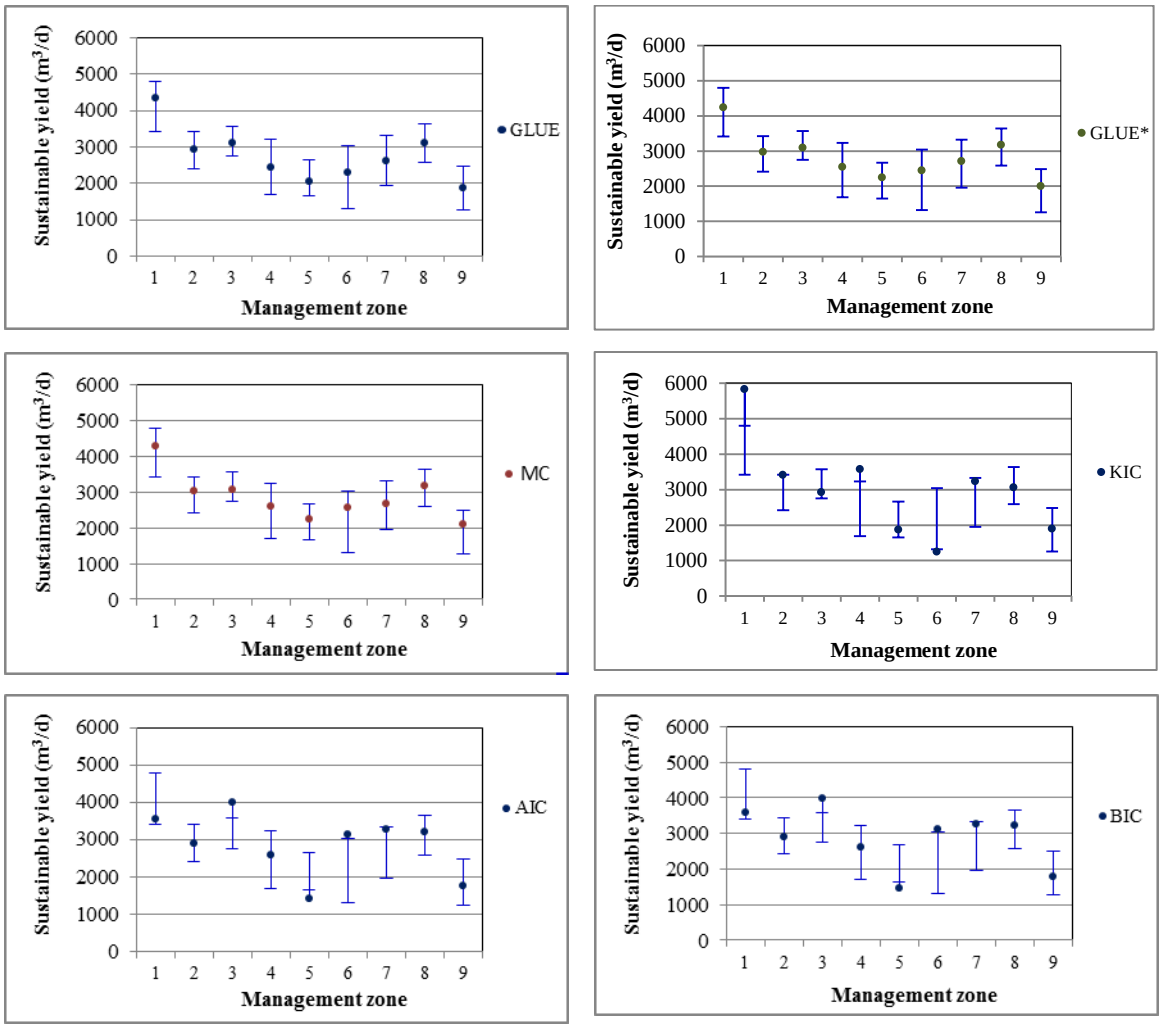


Figure 4-7 The statistics of the sustainable yield estimates from different model averaging methods in each management area. Q1 and Q3 represent the first and third quartiles, respectively

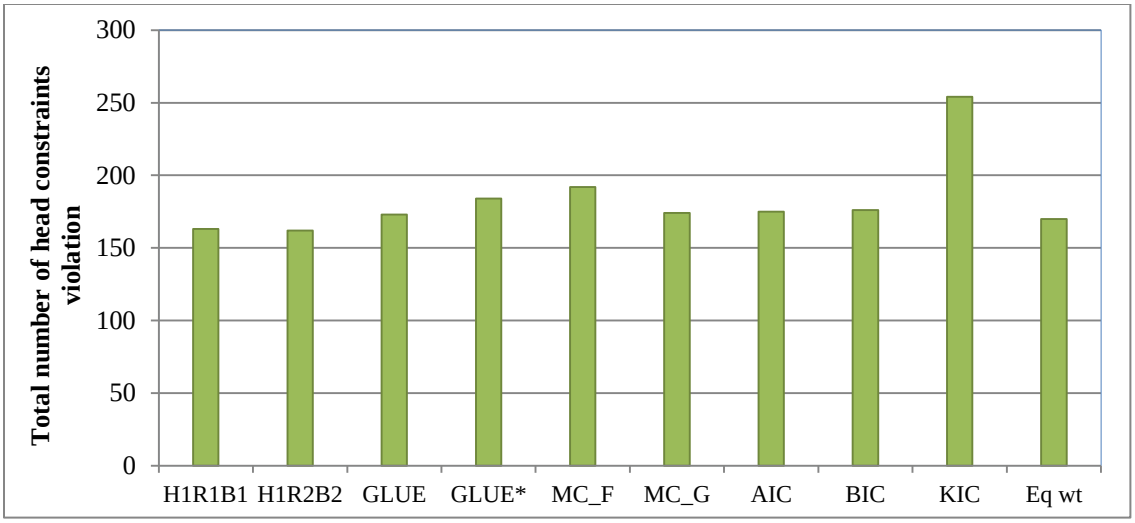


Figure 4-8 Number of head constraints violation in all 12 models using different model averaging pumping rates

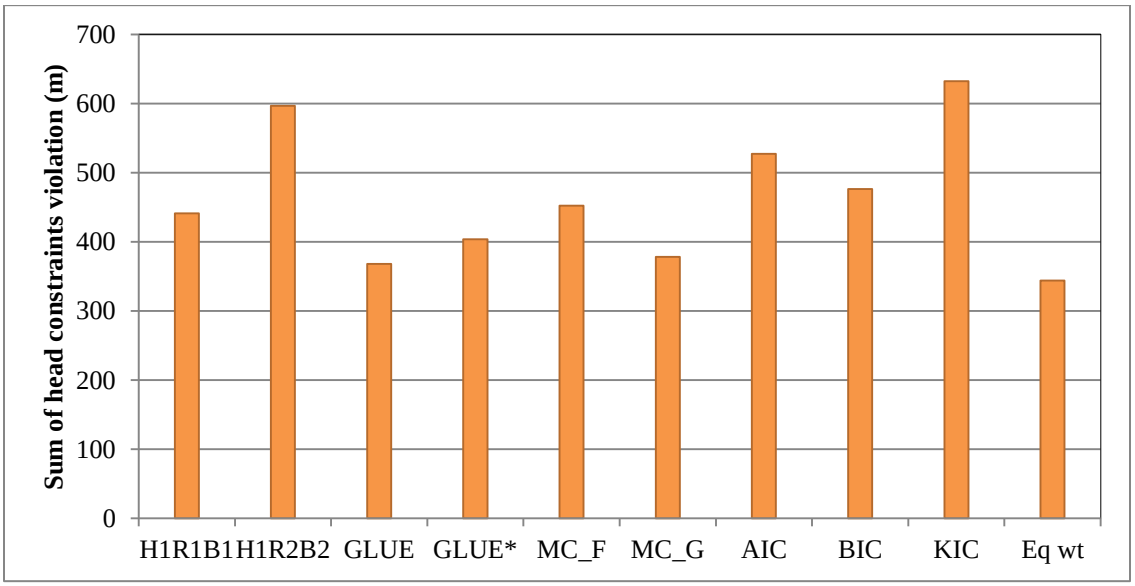


Figure 4-9 Total count of head constraints violations in all 12 models using the single best model and different model averaging pumping rates

Figure 4-10 shows the head constraint violations that occurred in each model when sustainable yield rates obtained from different weighted averaging were substituted into the individual alternative models. It can be seen that all H3 hydrogeologic models which are the less plausible models as determined from model fit from the SSWR and ARM of hydraulic head residuals, always produce a large number of head constraint violations.

For the management problem of this case study, if all twelve models are equally likely, the equal weight method appears to be the most appropriate approach since it produces the smallest sum of constraint violations in both the entire area and the critical area; the GLUE weight method is the second most appropriate approach; the MCDM method is the third, and the KIC method seems to be the least useful method because it produces the largest sum of constraint violations. This last finding indicates that basing pumping strategies on KIC is very likely to lead to overpumping. The decision-making could be carried out with a high degree of confidence if the maximum rates obtained from any model-averaging method caused no violation or fewer violations.

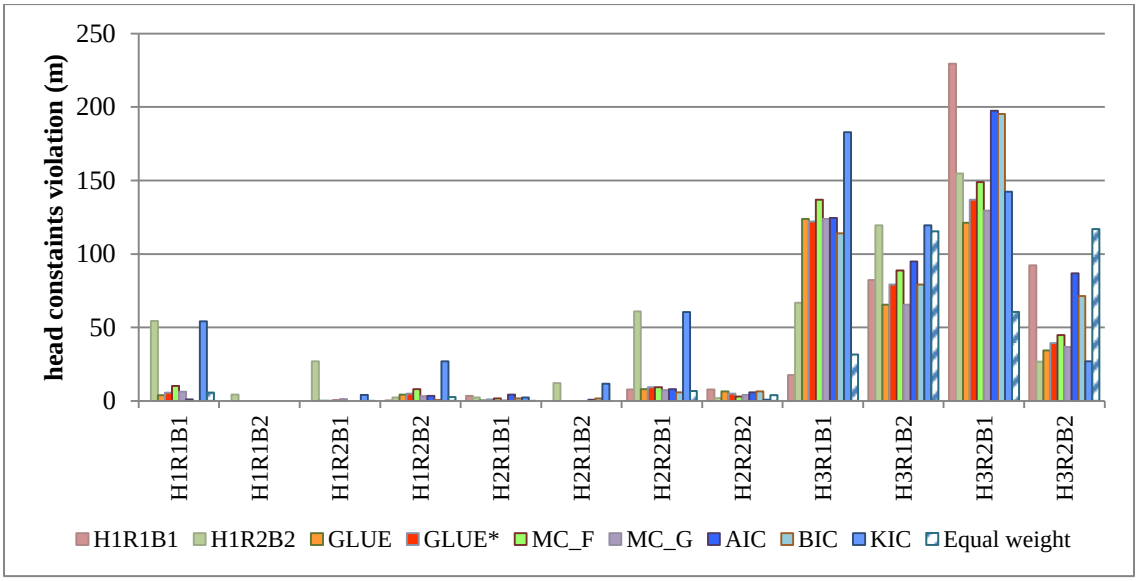


Figure 4-10 Sum of head constraints violations (m) in all 12 models using different model averaging pumping rates

4.4 Summary

The 12 calibrated models were applied in a management context to maximize total groundwater extraction rates for the Phu Thok aquifer in areas at high risk of water resource over-exploitation. Groundwater simulation models were linked with optimization techniques to determine optimal sustainable yields while ensuring that the drawdowns at control locations were less than specified limits. The uncertainty associated with the (typically fixed) spatial distributions of hydrogeologic properties, recharge rates, and boundary conditions was assessed to determine its significance in the decision making. Results for the individual models show that sustainable yield estimates vary substantially across models, and in this test case, there is clear utility in investigating the range of plausible conceptual models. Because of this variation, the recognition of conceptual model uncertainty in a management context is strongly recommended to help decision makers select and implement robust sustainable management strategies for the Phu Thok aquifer system.

Different model-averaging methods (GLUE, AICMA, MLBMA, Multi-Criteria, and equal weight) were assessed in terms of their ability to assist model uncertainty analysis in sustainable yield estimation. Two methods were used to examine these averaging approaches. First, the 90 percent prediction interval and inter-quartile ranges were used to compare the individual predictive distributions of the model averaging methods with the assumption that an appropriate model averaging techniques will provide a total sustainable yield that falls within 90% prediction limits. Sustainable yield values from both GLUE and MCDM are captured by the 90 percent confidence interval and the inter-quartile ranges of the ensemble estimations more than other methods. These results confirm that relying on a best single model, as typically results from use of AIC, AICc, BIC, or KIC is prone to produce an over-confident estimation of prediction uncertainty. It also implies that using information criterion-based model averaging for this case study

should likely be avoided, since these methods are effectively choosing one model above all others; this approach is shown to be unjustified for the models used here.

The second method used to assess model averaging approaches involved comparison of the sum of head violations caused by substitution of model-averaging pumping rates. For the management problem of this case study, MCDM, GLUE, or equal weight averaging are the most appropriate approaches since they produce the smallest sum of constraint violations, while the IC-based methods are less appropriate, for similar reasons to above: that the IC-based averaging schemes are under-conservative in their estimates of predictive uncertainty. Using SSWR with GLUE weight and equal weight can achieve a performance comparable to using MCDM and better performance than any information criteria. Therefore, these two methods are recommended based on these result since they are much simpler to implement and much less computationally demanding than more sophisticated approaches.

In this study, GLUE and combined-MCDM methods were favored over information-criterion-based approaches (AIC, AICc, BIC, and KIC) for model-averaging, since the latter were found to select only one or a few models. The selected models do not lead to better results in either calibration or evaluation than those of the discarded models. Moreover, the hydrogeologic and hydraulic data in this study area is too limited to justify the elimination of the other 10 models.

Chapter 5

Joint Assessment of Conceptual and Parameter Uncertainty

Chapter 4 proposed and assessed multi-model approaches to explicitly account for the predictive uncertainty in sustainable yield-estimation arising from conceptualizing hydrostratigraphy, recharge, and boundary conditions. In this chapter, these approaches are extended to include uncertainties arising from unknown model parameters. A joint estimation of predictive uncertainty, including parameter and conceptual model uncertainty, is presented. Most importantly, uncertainties obtained using model-averaging approaches are compared with results from parameter uncertainty to quantify the contribution of conceptual model and parameter uncertainty as applied to groundwater management models.

5.1 Estimating Sustainable Yields under the Impact of Parameter Uncertainty

The earlier calibration of the twelve models demonstrates that it is possible to obtain an effectively equivalent calibration performance for all 12, and thus these models can be accepted and used for predictive simulations. The purpose of parameter uncertainty analysis is to quantify the uncertainty in sustainable yields predicted by the 12 calibrated models when this results from uncertainty in the models' parameters. The calibrated models were used to produce multiple realizations of random parameters from feasible space using a simple Monte Carlo approach. The Monte Carlo sampling strategy assumes independence between parameter sets and randomly samples parameter values from their feasible space or posterior parameter distributions (USEPA, 1997). The stochastic MODFLOW of Groundwater Vista was used to generate parameter realizations for each model following a normal distribution and covariance variables. By comparing simulated and measured hydraulic heads, each set of parameter values is assigned likelihood values. The total of realizations is then split into behavioral and non-behavioral parameter combinations based on a cutoff threshold. For each model, the 12 behavioral parameter sets that have the smallest SSWR realizations are then selected and combined with optimization models to assess uncertainty in sustainable yield estimation.

Parameter uncertainty incorporated in the optimization models is conveyed to the sustainable yield estimates by sampling model parameters that preserve calibration. This study assumes that the structural parameterizations, e.g., the geometry of recharge distribution zones and hydraulic conductivity zones, used in the inverse calibrations for all models are known. Only parameters such as hydraulic conductivity and recharge rate values for each conceptual model are uncertain. The range of uncertainty for these parameters initially is the 95% confidence interval, as estimated from PEST during the calibration process. Both the magnitude of hydraulic conductivities and recharge rates

were sampled randomly from a normal distribution based upon these confident limits. Unfortunately, some of the 95% confidence interval limits correspond to unrealistic parameter values and are outside of the desired parameter ranges, sometimes by several orders of magnitude. These results point towards possible violations of underlying assumptions in PEST regarding the input data, which is assumed to have normally distributed errors and no biases, and the model, which is assumed to be approximately linear in the vicinity of the calibrated solution. Furthermore, the high dimensionality of the problem can impede PEST's ability to compute valid confidence limits. Thus, these confidence limits cannot be used to indicate lognormal standard deviations of parameters. Instead, prior knowledge regarding parameter values obtained from other studies is used to estimate lognormal standard deviations. Therefore, the upper and lower limits for hydraulic conductivity parameters were set to 2 orders of magnitude above and below the optimized values. For recharge rate, the confidence limit was set to ±40 mm/year, based on experience from previous findings.

