EXPERT WITNESS STATEMENT

prepared by:
Paula Cutillo, Ph.D.
Hydrogeologist

United States Department of the Interior
National Park Service
Water Resources Division
Fort Collins, Colorado

August 4, 2006

prepared for:

An Administrative Hearing before the State Engineer
State of Nevada
Department of Conservation and Natural Resources
Division of Water Resources
Carson City, Nevada
September 11-29, 2006
In the Matter of Water-Rights Applications
No. 53987-53992 and 54003-54030
Filed by the Southern Nevada Water Authority
In Spring Valley Hydrographic Area (#184)

Water-right applications No. 53987-53992 and 54003-54030 were filed by the Southern Nevada Water Authority (SNWA) for ground water in Spring Valley hydrographic area. Documents submitted by SNWA in support of these applications include Exhibit 507 titled "FEMFLOW3D Version 2.0, A Finite-Element Program for the Simulation of Three-Dimensional Groundwater Systems" and Exhibit 508 titled "Development and Use of a Groundwater Model for the Spring Valley Area."

Exhibit 508 Appendix A is a transient finite-element ground-water flow model of the Spring Valley area developed for the Southern Nevada Water Authority (SNWA) by Mr. Timothy Durbin using the computer program FEMFLOW3D, Version 2.0. FEMFLOW3D Version 2.0 is an updated version of FEMFLOW3D Version 1.0 (Durbin and Bond, 1998), but is not a public domain finite-element code developed by the U.S. Geological Survey (e.g., MODFE (Torak, 1993)).

Exhibit 508 (Page 8-1) states "The purpose of the groundwater model of the Spring Valley area is to provide a decision-making framework for managing and monitoring groundwater development in Spring Valley." The steady-state model calibration indicates good agreement with pre-development (1960's and earlier) ground-water level observations (Exhibit 508, Figure 8-3). The ground-water model simulates historic and proposed agricultural and community consumptive use within the model domain from 1955 to 2090. The model's transient simulation, however, does not include the proposed ground-water withdrawals that are the subject of the Spring Valley administrative hearing. A review of Exhibits 507 and 508 indicates that the ground-water flow model developed by Mr. Durbin can be run at this time with the proposed ground-water withdrawals to simulate stress upon the ground-water system and to predict changes in hydraulic head. The attendant results, although preliminary, provide valuable information for decision makers at this time.

The purpose of this report is to present the results of new transient simulations produced by including SNWA's proposed ground-water withdrawals in the FEMFLOW3D Spring Valley ground-water flow model. To simulate the effects of the proposed withdrawals, the sv_model_ts.FLX input file (Exhibit 508 Appendix A) was modified to include the 19 proposed basin-fill and carbonate wells (Figure 1). The same general method used by Mr. Durbin to simulate ground-water pumping was followed to add the new wells: wells were represented within the model domain as specified fluxes; three nodes nearest to each proposed point of diversion were identified in the input file; and the pumping rate for each set of specified flux nodes was assigned in a corresponding table in the input file. Pumping was divided evenly among the three nodes representing each well. Basin-fill wells were assigned in the topmost 'Upper Valley Fill' layer in the model, and carbonate wells were assigned in the first 'Lower Carbonate Rocks' layer in the model. Ground-water withdrawals at each of the proposed points of diversion were started in year 2015 of each simulation. The new wells were pumped continuously at the maximum proposed rate (i.e., 6 cfs for basin-fill wells and 10 cfs for carbonate wells) until the end of the 135-year simulation. The only input file modified was the .FLX file. The model converged for all simulations without modification of time step size or convergence criteria.

The new transient simulations calculate change in head due to historical and proposed ground-water withdrawals over the period 1955 to 2090. Figure 2 summarizes drawdown observed under several pumping scenarios at six preexisting hydrograph sites within the model domain. The location of the hydrograph sites are shown in Figure 1. The hydrograph at a particular site is the weighted average head for the nodes associated with the site (Exhibit 507, Page 3-15), and are assigned for this model in the sv_model_ss.HED input file (Exhibit 508 Appendix A).

Model results indicate that SNWA's proposed ground-water withdrawals produce drawdown of up to 200 ft in the area of the pumping wells after 75 years of pumping. The cone of depression produced by the proposed pumping grows beyond the hydrographic boundary of Spring Valley despite the presence of fault-bounded structural compartments within the model domain. Model results also show that elimination of the carbonate wells results in 30% to 50% less drawdown at the Lehman Creek, Spring/Hamlin Valley Divide and Shoshone Ponds hydrograph sites (Figure 2).

The 2-D fault mesh incorporated within the numerical model compartmentalizes flow. This structure produces high hydraulic gradients near several compartment boundaries and significant head differences across these compartment boundaries under steady-state conditions (Exhibit 508, Figure 8-10). When the system is stressed due to the proposed ground-water withdrawals, significant offsets in drawdown occur at these same locations. In contrast, hydraulic head and drawdown contours are more or less continuous within the domains of the finite-difference MODFLOW models described in Exhibits 2001 and 3001.

Regardless of the different conceptual models upon which each finite-element and finite-difference model is based, the results of the new FEMFLOW3D transient simulations indicate that drawdown due to pumping calculated by all three numerical models appears to be of the same order of magnitude (Figure 2) (Exhibit 2001; Exhibit 3001). As noted by SNWA in Exhibit 502, the lack of available data on aquifer properties creates prediction uncertainty in regard to the results of numerical models. It should be further emphasized that the fact that a model has been calibrated does not mean that the model is "correct" (e.g., Bredehoeft, 2003). Further analysis is needed to determine which model is best supported by observational data.

The results of the FEMFLOW3D Spring Valley ground-water model, as well as the results of the MODFLOW models submitted for the purpose of this hearing, provide first-order approximations of drawdown that may occur due to the proposed ground-water withdrawals. While all of the numerical models would benefit from continued development and calibration as new data are acquired, their results provide valuable information that should be considered by the interested parties and decision-makers. In time, one of the models may provide a more acceptable fit to observed data, but until that data is acquired and the necessary analyses performed to make such a determination, all of the calibrated models can be considered plausible approximations of the real ground-system (e.g., Poeter & Anderson, 2005). In using this approach, alternative conceptual models can be simultaneously evaluated to gain insight into potential changes in aquifer storage, discharge and/or recharge due to pumping.

REFERENCES CITED

Bredehoeft, J.D., 2003. From models to performance assessment: the conceptualization problem: *Ground Water* 41(5), 571-577.

Durbin, T. J. and L. D. Bond, 1998. FEMFLOW3D: A finite-element program for the simulation of three-dimensional aquifers, Version 1.0: *U.S. Geological Survey Open File Report 97-810.*

Poeter, E. and D. Anderson, 2005. Multimodel ranking and inference in ground water modeling: *Ground Water* 43(4), 597-605.

Torak, L.J., 1993. A MODular Finite-Element model (MODFE) for areal and axisymmetric ground-water-flow problems, part 1--model description and user's manual: *U.S. Geological Survey Techniques of Water-Resources Investigations*, book 6, chap. A3.

Paula A. Cutillo, Ph.D.

8/3/06

Date

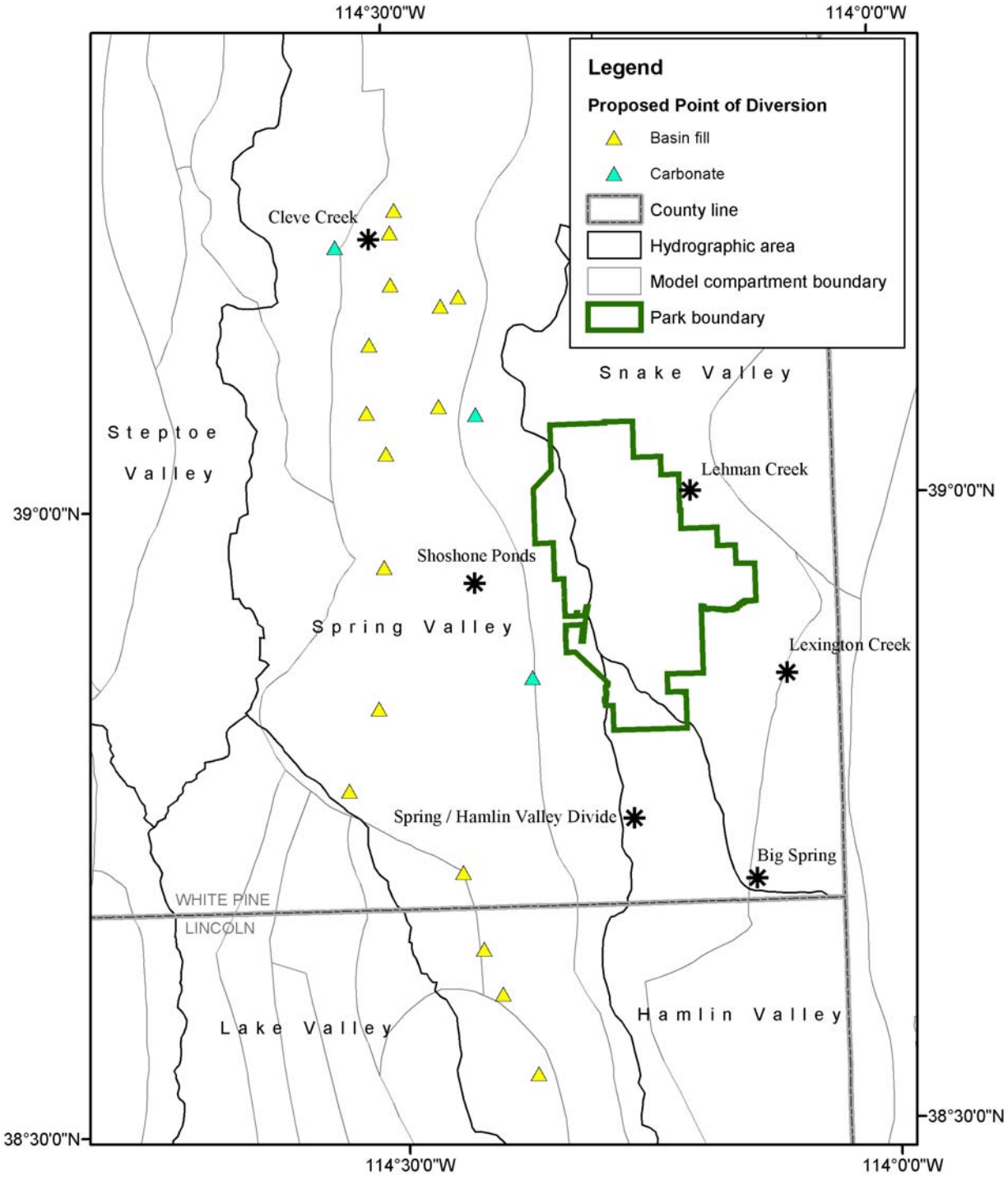

Figure 1. Map showing the location of select model hydrograph sites (asterisks) which represent an area within the model domain over which the change in head due to pumping was observed for each simulation. SNWA's proposed points of diversion, and Great Basin National Park, hydrographic area and model compartment boundaries are also shown for reference.

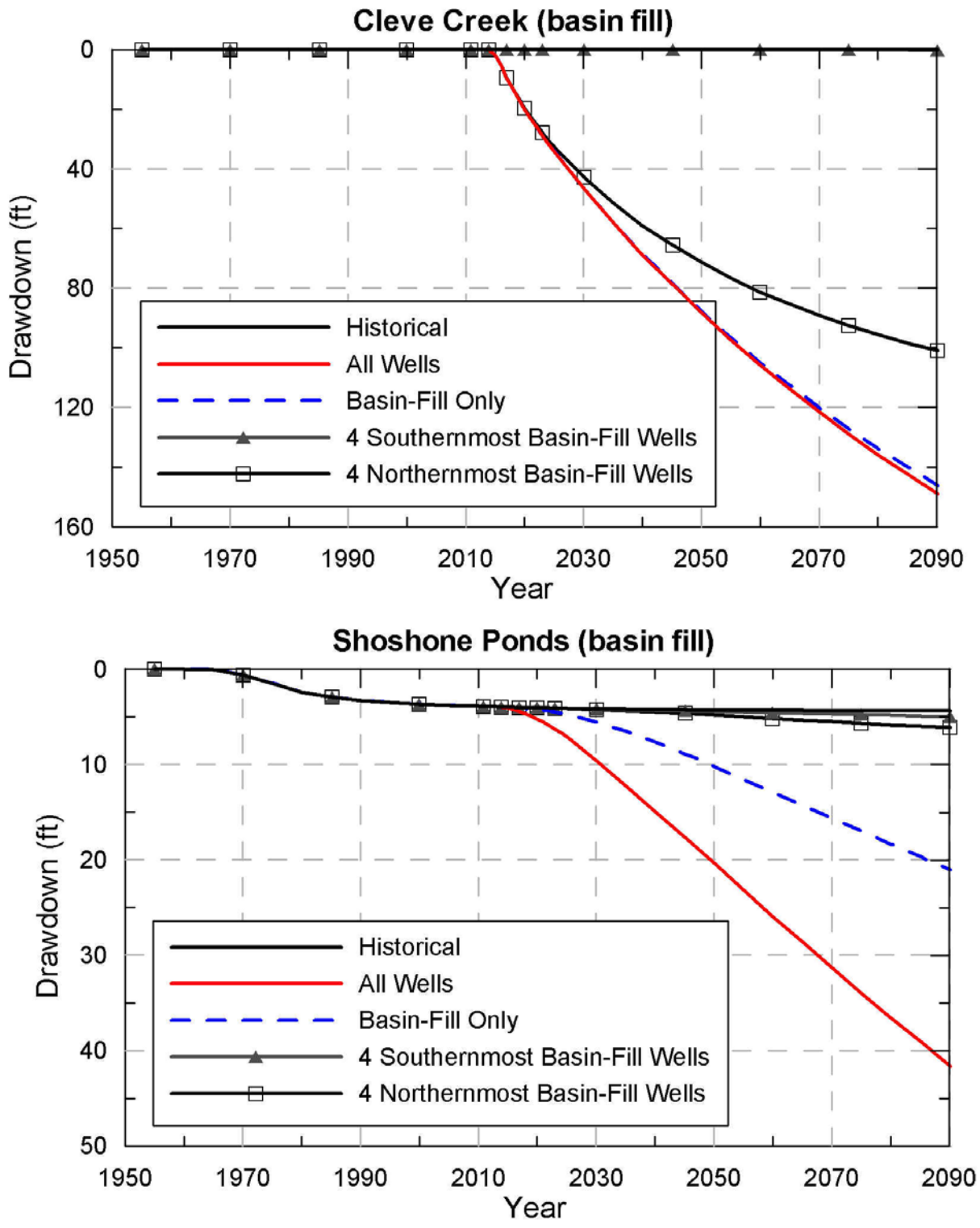

Figure 2a. Drawdown over time at hydrograph sites shown in Figure 1 for five pumping scenarios. Note that the scale of the vertical axis varies for each chart. The Historical simulation is the original simulation run by Mr. Durbin and includes historic and existing consumptive use, but does not include SNWA's proposed ground-water withdrawals. The remaining simulations include from 4 to 19 of SNWA's proposed points of diversion.

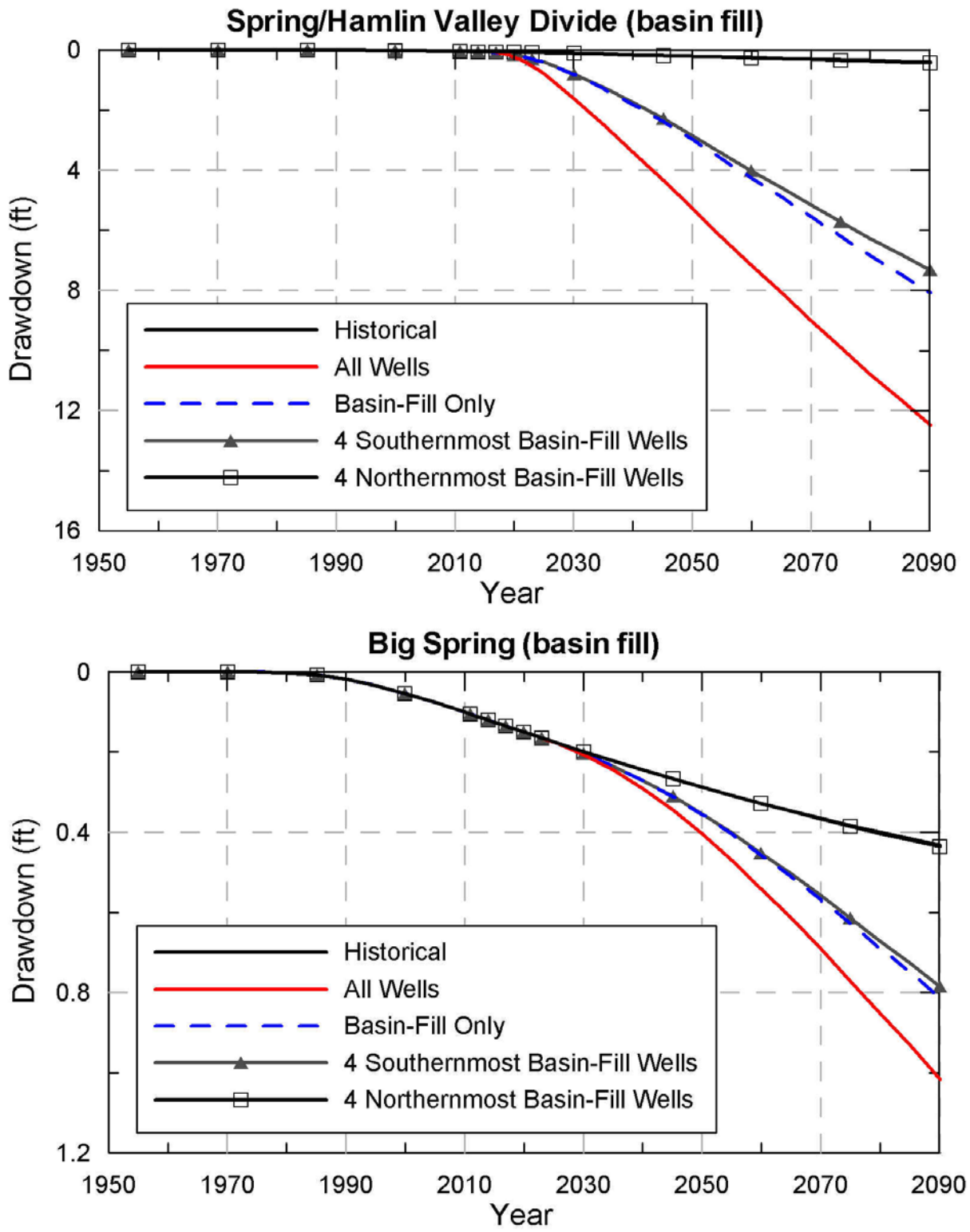

Figure 2b. Drawdown over time at hydrograph sites shown in Figure 1 for five pumping scenarios. Note that the scale of the vertical axis varies for each chart. The Historical simulation is the original simulation run by Mr. Durbin and includes historic and existing consumptive use, but does not include SNWA's proposed ground-water withdrawals. The remaining simulations include from 4 to 19 of SNWA's proposed points of diversion.

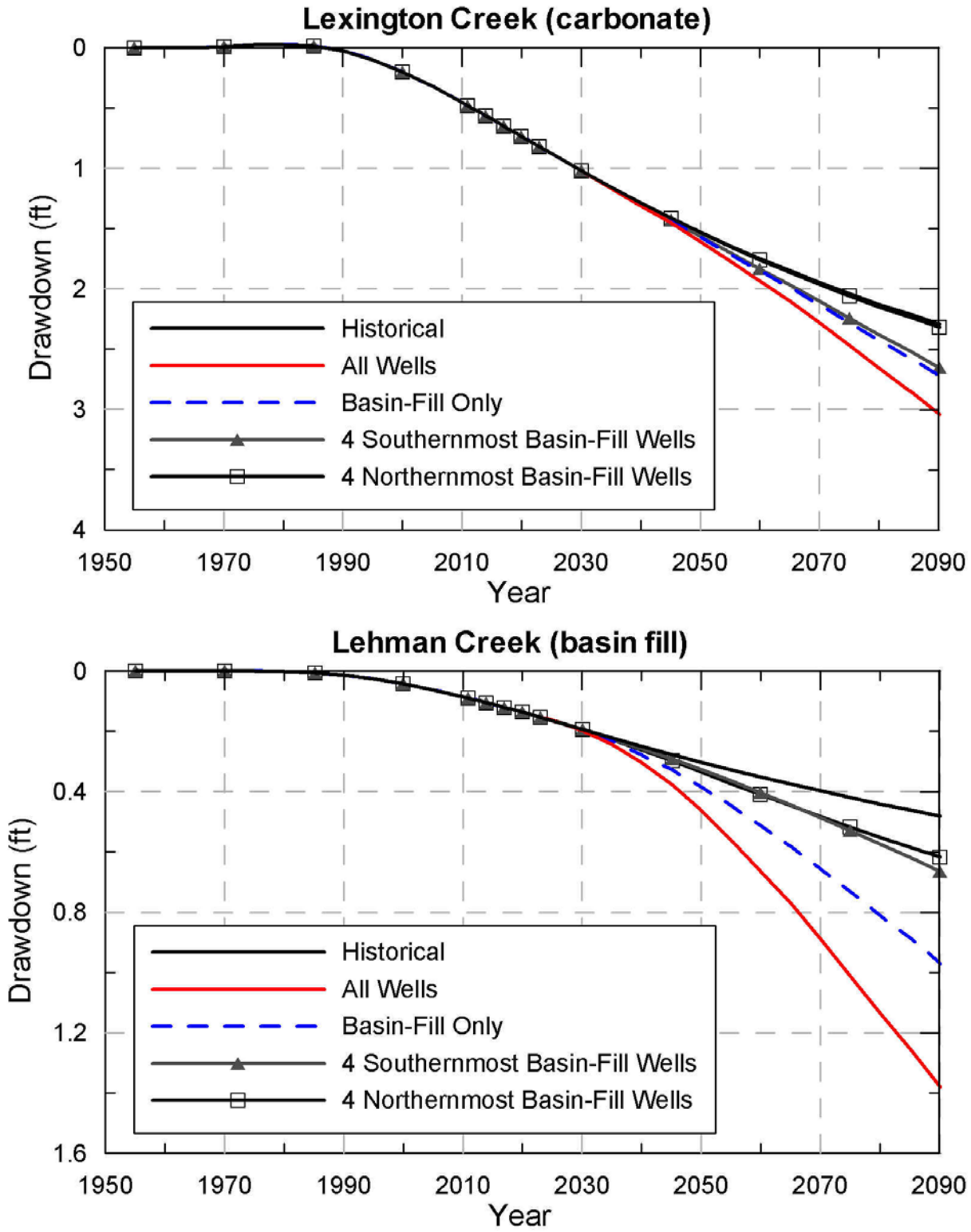

Figure 2c. Drawdown over time at hydrograph sites shown in Figure 1 for five pumping scenarios. Note that the scale of the vertical axis varies for each chart. The Historical simulation is the original simulation run by Mr. Durbin and includes historic and existing consumptive use, but does not include SNWA's proposed ground-water withdrawals. The remaining simulations include from 4 to 19 of SNWA's proposed points of diversion.

ground water

Issue Paper/

From Models to Performance Assessment: The Conceptualization Problem

by John D. Bredehoeft[1]

Abstract

Today, models are ubiquitous tools for ground water analyses. The intent of this paper is to explore philosophically the role of the conceptual model in analysis. Selection of the appropriate conceptual model is an a priori decision by the analyst. Calibration is an integral part of the modeling process. Unfortunately a wrong or incomplete conceptual model can often be adequately calibrated; good calibration of a model does not ensure a correct conceptual model. Petroleum engineers have another term for calibration; they refer to it as *history matching*. A caveat to the idea of history matching is that we can make a prediction with some confidence equal to the period of the history match. In other words, if we have matched a 10-year history, we can predict for 10 years with reasonable confidence; beyond 10 years the confidence in the prediction diminishes rapidly. The same rule of thumb applies to ground water model analyses. Nuclear waste disposal poses a difficult problem because the time horizon, 1000 years or longer, is well beyond the possibility of the history match (or period of calibration) in the traditional analysis. Nonetheless, numerical models appear to be the tool of choice for analyzing the safety of waste facilities. Models have a well-recognized inherent uncertainty. Performance assessment, the technique for assessing the safety of nuclear waste facilities, involves an ensemble of cascading models. Performance assessment with its ensemble of models multiplies the inherent uncertainty of the single model. The closer we can approach the idea of a long history with which to match the models, even models of nuclear waste facilities, the more confidence we will have in the analysis (and the models, including performance assessment). This thesis argues for prolonged periods of observation (perhaps as long as 300 to 1000 years) before a nuclear waste facility is finally closed.