Stochastic MODFLOW, within the Groundwater Vistas software, was used to generate 1,000 Monte Carlo samples of the calibrated hydraulic conductivity and recharge parameters for each model. One thousand samples were assumed to be sufficient for convergence, based on preliminary results and typically available computational resources. The plausibility of each realization was assessed to determine whether a sample was used in the uncertainty analysis, by evaluating calibration criteria. The best realizations were selected, "best" meaning those model samples that acceptably matched the observed and simulated head residuals. In this case, head residual values and acceptable ranges for calibration statistics are defined as 1) an absolute residual mean (ARM) lower than 3 meters, 2) a residual standard deviation (also called a root mean squared (RMS)) lower than 3 meters, and 3) a residual standard deviation divided by the range in heads for all targets (the normalized root mean square error (NRMS)) lower than 0.1 (10%).

All realizations were assigned weights. In a standard Monte Carlo simulation, all of these realizations, regardless of SSWR values, would have equal weight. However, it is reasonable to assume that realizations that more closely match the hydraulic head values should receive higher weight than flow realizations that only poorly match the given target head. Thus, the GLUE approach is implemented here in Monte Carlo simulations for weighting each realization based on its likelihood measure. This approach also provides consistency between the assessment of predictive and conceptual uncertainty. The model weight for each realization can be calculated (with a shape factor of N=1) from the inverse of the SSWR. The weights are then normalized to obtain the cumulative distribution function sum up to one.

Model weights for all realizations representing parameter uncertainty are listed in Appendix D. Typically, each realization receives model weights in the ranges of 0.075-0.09, indicating that all realizations match the hydraulic head targets at approximately the same level.

The formulations of groundwater optimization models are described in Chapter 4. Sustainable yield estimations were made for the 12 realizations (which meet our calibration criteria) to obtain estimated prediction intervals for each ensemble of alternative models. When the GLUE approach is used in Monte Carlo simulations and optimizations, the weighted mean and standard deviation of the sustainable yield can be obtained.

Summary statistics of total sustainable yield obtained from each alternative model that performed parameter uncertainty analysis are shown in Table 5-1, and the details of these statistics for each management area is presented Appendix E.

Table 5-1 Statistics of the total sustainable yield (m^3/d) obtained from a set of realizations for each alternative model within parameter uncertainty analysis

	Max	Min	Average	Standard deviation (σ)	Varaince (σ^2)
H1R1B1	30605	21839	27013	2720	7.40E+06
H1R1B2	35561	26620	31023	2215	4.91E+06
H1R2B1	36157	27711	31501	2482	6.16E+06
H1R2B2	43628	31331	35571	3646	1.33E+07
H2R1B1	30573	17308	25111	4074	1.66E+07
H2R1B2	19197	15655	17662	960	9.22E+05
H2R2B1	30393	18089	26573	3829	1.47E+07
H2R2B2	36295	29545	33149	2043	4.17E+06
H3R1B1	29267	16984	24688	3658	1.34E+07
H3R1B2	35124	20763	29762	5029	2.53E+07
H3R2B1	26184	19457	22730	2253	5.08E+06
H3R2B2	29379	15892	24035	3547	1.26E+07

The adequacy of using a single model and an associated parameter uncertainty analysis to describe the model's total uncertainty was assessed by comparing the predictive uncertainties as computed from the 12 optimization models for each alternative model. Figures 5-1 and 5-2 compare cumulative probability distributions of the sustainable yield obtained from using each model with parameter uncertainty assessment in the entire area and in each management area, respectively. The GLUE model-averaging method is also included in these figures to compare the range of prediction uncertainty. These figures show that significant differences in the estimations, even when accounting for parameter uncertainty. The most likely estimated values (50^{th} percentile) and the prediction intervals drastically differ among models. When considering only a single model, the prediction ranges are relatively narrow compared to the prediction range obtained from the GLUE model averaging method.

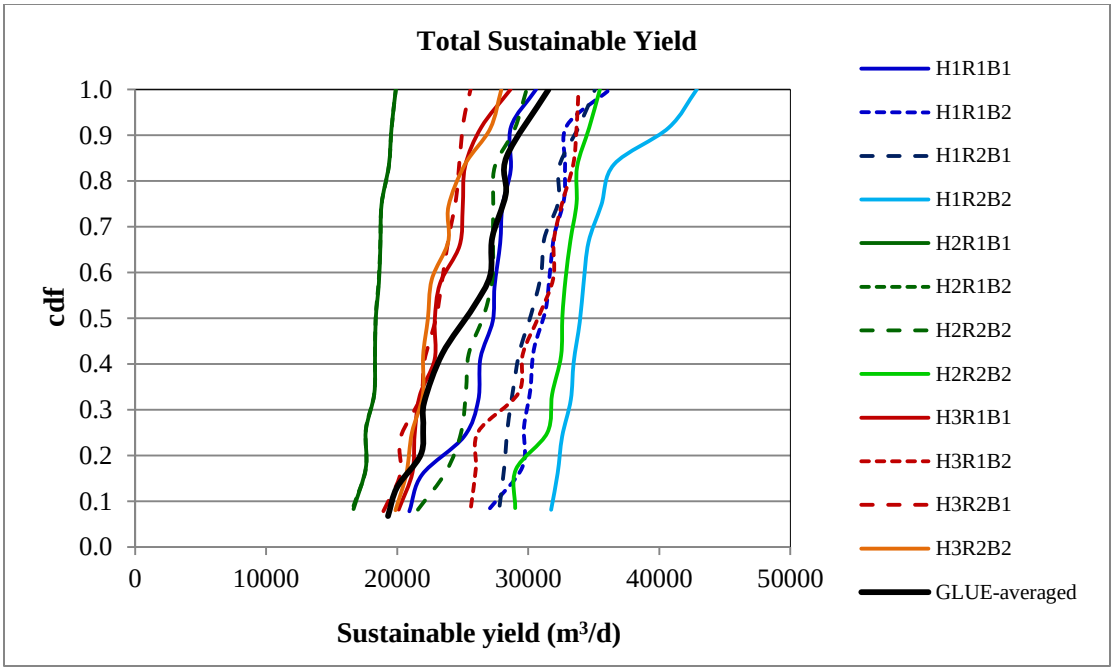


Figure 5-1 Cumulative distribution function for total sustainable yield estimation derived from ensemble of 12 realizations

In addition to the variation in total yield, Figure 5-2 shows the variation of sustainable yields in each management area. The relative 95% prediction intervals for sustainable yield estimates derived from the alternative optimization models vary considerably in each management area. Management area 6 has the highest variation in the sustainable yield estimated, and areas 7 and 9 have the second and third highest, respectively. It is noticeable that the relative prediction ranges from individual models are sometimes wider and sometimes narrower in each management area. The wide variation may indicate the 12 alternative models have very different conceptualizations of these areas. Moreover, the prediction range of the averaged-model is considerably larger than that of any individual model, attesting to the formers' superior performance. It may be interpreted that the influence of conceptual model uncertainty is higher than that of parameter uncertainty, but the meaning of this difference is difficult to gauge. These results suggest

that for a complex, data-poor system such as the one studied here, a single model cannot adequately assess predictive uncertainty.

Table 5-2 shows the calculated 95% prediction intervals and relative 95% prediction interval for the total sustainable yields computed by the ensemble of 12 models with parameter uncertainty assessment and the GLUE model-averaging uncertainty assessment. The lower limit of the 95% prediction interval corresponds to the prediction at the 2.5% level of cumulative probability, and the upper limit corresponds to the prediction at the 97.5% level of cumulative probability. The relative 95 % prediction interval of the total sustainable yields estimates for individual models vary from 0.12 to 0.37. These finding can be compared to the relative uncertainty of the total sustainable yields derived from conceptual model uncertainty (0.47). The prediction intervals obtained from models H1R2B2 and H2R2B2 are much higher than the average predictions of total sustainable yield, whereas model H2R1B2's range is much lower than the averaged one. These results likely suggest that using a single best model underestimates the predictive uncertainty and may produce misleading results.

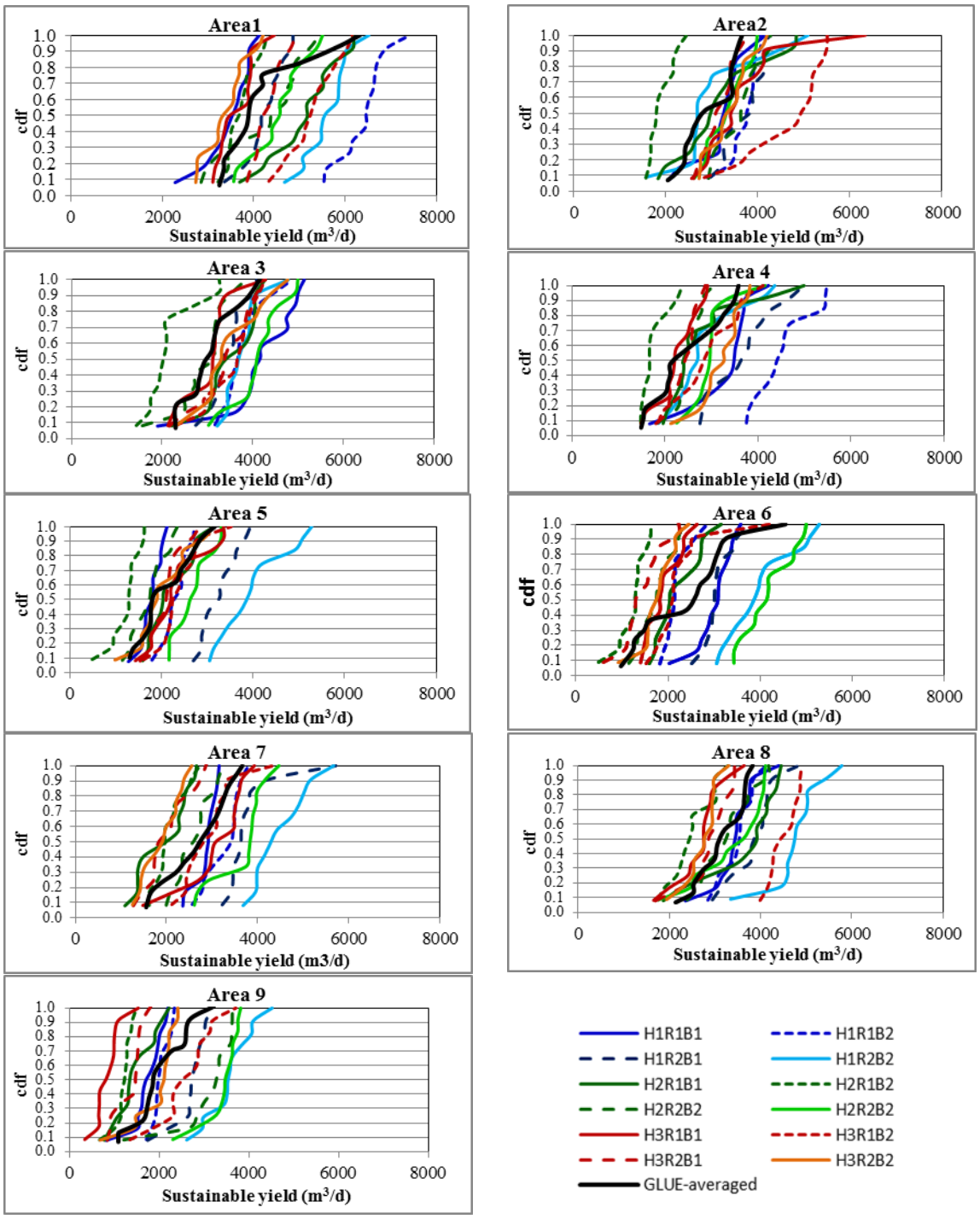


Figure 5-2 Cumulative distribution function for sustainable yield estimation derived from ensemble of 12 realizations in each groundwater management area

Table 5-2 The 95% prediction intervals and relative 95% intervals of the total sustainable yield computed by parameter uncertainty and conceptual model uncertainty assessment

Model	Parameter uncertainty assessment				
	Lower limit	Upper limit	Average	Width of interval	Relative 95% interval
H1R1B1	21195	30068	26585	8874	0.33
H1R1B2	27757	35323	31377	7566	0.24
H1R2B1	27913	34762	30726	6849	0.22
H1R2B2	31881	42265	35143	10384	0.30
H2R1B1	20104	29477	25085	9373	0.37
H2R1B2	16766	18963	17908	2197	0.12
H1R2B1	22173	29621	26308	7448	0.28
H2R2B2	29016	35240	32502	6224	0.19
H3R1B1	20401	28033	23611	7631	0.32
H3R1B2	25734	33791	30406	8057	0.26
H3R2B1	25734	33794	30406	8061	0.27
H3R2B2	20097	27714	23218	7616	0.33
Model	Conceptual model uncertainty assessment				
	Lower limit	Upper limit	Average	Width of interval	Relative 95% interval
GLUE* averaged	19297	31015	24744	11718	0.47

Figure 5-3 shows that the contributions of predictive uncertainty in each management area are more apparent than that in the entire area. Again, the impact of a conceptual model in each management area is higher than that of parameter uncertainty. The relative 95% prediction interval obtained from conceptual model uncertainty varies between 0.5 and 1.4, whereas parameter uncertainty varies between 0.2 and 1.1

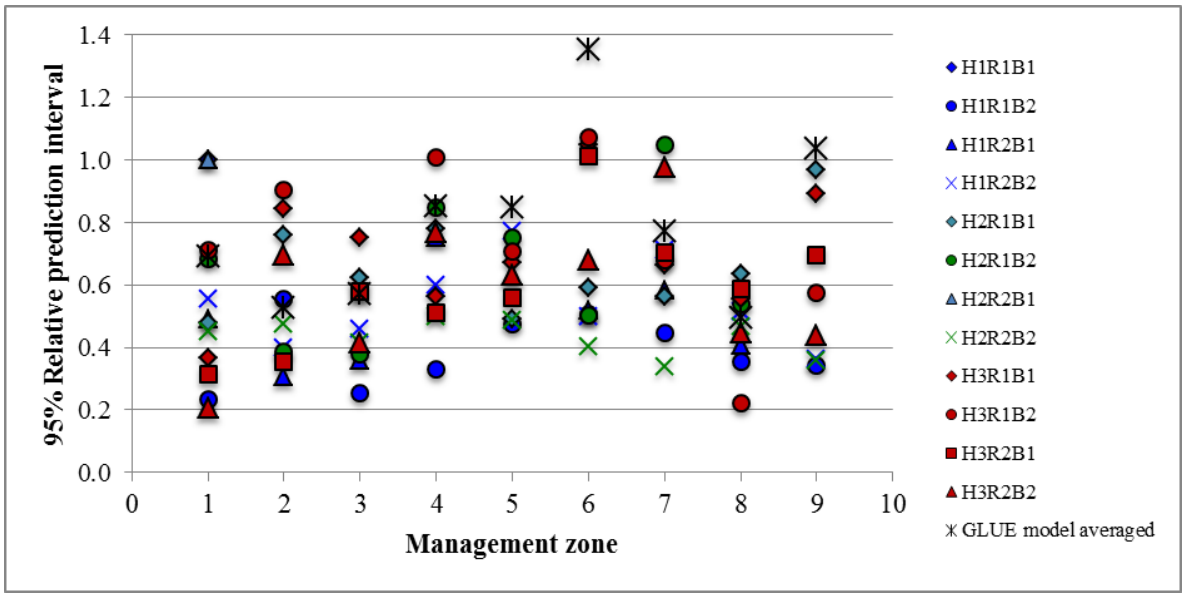


Figure 5-3 Relative 95% prediction intervals for 12 models with parameter assessment and GLUE model-averaging method in each management area

5.2 Influence of Conceptual and Parameter Uncertainty

The previous section demonstrates that focusing on parameter uncertainty analysis based on a single selected simulation/optimization model is likely to underestimate the total uncertainty, which should also include the conceptual deficiencies of the specific model. It also indicates that neither parameter nor conceptual uncertainty can be treated as negligible, as neither dominates. Improving the uncertainty bounds in model prediction requires the integration both conceptual model and parameter uncertainty. Typical, to do so, the final weights are calculated and normalized to one. Then predictive statistics (the mean, standard deviation, variance, and predictive interval) are evaluated to improve the robustness of model prediction.

Final weights for each model include both conceptual model-averaged weights and Monte Carlo realization weights assigned to the 144 optimization models. A model-averaged prediction can be derived as a weighted average of the predictions from each model in the ensemble set. The model-averaged (averaged over conceptual models) predictions can be calculated as

$$\overline{Z}_{ij} = \sum_{j=1}^{12} \sum_{i=1}^{12} w_i Z_{r,j} w_j \qquad (5\text{-}1)$$

where the final weights w_{ij} are the product of the probability assigned to the alternative models *i* and the GLUE flow weights assigned to each MCDM subset. The final weights are applied to the values of the ensemble sustainable yield for each realization, and an empirical cumulative distribution function (CDF) is computed from the final weights.

Because it was found in Chapter 4 that AICMA and MLBMA weightings are likely inappropriate, only GLUE with uniform prior and subjective prior probabilities, MCDM methods, and equal weights were used to determine the contribution of both conceptual model and parameter uncertainty in sustainable yield estimation.

The weights ($Z_{r,j}$) are then used to combine the CDFs in Figures 5-1 and 5-2 to get one combined CDF for the estimated sustainable yields. The averaged sustainable yield predictions were calculated using Equation (5-1) and are given in Figure 5-4. Predictive intervals (90%) and mean predicted values based on cumulative predictive distribution obtained from GLUE methodology for sustainable yield estimated obtained from 12 realization models parameters in each management area are shown in Appendix F.

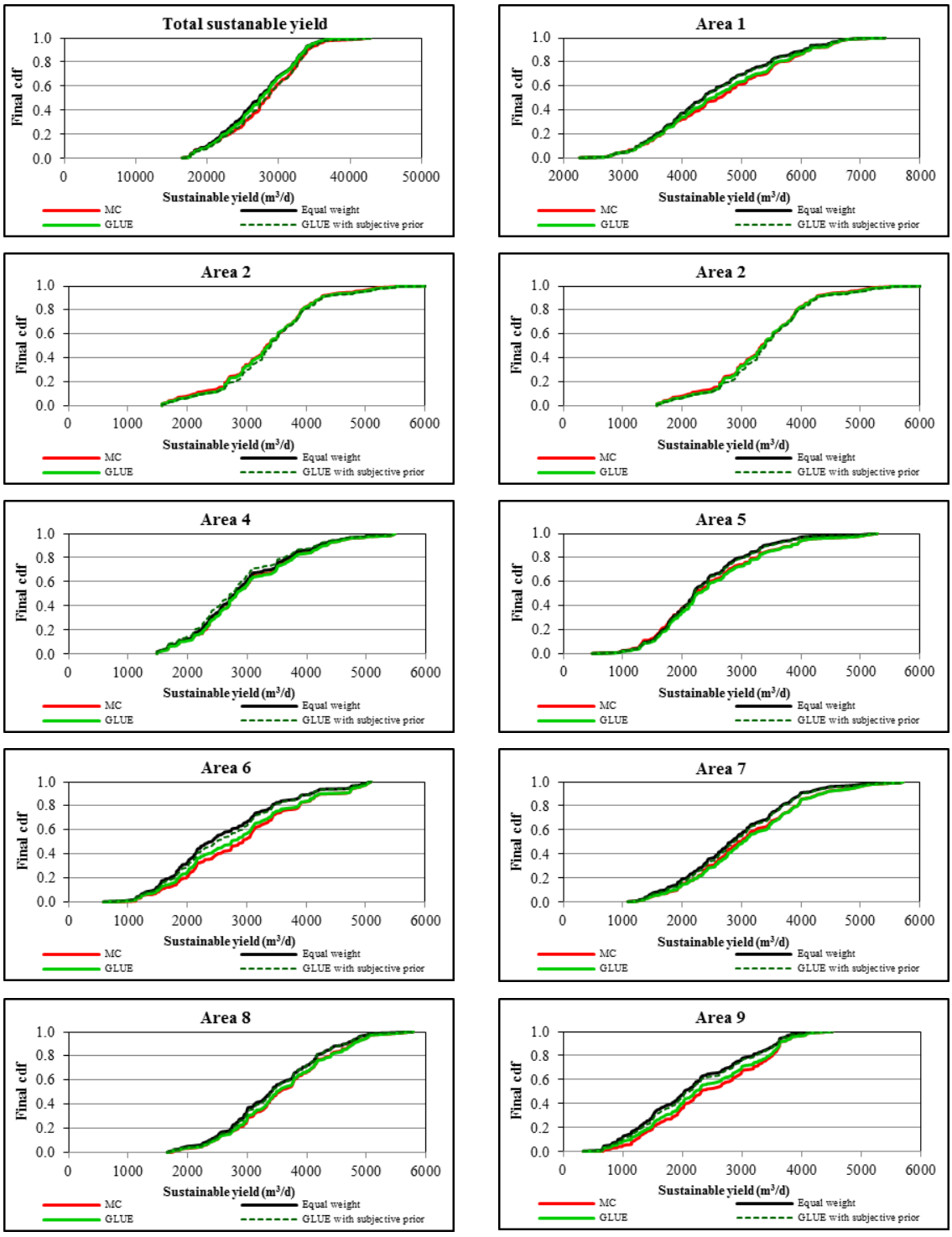


Figure 5-4 Final cumulative probability distributions for sustainable yield estimates that account for parameter and conceptual model uncertainty in each management area

When combining the impact of both conceptual model and parameter uncertainty, these CDFs show the larger range of optimal yield. Comparing the CDFs for different model-averaging weights shows very similar ranges in the sustainable yields using these criteria. The differences in CDF can be seen in areas 6 and area 9, and indicate the high variation in conceptual models. In these areas the uncertainty across the different models is reflected in CDF. For this study, the four criteria can yield approximately the same results and cannot distinguish among these criteria. Therefore, the MCDM CDF is used for further analysis.

The generated set of 12 ensemble members in the model averaging approach was used to address the question of whether the application of multiple conceptual model approaches can help to improve the prediction of sustainable yields by accounting for model uncertainty. Typically, the discrepancy in the ensemble set for each output variable comparing with observed data demonstrates the incompetence of the individual models in capturing the observations. In this comparison, the mean values (50^{th} percentile) of the integrated model weights and realization weights can be the representative of the optimal rates. Considering total uncertainty is expected to capture some of these deficiencies.

Figure 5-5 shows the contribution of parameter uncertainty associated with the best models (models H1R1B1 and H2R2B2), GLUE with subjective prior probability, MCDM model-averaging methods, and equal weights to the joint assessment of parameter and conceptual model uncertainties. As can be seen in this figure, the prediction ranges of between 25^{th} and 75^{th} percentile obtained from models H1R1B1 and H1R2B2 fall considerably outside the prediction ranges of the total model uncertainty. Unlike with parameter uncertainty analysis, the large degree of concurrence of the predictive uncertainty interval is observed from model-averaging methods. For GLUE and MCDM methods, the predictive intervals from a fractional overlap interval in 60% of the all pumping rates compared to a fractional overlap of 20%, and 30% of all pumping wells for individual parameter uncertainty analysis. The parameter uncertainty analysis is thus

not adequate to account for predictive uncertainty due to both parameter and conceptual model uncertainty.

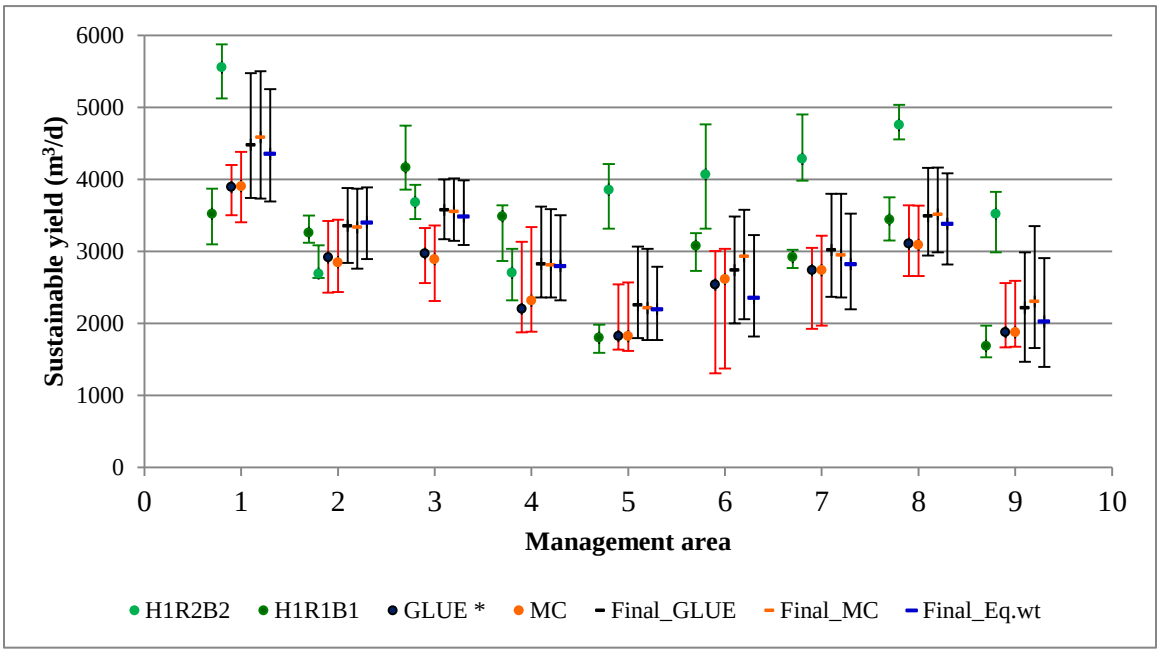


Figure 5-5 Comparison of predictive intervals (25th to 75th percentile) and most likely estimated values (50th percentile) obtained from the best single model, GLUE, MCDM model-averaging methods, and combined parameter and conceptual model uncertainty

Considering parameter and conceptual model uncertainty, the optimal pumping strategy will not be determined by one single ‘‘best case’’ simulation/optimization. Rather, there can be parts of one model that impact pumping in one area and parts of another model that impact pumping in another area; thus the optimal solution will meet each constraint in every part of all ensemble models. From optimization results, there are 156 sets of pumping rates obtained from predictive uncertainty assessments. However, we need to find the one solution for pumping rates. The unique solution may produce head constraint violations. To quantify how much water can be safely pumped from each management area without violation or with an acceptable degree of constraint violation, a unique set of pumping rates obtained from the 30th, 35th,40th, 45th and 50th percentiles of each pumping

well (in total 155 pumping wells) were selected and substituted into 156 simulation models.

Ideally, one would want to pump sufficient water to satisfy increased demand; however, for both conceptual and parameter uncertainty analysis, less water must be pumped so that head constraints violations are avoided in all models. Therefore, the optimal yields of 14,400 m^3/d are obtained to meet all constraints (Figure 5-6). This artificial reduction in pumping rates would result in a conservative management policies in terms of the total amount of pumping rates needed to satisfy all models. The extensive management model will consider constraints that include head predictions from individual models; in turn, this method would exaggerate the influence from less significant models. One can avoid this problem while still considering the model uncertainty. From a practical perspective, very small violations may be ignored in some areas. In such circumstances, one needs to find the best compromise between the increased pumping and the level of violation. For example, to find the solutions with a 90% reliability of no violation occurring, the decision maker can select an optimal yield of 16,210 m^3/d derived from the 40^{th} percentile solutions.

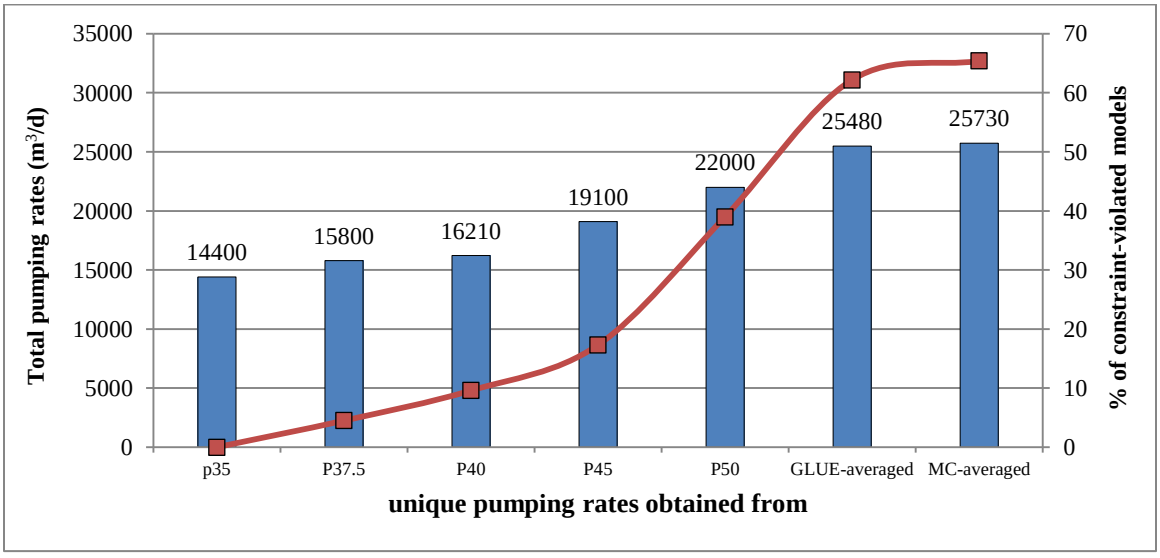


Figure 5-6 Comparison of optimal yields and percentage of constraints violations models

To find a solution with 95% reliability, one can select the optimal yield of 15,800 m^3/d for the TPA. These results can provide information to the decision makers as well as the degree of uncertainty to be expected when specific rates are used as well as where a specific head violation occurs, and where head constraints can be relaxed. This methodology can be applied with minimal computational expense.

If considering only conceptual model uncertainty, the optimal rates of 25,480 and 25,730 m^3/d obtained from GLUE and MCDM-averaged will result in 62 and 64% of models containing head violations, respectively. If considering the two best models (H1R1B1 and H1R2B2), the pumping rate of 28,000 m^3/d obtained from this model will result in 90% of models having head violations. The results in this section demonstrate the value of combining the impacts of conceptual and parameter uncertainty for groundwater management problems. Because assessing conceptual model uncertainty and parameter uncertainties individually results overestimation of yields, they should be considered together.

5.3 Summary

This chapter has evaluated multi-model ensemble approaches to assessing the impact of conceptual models and parametric uncertainty on reliable optimal groundwater resource allocation. All calibrated models were used to produce multiple realizations of random parameters from feasible space using a simple Monte Carlo approach. For each model, the 12 behavioral parameter sets that have the smallest SSWR realizations were then selected and combined with optimization models to assess uncertainty in sustainable yield estimation. The multi-model approaches of GLUE, equal weights, and MCDM criteria were integrated with GLUE weights for each realization to evaluate the total prediction uncertainty associated with both parameter uncertainty and model error.

The contribution of conceptual model uncertainties varied between 50 %and 135 % of the relative 95% predictive uncertainty in each management area, whereas the contribution of parameter uncertainty varied between 23 % and 103 % of the relative 95% predictive uncertainty in each management area. The multi-model averaging methods produce a higher percentage of predictive coverage than do any individual models.

Although accounting for all sources of uncertainty is very important in predicting environmental and management problems, the available techniques used in the literature may be too computationally demanding and, in some cases, unnecessarily complex, particularly for data-poor systems. The methods presented here to account for the main sources of uncertainty provide the required practical and comprehensive uncertainty analysis and can be applied to other case studies to provide reliable and accurate predictions for groundwater management applications, without computational expense.

The results demonstrate the significance of combining conceptual and parameter uncertainty in groundwater management problems. Incorporation of conceptual model uncertainty and parameter uncertainty results in 28 % and 54 % decreases in optimal yields if only conceptual model and parameter uncertainty were assessed, respectively.

The results of this chapter strongly support including both conceptual model uncertainty and parameter uncertainty in groundwater management practice. The uncertainty analysis can be used not only to reduce the uncertainty in allocation strategies, but also to make more informed and reliable decisions. In this case study, simple model-averaging techniques such as equal weights are shown to perform quite similar to GLUE and Multi-Criteria-Decision Making approaches and that these criteria are preferable to model selection criteria such AIC, AICc, BIC and KIC.

Chapter 6

Conclusions and Recommendations

6.1 Summary of Conclusions

This thesis presents an uncertainty assessment approach to explicitly account for conceptual model uncertainty inherent in groundwater management decisions. The difficulty of fully characterizing the subsurface environment makes uncertainty an integral component of groundwater flow models, however, reliable prediction of groundwater management models is vital for managing groundwater resources. Thus, to achieve an acceptable level of confidence in model predictions, the main sources of uncertainties need to be quantified. This thesis has described the development of the three-dimensional physically based groundwater flow models with various sources of conceptual model uncertainty, focusing on the Thaphra area (TPA), Northeast Thailand.