Introduction—Models

Models play a key role in the analysis of many, if not most, ground water problems. They are especially important in predicting the behavior of nuclear waste facilities far into the future. The Waste Isolation Pilot Plant (WIPP, a geologic repository for transuranic wastes in New Mexico) was recently opened and is receiving nuclear weapons waste. Yucca Mountain (the proposed high-level nuclear waste repository in Nevada) is near the licensing stage. Hydrogeological models play a key role in assessing the safety of these facilities.

The purpose of this paper is to discuss philosophically the use of models in making predictions. Many of these ideas have been expressed elsewhere, yet they seem worth restating. In particular, I want to examine the role that the conceptual model plays in analysis. I take a historical perspective in developing these ideas.

In the 19th century, various laws that describe the movement of heat, electricity, and ground water through a continuum were derived. Of special concern to those of us that investigate ground water is Darcy's law. By applying the principle of conservation of mass and incorporating Darcy's law as a constitutive relationship, we can derive a partial differential equation that describes the hydraulic head throughout a porous medium. Once the head is determined, we can apply Darcy's law to derive the ground water flow vectors throughout the system. These principles form the basis for all ground water flow and transport models.

For many ground water problems with simple geometry, simple parameter distributions, and simple boundary conditions analytical solutions to the mathematical problem can be derived. Because there is an analogy between the flow of ground water and the flow of both electricity and heat, we often can find the mathematical solution for the appropriate boundary value problem in the literature on

[1]The Hydrodynamics Group, 127 Toyon Lane, Sausalito, CA 94965; jdbrede@aol.com

heat flow and/or electrical flow. The analogy between heat flow and ground water flow was enriched in the 1930s and 1940s when Theis (1935) suggested that transient ground water flow was analogous to unsteady heat flow, and Jacob (1940) derived the transient ground water flow equation from first principles. Ground water in the 1940s and 1950s went through a period when various boundary value problems were solved for pumping wells; numerous pumping test procedures were developed. Many of the pumping test solutions could be found in the classical literature on heat flow. Carslaw and Jaeger (1959) was on the shelf of most serious ground water hydrologists. Many of the classical solutions involved numerical integration of mathematical functions; the digital computer enhanced the capability to numerically integrate these functions.

The Conceptualization Problem

The various pumping test solutions involve different conceptual models of the well and its geologic environment. For example, the response of a well pumping at a constant rate from an extensive confined aquifer can be analyzed as if (1) the overlying and underlying beds are impermeable (the Theis solution), (2) the overlying and underlying beds are leaky without storage (the leaky aquifer solution), or (3) the overlying and underlying beds are leaky with storage (the modified leaky aquifer solution, Hantush 1964). A pumping test in the Dakota Sandstone at Wall, South Dakota, illustrates the point (Bredehoeft et al. 1983). Different investigators fit the Wall data to both the Theis solution and the modified leaky aquifer solution; the data fit either solution equally well. We obtain a different answer depending on the conceptual model chosen; the predictions of long-term future well performance will be different depending on which model is selected.

Usually the conceptual model chosen for analysis is an a priori decision of the analyst. Sometimes the fit of the data to the analytical solution will suggest that the conceptual model is inappropriate, but more often than not the data will fit more than one conceptual model equally well, as was the case at Wall, South Dakota. My point is that we can choose the wrong conceptual model, fit the data, and get a wrong answer. In the 1940s and 1950s, hydrogeologists did not call these solutions to pumping test model analyses, but we might today.

We cannot overemphasize the role of the choice of the conceptual model in any analysis of a ground water system. A wrong conceptual model invariably leads to poor predictions, no matter how well the model is fit to the data. Time and again, the errors in prediction revolve around a poor choice of the conceptual model (Konikow and Bredehoeft 1992; Oreskes and Belitz 2001). Modeling invariably involves simplifying the real system into a conceptual model that can be analyzed; that conceptual model must capture the essence of the problem. Choosing the appropriate conceptual model is usually a matter of professional judgment. It is how we conceptualize the problem that separates good analysts from poor ones, especially today when anyone can run codes such as MODFLOW.

We tend to regard our conceptual model as immutable. Yet I remember one of my mentors as a young geologist, N.W. Bass, said to me: "A geologic report is always a progress report." I continue to reflect on this remark. What we choose as a conceptual model is a function of the status of knowledge in science. For example, plate tectonics changed geology and changed our conceptual model of tectonics. Theis (1935) and Jacob (1940) changed ground water hydrology by introducing the transient theory of ground water flow. Finding chlorine–36 at depth in Yucca Mountain has caused the community to rethink transport through a fractured unsaturated zone. As yet, there is no consensus on the appropriate conceptual model for transport within Yucca Mountain (National Research Council 2001). The point is that our conceptual model changes with advances in the science; concepts are by no means static.

Models

At the time well tests were the standard tool for analysis for hydrogeologists, it was apparent to some individuals that it would be of great value to find a procedure to solve the more global problem of flow through a porous medium with varying parameter distributions and complex boundary conditions. In other words, to treat flow in an entire aquifer. A group at the U.S. Geological Survey, led by Herb Skibitski, developed the resistor-capacitor electrical network as an *analog model* for ground water flow. This introduced into ground water the idea of a *model* of an aquifer system.

A parallel effort was under way in the petroleum industry where reservoir engineers simulated flow in realistic hydrocarbon reservoirs. A petroleum reservoir is often more complex than saturated ground water systems because of the presence of multiphases—oil, gas, and water. The petroleum engineering effort to model a reservoir used the digital computer. Some of the best applied mathematicians of the 1950s and 1960s worked on developing numerical methods to solve the equations of flow in porous media. The petroleum industry referred to both the computer codes and the models of specific fields as *reservoir simulators.* The term used to describe the analyses was *reservoir simulation* rather than modeling.

As digital computers grew in power, the analog methods of the 1950s and 1960s used in ground water were replaced by digital computer methods in the 1970s. As digital computers became more powerful and less costly, modeling became widely used. With the power of today's PCs, models of ground water systems are now commonplace.

The digital computer codes had the added benefit that solute transport also could be modeled. In a general way, sets of partial differential equations could be solved simultaneously. The analog models dealt only with the solution of the ground water flow equation. The digital computer added new dimensions to modeling. A whole industry has grown up that produces models of ground water flow and transport that are easily implemented. There are a number of pre- and post-processors for MODFLOW and MT3D, the most common of the ground water flow and transport codes. The pre- and post-processors make modeling relatively easy, and enable very large problems, involving big grids, to be modeled. Without the pre- and post-processors, large grids are exceedingly difficult to implement—for all practical purposes, they become intractable.

Calibration

An integral part of the modeling procedure is calibration. Calibration involves fitting the model output to a set of observations. Hopefully, at some point in the process, the model parameters are adjusted so that an *adequate* fit to the observations is achieved. Originally, calibration was a trial-and-error procedure. In recent years, the process of adjusting the parameters to achieve an adequate fit has been automated.

Numerical measures of the goodness-of-fit between the observations and the model predictions have been devised. The numerical measures provide the appearance that judging the adequacy of fit during calibration is no longer subjective. However, in the end, what constitutes an adequate fit is a subjective decision. Statistical measures of the goodness-of-fit can be calculated, but the question of whether a model is calibrated is a decision left to the analyst.

There are problems in the calibration procedure. As suggested by the previous discussion of pumping test analysis, the calibration commonly does not test our conceptual model. *In other words, a model involving a wrong or incomplete conceptual model can be adequately calibrated.* It is generally conceded that a model, even if it is well calibrated, is nonunique; another parameter set might result in an equally good calibration (Bethke 1992).

Post Audits

Since models have now been around for several decades, it is possible in certain limited instances to evaluate their performance. Predictions were made that can now be compared to what happened to a particular system. Many audits do not really test the adequacy of the model because what took place with the real system was not a scenario that was analyzed initially. Typically, pumping followed a different pattern than anticipated.

There are a limited number of post audits of model predictions; they are not reassuring. Many models did not provide good predictions (Anderson and Woessner 1992; Konikow and Bredehoeft 1992). Many models suffered from a conceptual omission: an important process was overlooked. In other cases, the range of parameters was much larger than was included in the model analysis. Models are known to have provided poor predictions, even models that were thought to have been well calibrated.

Validation

Validation is a term promoted by the nuclear waste community. Different people variously define validation; there is no consensus on what it means. Furthermore, in most cases, the goal of calibration and validation are the same: In both cases, we seek to create the best possible representation of the system. We as a community have formulated restrictive, and rather special, definitions of what it means to validate a code.

Recognizing that the traditional history match was impossible, the nuclear waste community set out to test different codes in situations where shorter histories of performance were available. They called this test of the models *validation.* This is only one of many specialized definitions of validation. This test of the codes was no different than the calibration procedure models normally undergo. If the model of a specified system could be adequately calibrated, the code was deemed validated. In many instances, we can substitute the words *well calibrated* for *validated* without changing significantly the author's meaning.

There are both pragmatic and philosophical grounds to avoid the idea of validation. The idea of validation (or invalidation) is deeply rooted in the philosophy of science. On philosophical grounds, Popper (1968) argued that scientific theory can be invalidated—not validated. Of course, Popper is not the only philosopher of science. Others, notably the pragmatists, of which John Dewey is perhaps the best known, argued that we learn from experience, observations, and mistakes (Menand 2001). The pragmatists argued we never find real truth, but we do get closer to understanding. Kuhn (1970) suggested that scientists try to make existing theory work until finally the evidence indicates that it does not; then they embrace a new theory. None of these philosophers argued that one could validate.

It is unfortunate that we have allowed the term *validation* to become a part of the model lexicon. Oreskes and Belitz (2001) summarize the status of validation:

> "The inherent uncertainties of models have been widely recognized, and it is commonly acknowledged that the term 'validation' is an unfortunate one, because its root—valid—implies a legitimacy that we are not justified in asserting. . . . But old habits die hard and the term persists. In formal documents of major national and international agencies that sponsor modeling efforts, and in the work of many modelers, 'validation' is still widely used in ways that assert or imply assurance that the model accurately reflects the underlying natural processes, and therefore provides a reliable basis for decision-making. This usage is misleading and should be changed. Models cannot be validated. The reasons why have been outlined in detail elsewhere (Konikow and Bredehoeft 1992; Oreskes et al. 1994)."

Reservoir Engineering: A Pragmatic Approach

The ground water community could take a lesson from petroleum reservoir engineering. The usual practice is to history match the reservoir simulator output to some temporal history of production. This is calibration in the ground water lexicon. Based on the match, a prediction of future performance is made, but one is cautious in extending that prediction much beyond a period equal to the production history. In other words, the rule of thumb is that, if we make a 10-year history match, we might be reasonably confident in predicting the next 10 years of performance; however, beyond 10 years, the confidence in a prediction greatly diminishes.

The reservoir engineering community makes no claims about the validity of the model. They simply imply: (1) we have a model that we think incorporates the appropriate physics and chemistry, including the appropriate parameter set, that matches an observed temporal history of reservoir performance; and (2) we will use that model to predict future reservoir performance. Furthermore, continued monitoring of the system is used to refine and improve the model.

Reservoir simulation is important in the petroleum industry. A small improvement in the fraction of petroleum

recovered from a reservoir can amount to many millions of dollars. It is worth looking to reservoir engineering practice as a guide to modeling ground water systems, especially systems involving high risk to society, such as nuclear waste repositories.

Many of the same techniques are used with most normal ground water models. In many water supply models, we have a history of the response of a ground water system to stress. One makes a model that reproduces the history (is calibrated), and then makes predictions of future performance. A well-known caveat is that if the system reaches a new state, the past history may be a poor analog for future performance. Perhaps an example is worth mentioning.