The three potential sources of conceptual model uncertainty considered here are different interpretations of hydrostratigraphic parameterization, different postulations of boundary conditions, and different techniques for recharge estimation. Although the 12 conceptual models cannot cover all potential sources of errors, they do cover a useful if limited

range, given that the overall uncertainty of model prediction cannot be absolutely assessed. The 12 alternative models were calibrated against head observations. Results demonstrate that these 12 models can attain an almost identical calibration performance, perhaps implying that head observations cannot clearly discriminate among different competing models.

Different model weighting approaches were used to rank the multiple conceptual models. Model weights obtained from AICMA (AIC, AICc) and MLBMA (BIC, and KIC) methods dominated in only one or two models, meaning that only one or two models are needed for making predictions. The GLUE or SSWR weights are more evenly distributed than those of IC-based model selection criteria.

All the calibrated models and the different model-averaging methods were individually validated against an independent data set obtained from two years of observation data in which pumping rates and recharge rates were different from those used for calibration. The validation data were used to evaluate the predictive ability of the individual models and model ensembles. The best results as evaluated by SSWR and ARM were obtained with a model ranked fifth by AIC, AICc, BIC, and GLUE but ranked second last by KIC. Overall, most of the model- averaging methods produce predictions closer to the calibration data than do the individual models. In the validation period, none of the model averaging methods produce predictions that are closer to the validation head data than individual model H1R2B2; however, their prediction quality is significantly better than that of the majority of the individual models. This finding indicates the strength of using model-averaging methods, which are able to make better predictions than a single model can, or at least to obtain predictions that are close to the best model.

Evaluating multiple models using different measures such as SSWR, and ARM in both calibration and validation periods can yield different ranking and weights. The results of the validation period are not consistent with those of the calibration period; thus, there is

no reason not to use the information obtained from validation period to evaluate model performance. Validation results can provide valuable information relevant to model performance, and can be included in the posterior probability as a subjective prior model probability instead of expert elucidation, which is not easily obtained in regular practice. Given incomplete knowledge of a system and data-poor aquifer systems, there are inevitably doubts about conceptual models that exist, arising from the judgments and conceptualizations made as part of a simulation model's development. Thus, it is reasonable to develop a set of performance measures, instead of relying on a single criterion.

To quantify the uncertainty associated with the (typically fixed) spatial distributions of hydrogeologic properties, recharge rates, and boundary conditions in management models, the 12 calibrated models were linked with optimization techniques to determine optimal sustainable yields while ensuring that the drawdowns at control locations were less than specified limits. Results for the individual models show that sustainable yield estimates vary substantially across models, and in this test case, there is clear utility in investigating the range of plausible conceptual models. Despite similar performances among the models calibrated, the total sustainable yield estimates vary substantially depending on the conceptual model used and range widely, by a factor of 0.6 in total, and even more extremely, in each management area by a factor of 4.

The comparison results convincingly demonstrate that simple model averaging achieves better performance than formal and sophisticated averaging methods such as AICMA and MLBMA, and produce a similar performance to GLUE and combined-multiple criteria averaging methods for both validation testing and management applications,.

Lastly, multi-model ensemble approaches were used to assess the impact of conceptual model and parametric uncertainty on management decisions. The GLUE and proposed Combined Multi-criteria multi-model approaches were used to compare the sustainable

yields estimations in the TPA. The contributions of conceptual model uncertainty to predictive uncertainty were rather significant, indicating the relevance of addressing this source of uncertainty in groundwater management. For the Thaphra test case, the sustainable yield uncertainty arising from conceptual model uncertainty varies between 47 and 110 % of the predictive uncertainty. The multi-model averaging methods produce a higher percentage of predictive coverage than do any individual models. The relative contribution of conceptual model uncertainty to different management zones provides useful information for decision-makers to use when implementing groundwater extraction policies for sustainable management of the TPA.

Model-averaging predictions have been shown to be capable of integrating multiple models for prediction in management models. Optimized operations based on the average predictions show more reliable management outcomes than those from one simulation model. However, the optimized operation is more conservative in order to avoid constraint violations elevated by considering many models. Incorporation of conceptual model uncertainty and parameter uncertainty results in a 28% and 54% decrease in optimal yields if only conceptual model and parameter uncertainty are assessed, respectively.

Groundwater models are imperfect representations of real world systems. Different models have specific strengths in capturing different aspects of real world processes. Multi-model averaging methods can take advantage of the diverse abilities in different competing predictions (Duan et al., 2007; Ajami et al., 2006). The use of multiple sets of model weights to generate model predictions is a method that emphasizes the strengths of individual models in capturing different observation periods. This is achieved by combining the calibration and validation data. Although we cannot prove here that the likelihood of this model being the most appropriate, we can use current knowledge and the reliable data we have to make robust decisions. Doing so will help to reduce the inherent uncertainty and improve decisions. This study has illustrated how the model

averaging method can be used to generate probabilistic optimization predictions from several competing individual predictions.

Although accounting for all sources of uncertainty is very important in predicting environmental and management problems, the techniques described/offered in the literature are too computationally demanding. The methods used to account for the main sources of uncertainty presented here provide the required practical and comprehensive uncertainty analysis and can be applied to other groundwater management applications, but with a reduced computational expense.

6.2 Recommendations for future research

Application of the proposed method to a real aquifer system has demonstrated the importance of conceptual model uncertainty analysis. In this case study, the alternative conceptual models provided fair approximations of the underlying groundwater system. In this data-poor system, model weights were calculated based on hydraulic heads and used for multi-model aggregation of variable different from heads. When model weights obtained solely from head targets were also used for optimal yield estimation in a management context, the information content of the heads was relatively low to discriminate between alternative conceptual models. This situation may regularly be found in real applications since an incomplete set of groundwater heads may be the only data available for the modeling task. If other sources of data, such as river discharge measurement, groundwater flow observations, or groundwater age data were considered, model weights and the relative importance of the data may vary.

The main conclusion of this study is that multi-model averaging should be considered as an important tool for quantifying predictive uncertainty in groundwater management practices. However, the only uncertainties considered in this study involve the conceptual

model and model parameter errors. Not covered are the quantifying uncertainty associated with scenario uncertainty for example, the effect of climate change, changes in groundwater development plans, and the application of the proposed methodology to a transient case. In addition, the uncertainty associated with optimization models, such as the layout of decision variables and constraints and the formulation of any optimization model should be conducted in the future.

Bibliography

Ajami, N., Duan, Q., Gao, X., and Sorooshian S. (2006) Multimodel combination techniques for hydrologic forecasting: Application to distributed model intercomparison project results, *J. Hydrometeorol*., 7(4), 755–768, doi:10.1175/JHM519.1.

Ajami, N., Duan, Q., and Sorooshian, S. (2007) An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resource Research*, 43, W01403, doi:10.1029/2005WR004745.

Ajami, N.K., and Gu, C., (2010) Complexity in microbial metabolic processes in soil nitrogen modeling: a case for model averaging, *Stoch Environ Res Risk Assess* (2010) 24:831–844 DOI 10.1007/s00477-010-0381-4

Akaike, H. (1974) A new look at statistical model identification, IEEE Transactions on Automatic Control, AC-19, 716-722.

Anderson, D. R. (2003) "Multimodel Inference Based on Kullback-Leibler Information." MODFLOW and MORE 2003: Understanding through Modeling

Anderson, M. P., and Woessner, W. W. (1992) *Applied Groundwater Modeling, Simulation of Flow and Advective Transport*, Academic Press, Inc.

Alley W M, Reilly T E, Franke O E (1999) Sustainability of groundwater resources. US Geological Survey Circular 1186, Denver, Colorado.

Aly, A.H., Peralta RC. (1999) Comparison of a genetic algorithm and mathematical programming to the design of groundwater cleanup systems. *Water Resource Research* 1999; 35(8):2415–25.

Anderson, M. P., and Woessner, W. W. (1992) *Applied Groundwater Modeling, Simulation of Flow and Advective Transport*, Academic Press, Inc.

Barlow, P. M., Ahlfeld D. P., et al. (2003) Conjunctive-Management Models for Sustained Yield of Stream-Aquifer Systems, *Journal of Water Resources Planning and Management* 129(1): 35-48.

Bear, J., and A.Verruitj (1987) Modeling Groundwater Flow and Pollution. Boston.

Beven, K.J. 2000. Uniqueness of place and process representations in hydrological modelling. *Hydrology and EarthSystem Sciences* 4, 203–213.

Beven, K. (2006) A manifesto for the equifinality thesis, *Journal of Hydrology*, 320(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K. (2009) Environmental Modelling: An Uncertain Future? An Introduction to Techniques for Uncertainty Estimation in Environmental Prediction. London: Routledge Publishing.

Beven, K., and A. Binley (1992) The future of distributed models: Model calibration and uncertainty prediction, *Hydrological Processes*, 6(5), 279–283, doi:10.1002/hyp.3360060305.

Beven, K., and J. Freer (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, 249(1–4), 11–29, doi:10.1016/S0022-1694(01)00421-8.

Bose, D. and Bose, B. (1995) Evaluation of alternatives for a water project using a multiobjective decision matrix, Water International, 20, 169-175.

Bredehoeft, J. (2002) The water budget myth revisited: Why Hydrogeolgists Model, *Ground Water*, 40(4), 340-345, DOI: 10.1111/j.1745-6584.2002.tb02511.x

Bredehoeft, J. (2003) From models to performance assessment: The conceptualization problem, *Ground Water*, 41(5), 571–577, doi:10.1111/j.1745-6584.2003.tb02395.x.

Bredehoeft, J. (2005) The conceptualization model problem—Surprise, *Hydrogeology Journal*, 13(1), 37–46, doi:10.1007/s10040-004-0430-5.

Brooks, RJ, Lerner D N, Tobias AM (1994) Determining the range of predictions of a groundwater model which arises from alternative calibrations. *Water Resources Research* 30(11): 2993-3000.

Burnham, K. P., and Anderson, D. R. (2002). Model Selection and Multiple Model Inference: A Practical Information-Theoretical Approach, Springer, New York.

Burnham, K. P., and Anderson, D. R. (2004). "Multimodel inference—Understanding AIC and BIC in model selection." *Sociol. Methods Res*, 33(2), 261-304.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., and Slooten, L. J. (2005). "Inverse problem in hydrogeology." *Hydrogeology*, 13, 206-222.

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

Carrera, J., and S. Neuman (1986b), Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, *Water Resources Research*, 22(2), 211–227.

Carrera, J., and S. Neuman (1986c), Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data, *Water Resources Research*, 22(2), 228–242.

Charoensri, P., Srisuk, K., Siriphocagit, V., Dhanesvanich, O., Charurat, A., and Nettasana, T. (2001). "Evaluation of Groundwater Resources Potential and Impact of groundwater Exploitation,Tha Phra Area, Amphoe Mueang, Changwat Khon Kaen, Thailand." Ground Water Division, Department of Mineral Resources.

Christensen, S., and Cooley, R. L. (1999). "Evaluation of prediction intervals for expressing uncertainties in groundwater flow model predictions." *Water Resources Research*, 35(9), 2627-2639.

Cooley, R.L. (1985) A comparison of several methods of solving nonlinear-regression groundwater-flow problems, Water Resource Research, 21(10):1525–1538

Cooley, R., and R. Naff (1990), Regression modeling of ground-water flow, U.S. Geol. Surv. Tech. Water Resource Investigation, Book 3, Chap. B4, 232 pp.

Cooley, R., L. Konikow, and R. Naff (1986) Nonlinear regression groundwater flow modeling of a deep regional aquifer system, Water Resource Research, 22(13), 1759–1778.

Czarnecki, J. B., Clark, B. R., Stanton, G.P. (2003) Conjunctive-use optimization model of the Mississippi River Valley alluvial aquifer of Southeastern Arkansas, U.S. Geological Survey Water-Resources Investigation Report 03-4230. 29 p.

Dagan, G. (1985) Stochastic modelling of groundwater flow by unconditional and conditional probabilities: The inverse problem, Water Resource Research, 21(1), 65–72.

Dagan, G., and S. Neuman (1997) Subsurface Flow and Transport: A Stochastic Approach, 1st ed., 256 pp., Cambridge Univ. Press, Cambridge, U. K.

Das, A. and B. Datta, (1999) Development models for sustainable use of coastal aquifer *J.Irrig*. Drain. Eng., 125, 112–121,

Das, A., and B. Datta. (2001) Application of optimization techniques in groundwater quantity and quality management: *Sadhana* vol. 26 part 4, 293-316..

Delhomme, J. (1979) Spatial variability and uncertainty in groundwater flow model parameters: A geostatistical approach, Water Resources Research, 15(2), 269–280.

Department of Groundwater Resources (2005) Assessment of groundwater resources in five provinces, Northeast Thailand, Department of Groundwater Resources, Thailand [in Thai]

Department of Groundwater Resources (2007) Aquifer Storage and Recovery Pilot Project in the Chi River Basin for the mitigation of flooding and drought conditions, Thaphra Sub-district, Khon Kaen Province, Department of Groundwater Resources, Thailand [in Thai]

Department of Groundwater Resources (2008) The Assessment and mitigation of groundwater level decline and saline intrusion in Thaphra Sub-district, Khon Kaen province and its vicinity area, Department of Groundwater Resources, Thailand [in Thai]

Department of Mineral Resources, Khon Kaen University, Khon Kaen Brewery Co., Ltd and Thai Namtip Co. (2001) Evaluation of groundwater resources potential and impact of groundwater exploitation, Thaphra area, Muang District, Khon Kaen province, Thailand Ground Water Division, Department of Mineral Resources, Thailand [in Thai]

Department of Mineral Resources (2002) The Proceedings of the symposium on geology of Thailand, Bangkok, Thailand

Department of Mineral Resources (1999) Geological map of Thailand, scale of 1:1,000,000, Geological Survey Division, Department of Mineral Resources, Bangkok, Thailand

Diks, C.G.H., Vrugt, J.A.(2010) Comparison of point forecast accuracy of model averaging methods in hydrologic applications, *Stoch Environ Res Risk Assess*, 24:809-820, Doi 10.1007/s00477-010-0378-z

Doherty, J. (2003) Groundwater model calibration using pilot points and regularization, *Ground Water*, 41(4), 170-177.

Doherty J (2004) PEST-Model-Independent-Parameter-Estimation user's manual Watermark Computing, Australia

Doherty, J. (2007) Model-Independent Parameter Estimation, Watermark Numerical Computing. Corinda (Australia). Available at http://www.sspa.com/PEST/index.shtml

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

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc*., Ser. B, 57(1), 45–97.

Dou, C., Woldt, W., Bogardi, I., and Dahab, M (1995) Steady state ground water flow simulation with imprecise parameters, *Water Resource Research*, 31(11), 2709-2719

Duan, Q., Ajami, N., Gao, X., and Sorooshian, S. (2007) Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Advances in Water Resources*, 30(5), 1371–1386, doi:10.1016/j.advwatres.2006.11.014.

Environmental Simulations, Inc. (ESI) (2007) Guides to using Groundwater Vista: Version 5, Environmental Simulations. Inc., 372 pp.

Feyen, L., Beven, K., De Smedt, F., and Freer J. (2001) Stochastic capture zone delineation within the Generalized Likelihood Uncertainty Estimation methodology: Conditioning on head observations, *Water Resource Research*, 37(3), 625–638.

Feyen L., Gorelick S.M. (2004) Reliable groundwater management in hydroecologically sensitive areas. Water Resour Res 40(7): W07408. doi:10.1029/2003WR003003

Feyen L., Gorelick S.M. (2005) Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. Water Resour Res 41(3):W03019. doi:10.1029/2003WR002901

Feyen, L., Ribeiro, P., De Smedt, F., and Diggle, P. (2002) Bayesian methodology to stochastic capture zone determination: Conditioning on transmissivity measurements, *Water Resources Research*, 38(9), 1164, doi:10.1029/2001WR000950.

Feyen, L., Gómez-Hernández, J., Ribeiro, P., Beven, K., and De Smedt, F. (2003) A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations, *Water Resource Research*, 39(5), 1126, doi:10.1029/2002WR001544.

Fienen, M., T. Clemo, and P. Kitanidis (2008) An interactive Bayesian geostatistical inverse protocol for hydraulic tomography, *Water Resource Research*, 44, W00B01, doi:10.1029/2007WR006730.

Fienen, M., Hunt R., Krabbenhoft, D., and Clemo, T. (2009) Obtaining parsimonious hydraulic conductivity fields using head and transport observations: A Bayesian geostatistical parameter estimation approach, *Water Resource Research*, 45, W08405, doi:10.1029/2008WR007431.