Ground water is being mined from the Denver Basin aquifers in the area just to the south of Metropolitan Denver. Water levels over much of the Denver Basin are declining at rates of 20 to 30 feet per year. So far, the aquifers are still artesian over much of the basin. The question arises as to what will happen as the artesian head is removed and the system becomes water table. Theory suggests that the rate of water level decline will slow as the aquifers pass to water table conditions. However, there are a number of complicating factors. The aquifers are composed of multiple lenticular sand bodies that are not continuous either vertically or across the basin. The layered nature of the sand bodies that make up the permeable portion of the aquifer restrict its drainage; a highly layered aquifer may drain quite differently than a massive thick sand. In addition, there are extreme drawdowns during summer pumping periods that cause the sand bodies to become unsaturated during the summer and then saturate again during the winter; this cyclic drawdown tends to trap air in the sands. There is a debate raging among concerned professionals over what the impact of the complications will be on the water table drainage of the Denver Basin aquifers. We can only speculate until there is some history of how the system responds under water table conditions.

Modeling as an Iterative Process

Good modeling is an iterative process. As new data are acquired, the model is revisited and adjusted (or recalibrated) so that the model predictions are consistent with all the data, including the new data. The model becomes a living tool for analysis. With this paradigm, the modeling strategy changes; it requires continued monitoring and model updating.

We see this strategy at work in many ground water problems. Many problems, especially where there is major concern over the water supply, have been modeled numerous times. I was recently in the Tampa Bay area where there have been three models of seawater intrusion built during the last decade; as digital computers increased in power, each model was more complex. The models progressed from two-dimensional cross sections to fully three-dimensional representations of the system. Each improved the representation of the system. A new flow model is also under construction for the area. The modeling continues because it provides new insights and increased confidence in understanding.

The iterative process is important in addressing the adequacy of the conceptual model. A mismatch between the model prediction and the observed data should raise the issue of conceptualization: Is the mismatch a result of parameter misadjustments or does it suggest that we rethink the conceptual model?

Nuclear Waste

That brings me to the licensing of nuclear waste facilities. The time horizons for these facilities are long—1000 to 10,000 years or longer. The usual model practice of matching a temporal history of system performance and then predicting for a more or less equal period is out of the question. The models are used to predict system performance well beyond an observable history. As modelers, we have to hope that we have (1) included all of the relevant processes in our conceptual model of the system, (2) described the appropriate boundary conditions that are operable through the time horizon of our prediction, and (3) captured the parameters, and their uncertainty, in our representation of the system. This is a tall order.

Performance Assessment

A nuclear waste facility is judged safe if the predicted exposure to radioactivity to an individual located near the boundary of the facility is below a set standard. To make the dose calculation, transport of radioactive components of the wastes is investigated along various exposure pathways. Transport of radioactive isotopes of concern, often by moving subsurface fluids, is predicted within various components of the repository system along the pathways of concern. Various models of the transport processes are linked to perform a performance assessment. The performance assessment is run stochastically so that a probabilistic prediction of the radioactive dose to the hypothetical individual of concern is computed. Ewing et al. (1999) review the use of performance assessment.

Performance assessment sounds obscure in the abstract; perhaps an example will illustrate the procedure. The Waste Isolation Pilot Plant (WIPP) is a salt mine, 2200 feet deep, in the Permian Basin of New Mexico, near Carlsbad, where nuclear waste created by the U.S. weapons programs is being buried. The original concept was that the Salado Salt in which the mine is built would deform plastically and encapsulate the buried nuclear waste within a period of several decades. There are problems with this concept. Once an exploratory mine was constructed, the salt was found to contain 1% to 3% interstitial brine—brine between the salt crystals. This brine migrates into the mine. During mining, the mine ventilation removes the moisture; however, once the mine is closed, the brine accumulates in the closed rooms. Under humid or partially wet conditions, steel drums containing much of the waste will react with the brine producing hydrogen gas, and cellulose in the waste will biodegrade, producing additional gas. Under these conditions, the repository becomes a pressurized, sealed mine in which the pore fluids (brine and gas) resist the plastic collapse of the salt. Finding 1% to 3% brine within the salt required a revised conceptual model for WIPP.

Further complicating WIPP are commercial grade potash deposits that overlie the mine, and oil and gas fields in the surrounding area. The oil and gas fields are believed

to extend beneath the repository. In evaluating the safety of the repository, the U.S. Environmental Protection Agency (EPA) insisted that the scenario of drilling into the repository be assessed. EPA directed that, for the assessment, the current rate of drilling in the area, using the current drilling technology, be extended throughout the 10,000-year time horizon of analysis. The attorney general of New Mexico challenged in court EPA's idea of extending current technology and current drilling frequency into the future. The U.S. Court of Appeals agreed with EPA that it was reasonable to use the current technology and frequency over the entire time horizon as a surrogate measure of the risk from an unknown future in which both drilling frequency and technology will undoubtedly change.

Most investigators thought that the Achilles' heel of WIPP was the human intrusion scenario. Extending the current drilling rate for the 10,000-year planning horizon means statistically that WIPP will be drilled into several times with a probability of 1.0. Using EPA's imposed conditions, there will be drilling hits into the repository.

Performance Assessment: A Cascade of Models

A number of models of the mine and its environment were linked into a single system: the *performance assessment model.* At the base of the pyramid of performance assessment models was a model of the near field; the actual mine, and the reservoir formed by the fluid-filled nuclear wastes. The basic fluid model of the mine describes the multiphase pressure environment within the mine (1) once the mine is sealed, (2) the salt deforms around the waste, and (3) the moist waste and steel drums produce gases. A submodel predicts the temporal concentration of radioactive chemical species of concern in the fluids contained within the waste. An additional submodel predicts the rock mechanics of the salt deformation in response to the fluid pressure in the repository.

The near-field model was embedded in a far-field model that represents the geologic setting that contains the mine. As explained previously, human intrusion through subsequent drilling into the facility is a major concern. Additional submodels of the performance assessment ensemble describe the exhumation of nuclear waste by subsequent drilling. There are models of how drilling through the repository waste brings waste to the surface.

The performance assessment model is operated in a stochastic mode so that a probabilistic prediction is generated. Performance assessment recognizes that the parameters of the models are incompletely known. Using a Latin Hypercube sampling procedure, the parameters of the various submodels are sampled from their expected distribution, although in many instances the assumed parameter distributions are highly uncertain. The idea is that, by running the performance assessment model with repeated sampling of the parameters, we can calculate a statistical distribution of the probable radioactive dose to the hypothetical individual of concern. It is possible through sensitivity analysis to identify the parameters that most control the predicted dose of radioactivity. It is, however, difficult to determine how errors are propagated through the suite of interconnected models (Konikow and Ewing 1999).

WIPP was judged safe largely on the basis of performance assessment. The WIPP performance assessment will form the template for future safety analyses of nuclear waste repositories in the United States.

Bredehoeft (1997, 1998) argued in the case of WIPP that certain human intrusion scenarios were inadequately examined. One of these scenarios was drilling with air. Drilling with air makes penetrating a highly pressurized repository much more hazardous. The weight of the drilling mud compensates for part or all of the high pressure in the repository; when drilling with air, there is no mud column to compensate for the pressure in the repository. Given high pressure in the repository, drilling with air exhumes more waste. A second scenario of concern was a leak in a reinjection brine well that created an extended hydraulic fracture. Such hydraulic fractures had occurred, associated with water flooding for oil recovery within the New Mexico portion of the Delaware Basin. A hydraulic fracture into the repository could create a fluid short circuit and potentially move large volumes of brine through the repository, leaching and transporting hazardous radionuclides. The U.S. Department of Energy and EPA viewed these scenarios as low probability events.

Many of the individuals concerned with the safety of WIPP believed that the human intrusion scenario dictated by EPA was unlikely. Because human intrusion is the most probable foreseeable failure scenario, these individuals felt the repository was inherently safe.

Linking a cascade of models compounds the calibration problems associated with each component model. Many of the models used for performance assessment at WIPP were theoretical and poorly calibrated. Extending the time horizon to 10,000 years further compounds the difficulties. The hope is that the statistical sampling of the important parameters in the performance assessment model will provide a probabilistic range of future outcomes. If 95% or 99% of these outcomes are within acceptable limits, the repository is judged to be safe. This approach does not address the problem that we may have overlooked something important in our conceptual model of the system. *Repeated sampling of a large parameter set may compensate for the uncertainty in the parameter values for the models used in performance assessment, but it does not compensate for wrong or incomplete conceptual models.*

Probabilistic performance assessment raises the issue of precision versus accuracy. The probabilistic approach may give the illusion that the modeler has quantified the error associated with the model. However, if darts are thrown at the wrong target, the spread of darts does not provide an assessment of whether the right target was chosen.

History Matching in Nuclear Waste Facilities

As indicated previously, the time horizon for formally predicting doses from nuclear waste facilities is 1000 to 10,000 years, or longer. Some of the longer lasting radioactive isotopes will persist well beyond 10,000 years. Given that the time horizon of the predictions is 10,000 years or longer, there is no opportunity for the traditional history match followed by a more or less equal period of prediction. Given the

current strategy, there is, at best, a set of experiments of limited duration to which the models can be calibrated.

Yucca Mountain

One of the principal tools for evaluating the suitability of Yucca Mountain as a repository will be performance assessment—performance assessment similar to that used at WIPP. There are additional complications at Yucca Mountain. The waste to be emplaced at Yucca Mountain will generate heat. At issue is whether the loading will be relatively dense, producing high temperatures within the host rock, temperatures above the boiling point of water, or whether the spent nuclear fuel will be distributed more widely, keeping the host rock temperatures below the boiling point of water. Many investigators have concluded that the higher temperatures greatly increase the uncertainty of how the ground water system within the mountain will respond, and are to be avoided.

At both WIPP and Yucca Mountain, there were surprises once mining allowed scientists/engineers to actually visit the underground. At WIPP, the salt observed in the underground contained 1% to 3% interstitial brine. The original concept was that the only brine in the salt was in vesicles contained within the salt crystals—about 0.5%. Finding interstitial brine meant that the facility would be moist, a fact that was not included in the original conceptual model. The brine at WIPP did not preclude using the facility as a repository; it greatly complicated the analysis of the safety of the facility. It increased the uncertainty of the prediction of performance. In the end, WIPP was still judged to be safe by EPA, as well as by much of the scientific community (National Research Council 1996).

At Yucca Mountain, water containing chlorine–36 that was derived from atmospheric testing of nuclear weapons was found in the underground drift. The chlorine–36 indicates a *fast-path* for moisture movement in the mountain, a path that is unpredicted by the conventional theory of transport in the unsaturated zone, even a fractured unsaturated zone. The task at Yucca Mountain is to predict transport in a fractured, unsaturated media, subjected to a heat load for a prolonged period—10,000 years.

Performance assessment is dependent on having a correct conceptual model of transport within the mountain. At Yucca Mountain, the appropriate conceptual model for simulating unsaturated transport in the fractured tuffs at the site is unclear. A recent study by the National Research Council (2001) concluded that there is no consensus within the hydrogeologic community of what the appropriate conceptual model is to describe transport in a fractured, unsaturated zone. The generally accepted theory does not predict the chlorine–36 movement. This lack of a clear conceptual model greatly increases the uncertainty associated with performance assessment at Yucca Mountain. Without a consensus on the appropriate conceptual model, predictions of future system performance become highly questionable, at best.

Where Are We as a Community of Modelers?

Models are useful in integrating and synthesizing our knowledge about hydrogeologic systems in a way that allows us to make predictions about the future performance of the system. Most of us regard models as our best tools for the task. However, anyone engaged in this processes recognizes its inherent uncertainty. Modelers also recognize a pervasive element of professional judgment in creating models and judging their effectiveness. To some extent, these ideas are embedded in what we generally refer to as model calibration. Unfortunately, model calibration may or may not adequately test our conceptual model. Too often, an incomplete conceptual model can pass the test of being calibrated. Too often, the models have proven to be incomplete or wrong. As hydrogeologists, we make mistakes.

Oreskes et al. (1994) summarize the uncertainty in modeling; they state, " . . . the establishment that a model accurately represents the 'actual processes occurring in a real system' is not even a theoretical possibility."

Probabilistic performance assessment does not overcome the inherent uncertainty in modeling. Performance assessment is conducted in a probabilistic mode to compensate for the uncertainties in the parameters (and perhaps the boundary conditions). As suggested previously, uncertainties in what are the appropriate conceptual models are not compensated for by probabilistic sampling of the parameter sets of wrong or incomplete conceptual models.

Oreskes and Belitz (2001) regard the conceptual model as the most difficult problem in modeling; they state:

> "Conceptualization is probably the most thorny issue in modeling. It is the foundation of any model, and everyone knows that a faulty foundation will produce a faulty structure. . . . Yet what to do about it remains a problem. Much attention in model assessment has focused on quantification of error, but how do we quantify the error in a mistaken idea? . . . It is uncertainty rooted in the foundations of our knowledge, a function of our limited access to and understanding of the natural world. Almost by definition, conceptual error cannot be quantified. We don't know what we don't know, and we can't measure errors that we don't know we've made."

Iterative modeling in which we continue to monitor and revise the models to fit new data provides the best opportunity to avoid errors, including errors of conceptualization. However, iterative modeling, while it improves our odds for success, is not foolproof; modeis still have an inherent uncertainty.

Given the inherent uncertainty associated with models, Oreskes and Belitz (2001) ask the relevant question: Are predictions necessary for policy decisions? Uncertainty associated with model predictions may make alternative strategies or complementary courses of action more reasonable for society. We should examine the alternatives.

A Return to History Matching

The closer we can approach the idea embedded in the reservoir engineering concept of history matching, the more confidence we have in predictions. We would like the period of the history match to approach as nearly as possible the length of the prediction—our rule of thumb for confidence in prediction discussed previously.

Yucca Mountain could be a case in point. At Yucca Mountain, it seems that nuclear wastes could be emplaced in a retrievable mode within the repository for a long period. Our uncertainty in the models of the basic processes

at Yucca Mountain argues strongly for a long period of observation.

The concept of Yucca Mountain could be changed to one of monitored retrievable storage for an indefinite period, perhaps 300 to 1000 years. A long period of monitoring of the facility could provide a history of performance to which the models could be repeatedly matched and improved. At the time that the models are demonstrated to reproduce the performance of the repository for a greatly extended period, society will be in a much stronger position to judge the suitability of the site as a permanent repository. I would urge that we rethink the nuclear repository at Yucca Mountain with the idea of keeping the repository open for observation for a prolonged and, for now, indefinite period.

The arguments for early closure of Yucca Mountain do not seem scientific, but rather political. Political considerations can often be changed by persuasive scientific arguments. Uncertainty associated with the predictions of the system behavior is a good reason not to be in a hurry to close the repository. Early closing of the repository may well be premature.

Acknowledgments

I wish to thank my colleagues Bill Alley, Mary Anderson, Ken Belitz, Lenny Konikow, and Naomi Oreskes for their insights in reviewing this paper.

References

Anderson, M.P., and W.W. Woessner. 1992. The role of the postaudit in model validation. *Advances in Water Resources* 15, 167–173.

Bethke, C. 1992. The question of uniqueness in geochemical modeling. *Geochimica et Cosmochimica Acta* 56, 4315–4320.

Bredehoeft, J.D., C.E. Neuzil, and P.C.D. Milly. 1983. Regional flow in the Dakota Aquifer: A study in the role of confining layers. U.S. Geological Survey Water-Supply Paper 2237.

Bredehoeft, J.D. 1997. *Air Drilling into WIPP.* Submitted to U.S. EPA Compliance Docket No. A–93–02—WIPP (prepared for New Mexico attorney general).

Bredehoeft, J.D. 1998. *Drilling with Mud and Air into WIPP—Revisited.* Submitted to U.S. EPA Compliance Docket No. A–93–02—WIPP (prepared for New Mexico attorney general).

Carslaw, H.S., and J.C. Jaeger. 1959. *Heat Conduction in Solids,* 2nd edition. Oxford, U.K.: Oxford University Press.

Ewing, R.C., M.S. Tierney, L.F. Konikow, and R.P. Rechard. 1999. Performance assessments of nuclear repositories: A dialogue on their value and limitations. *Risk Analysis* 19, 933–958.

Hantush, M.S. 1964. Hydraulics of wells. In *Advances in Hydroscience,* vol. 1, ed. V.T. Chow, 281–432. New York: Academic Press.

Jacob, C.E. 1940. On the flow of water in an elastic artesian aquifer. *Transactions of the American Geophysical Union* 2, 585–586.

Konikow, L.F., and J.D. Bredehoeft. 1992. Groundwater models cannot be validated. *Advances in Water Resources* 15, 75–83.

Konikow, L.F., and R.C. Ewing. 1999. Is a probabilistic performance assessment enough? *Ground Water* 37, no. 4: 481–482.

Kuhn, T.S. 1970. *The Structure of Scientific Revolution,* 2nd edition. Chicago: University of Chicago Press.

Menand, L. 2001. *The Metaphysical Club: A Story of Ideas in America.* New York: Farrar, Straus, and Giroux.

National Research Council. 1996. *The Waste Isolation Pilot Plant: A Potential Solution for the Disposal of Transuranic Waste.* Committee on the Waste Isolation Pilot Plant. Washington, D.C.: National Academy Press.

National Research Council. 2001. *Conceptual Models of Flow and Transport in the Fractured Vadose Zone.* Washington, D.C.: National Academy Press.

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

Oreskes, N., and K. Belitz. 2001. Philosophical issues in model assessment. In *Model Validation: Perspectives in Hydrological Sciences,* ed. M.G. Anderson and P.D. Bates, 23–41. New York: John Wiley and Sons.

Popper, K.R. 1968. *The Logic of Scientific Discovery.* New York: Harper and Row.

Theis, C.V. 1935. The relation between lowering the piezometric surface and the rate and duration of discharge of a well using groundwater storage. In *Transactions of the American Geophysical Union,* 16th annual meeting, part 2; 519–524. Washington, D.C.: AGU.

Editor's Note: We invited Dr. Bredehoeft to contribute an issue paper on a topic of his choice to mark our 40th anniversary celebration this year of the publication of the first issue of *Ground Water.* Dr. Bredehoeft is a former editor-in-chief of the journal.

---

ground water

Multimodel Ranking and Inference in Ground Water Modeling

by Eileen Poeter[1] and David Anderson[2]

Abstract

Uncertainty of hydrogeologic conditions makes it important to consider alternative plausible models in an effort to evaluate the character of a ground water system, maintain parsimony, and make predictions with reasonable definition of their uncertainty. When multiple models are considered, data collection and analysis focus on evaluation of which model(s) is(are) most supported by the data. Generally, more than one model provides a similar acceptable fit to the observations; thus, inference should be made from multiple models. Kullback-Leibler (K-L) information provides a rigorous foundation for model inference that is simple to compute, is easy to interpret, selects parsimonious models, and provides a more realistic measure of precision than evaluation of any one model or evaluation based on other commonly referenced model selection criteria. These alternative criteria strive to identify the true (or quasi-true) model, assume it is represented by one of the models in the set, and given their preference for parsimony regardless of the available number of observations the selected model may be underfit. This is in sharp contrast to the K-L information approach, where models are considered to be approximations to reality, and it is expected that more details of the system will be revealed when more data are available. We provide a simple, computer-generated example to illustrate the procedure for multimodel inference based on K-L information and present arguments, based on statistical underpinnings that have been overlooked with time, that its theoretical basis renders it preferable to other approaches.

Introduction

Sparse subsurface data cause us to be uncertain of the exact nature of ground water system structure and components. Consequently, it is a best, although not always customary, practice to evaluate multiple models of a ground water system before making predictions of system behavior. Alternative models include variations in the structure of hydrogeologic units, boundary conditions, and parameter fields. Each alternative model must be calibrated (i.e., parameter values adjusted to obtain the best fit to the observed data, e.g., using nonlinear least squares) before models can be compared (Poeter and Hill 1997). The advent of high-speed computing and robust inversion algorithms makes calibration of multiple models feasible.

We often find that prediction uncertainty is larger across the range of potential models than that which arises from the misfit and insensitivity of any one optimized model, even to the extent that confidence intervals on predictions from some of the models may not include the values predicted by others. This raises the question of whether to select the best model and use those predictions and confidence intervals for decision and design or to weight all the models and calculate model-averaged predictions and intervals. If one model is clearly superior to the rest, it is reasonable to use that model for prediction, but its uncertainty should be evaluated using the entire set of candidate models. If one model is not clearly superior, then it is reasonable to weight all predictions. If the alternative models yield substantially different results for the prediction of interest such that a reasonable decision is untenable, then additional data should be collected to develop better models.

A more representative model of ground water system behavior (1) exhibits no consistent spatial or temporal

[1]Corresponding author: International Ground Water Modeling Center, Department of Geology and Geological Engineering, Colorado School of Mines, 1516 Illinois Street, Golden, CO 80401; (303) 273-3829; fax (303) 384-2037; epoeter@mines.edu

[2]Applied Information Company, 707 Breakwater Drive, Fort Collins, CO 80525; (970) 229-0255; fax (970) 229-0255; aicanderson1@comcast.net

Received July 2004, accepted October 2004.

pattern in the weighted residuals; (2) results in reasonable estimated parameter values (e.g., hydraulic conductivity of gravel is higher than that of silt and falls within the range of values that might be expected for gravels); and (3) has better fit statistics for the same data while maintaining parsimony (i.e., balancing the bias vs. variance trade-off or the trade-off between underfitting and overfitting). There is a general agreement that considerable mental effort, training, and experience are required to define a set of reasonable models (Bredehoeft 2003; Neuman and Wierenga 2003). However, the profession has not agreed upon a procedure for ranking or weighting models (Carrera and Neuman 1986; Neuman and Wierenga 2003; Ye et al. 2004).

We have several objectives. First, we call attention to the famous geologist Chamberlin's (1890) call for "multiple working hypotheses" as a strategy for rapid advances in understanding applied and theoretical problems. Each hypothesis or conceptualization is represented by a mathematical model, which gives rigor to the procedure, then data collection and analysis focus on which model is the best, that is, most supported by the data. Second, we introduce a simple and effective approach for the selection of a best model: one that balances underfitting and overfitting (i.e., maintains parsimony). Third, we provide an effective method for making formal multimodel inference, including prediction, from all models in a candidate set. Finally, we present a computer-generated example to illustrate the method, and we comment on alternative approaches.

Model Ranking and Inference from the Best Model

Multiple Working Hypotheses

Ideally, understanding in science comes from strict experimentation. Here, causation can be identified and interactions can be explored. In most cases, an array of practical considerations prevent experimentation in ground water studies. At the opposite extreme are studies that are merely descriptive. Here, progress in understanding is slow and risky. Lack of causation, and other issues, makes this a relatively poor approach. Between these extremes lie studies that can be termed "observational," where inference is model based. One attempts to extract the information in the data using a model. Many ground water problems are in this observational category, and inferences are inherently model based, thus the need for multimodel inference in ground water modeling.

Given a well-defined ground water problem, with extensive thoughtful consideration a hydrologist can conceptualize R hypotheses concerning the system and the questions to be asked. R might range from two to three to perhaps a few dozen or even 100s in cases where statistical techniques are used to generate realizations. Given a good set of data, hypotheses, and models, an investigator can ask, "which hypothesis is most supported by the data?" This is the model selection problem and the heart of Chamberlin's strategic approach. Model selection is a fundamental part of the data analysis. Approaches to optimal inference for one model and data set are known (e.g., least squares or maximum likelihood methods). The central issue is "which model to use?"

Model Selection

A large effort has been spent on a coherent theory of model selection over the past 30 years. We will not review this material in detail as it is covered in a number of books (e.g., Linhart and Zucchini 1986; McQuarrie and Tsai 1998; Burnham and Anderson 2002), research monographs (e.g., Sakamoto et al. 1986), and hundreds of journal papers (e.g., deLeeuw 1992). Instead, we briefly outline the approach we recommend.

The starting point for effective model selection theory is Kullback-Leibler (K-L) information, $I(f,g)$ (Kullback and Leibler 1951). This is interpreted as the information, I, lost when full truth, f, is approximated by a model, g. Given a set of candidate models g_i, one might compute K-L information for each of the R models and select the one that minimizes information loss—that is, minimize $I(f,g)$ across models. This is a compelling approach. However, for ground water models, K-L information cannot be computed because the truth and the optimal effective parameters (e.g., hydraulic conductivities, boundary heads, and fluxes) are not known (Anderson 2003).