Foglia, L., Mehl, S. W., Hill, M. C., Perona, P., and Burlando, P. (2007) Testing Alternative Ground Water Models Using Cross-Validation and Other Methods. *Ground Water*, 45(5), 627-641.

Foglia, L., Hill, M., Mehl, S., and Burlando, P. (2009) Sensitivity analysis, calibration, and testing of a distributed hydrological model using error based weighting and one objective function, Water Resource Research, 45,W06427, doi:10.1029/2008WR007255.

Gaganis, P., and Smith, L. (2006) Evaluation of the uncertainty of groundwater model predictions associated with conceptual error: A per-datum approach to model calibration." *Advances in Water Resources*, 29, 503-514.

Gallagher, M., and Doherty, J. (2007) Parameter estimation and uncertainty analysis for a watershed model, *Environ. Modell. Software*, 22(7), 1000–1020, doi:10.1016/j.envsoft.2006.06.007.

Georgakakos, K., Seo, D., Gupta, H., Schaake, J., and Butts, M. (2004) Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *Journal of Hydrology*, 298(1–4), 222–241, doi:10.1016/j.jhydrol.2004.03.037.

George, E. I. (1999) Sampling considerations for model averaging and model search. Invited discussion of "Model Averaging and Model Search, by M. Clyde. In Bayesian Statistics 6 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 175–177. Oxford Univ. Press.

Gharbi, A., and Peralta, R. C. (1994) Integrated embedding optimization applied to Salt Lake Valley aquifers." *Water Resources Research*, 30(3), 817-832.

Ghosh, J., M. Delampady, and T. Samanta (2006) An Introduction to Bayesian Analysis—Theory and Methods, 1st ed., 352 pp., Springer, New York.

Goldberg, D.E. (1989) Genetic algorithms in search, optimization and machine learning, Addison-Wesley-Longman, Reading, MA.

Gómez-Hernández, J. (2006) Complexity, *Ground Water*, 44(6), 782–785, doi:10.1111/j.1745-6584.2006.00222.x.

Gómez-Hernández, J., Sahuquillo, A., and Capilla, J. (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data: I. Theory, *Journal of Hydrology*, 203(1–4), 162–174, doi:10.1016/S0022-1694(97)00098-X.

Harbaugh A. W., Banta, E. R., Hill, M. C., McDonald, M. G. (2000) MODFLOW–2000, the US Geological Survey Modular Ground-Water Model-User guide to modularization concepts and the Ground-Water Flow Process US Geological Survey Open-File Rep 00-92

Harrar, W., Sonnenberg, T., and Henriksen, H. (2003) Capture zone, travel time, and solute transport predictions using inverse modelling and different geological models, *Hydrogeology Journal*, 11(5), 536–548, doi:10.1007/s10040-003-0276-2.

Hassan, A., H. Bekhit, and J. Chapman (2008) Uncertainty assessment of a stochastic groundwater flow model using GLUE analysis, *Journal of Hydrology*, 362(1–2), 89–109, doi:10.1016/j.jhydrol.2008.08.017.

Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57(1), 97–109, doi:10.1093/biomet/ 57.1.97.

Hays, P. D. (2000) Sustainable-yield estimation for the Sparta Aquifer in Union County, Arkansas, U.S. Geological Survey Water-Resources Investigation Report 99-4274. 17 p.

Hill, M.C. (2006) The practical use of simplicity in developing ground water models, Ground Water, 44(6), 775–781, doi:10.1111/j.1745-6584.2006. 00227.x.

Hill, M.C., Banta, E.R., Harbaugh, A.W., Anderman E.R. (2000) MODFLOW-2000, The U.S. Geological Survey modular ground-water model-User guide to the observation, sensitivity, and parameter-estimation processes. U.S.Geological Survey Open-File Report 00-184.

Hill, M.C., Cooley, R., and Pollock, D. (1998) A controlled experiment in ground water flow model calibration, *Ground Water*, 36(3), 520–535, doi:10.1111/j.1745-6584.1998.tb02824.x.

Hill, M.C. (1998) Methods and guidelines for effective model calibration: with application to UCODE, a computer code for universal inverse modeling, and MODFLOWP, a computer code for inverse modeling with MODFLOW: U.S. Geological Survey Water-Resources Investigations Report 98-4005

Hill, M.C., Tiedeman, C. R. (2007) Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty, John Wiley & Sons, Inc

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999) Bayesian model averaging: A tutorial, *Stat. Sci*., 14(4), 382–401.

Højberg, A., and Refsgaard, J. (2005) Model uncertainty—Parameter uncertainty versus conceptual models, *Water Sci. Technol*., 52(6), 177–186.

Hsu, K., Moradkhani, H., and Sorooshian, S. (2009) A sequential Bayesian approach for hydrologic model selection and prediction, *Water Resource Research*, 45, W00B12, doi:10.1029/2008WR006824.

Huang, C. and Mayer, A. S. (1997) Pump-and-treat optimization using well locations and pumping rates as decision variables, *Water Resource Research*, Vol. 33, pp. 1001–1012.

Hunt, R., J. Doherty, and M. Tonkin (2007) Are models too simple? Arguments for increased parameterization, *Ground Water*, 45(3), 254–262, doi:10.1111/j.1745-6584.2007.00316.x.

Hurvich, C.M., and C-L. Tsai (1989) Regression and time series model selection in small sample, Biometrika, 76(2), 99-104.

Ijiri, Y., H. Saegusa, A. Sawada, M. Ono, K. Watanabe, K. Karasaki, C. Doughty, M. Shimo, and K. Fumimura (2009) Evaluation of uncertainties originating from the different modeling approaches applied to analyze regional groundwater flow in the Tono area of Japan, *J. Contam. Hydrol*., 103(3–4), 168–181, doi:10.1016/j.jconhyd.2008.10.010.

James, A. L., and Oldenburg, C. M. (1997) Linear and Monte Carlo Uncertainty Analysis for Subsurface Contaminant Transport Simulation, *Water Resources Research*, 33(11), 2435-2508.

Japakasetre T (1985) Review on rock salt and potash exploration in Northeast Thailand conference on geology and mineral resources development of the Northeast, Khon Kaen, Thailand: 135-147

Jensen, J. (2003) Parameter and uncertainty estimation in groundwater modelling, Ph.D. thesis, Dep. of Civ. Eng., Aalborg Univ., Aalborg, Denmark.

Kashyap, R.L. (1982) Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(2), 99-104.

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

Konikow, L. F., and Bredehoeft, J. (1992) Ground-water models cannot be validated, *Advances in Water Resources*, 15, 75-83.

Krzysztofowicz, R. (1999), Bayesian theory of probabilistic forecasting via deterministic hydrologic model, *Water Resource Research*, 35(9), 2739–2750.

Kuchanur, M.P. (2006) Simulation-Optimization methodologies to estimate groundwater availability, Doctoral Dissertation, Texas A&M University-Kingsville, 188 pp

Kunstmann, H. (2002) Conditional first-order second-moment method and its application to the quantification of uncertainty in groundwater modeling, *Water Resources Research*, 38(4), 1035, 10.1029/2000WR000022.

LaVenue, A., RamaRao, B., de Marsily, G., and Marietta, M. (1995) Pilot points methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 2. Application, *Water Resource Research*, 31(3), 495–516.

Li, X., and Tsai, F. (2009) Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod, *Water Resource Research*, 45, W09403, doi:10.1029/2008WR007488.

Madigan, D., and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89(428), 1535-1546.

Mantovan, P., and Todini, E. (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, 330(1–2), 368–381, doi:10.1016/j.jhydrol.2006.04.046.

McDonald, M. G., and Harbaugh, A. W. (1988) A Modular three-dimensional finite-difference ground-water flow model. U.S. Geological Survey Techniques of Water Resources Investigations, Book 6, Chapter A1. Available at http://water.usgs.gov/pubs/twri/twri6a1.

McKinney, D.C. and Lin, M., (1994) Genetic algorithms solution of groundwater management models, *Water Resource Research*, Vol. 30, pp. 1897–1906.

McKee, P. W., Clark, B. R., Czarnecki, J.B. (2004) Conjunctive-Use Optimization Model and Sustainable-yield Estimation for the Sparta Aquifer of Southeastern Arkansas and North-Central Louisiana, U.S. Geological Survey Water-Resources Investigations Report o3-4231, 30 p.

McLaughlin, D., and E. Wood (1988) A distributed parameter approach for evaluating the accuracy of groundwater model predictions: 2. Application to groundwater flow, *Water Resource Research*, 24(7), 1048–1060.

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

McPhee, J., and Yeh, W.W.-G. (2004) Multiobjective Optimization for Sustainable Groundwater Management in Semiarid Regions, *J. Water Resources Planning and Management*, DOI: 10.1061/(ASCE)0733-9496(2004)130:6(490)

Meesuk A, Suteethorn V, Chairangsee C, Wongprayoon T (1998) IGCG 350 Special field trip on Mesozoic Red Beds of Northeast Thailand. Bangkok: Geological Survey Division, Department of Mineral Resources

Meyer, P., M. Ye, S. Neuman, and K. Cantrell (2004) Combined estimation of hydrogeologic conceptual model and parameter uncertainty, Rep. NUREG/CR-6843 PNNL-14534, U.S. Nucl. Regul. Comm., Washington, D. C.

Meyer, P., M. Ye, M. Rockhold, S. Neuman, and K. Cantrell (2007) Combined estimation of hydrogeologic conceptual model parameter and scenario uncertainty with application to uranium transport at the Hanford site 300 area, Rep. NUREG/CR-6940 PNNL-16396, U.S. Nucl. Regul. Comm., Washington, D. C.

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-bmaisnt-mc.pdf.

Morgan, D.R., Eheart, J.W., Valocchi, A.J. (1993) Aquifer remediation design under uncertainty using a new chance constrained programming technique. Water Resour Res 29(3):551–561

Montanari, A. (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall–runoff simulations. *Water Resources Research*, 41, W08406. doi:10.1029/2004WR003826.

Moore, C., and J. Doherty (2005) Role of the calibration process in reducing model predictive error, *Water Resource Research*, 41, W05020, doi:10.1029/2004WR003501.

Morales-Casique, E., Neuman, S.P., and Vesselinov, V.V. (2010) Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows, *Stoch Environ Res Risk Assess*, 24:863-880, Doi 10.1007/s00477-010-0383-2

Morgan, M. G., and Henrion, M. (1990) Uncertainty a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, New York.

Morse, B., G. Pohll, J. Huntington, and R. Rodriguez (2003) Stochastic capture zone analysis of an arsenic-contaminated well using the generalized likelihood uncertainty estimator (GLUE) methodology, *Water Resource Research*, 39(6), 1151, doi:10.1029/2002WR001470.

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

Nettasana, T. (2002) Groundwater Balance Study in Thaphra sub-district, Muang District, Khon Kaen Province and its Adjacent Area, ISBN 974-367-175-7, Khon Kaen University.

Nettasana T., Criag, J.., and Tolson, B. (2012) Conceptual and Numerical Models for Sustainable Groundwater Management in the Thaphra Area, Chi River Basin, Thailand, *Hydrogeology Journal*.

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

Neuman, S., and P. Wierenga (2003) A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, Rep. NUREG/CR-6805, U.S. Nucl. Regul. Comm., Washington, D. C.

Oreskes, N.K. Shrader-Frechette and K.Belitz (1994) Verification, validation, and confirmation of numerical models in the earth sciences, *Science* 263:641-646.

Pappenberger, F., and K. Beven (2006) Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resource Research*, 42, W05302, doi:10.1029/2005WR004820.

Pasquier, P., and D. Marcotte (2006) Steady- and transient-state inversion in hydrogeology by successive flux estimation, *Advances in. Water Resources*, 29(12), 1934–1952, doi:10.1016/j.advwatres.2006.02.001.

Poeter, E. (2007) All Models are Wrong, How Do We Know Which are Useful? *Ground Water*, 45(4), 390-391.

Poeter, EP, Hill MC, Banta ER, Mehl SW (2005) Ucode_2005 and three post-processors-computer codes for universal sensitivity analysis, inverse modeling, and uncertainty evaluation. U.S. Geological Survey techniques and Methods Report TM6A-11.

Poeter, E., and Anderson, D. (2005) Multimodel ranking and inference in ground water modelling, *Ground Water*, 43(4), 597–605, doi:10.1111/ j.1745-6584.2005.0061.x.

Poeter, E., and Hill, M. (1997) Inverse models: A necessary step in groundwater modeling, *Ground Water*, 35(2), 250–260, doi:10.1111/j.1745-6584.1997.tb00082.x.

Poeter, E., and Hill, M. C. (2007) MMA, A Computer Code for Multi-Model Analysis:U.S. Geological Survey Techniques and Methods 6-E3.

Poeter, E., and S. McKenna (1995), Reducing uncertainty associated with ground-water flow and transport predictions, *Ground Water*, 33(6), 899–904, doi:10.1111/j.1745-6584.1995.tb00034.x.

Pohlmann, K., Ye, M., Reeves, D., Zavarin, M., Decker, D., and Chapman, J. (2007) Modeling of Groundwater Flow and Radionuclide Transport at the Climax Mine sub-CAU. Nevada Test Site, DRI Publication 45226, DOE/NV/26383-06, Nevada Site Office, National Nuclear Security Administration. Las Vegas, Nevada: U.S. Department of Energy.

Qahman K., Larabi A., Ouazar D., Naji A., and Cheng A. H.-D.(2005) Optimal and Sustainable Extraction of Groundwater in Coastal Aquifers, *Stochastic Environmental Research and Risk Assessment Journal*, Volume 19 (2).

Rambaugh J O, Rambough D B (2007) Guide to using Groundwater Vistas: version 5 Environmental Simulations, New York

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*, 25, 111-163.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski ,M. (2005) Using bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev; 113:1155–74.

Raftery, A., and Y. Zhang (2003) Discussion: Performance of Bayesian model averaging, *J. Am. Stat. Assoc*., 98(464), 931–938.

RamaRao, B., A. LaVenue, G. de Marsily, and M. Marietta (1995) Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments, *Water Resource Research*, 31(3), 475–493.

Refsgaard, J., J. Van der Sluijs, A. Højberg, and P. Vanrolleghem (2005) Harmoni-CA guidance uncertainty analysis, technical report, Eur. Comm., Brussels.

Refsgaard, J., J. Van der Sluijs, J. Brown, and P. Van der Keur (2006) A framework for dealing with uncertainty due to model structure error, *Advances in Water Resources*, 29(11), 1586–1597, doi:10.1016/j.advwatres.2005. 11.013.

Refsgaard, J., J. Van der Sluijs, A. Højberg, and P. Vanrolleghem (2007) Uncertainty in the environmental modelling process—A framework and guidance, *Environ. Modell. Software*, 22(11), 1543–1556, doi:10.1016/j. envsoft.2007.02.004.

Reeves, D.M., Pohlmann, K.F., Pohll, G.M., and Chapman, J.B. (2010) Incorporation of conceptual and parametric uncertainty into radionuclide flux estimates from a fractured granite rock mass, *Stoch Environ Res Risk Assess*, 24:899-915, Doi 10.1007/s00477-010-0385-0

Renard, P. (2007) Stochastic hydrogeology: What professionals really need? *Ground Water*, 45(5), 531–541, doi:10.1111/j.1745-6584.2007. 00340.x.

Robert, C. (2007) The Bayesian Choice—From Decision-Theoretic Foundations to Computational Implementation, 2nd ed., 577 pp., Springer, New York.

Rojas, R., Feyen, L., and Dassargues, A. (2008a) Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resource Research*, 44, W12418, doi:10.1029/2008WR006908.

Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., and Dassargues, A. (2008b) Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, *Journal of Hydrology*, 394, Issues 3-4, 416-435.

Rojas, R., Feyen, L., and Dassargues A. (2009) Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling, *Hydrol. Processes*, 23(8), 1131–1146, doi:10.1002/hyp.7231.

Rojas, R., O. Batelaan, L. Feyen, and A. Dassargues (2010) Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal–north Chile, *Hydrol. Earth Syst. Sci*., 14(2), 171–192.

Royal Irrigation Department (RID) (2009) Hydrological data, Hydrological Division, Office of Hydrology and Water management, Royal Irrigation Department, Bangkok.

Rubin, Y. (2003) Applied Stochastic Hydrogeology, 1st ed., 416 pp., Oxford Univ. Press, New York.

Rubin, Y., and G. Dagan (1987) Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 1. Theory, Water Resource Research, 23(7), 1185–1192.