Akaike (1973, 1974) provided a simple way to estimate expected K-L information, based on a bias-corrected, maximized log-likelihood value. This was a major breakthrough (Parzen et al. 1998). Soon thereafter, better approximations to the bias were derived (Sugiura 1978; Hurvich and Tsai 1989, 1994) and the result, of relevance here, is an estimator Akaike Information Criterion (AICc) of twice the expected K-L information loss

$$\mathrm{AICc} = n \log(\sigma^2) + 2k + \left(\frac{2k(k+1)}{n-k-1}\right) \qquad (1)$$

where σ^2 is the estimated residual variance, n is the number of observations, and k is the number of estimated parameters for the model. Here, the estimator of σ^2 = WSSR/n, where WSSR is the weighted sum of squared residuals. The second term accounts for first-order bias, and the third term accounts for second-order bias resulting from a small number of observations. This is a precise mathematical derivation, with the third term depending on the assumed distribution of residuals, in this case, normally distributed error. Accounting for second-order bias is important when $n/k < 40$, which is typical of ground water models. The aforementioned expression applies to analyses undertaken by a least squares approach; similar expressions are available for those using maximum likelihood procedures (Akaike 1973). AICc is computed for each of the models; the model with the lowest AICc value is the best model, and the remaining models are ranked from best to worst, with increasing AICc values.

As parameters are added to a model, accuracy and variance increase (fit improves, while uncertainty increases). Use of AICc selects models with a balance between accuracy and variance; this is the principle of parsimony. Prediction can be further improved by basing inference on all the models in the set (multimodel inference, as discussed later).

Delta Values

Calculation of the AICc values can be posed so as to retain or omit values that are constant across models (e.g., multinomial coefficients) and are affected by the number of observations; thus, it is essential to compute and use simple differences

$$\Delta_i = \mathrm{AICc}_i - \mathrm{AICc}_{\min} \tag{2}$$

for each model, i, in the set of R models, where $\mathrm{AICc}_{\min}$ is the minimum AICc value of all the models in the set. These values are on an information scale (–log[probability]), free from constants and sample size issues. A Δ_i represents the information loss of model i relative to the best model. As discussed by Burnham and Anderson (2002, p. 70–72 and particularly 78), models with $\Delta_i <$ 2 are very good models, while models with $4 < \Delta_i <$ 7 have less empirical support. In most cases, models with Δ_i greater than ~10 can be dismissed from further consideration.

Model Probabilities

Simple transformation yields model probabilities or Akaike weights (also referred to as posterior model probabilities)

$$w_i = \frac{\exp^{-0.5\Delta_i}}{\sum_{j=1}^{R} \exp^{-0.5\Delta_j}} \tag{3}$$

where w_i is the weight of evidence in favor of model i being the best model in the sense of minimum K-L information loss. These weights are also useful in multimodel inference as discussed later.

Evidence Ratios

It is convenient to take ratios of the model probabilities for models i and j as w_i/w_j and call these evidence ratios. These are most useful when i is the best model and j is another model of interest because they can be used to make statements such as "there is 'w_i/w_j' times more evidence supporting the best model."

Example Problem

Our goal is to illustrate model evaluation first by calibrating a set of simple (coarse versions of the "truth") ground water models of a synthetic (known) system (as defined by a generating model), then making multimodel inference of predictions. The alternative models used for the example are simplistic relative to models of field sites using only zonation variations generated by a geostatistical simulator. We do not offer this as a desired approach to model development, only as a method for generating models to demonstrate the procedure. Each coarse model is calibrated by weighted least squares nonlinear regression under the initial pumping condition using 20 head observations and 1 base flow observation. Then, we rank and determine weights for the models. In the predictive stage, additional pumping is simulated at another location and head is predicted at 20 locations (offset from the calibration data locations), while two flows are also predicted. In a subsequent section, we illustrate multimodel inference of the predicted heads and flow rates and compare them to the known predictions simulated by the generating model.

Synthetic Model

A two-dimensional, unconfined steady-state system is synthesized with a model domain 5000 m in the east-west direction and 3000 m north-south direction (Figure 1). The aquifer is assigned boundary conditions as follows:

- A no-flow boundary is defined on the northern, western, and southern borders, and the aquifer base at −10 m.
- A 10-m-wide river, in the center of the watershed, ranges in stage from 20 to 5 m and is underlain by 5-m-thick sediments with their base at an elevation of 5 m. Rivers are represented as a head-dependent flux boundaries using the MODFLOW-2000 (Harbaugh et al. 2000) river package.
- A 10-m-wide river also bounds the east edge with a stage of 5 m, and 5 m of sediments with their base at 0 m.
- A recharge of 8×10^{-4} m/d is applied uniformly to the top of the model, constituting all the inflow to the system.
- A well pumps 2000 m^3/d at $x = 2050$ and $y = 550$.

True heads and flows are generated using a synthetic heterogeneous model with five zones of hydraulic conductivity (K), and a grid of 250 × 150 cells, each 20 × 20 m (Figure 1). The model grid used for calibration and prediction consists of 50 × 30 cells, each 100 × 100 m (Figure 2). The "true" hydraulic conductivity distribution (Figure 1) includes five zones, with values ranging from 1 to 25 m/d. Vertical hydraulic conductivity of the east-west–oriented riverbed is 0.2 m/d, while that of the north-south riverbed is 0.1m/d.

Alternative Models

In practice, alternative models should be developed based on careful consideration of the uncertainties associated with understanding of the site hydrology and their representation by the simulation software. For the purpose of illustrating the model evaluation procedure, we generate alternative models by varying the number and

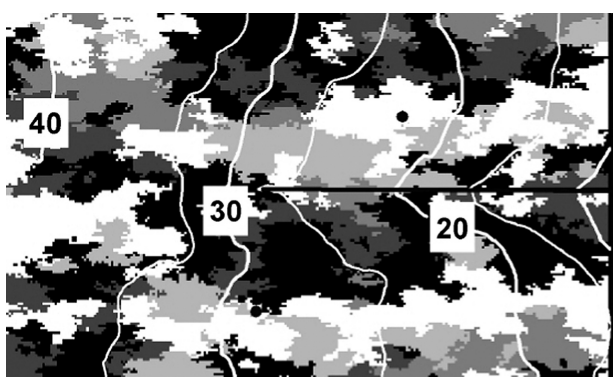

Figure 1. True heterogeneity and head distribution for the synthetic model under hydraulic conditions used to generate calibration data.

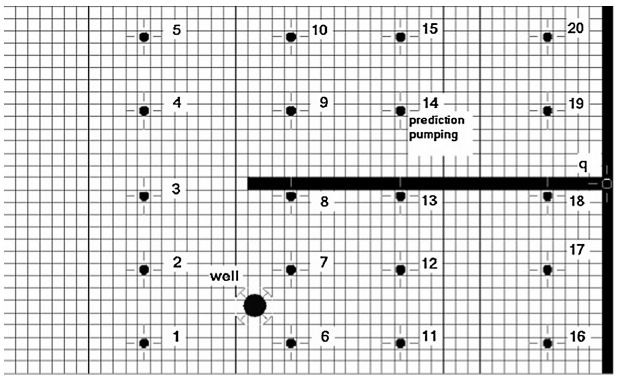

Figure 2. Coarse grid showing rivers (bold lines), observation locations (heads: dots, flux: bracket), and pumping well location.

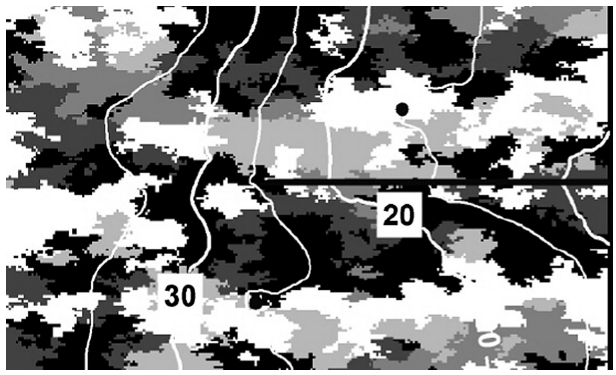

Figure 3. Head distribution in the synthetic model under hydraulic conditions for prediction.

distribution of hydraulic conductivity zones. Ten sequential indicator simulations (A through J) are generated on the fine grid using GeoStatistical LIBrary (GSLIB) (Deutsch and Journel 1992), the indicator variograms of the synthetic hydrogeologic units, and honoring 144 points of known lithologic type taken from the generating model on a regular grid. Each realization was partitioned into 2 two-zone (e.g., 2A-2J and 2AL-2JL: L indicates that bias is toward low-K material because zone 3 is included with zones 1 and 2 rather than zones 4 and 5), 1 three-zone (3A-3J), 2 four-zone (4A-4J and 4AL-4JL; for L models, zone 3 is included with zone 2 rather than zone 4), and 1 five-zone model (5A-5J); in addition, a homogeneous model was evaluated, resulting in a total of 61 models. In field application the diversity of models will be much greater, including variations of boundary conditions, geologic structure and unit thicknesses, as well as the use of alternative code features to represent features of the ground water system (e.g., in MODFLOW using constant head cells vs. drains, rivers, or streams to simulate communication with surface water).

Calibration data include 20 head observations on a regular grid from the generating model with a hypothesized standard deviation of 0.02-m measurement error and a base flow observation to the central tributary of 6188 m^3/d with a standard deviation of 58 m^3/d. These standard deviations needed to be increased by a factor of 38 to account for model error and obtain a calculated error variance of 1.0. MODFLOW (Harbaugh et al. 2000; Hill et al. 2000) is used to simulate heads and flows for each model and to estimate a value for K of each zone and the uniform recharge rate; using weights calculated as the inverse of the measurement variance resulted in a dimensionless weighted sum of squared residuals (WSSR). The calibrations require a few seconds on a 3-GHz Pentium 4 PC.

Predictions of flow to both the central tributary and the eastern river and heads at 20 locations, each 200 m upgradient of the calibration data locations, are made while simulating additional pumping of 3000 m^3/d at x = 3250 m and y = 2150 m. Head distribution in the generating model for the predictive conditions is illustrated in Figure 3.

Evaluation Software

J_MMRI is used to evaluate example models. J_MMRI is an early-stage application of the JUPITER (Joint Universal Parameter IdenTification and Evaluation of Reliability) application programming interface (API), which is currently under development through cooperation of the USGS and U.S. EPA (Poeter et al. 2003). The API provides researchers with open-source program modules and utilities that undertake universal basic tasks required for evaluating sensitivity, assessing data needs, estimating parameters, and evaluating uncertainty, so researchers can focus on developing methods without "reinventing the wheel," while providing practitioners with public domain software to facilitate the use of the new techniques. J_MMRI collects soft information about each model including (1) model structure: dimensionality, complexity of processes, method of parameter generation/degree of regularization, model representation of features, number/size of model cells/elements, and length/mass/time units; (2) residual distribution: spatial, temporal, and randomness; (3) feasibility of optimal parameter values: absolute and relative; (4) objective function: weighted sum of squares and log likelihood; (5) model selection statistics (e.g., AICc, Bayesian information criterion [BIC], Hannan and Quinn's criterion [HQ], and Kashyap's information criterion [KIC]); (6) residual quality: Gaussian character, degree of spatial bias, and similarity to data error; and (7) parameter correlation/certainty. This information is analyzed and organized to facilitate subjective evaluation of the models and provide quantitative model ranking and weighting measures.

Model Ranks

Models were discarded from consideration if the regression did not converge in 20 iterations (two models), or K of a lower-zone number (finer grained material) exceeded the K of a higher-zone number (13 models), leaving 46 of the 61 models for ranking and weighting. It is preferable to include a defensible number of plausible models; however, some models yield unreasonable relative values or cannot be used because they do not converge; thus, the results are not valid for use in further computation. Examinations show that these situations typically occur when the connectivity of hydraulic

conductivity units differ significantly from the true conditions (Poeter and McKenna 1995). For example, if a discontinuous high-K field unit is represented in a model by a continuous unit, then a low-K value may be estimated for the high-K unit in order to compensate for too much continuity. Model selection statistics are given for the best 18 models in Table 1. The number of parameters varied from only three (K for one zone, recharge rate, and σ^2) to seven (K for five zones, recharge rate, and σ^2). σ^2 is counted as a parameter because formally, the likelihood function in the case of normal errors reads as $L(B, \sigma^2|X,g)$ and means "the likelihood of the (unknown) vector of β_s, and, σ^2, given the data (X) and the model (g)." From the AICc scores, the Δ_i values, and weights, model 4F is the best model, 2J ranks second, and models 5J, 4FL, and 3F have less support, while a number of models have weights of a few percent. The remaining models have relatively little empirical support. Most of the 10 five-zone (seven-parameter) models, are not retained based on unreasonable relative parameter values. Although there are only 21 observations, the more complex models receive high ranks, likely due to the fact that all the geostatistical simulations were well conditioned so the complex models capture the zones well. With less conditioning, simpler models may do a better job of capturing the gross connectivity.