Ruksaskulwong M, Munjai D, and Janchat A (2007) General geology for Khon Kaen province (5541 I), Ban Hin Tank (5541 II) scale 1:50,000, Conferences on Geology Investigation, Bureau of Geology, Department of Mineral Resources, Bangkok, Thailand

Schwarz, G. (1978) Estimating the dimension of a model. Annals of Statistics 6, no. 2: 461–464.

Seifert, D., Sonnenberg, T., Scharling, P., and Hinsby, K. (2008) Use of alternative conceptual models to assess the impact of a buried valley on groundwater vulnerability, *Hydrogeology Journal*, 16(4), 659–674, doi:10.1007/s10040-007-0252-3.

Shafike, N., Duckstein, G.L., and Maddock, T., (1992) Multicriterion analysis of groundwater contamination management, *Water Resources Bulletin*, 28(1), pp. 33-43.

Srisuk, K. (1994) Genetic characteristics of the groundwater regime in the Khon Kaen Drainage Basin, Northeast Thailand PhD in Geology, University of Alberta, Canada

Srisuk K., Toth, J. (1994) Groundwater salinity and three-dimensional groundwater flow model at Ban Nong Khai Nun, Khon Kaen, Northeast Thailand in Thanvarachorn P, Hokjaroen S, and Youngme W (Eds) Conference on Geology and Mineral Resources Development of the Northeast, Thailand: 186-216

Suttayarak N (1985) Review on geology of Khorat Plateau, Proc of the Conf on Geology and Mineral Resources Development of the Northeast Thailand, Khon Kaen

Seifert, D., T. Sonnenberg, P. Scharling, and K. Hinsby (2008), Use of alternative conceptual models to assess the impact of a buried valley on groundwater vulnerability, *Hydrogeology Journal*, 16(4), 659–674, doi:10.1007/s10040-007-0252-3.

Singh, A., Minsker, B.S .(2008) Uncertainty-based multiobjective optimization of groundwater remediation design. Water Resour Res 44(2):W02404. doi:10.1029/2005WR004436

Singh, A., S. Mishra, and G. Ruskauff (2010), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, doi:10.1111/j.1745-6584.2009.00642.x.

Sorensen, D., and Gianola, D. (2002) Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics, vol. 1, 1st ed., 740 pp., Springer, New York.

Swaify, El., and D.S. Yakowitz, (1998) Multiple objective decision making for land, water, and environmental management proceedings of the First International Conference on Multiple Objective Decision Support Systems (MODSS) for Land, Water, and Environmental Management: Concepts, Approaches, and Applications,Boca Raton, Fla. ; London : Lewis Publishers c1998

Szidarovsky, F., Gershon, M., and Duckstein, L. (1986) Techniques for multiobjective decision making in systems management. Amsterdam, Elsevier.

Tiedeman, C., Goode, D., and Hsieh, P. (1997) Numerical simulation of ground-water flow through glacial deposits and crystalline bedrock in the Mirror Lake area, Grafton County, New Hampshire, U.S. Geol. Surv.Prof. Pap., 1572.

Tiedeman, C., Goode, D., and Hsieh, P. (1998) Characterizing a ground water basin in a New England mountain and valley terrain, *Ground Water*, 36(4), 611–620, doi:10.1111/j.1745-6584.1998.tb02835.x.

Tiedeman, C., Hill, M., D'Agnese, F., and Faunt, C. (2003) Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system, *Water Resources Research*, 39(1), 1010, doi:10.1029/2001WR001255.

Tiedeman, C., D. Ely, M. Hill, and G. O'Brien (2004) A method for evaluating the importance of system state observations to model predictions, with application to the Death Valley regional groundwater flow system, *Water Resources Research*, 40, W12411, doi:10.1029/2004WR003313.

Tierney, L. (1994), Markov chains for exploring posterior distributions, Ann. Stat., 22(4), 1701–1728.

Tkach, R.J., and Simonovic, S., (1997) A new approach to multi-criteria decision making in water resources, *Journal of geographic information and decision analysis* 1(1):25-43.

Tonkin, M., Tiedeman, C., Ely, D., and Hill, M. (2007) OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics, U.S. Geol. Surv. Tech. Methods, 6-E2.

Troldborg, L., Refsgaard, J., Jensen, K., and Engesgaard, P. (2007) The importance of alternative conceptual models for simulation of concentrations in a multi–aquifer system, *Hydrogeology J.*, 15(5), 843–860, doi:10.1007/s10040-007-0192-y.

Tsai, F., and X. Li (2008) Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, Water Resources Research, 44, W09434, doi:10.1029/2007WR006576.

Tsai, F., and X. Li (2010) Reply to comment by Ming Ye et al. on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window", Water Resources Research, 46, W02802, doi:10.1029/2009WR008591.

Tung, Y.K. (1986) Groundwater management by chance-constrained model. J Water Resour Plan Manag - ASCE 112(1):1–19.

Usunoff, E., Carrera, J., and Mousavi, S. F. (1992) "An Approach to design of experiments for discriminating among alternative conceptual models." *Advances in Water Resources*, 15, 199-214.

U.S. Environmental Protection Agency (1997) Guiding Principles for Monte Carlo Analysis, Risk Assessment Forum. EPA/630/R-97/001.

U.S. Environmental Protection Agency (2009) Secondary drinking water regulations: guidance for nuisance chemicals. EPA816-F-10-079, available at http://water.epa.gov/drink/contaminants/secondarystandards.cfm (last accessed August 01/2011)

Van der Sluijs, J. (2005) Uncertainty as a monster in the science-policy interface: Four coping strategies, *Water Sci. Technol*., 52(6), 87–92.

Vecchia A.V., and R.L. Cooley (1987) Simultaneous confidence and prediction intervals for non-linear regression models with application to a groundwater flow model, *Water Resources Research*, 23(7): 1237-1250.

Vogel, R.M., Batchelder, R., and Stedinger, J.R. (2007) Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) method. *Water Resources Research* 44, WOOB06.doi:10.1029/2008WR006822, 2008.

Vrugt, J., and B. Robinson (2007) Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resources Research*, 43,W01411, doi:10.1029/2005WR004838.

Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., and Robinson, B.A. (2009) Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, *Stochastic Environmental Research and Risk Assessment*, 23(7), 1011-1026, doi:10.1007/s00477-008-0274-y.

Vrugt, J., M. Clarck, C. Diks, Q. Duan, and B. Robinson (2006) Multiobjective calibration of forecast ensembles using Bayesian model averaging, *Geophys. Res. Lett.*, 33, L19817, doi:10.1029/2006GL027126.

Wagner, B.J., Gorelick, S.M. (1987) Optimal groundwater quality management under parameter uncertainty. Water Resour Res 23(7):1162–1174

Wagner, B.J., Gorelick, S.M. (1989) Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: from data to design. Water Resour Res 25(10):2211–2225

Wagner, B. (1995) Sampling design methods for groundwater modeling under uncertainty, *Water Resources Research*, 31(10), 2581–2591.

Walker, W., and Marchau, V. (2003) Dealing with uncertainty in policy analysis and policy making, *Integr. Assess.*, 4(1), 1–4.

Wasserman, L. (2000) Bayesian model selection and model averaging, *J. Math. Psychol.*, 44(1), 92–107, doi:10.1006/jmps.1999.1278.

Wintle, B.A., McCarthy, M.A., Volinsky, C.T. (2003) The use of Bayesian Model Aveaging to better represent uncertainty in Ecological models, *Conservation Biology*, volume 17 (6), 1579-1590.

Wöhling, T., and Vrugt J. (2008) Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models, *Water Resources Research*, 44, W12432, doi:10.1029/ 2008WR007154.

Wongsawat S., Dhanesvanich O, Tanthakasem S, Kontos N, Rojanajan S, Panjasutharos S, Prayoonchat C, and Nettasana T (1989a) Groundwater availability map of Khon Kaen Province, Northeast Thailand, Groundwater Division, Department of Mineral Resources and Thai-Australian Northeast Village Water Resource Project

Wongsawat S, Srisuk K, Rojanajan S, Tantikaseme S (1989b) Groundwater resource evaluation pilot project, progress report Thai-Australian Northeast Village Water Resource Project, Khon Kaen, Thailand

Ye, M., Neuman, S., and Meyer, P. (2004) Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, Water Resources Research, 40, W05113, doi:10.1029/2003WR002557.

Ye, M., Neuman, S., Meyer, P., and Pohlmann, K. (2005) Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff, *Water Resources Research*, 41, W12429, doi:10.1029/2005WR004260.

Ye, M., Pohlmann, K., Chapman, J., and Shafer, D. (2006) On evaluation of recharge model uncertainty: A priori and a posterior, in 2006 International High Level Radioactive Waste Management Conference, American Nuclear Society, Las Vegas Nevada US.

Ye, M., Pohlmann, K., and Chapman J. (2008a) Expert elicitation of recharge model probabilities for the Death Valley regional flow system, *J. Hydrol*., 354(1–4), 102–115, doi:10.1016/j.jhydrol.2008.03.001.

Ye, M., Meyer, P., and Neuman, S. (2008b) On model selection criteria in multimodel analysis, *Water Resources Research*, 44, W03428, doi:10.1029/ 2008WR006803.

Ye, M., Pohlman, K., Chapman, J., Pohll, G., and Reeves, D. (2010a) A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, doi:10.1111/j.1745-6584.2009.00633.x, in press.

Ye, M., D. Lu, S. Neuman, and P. Meyer (2010b) Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li, *Water Resources Research*, 46, W02801, doi:10.1029/2009WR008501.

Yeh, W. W.-G. (1986) Review of Parameter Identification Producres in Groundwater Hydrology: The Inverse Problem, *Water Resources Research*, 22(2), 95-108

Zeleny, M. (1982) Multiple criteria decision Making. New York, McGraw-Hill Book Company.

Zadeh, L.A. (1965) Fuzzy Sets Information Control 8:338-359.

Zheng, C., and Wang, P. (1999) MT3DMS, A Modular Three-Dimensional Multi-Species Transport Model for Simulation of Advective, Dispersion and Chemical Reactions of Contaminants in Groundwater Systems; Documentation and User's Guid, U.S. Army Engineer Research and Development Center Contract Report SERDP-99-1, Vicksburg, MS. Available at http://hydro.geo.ua.edu/mt3d/.

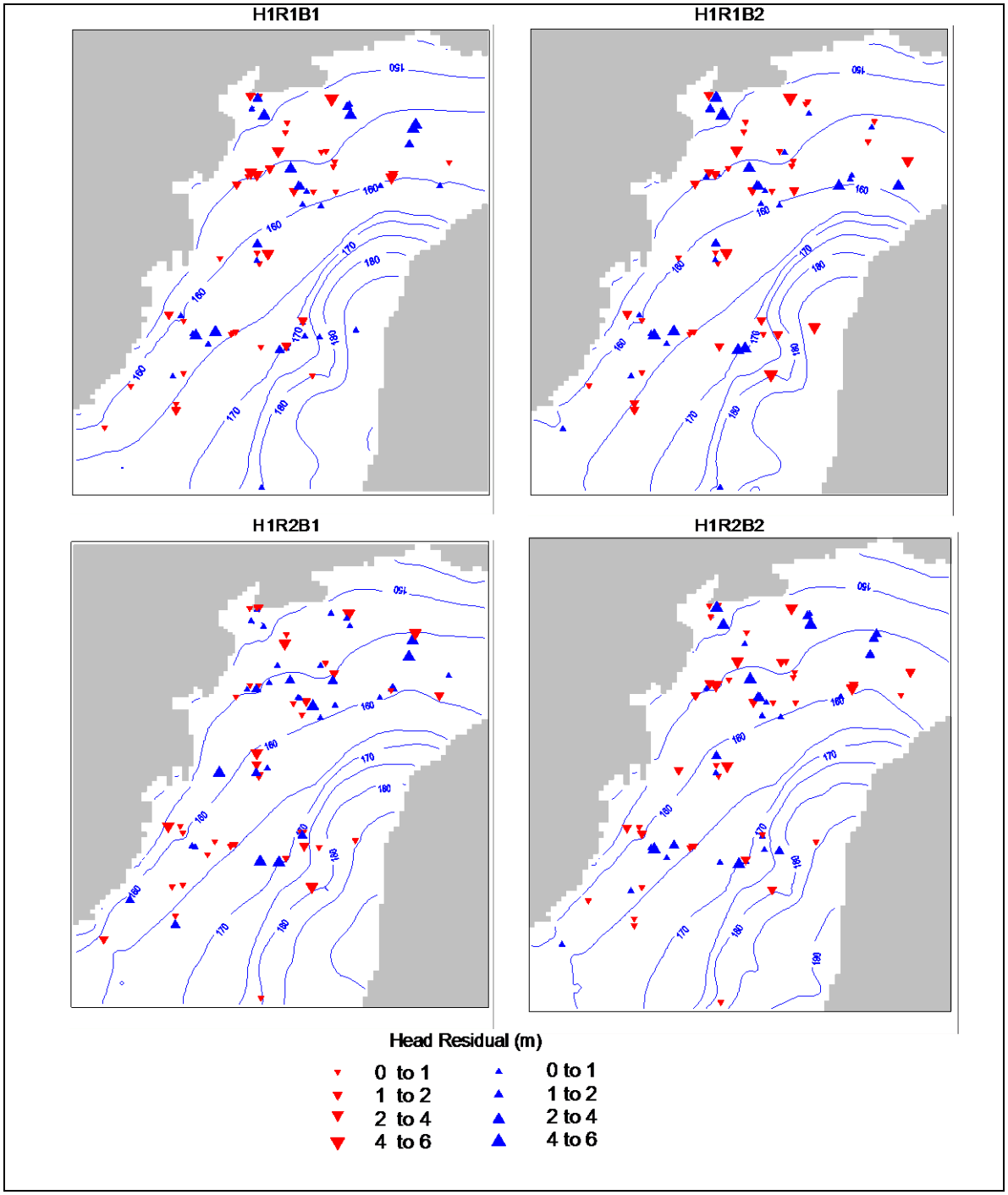
Zheng, C., and Wang, P. P. (2003) MGO: A Modular Groundwater Optimizer incorporating MODFLOW and MT3DMS; Document and User's Guide, The university of Alabama and Groundwater Systems Research Ltd. (Available at http://www.frtr.gov/estcp/source_codes.htm).