Alternative Model Selection Criteria

We recommend approaches based on K-L information (e.g., AICc) for both model selection and multimodel inference. These methods are based on the concept that models are approximations (i.e., there are no true models of field systems) and select models with more parameters (structure) as the number of observations increase. That is, in complex systems, smaller effects are identified as the number of observations increase.

There are many other criteria for model selection (McQuarrie and Tsai 1998), and we offer brief comments on some of the alternatives. The BIC (Schwarz 1978), HQ (Hannan and Quinn's 1979) criterion, and KIC (Kashyap 1982) have been suggested for selection of ground water models (Carrera and Neuman 1986; Neuman 2003; Neuman and Weirenga 2003; Ye et al. 2004). These criteria are similar in form to AICc and are as follows

$$\mathrm{BIC} = n \log(\sigma^2) + k \log(n) \quad (4)$$

$$\mathrm{HQ} = n \log(\sigma^2) + ck \log(\log(n)) \quad \text{where, } c > 2 \quad (5)$$

$$\mathrm{KIC} = n \log(\sigma^2) + k \log\left(\frac{n}{2\pi}\right) + \log|X^{\mathrm{T}}\omega X| \quad (6)$$

where, $|X^{\mathrm{T}}\omega X|$ is the determinant of the Fisher information matrix, X is the sensitivity matrix, X^{T} is its transpose, and ω is weight matrix.

We do not recommend these procedures as they assume that the true (or quasi-true) model exists in the set of candidate models (Burnham and Anderson 2004), and their goal is to identify this model (as n approaches infinity, probability converges to 1.0 for the true model). These criteria strive for consistent complexity (constant k) regardless of the number of observations. In practice, these criteria can perform similarly to AICc; however, their theoretical underpinnings are philosophically weak. McQuarrie and Tsai (1998) give a readable account of this issue, as do Burnham and Anderson (2002, sections 6.3 and 6.4). Deeper insights are provided in Burnham and Anderson (2004).

Recall that as the number of estimated parameters increases, bias decreases but variance increases (i.e., precision decreases, error bars are larger). The alternative criteria approach the "true model" asymptotically (i.e., as the number of observations increase). However, in most ground water models, the number of observations is small relative to the number of parameters estimated, and these criteria tend to select models that are too simple (i.e., underfitted). Thus, they tend to select for less bias and greater certainty, which threatens to capture a precise but inaccurate answer. We argue that it is preferable to select the model that provides the best approximation to reality for the number of observations available.

A final comment is that AICc and BIC can be derived under either a Bayesian or a frequentist framework. Thus, an argument for or against a criterion should not be based on its Bayesian or frequentist lineage. Rather, one must ask if the true (or quasi-true) model can be expected to be in the set of candidate models in a particular discipline. If so, then criteria such as BIC, HQ, and KIC should be used. In cases where models are merely approximations to complex reality, AICc is preferable (Burnham and Anderson 2002). In addition, AICc has a cross-validation property that is important and stems from its derivation (Stone 1977).

Ranks and model probabilities (weights) for the best 18 models based on AICc are presented in Table 2. BIC

Table 1
Statistics for the 18 Best Models[1] (n = 21 in all)

ID**	WSSR	σ^2	k	AICc	Δ_i	w_i
4F	9.40	0.45	5	1.1	0.0	0.2585
2J	13.67	0.65	3	1.5	0.4	0.2155
5J	8.03	0.38	6	2.4	1.3	0.1356
4FL	10.41	0.50	5	3.3	2.1	0.0884
3F	12.82	0.61	4	3.6	2.5	0.0734
3D	13.67	0.65	4	5.0	3.9	0.0374
2H	16.53	0.79	3	5.5	4.3	0.0294
2F	16.68	0.79	3	5.7	4.5	0.0267
5F	9.38	0.45	6	5.7	4.6	0.0262
2A	17.75	0.85	3	7.0	5.8	0.0139
3G	15.25	0.73	4	7.3	6.2	0.0119
2GL	18.11	0.86	3	7.4	6.3	0.0112
2E	18.14	0.86	3	7.4	6.3	0.0111
2C	18.61	0.89	3	8.0	6.8	0.0085
2B	18.71	0.89	3	8.1	6.9	0.0080
4GL	13.17	0.63	5	8.2	7.1	0.0075
2FL	19.17	0.91	3	8.6	7.5	0.0062
4D	13.52	0.64	5	8.8	7.6	0.0057

[1] The remaining 28 models had essentially zero weight ($<5\times10^{-03}$) and are not shown.
** See Alternative Conceptual Models Section for description of model IDs.

Table 2
Weights in Rank Order[1]

Model[2]	BIC	Model[2]	HQ	Model[2]	KIC
5J	0.3034	5J	0.4280	5J	0.4549
4F	0.2638	4F	0.2473	4F	0.1884
2J	0.1086	4FL	0.0845	5F	0.0795
4FL	0.0902	5F	0.0828	4FL	0.0717
5F	0.0587	2J	0.0449	2J	0.0695
3F	0.0464	3F	0.0289	3F	0.0276
3D	0.0237	3D	0.0147	3D	0.0165
2H	0.0148	4GL	0.0072	5G	0.0116
2F	0.0134	2H	0.0061	4GL	0.0103
4GL	0.0077	2F	0.0056	2H	0.0074
3G	0.0075	4D	0.0055	2F	0.0067
2A	0.0070	4N	0.0049	5A	0.0066
4D	0.0058	3G	0.0047	3G	0.0065
2GL	0.0057	5B	0.0044	5B	0.0052
2E	0.0056	5G	0.0043	4D	0.0051
4N	0.0052	5A	0.0034	4N	0.0045
2C	0.0043	2A	0.0029	2A	0.0036
2B	0.0040	2GL	0.0023	5D	0.0035

[1] Top 18 ranked models, remaining models had very low weights.
[2] See "Alternative Conceptual Models" section for description of model labels

and HQ produce results similar to AICc for this particular example, where n = 21 and k ranges from only three to seven parameters. The same model is ranked highest by all three measures. The same seven models occupy the top seven ranks (constituting 89%, 93%, and 91% of the weight for BIC, HQ, and AICc, respectively) although in slightly different order. At lower ranks, there is more variation.

Multimodel Inference

The traditional approach to data analysis has been to find the best model, based on some criteria or test result, and make inferences, including predictions and estimates of precision, conditional on this model (as if no other models had been considered). In hindsight, this strategy is poor for a number of reasons. Often, the best model is not overwhelmingly best; perhaps, the weight for the best model is only 0.25 as in Table 1. Thus, there is nonnegligible support for other models. In this case, confidence intervals estimated using the best model are too narrow, and multimodel inference is desirable.

Model Averaging

Model averaging allows estimation of optimal parameter values and predictions from multiple models. Both are calculated in a similar manner; however, we discuss model averaging of predictions first because it is straightforward due to the fact that the same items are predicted using each model, whereas each model may not have the same parameters.

In the example, the best model, 4F, has an AICc weight of only 0.26. This value reflects substantial model uncertainty. If a predicted value differs markedly across the models (i.e., the $\hat{y}$ differs across the models i = 1, 2,…, R), then it is risky to base prediction only on the selected model. An obvious possibility is to compute an estimate of the predicted value, weighting the predictions by the model weights (w_i). This can be done under either a frequentist or Bayesian paradigm. Here, we take the frequentist approach, using K-L information because it is easy to compute and effective in application. If no single model is clearly superior, one should compute model-averaged predictions as

$$\hat{\bar{y}} = \sum_{i=1}^{R} w_i \hat{y}_i \tag{7}$$

where $\hat{y}_i$ is the predicted value for each model i, and $\hat{\bar{y}}$ denotes the model-averaged estimate.

For the estimated regression parameter, $\hat{\beta}_j$, we average over all models where $\hat{\beta}_j$ appears

$$\hat{\bar{\beta}}_j = \sum_{i=1}^{R'} w'_i \hat{\beta}_{j,i} \tag{8}$$

Thus, the model weights must be recalculated to sum to 1 for the subset of models, R', that include $\hat{\beta}_j$. When possible, one should use inference based on the subset of models that include $\hat{\beta}_j$ via model averaging because this approach has both practical and philosophical advantages. Where a model-averaged estimator can be used, it appears to improve accuracy and estimates of uncertainty, compared to using $\hat{\beta}_j$ from the selected best model (Burnham and Anderson 2002, section 7.7.5). Parameter averaging is rarely useful for ground water modeling because use of an average parameter value in a particular model construct is not appropriate. However, model-averaged parameter values could provide a range of values for a material type given its multiple representations in alternative models.

Unconditional Variance

Unconditional variance is calculated from multiple models for either parameter values or predictions as shown here for predictions

$$\hat{\bar{v}}\mathrm{ar}(\hat{\bar{y}}) = \left[\sum_{i=1}^{R} w_i \left[\hat{\mathrm{v}}\mathrm{ar}(\hat{y}_i \middle| \mathrm{model}_i) + (\hat{y}_i - \hat{\bar{y}})^2\right]^{0.5}\right]^2 \tag{9}$$

This expression allows for model selection uncertainty to be part of precision because the first term represents the variance, given one calibrated model, and the second term represents the among-model variance, given the set of models. This variance should be used whether the prediction is model averaged or not. The standard deviation is merely the square root of the unconditional variance. Thus, approximate 95% confidence intervals can be expressed using the usual procedure

$$95\%\ \text{confidence intervals on}\ \hat{\bar{y}} = \hat{\bar{y}} \pm 2\sqrt{\hat{\bar{v}}\mathrm{ar}(\hat{\bar{y}})} \tag{10}$$

If a $\hat{\beta}_j$ is to be averaged across models where it appears, the number of models (R') and the recalculated model weights (w'_i) must be used in expressions

equivalent to Equations 9 and 10, with $\hat{\bar{\beta}}$ replacing $\hat{\bar{y}}$, and $\hat{\beta}_j$ replacing $\hat{y}_i$.

Extended Example

None of the models considered for the example problem is clearly "the best" as indicated by the AICc weight of 0.26 for the model with the highest rank. The evidence ratio for the best and second best models indicates that the best model is only 1.2 times more likely than the second model, given the evidence, indicating a lack of strong support for the best model. A model needs to have a weight greater than ~0.95 before considering it as the best model and bypassing the multimodel averaging process.

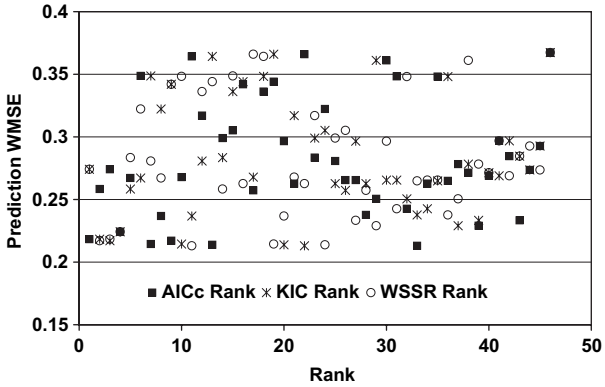

Figure 4. Relationship between AICc, WSSR, and KIC ranks and predictive WMSE.

Predictive Quality of the Best Model and the Multimodel Average

Desirable predictive models are those with small weighted mean square error (WMSE) between their predictions and those of the generating model. The prediction locations are 200 m upgradient (left) of the calibration locations, which are shown in Figure 2. Weighting by the inverse of measurement variance, using the same weights for heads and flows as used in the calibration to account for differences in magnitude and units of measurements, predictive WMSE is calculated. WMSE is the sum of the mean weighted squared differences between the 20 heads and 2 flows predicted by the alternative models and the true heads and flows simulated by the generating model with the additional pumping. The correlation between AICc and KIC model ranks and that WMSE for the 46 retained models illustrates the best fit to calibration data does not assure the most accurate predictions at all locations in the model (Figure 4). This is also illustrated by the relationship of the WSSR for the calibration and the WMSE for predictions (Figure 4). It has been noted that ground water models with the best fit to calibration data will not necessarily produce the most accurate predictions (Yeh and Yoon 1981; Rushton et al. 1982).

Rigorous experimental comparison of the alternative model ranking criteria requires evaluation of many different systems and numerous realizations of observation sets that is beyond the scope of this paper. Such an exercise would only reveal empirical value of the alternative methods because their theoretical underpinning is not well founded, as we know it is impossible to include the true model of a ground water system in the set of models.

Predictions at most locations are fairly accurate and readily captured by the linear confidence intervals of most of the alternative models. Model-averaged head predictions and their Scheffe confidence intervals are presented in Figure 5 for each individual model at locations , and 10 (located 200 m upgradient of the calibration points with the same ID in Figure 2). At location 7, nearly all models underestimate head, and confidence intervals of the top four models do not capture the truth. Model averaging (Equations 9 and 10) increases the confidence intervals and captures the truth (i.e., the value predicted by the generating model) (Figure 5a). Although large numbers of simulations would be needed to make a rigorous statistical evaluation, the practical similarity of the approaches is illustrated by noting the following: of the 22

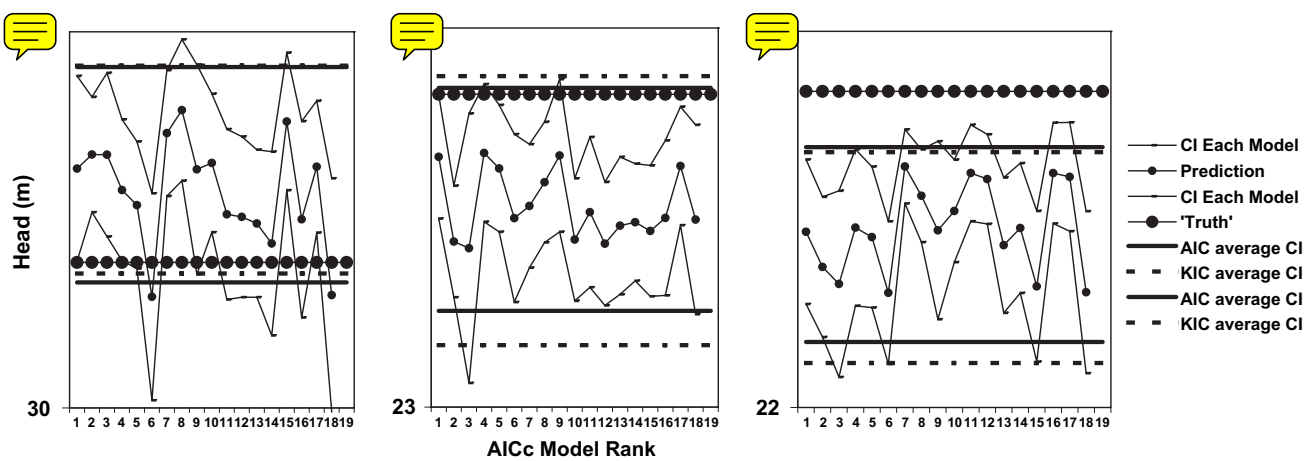

Figure 5. Predictions for location numbers , , and 8 (see Figure 2 for locations) are examples of locations where the models tend to (a) underestimate, (b) overestimate, and (c) inaccurately estimate the head, respectively, under the new pumping conditions. The predicted value and linear individual confidence intervals (CI) are shown for each model, as are model-averaged confidence intervals based on AICc and KIC.

predictions made for this example, 21 were captured by the model-averaged intervals. The predicted head at location 1 (Figure 5b) tends to be overestimated, and some of the best ranked models barely capture the truth in the confidence intervals based on individual model variance, but model averaging clearly captures the truth. Head at location 8 (Figure 5c) is not predicted successfully by any of the models and this is so consistent that it is not captured by model averaging. Although large numbers of simulations would be needed to make a rigorous statistical evaluations, the practical similarity of the approaches is illustrated by noting the following: of the 22 predictions made for this example, 12 were captured by the model-averaged intervals.

In this example, predictions generated by the alternative models vary considerably in some locations and not in others. This is illustrated in Figure 6 where the difference between the high and low head of all 46 models is displayed as a function of location. It is also indicated by noting that model averaging increases the confidence intervals indicated by the best model on 1 of the 22 predictions by less than 25% and another by 166%, with an average increase of 72% and a median of 64%. This variability serves to increase model-averaged variance through the second term of Equation 9, which is carried forward to confidence intervals in Equation 10. Field applications are likely to exhibit more striking variation in models including differences in geometry and boundary conditions, hence more significant shift of prediction and broadening of confidence intervals as a result of model averaging.

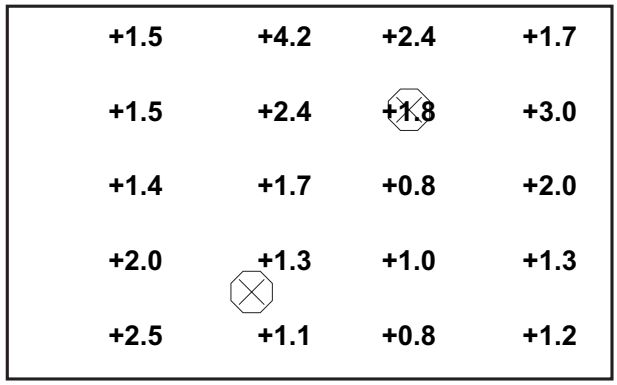

Figure 6. Total head drop across the model is 35 m, while the difference between the highest and lowest predicted head of the 46 models at a given location ranges from 0.8 to 4.2 m.

Summary and Conclusions

Given our uncertainty of site conditions, hydrologists should routinely consider several, well-thought-out models to maintain an open mind about the system. Generally, inferences should stem from multiple plausible models (multimodel inference) because it yields more robust predictions and a more "honest," realistic measure of precision. Modelers should be keenly aware of the fact that even multimodel inference, which provides greater consideration for uncertainty, is vulnerable to yielding poor predictions if fundamentally important processes are not included in the model, predictive locations and/or conditions differ substantially from those of the calibration, or the prediction horizon is large relative to the calibration time frame as discussed by Bredehoeft (2003).

Multimodel ranking and inference approaches based on K-L information, such as the AICc measure presented here, are simple to compute, easy to interpret, and provide a rigorous foundation for model-based inference. Approaches based on K-L information view models as approximations of the truth, and assume (1) a true model does not exist and cannot be expected to be in the set of models and (2) as the number of observations increases, one can uncover more details of the system; thus, AICc will select more complex models when more observations are available. Alternative model selection criteria (e.g., BIC, HQ, and KIC) seek to identify the true (or quasi-true) model with consistent complexity as the number of observations goes to infinity. These alternatives are based on the assumption that reality can be nearly expressed as a model and that this quasi-true model is in the set. Although these measures may perform similarly in application, it is unreasonable to assume that one would ever include the true or quasi-true model in the set of alternative ground water models; thus, approaches based on K-L information such as AICc are the preferable model ranking and inference criterion.

Acknowledgments

The final manuscript was improved based on suggestions of Neil Blanford, Steffen Mehl, and Jim Yeh. Their reviews are greatly appreciated. U.S. EPA and USGS provided funding for this work, but responsibility for its content lies with the authors.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* 19, 716–723.

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, ed. B.N. Petrov, 267–281. Budapest, Hungary: Akademiai Kiado.

Anderson, D.R. 2003. Multi-model inference based on Kullback-Leibler information. In *Proceedings of MODFLOW and More 2003: Understanding Through Modeling, IGWMC*, 366–370.

Bredehoeft, J.D. 2003. From models to performance assessment: The conceptual problem. *Ground Water* 41, no. 5: 571–577.

Burnham, K.P., and D.R. Anderson. 2004. Multi-model inference: Understanding AIC and BIC model selection. *Sociological Methods and Research* 33, no. 2: 261–304.

Burnham, K.P., and D.R. Anderson. 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Carrera, J., and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research* 22, no. 2: 199–210.

Chamberlin, T.C. 1890. The method of multiple working hypotheses. *Science* 15: 93–98. Reprinted 1965, *Science* 148: 754–759.

deLeeuw, J. 1992. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In *Breakthroughs in Statistics*, vol. 1, ed. S. Kotz and N.L. Johnson, 599–609. London, U.K.: Springer-Verlag.

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

Hannan, E.J., and B.G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series* B 41, no. 1: 190–195.

Harbaugh, A.W., E.R. Banta, M.C. Hill, and M.G. McDonald. 2000. MODFLOW-2000 the U.S.Geological Survey modular ground water model—User guide to modularization concepts and the ground water flow process. USGS Open-File Report 00–92. Reston, Virginia: USGS.

Hill, M.C., E.R. Banta, A.W. Harbaugh, and E.R. Anderman. 2000. MODFLOW-2000, the U.S. Geological Survey modular ground water model—User guide to the observation, sensitivity, and parameter-estimation processes and three post-processing programs. USGS Open-File Report 00–184. Reston, Virginia: USGS.

Hurvich, C.M., and C.-L. Tsai. 1994. Autoregressive model selection in small samples using a bias-corrected version of AIC. In *Engineering and Scientific Applications*, vol. 3, ed. Bozdogan, H. 137–157. Proceedings of the First U.S./Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Dordrecht, Netherlands: Kluwer Academic Publishers.

Hurvich, C.M., and C-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76, no. 2: 297–307.

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, no. 2: 99–104.

Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.

Linhart, H., and W. Zucchini. 1986. *Model Selection*. New York: John Wiley and Sons.

McQuarrie, A.D.R., and C.-L. Tsai. 1998. *Regression and Time Series Model Selection*. Singapore: World Scientific Publishing Company.

Neuman, S.P. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* 17, no. 5: 291–305.

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

Parzen, E., K. Tanabe, and G. Kitagawa, eds. 1998. *Selected Papers of Hirotugu Akaike*. New York: Springer-Verlag.

Poeter, E., M. Hill, J. Doherty, J.E. Banta, and J. Babendreier. 2003. JUPITER Project—Joint Universal Parameter IdenTification and Evaluation of Reliability, Fall AGU Meeting. Abstract H12G-03. *Eos* 84, no. 46: F608–F609.

Poeter, E.P., and M.C. Hill. 1997. Inverse methods: A necessary next step in ground water modeling. *Ground Water* 35, no. 2: 250–260.

Poeter, E.P., and S.A. McKenna. 1995. Reducing uncertainty associated with groundwater flow and transport predictions. *Ground Water* 33, no. 6: 899–904.

Rushton, K.R., E.J. Smith, and L.M. Thomlinson. 1982. An improved understanding of flow in a limestone aquifer using field evidence and a mathematical model. *Journal of the Institute for Water Engineers and Scientists*. 36, no. 5: 369–383.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. *Akaike Information Criterion Statistics*. Tokyo, Japan: KTK Scientific Publishers.

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

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*, Series B 39, no. 1: 44–47.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* A7: 13–26.

Ye, M., S.P. Neuman, and P.D. Meyer. 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research* 40, no. 5: W05113, 1–19.

Yeh, W.W.-G., and Y.S. Yoon. 1981. Aquifer parameter identification with optimum dimension in parameterization. *Water Resources Research* 17, no. 3: 664–672.

A Modular Finite-Element Model (MODFE) for Areal and Axisymmetric Ground-Water-Flow Problems, Part 1: Model Description and User's Manual

U.S. Geological Survey, Techniques of Water-Resources Investigations, Book 6, Chapter A3
by Lynn J. Torak

This report is available as a pdf below

Preface

The series of manuals on techniques describes procedures for planning and executing specialized work in water-resources investigations. The material is grouped under major subject headings called "Books" and further subdivided into sections and chapters. Section A of Book 6 is on ground-water modeling. The unit of publication, the chapter, is limited to a narrow field of subject matters. This format allows flexibility in revision and publication as the need arises.

Chapters 6A3, 6A4, and 6A5 are on the use of a particular transient finite-element numerical method for two-dimensional, ground-water-flow problems. These Chapters (6A3, 6A4, and 6A5) correspond to reports prepared on the finite-element model given the acronym MODFE and designated as parts 1, 2, and 3, respectively. Part 1 is on "model description and user's manual," part 2 is on "derivation of finite-element equations and comparisons with analytical solutions," and part 3 is on "design philosophy and programming details."

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Table of Contents

PDF Files

Attachment 4

Data files are available upon request.
Please Contact: Paula Cutillo
National Park Service
Water Resources Division
1201 Oakridge Drive
Fort Collins, CO 80525
Phone: 970-225-3537
Email: Paula_Cutillo@nps.gov