Appendices

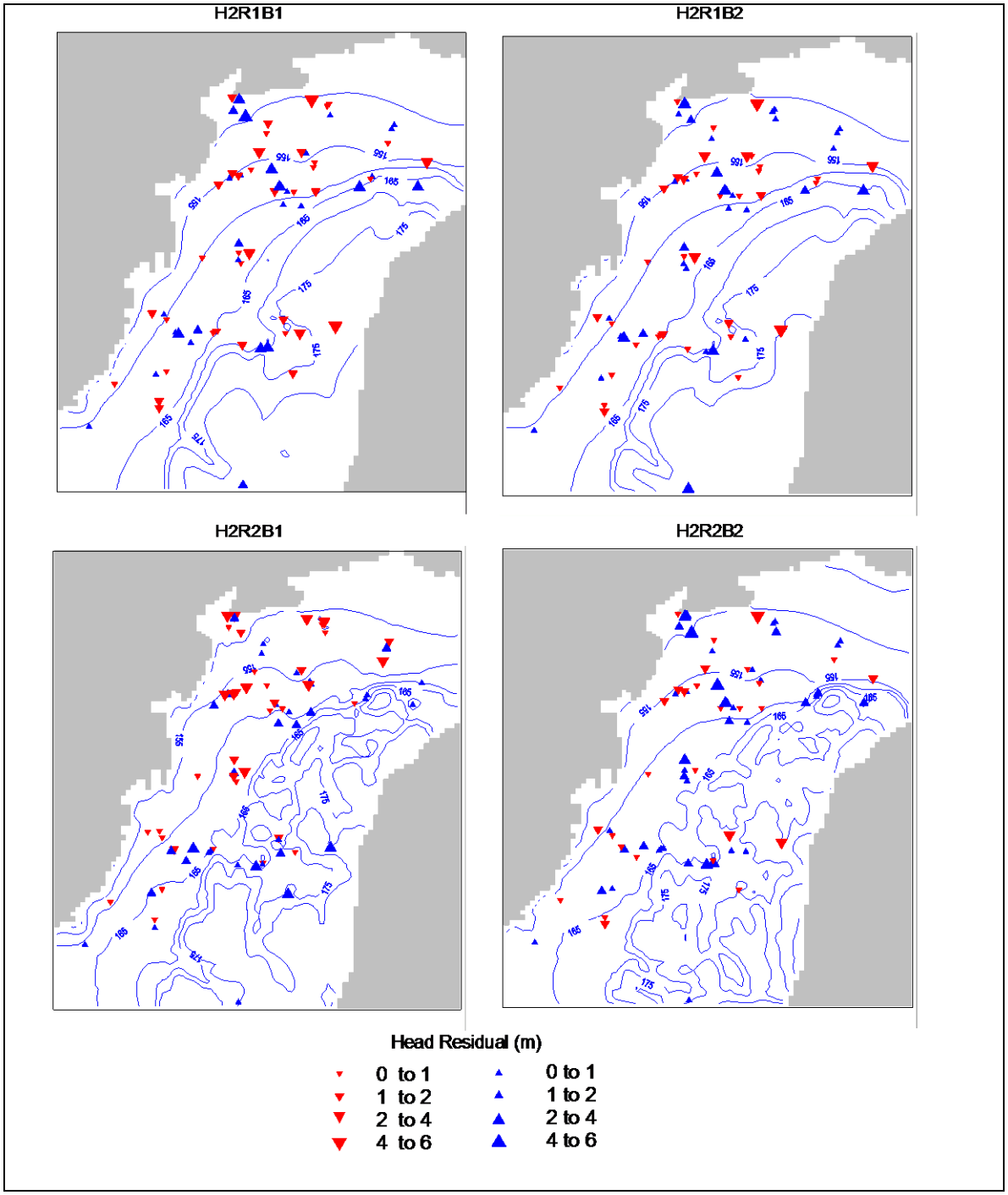
Appendix A

Scatter plots of heads residual and contours of hydraulic heads at model layer 4 in TPA

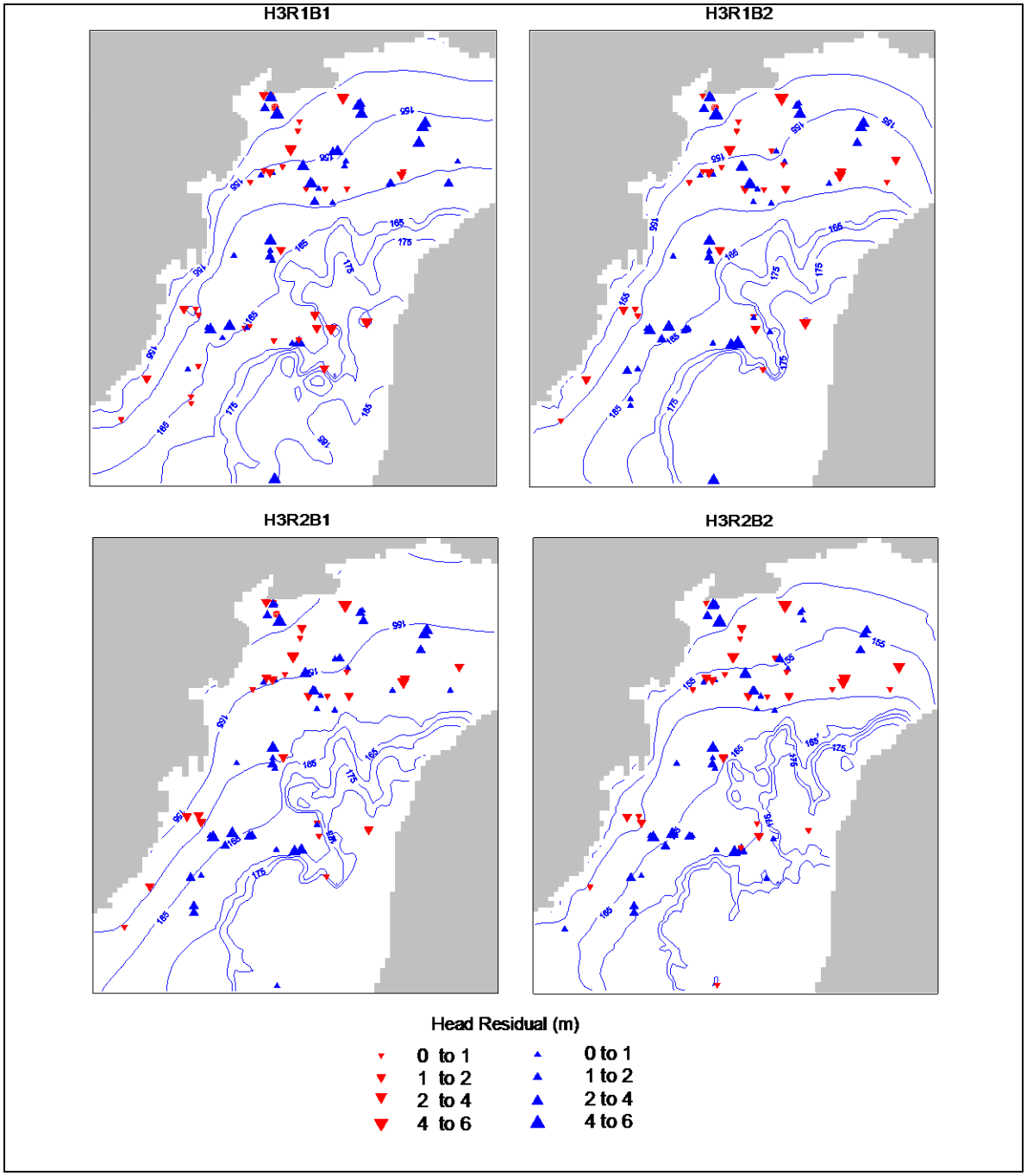
A-1 Scatter plots of heads residual and contours of hydraulic heads at model layer 4 in TPA simulated by model H1R1B1, H1R1B2, H1R2B1, and H1R2B2



A-2 Scatter plots of heads residual and contours of hydraulic heads at model layer 4 in TPA simulated by model H2R1B1, H2R1B2, H2R2B1, and H2R2B2

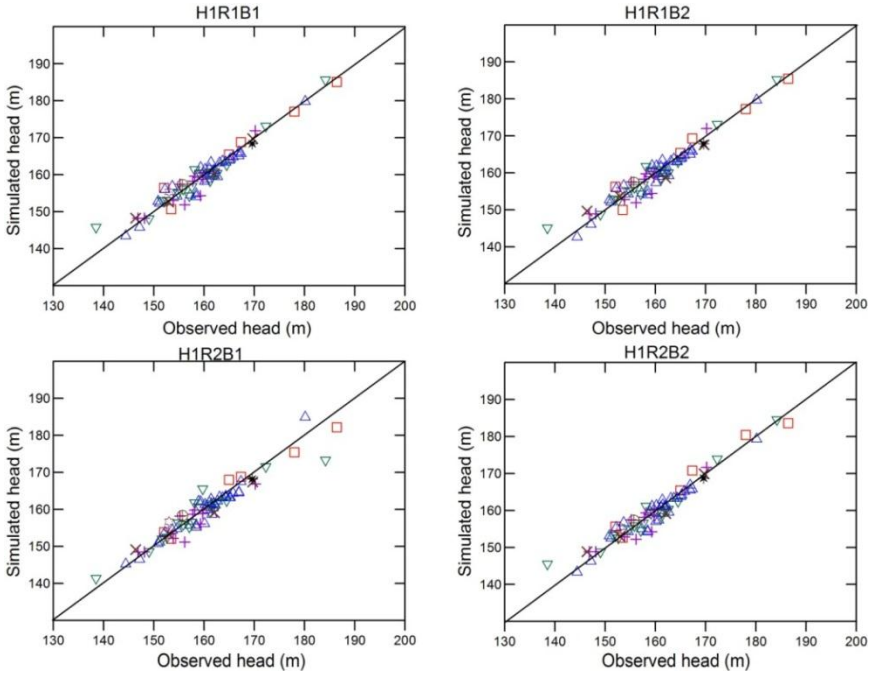


A-3 Scatter plots of heads residual and contours of hydraulic heads at model layer 4 in TPA simulated by model H3R1B1, H3R1B2, H3R2B1, and H3R2B2

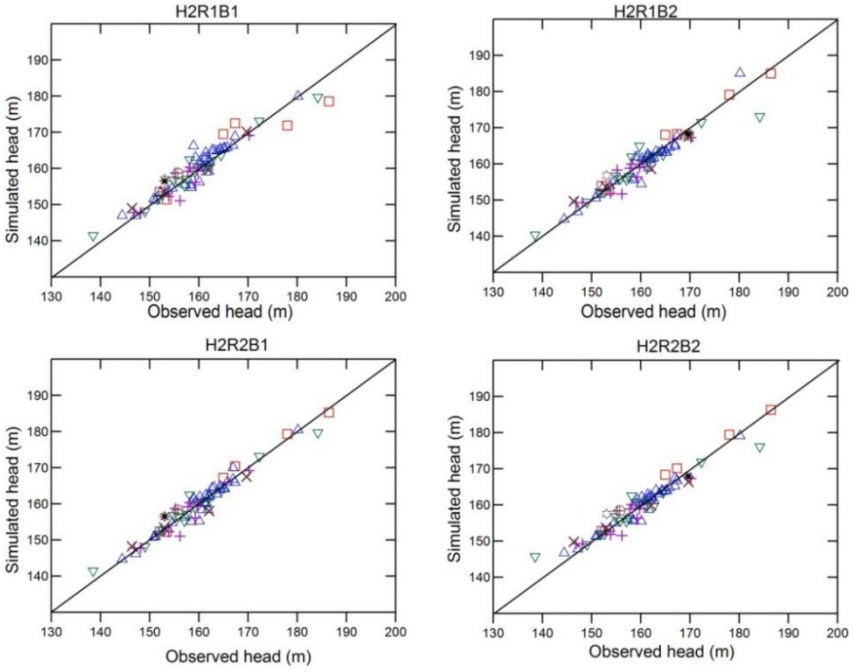


Appendix B
Observed vs. simulated heads for the groundwater model of the TPA

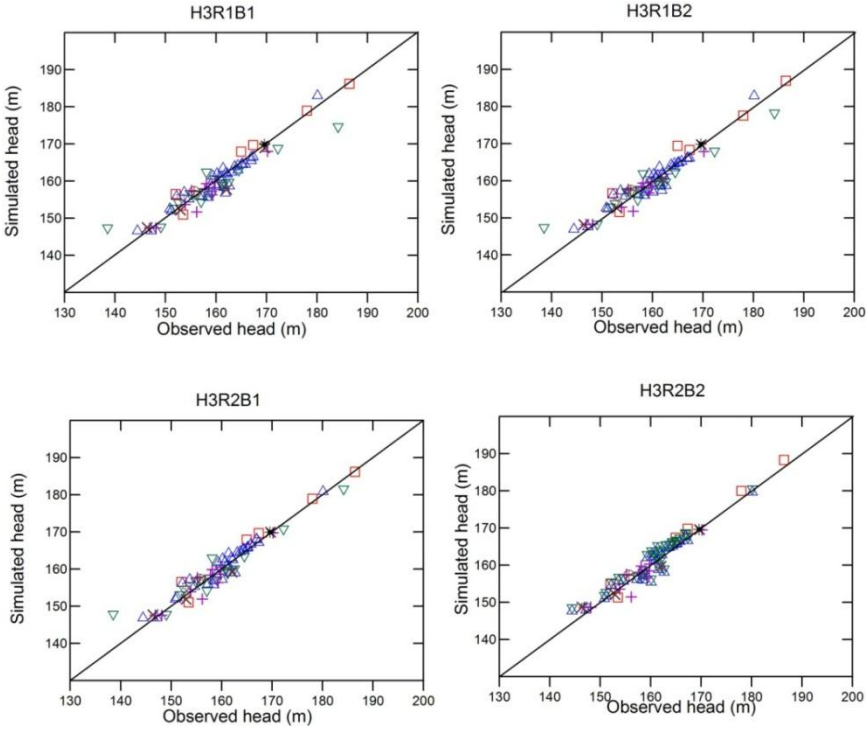
B-1 Observed vs. simulated heads for model H1R1B1, H1R1B2, H1R2B2, and H1R2B2



B-2 Observed vs. simulated heads for model H2R1B1, H2R1B2, H2R2B1, and H2R2B2



B-3 Observed vs. simulated heads for model H3R1B1, H3R1B2, H3R2B2, and H3R2B2



Appendix C

Optimized yields in each well for twelve alternative models (unit: m^3/d)

Name	Zone	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2
DGR2	1	245	58	87	168	310	45	110	155
F0635	1	336	36	561	600	323	581	219	20
F0636	1	367	87	561	400	65	523	387	180
GE118	1	200	361	155	310	323	323	371	71
PWDF	1	20	387	142	168	52	90	374	76
S0894	1	48	142	87	90	155	200	20	26
S0915	1	90	97	65	194	800	284	271	31
S1925	1	45	148	32	232	245	26	39	310
T04	1	310	323	90	181	103	619	187	206
T05	1	142	181	181	142	490	671	323	45
T06	1	361	103	94	123	774	490	26	181
T07	1	136	348	348	149	723	90	323	206
T09	1	129	871	600	400	20	26	297	284
T10	1	58	503	142	181	727	20	129	232
T11	1	20	65	129	45	20	52	71	65
T12	1	39	400	84	129	26	20	65	710
T13	1	335	258	348	136	465	20	168	194
T14	1	297	523	129	116	361	233	181	97
U833	1	168	968	361	187	323	675	245	500
B82	2	271	523	936	187	20	297	542	161
C30	2	416	26	26	284	77	39	123	483
DOH1	2	20	58	232	407	297	748	65	20
DOH2	2	474	168	348	387	271	284	52	161
DP251	2	174	52	52	39	361	136	77	77
DP321	2	129	61	103	200	26	194	103	129
F0234	2	155	20	387	374	322	58	141	581
MP1	2	387	542	367	71	26	90	20	329
P14	2	123	148	20	77	116	168	271	581
S0913	2	32	123	361	361	135	90	116	64
S1927	2	58	39	39	187	142	129	26	348
S1928	2	110	387	94	187	52	263	168	383
THK3	2	58	23	52	39	20	20	187	32
TP06	2	232	200	116	187	20	20	200	20
Y1984	2	65	903	232	374	639	39	323	270
Y1985	2	155	48	55	200	32	58	20	20
BR13	3	323	20	200	20	20	142	77	69
BR15	3	103	77	32	39	38	65	20	109
BR17	3	65	20	39	52	26	58	84	20
C28	3	484	87	194	264	116	20	310	232

Name	Zone	H3R1B1	H3R1B2	H3R2B1	H3R2B2
DGR2	1	335	20	68	97
F0635	1	58	323	123	194
F0636	1	503	323	310	400
GE118	1	284	561	32	97
PWDF	1	431	270	181	97
S0894	1	40	20	20	45
S0915	1	626	20	200	200
S1925	1	58	136	77	61
T04	1	194	297	94	45
T05	1	81	110	20	374
T06	1	168	81	180	161
T07	1	194	20	52	23
T09	1	23	475	336	81
T10	1	77	161	361	968
T11	1	161	45	245	374
T12	1	77	77	77	200
T13	1	32	187	148	116
T14	1	155	212	374	71
U833	1	348	20	348	600
B82	2	20	284	71	232
C30	2	445	271	23	74
DOH1	2	71	219	116	58
DOH2	2	110	271	65	68
DP251	2	71	161	55	20
DP321	2	42	77	65	90
F0234	2	374	387	310	194
MP1	2	100	323	97	55
P14	2	229	400	134	587
S0913	2	278	286	20	84
S1927	2	142	20	20	39
S1928	2	161	71	200	54
THK3	2	65	52	39	20
TP06	2	20	90	800	68
Y1984	2	76	400	380	361
Y1985	2	81	103	20	45
BR13	3	123	20	142	200
BR15	3	142	20	45	20
BR17	3	94	20	45	82
C28	3	58	320	84	20

Name	Zone	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2
DGR3	3	252	58	90	81	323	26	39	174
DP329	3	61	542	194	58	103	290	90	368
DP495	3	323	523	297	383	97	380	161	84
F0598	3	90	71	129	161	367	20	123	252
GF99	3	110	84	68	71	20	20	181	52
KTV	3	81	45	87	20	407	39	39	81
P25	3	542	80	207	168	65	83	26	245
PW95	3	97	20	174	174	77	20	374	142
PWA1	3	39	20	90	58	26	26	128	77
PWA2	3	65	20	29	32	103	58	20	371
RTB169	3	148	372	65	91	368	39	238	242
RTB177	3	20	20	465	20	110	90	26	402
S0472	3	48	26	387	20	65	20	26	20
S0702	3	258	187	32	20	310	78	20	65
S0895	3	374	77	97	154	336	323	26	174
S1960	3	445	400	36	84	155	258	294	39
TP01	3	142	26	90	155	258	52	39	177
TP08	3	26	128	161	39	323	116	168	32
Z91	3	58	20	36	138	206	77	310	242
BR01	4	52	116	161	52	174	122	26	32
BR02	4	26	168	20	271	20	45	77	135
BR03	4	90	245	400	28	20	185	116	63
BR04	4	97	97	97	110	20	20	39	20
BR05	4	181	61	136	142	70	20	136	39
BR06	4	90	336	61	258	65	20	194	142
BR07	4	77	232	20	20	26	26	148	374
BR08	4	354	155	51	174	161	20	103	307
BR09	4	52	29	20	181	20	20	26	115
BR11	4	103	20	65	168	122	20	39	128
BR12	4	96	86	187	167	58	50	71	128
BR14	4	20	71	110	20	20	20	45	168
BR16	4	26	39	87	161	45	20	122	77
BR18	4	20	136	97	20	103	20	20	123
BR19	4	65	77	83	174	252	38	20	245
BR21	4	39	180	20	32	26	38	20	97
BR22	4	200	20	52	20	70	51	26	26
BR23	4	65	174	20	103	20	20	45	52
BR24	4	20	65	119	96	103	38	20	103
BR25	4	116	71	200	20	20	20	20	32

Name	Zone	H3R1B1	H3R1B2	H3R2B1	H3R2B2
DGR3	3	74	20	55	96
DP329	3	115	323	61	110
DP495	3	45	374	61	321
F0598	3	20	297	593	336
GF99	3	20	77	20	20
KTV	3	215	97	71	20
P25	3	598	187	181	116
PW95	3	54	74	264	361
PWA1	3	46	52	26	71
PWA2	3	107	39	20	39
RTB169	3	20	123	284	65
RTB177	3	207	48	33	71
S0472	3	150	142	90	77
S0702	3	46	271	194	94
S0895	3	913	168	87	29
S1960	3	45	97	29	65
TP01	3	94	161	200	600
TP08	3	61	20	90	129
Z91	3	26	190	58	148
BR01	4	28	87	194	161
BR02	4	20	81	87	26
BR03	4	23	26	57	97
BR04	4	71	26	20	135
BR05	4	32	65	116	20
BR06	4	13	116	32	183
BR07	4	20	65	97	20
BR08	4	23	77	245	71
BR09	4	103	23	100	180
BR11	4	45	65	65	20
BR12	4	64	199	155	29
BR14	4	20	52	20	20
BR16	4	45	20	20	20
BR18	4	51	52	26	20
BR19	4	20	39	65	179
BR21	4	32	39	161	26
BR22	4	20	58	20	58
BR23	4	57	32	23	100
BR24	4	20	71	97	65
BR25	4	87	77	26	30

Name	Zone	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2
BR26	4	136	90	168	20	20	20	58	39
BR27	4	39	58	174	20	20	38	97	58
DP327	4	245	581	407	116	212	245	39	600
P23	4	97	245	100	310	129	84	141	77
P26	4	39	77	523	97	245	284	187	39
TP02	4	161	155	103	310	26	103	167	58
D1724	5	26	48	581	574	20	65	142	600
DGR1	5	55	103	97	20	161	65	58	561
F0233	5	297	58	361	419	310	161	52	187
F0943	5	36	39	23	20	142	65	129	84
F0999	5	81	65	129	194	20	156	207	103
P15	5	252	123	400	174	212	20	361	484
P16	5	129	110	387	374	323	155	142	109
P27	5	64	194	148	65	154	26	97	116
P34	5	20	155	20	168	20	116	135	70
P53	5	29	103	65	20	45	181	20	20
P71	5	39	284	387	251	361	315	45	25
RTB091	5	84	194	484	284	400	20	52	20
S0292	5	180	58	32	123	232	103	136	20
S1568	5	20	348	52	181	187	187	200	20
C59	6	20	103	581	200	20	20	20	271
DQ225	6	96	222	258	20	264	20	374	32
GF100	6	58	81	387	277	155	20	136	226
PW91	6	161	39	81	142	361	20	271	194
PW92	6	323	20	87	168	142	173	97	468
PWD1	6	355	39	20	310	187	97	148	81
S0309	6	20	20	61	387	129	161	116	97
S0499	6	348	20	297	136	129	150	129	371
S0875	6	90	61	174	284	310	200	310	387
S0969	6	232	194	45	200	187	116	187	452
S1618	6	245	155	65	103	400	91	142	129
S1947	6	1000	39	387	374	206	20	336	387
S1949	6	26	90	97	361	20	134	45	348
SC	6	323	77	77	20	180	26	194	748
TX9	6	23	77	77	103	219	20	84	355
GF98	7	123	194	116	148	194	52	323	310
KP	7	71	77	600	561	52	142	71	83
P17	7	581	20	561	336	142	387	136	361

Name	Zone	H3R1B1	H3R1B2	H3R2B1	H3R2B2
BR26	4	174	26	36	20
BR27	4	74	32	77	366
DP327	4	232	70	87	42
P23	4	177	106	168	39
P26	4	20	20	84	174
TP02	4	20	20	48	32
D1724	5	76	161	20	87
DGR1	5	271	83	590	52
F0233	5	129	20	168	81
F0943	5	90	39	168	387
F0999	5	20	20	194	187
P15	5	99	58	129	160
P16	5	94	561	181	65
P27	5	89	90	181	26
P34	5	152	20	103	20
P53	5	20	45	115	77
P71	5	75	74	39	20
RTB091	5	20	39	20	94
S0292	5	464	97	187	65
S1568	5	166	77	219	20
C59	6	20	81	29	74
DQ225	6	20	26	20	57
GF100	6	97	84	51	20
PW91	6	20	116	20	45
PW92	6	105	26	57	71
PWD1	6	106	58	20	23
S0309	6	864	84	93	29
S0499	6	331	155	71	581
S0875	6	88	400	71	187
S0969	6	360	97	39	161
S1618	6	78	87	200	116
S1947	6	190	181	77	65
S1949	6	20	23	58	20
SC	6	20	26	84	90
TX9	6	20	36	74	87
GF98	7	87	39	90	42
KP	7	1548	374	20	61
P17	7	216	161	241	903

Name	Zone	H1R1B1	H1R1B2	H1R2B1	H1R2B2	H2R1B1	H2R1B2	H2R2B1	H2R2B2
PWLR	7	155	52	61	387	90	39	97	194
S0649	7	800	219	387	155	245	323	103	210
S0971	7	348	374	103	155	123	348	100	187
S1377	7	181	297	97	26	20	174	142	480
S1378	7	142	387	181	42	154	310	174	20
S1649	7	48	367	87	348	161	284	52	129
S1932	7	148	20	55	32	45	90	48	219
TO2	7	336	384	161	465	232	20	387	355
TO3	7	426	839	348	426	97	225	20	1000
ARD2	8	271	336	310	20	516	238	387	148
BHP	8	219	194	168	334	232	20	258	371
F0162	8	36	181	345	20	200	90	136	155
F0279	8	74	23	20	110	77	109	129	309
GF97	8	45	20	97	180	206	20	123	20
S0297	8	129	187	161	103	52	52	71	403
S0311	8	523	77	35	484	52	251	155	123
S0916	8	426	100	600	310	271	20	194	77
S0938	8	77	613	561	581	310	523	523	452
S1159	8	387	181	400	400	336	194	123	403
S1367	8	542	20	400	77	297	26	20	168
S1599	8	173	387	84	110	400	323	71	181
S1600	8	20	374	129	348	65	194	161	339
S1701	8	52	20	55	52	361	45	232	129
S1931	8	245	348	465	561	129	39	161	348
CTC	9	90	22	97	155	387	168	375	374
D1723	9	84	20	87	84	58	52	136	936
DQ224	9	110	348	400	116	136	45	71	258
F0941	9	219	77	58	271	136	20	110	103
F0942	9	207	200	400	374	207	135	361	110
F0997	9	374	374	42	542	52	71	232	258
F0998	9	168	323	181	562	336	232	374	123
F1303	9	58	123	129	97	142	271	251	84
S0248	9	87	161	32	194	32	20	129	330
S1926	9	103	58	20	65	39	52	20	168
TP10	9	232	200	180	200	323	20	200	468
Sum		28136	27771	29134	29296	28257	20713	22926	32880

Name	Zone	H3R1B1	H3R1B2	H3R2B1	H3R2B2
PWLR	7	354	387	65	39
S0649	7	746	387	581	297
S0971	7	21	58	58	400
S1377	7	61	87	42	84
S1378	7	70	374	68	187
S1649	7	78	57	20	94
S1932	7	20	20	58	23
TO2	7	174	200	541	32
TO3	7	290	445	87	45
ARD2	8	103	245	361	168
BHP	8	503	284	297	148
F0162	8	36	174	103	123
F0279	8	284	142	77	400
GF97	8	68	503	161	194
S0297	8	62	36	61	81
S0311	8	216	39	39	42
S0916	8	45	52	110	387
S0938	8	405	155	387	400
S1159	8	737	181	142	103
S1367	8	79	77	26	194
S1599	8	677	52	254	84
S1600	8	158	116	20	284
S1701	8	100	97	74	23
S1931	8	168	368	426	361
CTC	9	97	348	97	29
D1723	9	77	110	71	68
DQ224	9	65	387	94	123
F0941	9	71	136	52	94
F0942	9	407	400	25	94
F0997	9	39	116	20	90
F0998	9	71	600	374	1000
F1303	9	97	110	65	71
S0248	9	42	77	97	29
S1926	9	90	74	103	77
TP10	9	84	200	90	97
Sum		23859	22912	19897	22579

Appendix D
Model weights of 12 realizations based on GLUE weight (with a shape factor of N=1)

Model	1	2	3	4	5	6	7	8	9	10	11	12
H1R1B1	0.075	0.075	0.076	0.077	0.084	0.086	0.087	0.087	0.088	0.088	0.088	0.090
H1R1B2	0.077	0.078	0.079	0.079	0.081	0.082	0.082	0.086	0.088	0.089	0.089	0.089
H1R2B1	0.077	0.079	0.080	0.080	0.081	0.082	0.086	0.086	0.087	0.088	0.088	0.088
H1R2B2	0.087	0.086	0.079	0.085	0.082	0.085	0.087	0.078	0.077	0.083	0.085	0.087
H2R1B1	0.055	0.085	0.089	0.088	0.085	0.087	0.084	0.086	0.078	0.085	0.089	0.089
H2R1B2	0.082	0.077	0.083	0.084	0.085	0.081	0.083	0.086	0.086	0.085	0.080	0.088
H2R2B1	0.084	0.086	0.080	0.087	0.082	0.077	0.085	0.086	0.082	0.083	0.084	0.085
H2R2B2	0.081	0.081	0.082	0.082	0.082	0.082	0.082	0.083	0.085	0.086	0.086	0.087
H3R1B1	0.081	0.085	0.088	0.087	0.084	0.087	0.086	0.085	0.068	0.076	0.085	0.088
H3R1B2	0.084	0.084	0.080	0.085	0.085	0.083	0.081	0.086	0.086	0.078	0.082	0.086
H3R2B1	0.077	0.088	0.087	0.089	0.086	0.075	0.074	0.088	0.089	0.082	0.078	0.087
H3R2B2	0.084	0.086	0.087	0.089	0.078	0.076	0.079	0.088	0.081	0.085	0.080	0.087

Appendix E
Statistics for sustainable yield estimated obtained from 12 realization models

Zone	H1R1B1		H1R1B2		H1R2B1		H1R2B1	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
1	3428	503	6426	504	4258	423	5617	484
2	3227	565	3715	313	3721	416	3088	926
3	4100	834	3808	468	3447	377	3773	395
4	3246	679	4552	587	3684	700	2864	729
5	1820	185	2313	270	3281	389	3996	718
6	3037	400	2194	284	3077	334	4155	754
7	2864	253	3283	375	3814	624	4498	575
8	3456	501	3508	360	3867	498	4793	572
9	1746	371	2039	188	2698	378	3557	511
Total	26923	2875	31838	2210	31847	1838	36342	3300

Zone	H2R1B1		H2R1B2		H2R2B1		H2R2B1	
	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
1	5368	729	3728	319	4679	933	4773	807
2	3028	892	1919	270	3543	396	3246	467
3	3700	620	2192	555	3043	428	3937	886
4	2681	958	1777	291	2686	589	2904	403
5	1339	201	1238	309	1949	383	2780	527
6	2252	593	1517	330	3080	497	3928	1031
7	1869	397	2075	338	2758	351	3701	740
8	3616	826	2514	455	3292	556	3398	741
9	1474	440	1199	206	3106	650	3403	418
Total	25327	2646	18160	825	28136	3067	32069	3995

	H3R1B1		H3R1B2		H3R2B1		H3R2B1	
Zone	Mean	StDev	Mean	StDev	Mean	StDev	Mean	StDev
1	3562	420	5233	562	4335	307	3888	434
2	3593	1013	4585	878	3169	351	2940	355
3	3016	631	3431	654	3443	610	3686	819
4	2247	465	2860	653	2410	282	2776	514
5	2290	631	2397	463	2155	377	2099	453
6	1881	425	2204	679	1404	397	1770	419
7	3009	672	2900	568	2066	537	2011	331
8	2698	503	4515	306	2894	495	2745	349
9	856	289	2599	606	1356	308	1943	479
Total	23151	3112	30724	3209	23231	2082	23858	3007

Appendix F

Predictive intervals (90%) and mean predicted values based on cumulative predictive distribution obtained from GLUE methodology

	H1R1B1			H1R1B2			H1R2B1			H1R2B2		
Area	P5	P50	P95	P5	P50	P95	P5	P50	P95	P5	P50	P95
1	2603	3468	4002	5573	6426	7138	3570	4258	4846	4896	5617	6261
2	2232	3240	3983	3188	3715	4066	3132	3721	4195	2109	3088	4789
3	2800	4187	5068	3246	3808	4660	2891	3447	4057	3335	3773	4402
4	2058	3284	3963	3782	4552	5463	2784	3684	4807	1997	2864	4217
5	1428	1775	2074	1882	2313	2699	2796	3281	3880	3098	3996	5173
6	2335	3000	3486	1862	2194	2719	2639	3077	3632	2941	4155	5057
7	2390	2871	3141	2680	3283	3718	3331	3814	4899	3832	4498	5398
8	1343	2053	2530	1674	2016	2361	1724	2253	2543	2547	2808	3250
9	892	1315	1641	1360	1578	1921	1630	1815	2017	1623	2309	3035
Total	20026	25192	28082	27157	29885	32860	26973	29350	32453	30218	33109	39261

	H2R1B1			H2R1B2			H2R2B1			H2R2B2		
Area	P5	P50	P95	P5	P50	P95	P5	P50	P95	P5	P50	P95
1	3948	5119	6162	3335	3728	4245	2910	4275	5317	3600	4550	5425
2	1936	3217	4804	1616	1919	2352	2962	3617	4267	2739	3351	3974
3	2622	3516	4156	1594	2192	3256	2047	2934	3674	3164	4133	4963
4	2007	2807	4546	1484	1777	2317	2064	2505	2916	2413	2987	3736
5	1656	2219	2939	721	1238	1620	1241	1749	2281	2171	2694	3305
6	1418	2264	3144	1139	1517	2057	2545	3258	3896	3433	4172	4959
7	1228	1933	2633	1619	2075	2633	2043	2613	3183	2650	3665	4313
8	772	1477	2102	1006	1244	1612	875	1461	1834	1271	1666	2150
9	720	1150	1678	678	896	1058	1616	2209	2660	2551	3098	3406
Total	19848	23701	26835	15130	16586	18149	21163	24620	27643	27246	30316	32653

	H3R1B1			H3R1B2			H3R2B1			H3R2B2		
Area	P5	P50	P95	P5	P50	P95	P5	P50	P95	P5	P50	P95
1	3128	3637	4180	4460	5289	6090	3895	4318	4856	2747	3397	4135
2	2735	3666	5160	3241	4688	5506	2659	3178	3708	2685	3405	4065
3	2199	3059	3844	2641	3545	4201	2345	3423	4187	2521	3479	4575
4	1565	2287	2891	1926	2936	3877	1956	2424	2824	2436	3194	3797
5	1616	2330	3355	1832	2410	3119	1663	2121	2750	1296	2061	3011
6	1436	1908	2476	1629	2250	3318	842	1471	2226	1226	1777	2292
7	1822	3077	3784	2285	2941	3797	1365	2041	2820	1341	1915	2464
8	1065	1342	1734	1901	2488	2816	713	1531	1997	829	1188	1714
9	653	906	1291	1252	1812	2417	548	933	1297	788	1016	1272
Total	19492	22213	25856	23836	28359	31716	18379	21440	23881	18725	21432	25490