A Controlled Experiment in Ground Water Flow Model Calibration

by Mary C. Hill[a], Richard L. Cooley[a], and David W. Pollock[b]

Abstract

Nonlinear regression was introduced to ground water modeling in the 1970s, but has been used very little to calibrate numerical models of complicated ground water systems. Apparently, nonlinear regression is thought by many to be incapable of addressing such complex problems. With what we believe to be the most complicated synthetic test case used for such a study, this work investigates using nonlinear regression in ground water model calibration. Results of the study fall into two categories. First, the study demonstrates how systematic use of a well designed nonlinear regression method can indicate the importance of different types of data and can lead to successive improvement of models and their parameterizations. Our method differs from previous methods presented in the ground water literature in that (1) weighting is more closely related to expected data errors than is usually the case; (2) defined diagnostic statistics allow for more effective evaluation of the available data, the model, and their interaction; and (3) prior information is used more cautiously. Second, our results challenge some commonly held beliefs about model calibration. For the test case considered, we show that (1) field measured values of hydraulic conductivity are not as directly applicable to models as their use in some geostatistical methods imply; (2) a unique model does not necessarily need to be identified to obtain accurate predictions; and (3) in the absence of obvious model bias, model error was normally distributed. The complexity of the test case involved implies that the methods used and conclusions drawn are likely to be powerful in practice.

Introduction

Regression has been a powerful tool for using data to test hypothesized physical relations and to calibrate models in many fields (Draper and Smith 1981; Seber and Wild 1989). Despite its introduction into the ground water literature in the 1970s (reviewed by McLaughlin and Townley 1996), regression has been used very little with numerical models of complicated ground water systems. The sparsity of data, nonlinearity of the regression, and complexity of the physical systems produce substantial difficulties. Obtaining tractable models that are sufficiently representative of the true system to yield useful results is arguably the most important problem in the field. The only options are improving the data, ignoring the nonlinearity, and(or) carefully ignoring some of the system complexity. Sparsity of data is a perpetual problem not likely to be alleviated at most field sites despite recent impressive advances in geophysical data collection and analysis (e.g., Hyndman and Gorelick 1996; Eppstein and Dougherty 1996). Methods that ignore nonlinearity are presented by, for example, Hoeksema and Kitanidis (1984) and Sun (1994, p. 182). The large changes in parameter values that occur in most nonlinear regressions of ground water problems after the first iteration, however, indicate that linearized methods are unlikely to produce satisfactory results in many circumstances. Thus, simplification related to parameterization appears to be the only potentially useful option, and is the mechanism considered in this work.

Defining a tractable but useful level of parameterization for ground water inverse problems has been an intensely sought goal, focused mostly on the representation of hydraulic conductivity or transmissivity. Suggested approaches vary considerably in complexity. The most complex parameterizations are cell- or pixel-based methods in which hydraulic conductivity or transmissivity varies from one finite-difference cell or other basic model entity to another, using prior information or regularization to stabilize the solution (for example, McLaughlin and Townley 1996; Clifton and Neuman 1982; Tikhonov and Arsenin 1977). Prior information and regularization produce similar penalty function terms in the objective function, but prior information needs to satisfy either classical or Bayesian assumptions, while regularization does not. Grid-scale parameterizations minimize user-imposed simplifications, but have the following problems: (1) heterogeneities smaller than the grid scale often are important, so use of grid-scale parameterization generally does not eliminate the scale problem; (2) more hydraulic-conductivity or transmissivity data than are available in most circumstances or often unrealistic assumptions about smoothness generally are needed; and (3) as presently developed, it is not straightforward to include knowledge about geologic structure into

[a]U.S. Geological Survey, P.O. Box 26034, MS 413, Lakewood, CO 80225. E-mail: mchill@usgs.gov (first author)

[b]U.S. Geological Survey, 12201 Sunrise Valley Dr., MS 411, Reston, VA 22092.

Received March 1997, accepted October 1997.

grid-scale methods to produce, for example, constraints that can reduce the need for so many measurements of hydraulic conductivity or transmissivity.

Simpler parameterizations include zonation, interpolation, or eigenvectors of the variance-covariance matrix of grid-scale parameters (for example, Jacobson 1985; Sun and Yeh 1985; Cooley et al. 1986; RamaRao et al. 1995; Reid 1996; D'Agnese et al. 1998, in press). Stochastic methods (for example, Gelhar 1993; Yeh et al. 1995; Kitanidis 1995) also generally fall into this category, although they share some of the characteristics of the grid-scale methods. These simpler parameterizations produce a more tractable problem, but it is not clear at what point the simplicity diminishes utility. The principle of parsimony (Box and Jenkins 1976; Parker 1994) suggests that simple models should be considered, but the perception remains that complex systems cannot be adequately represented using parsimonious models. For example, Gelhar (1993, p. 341) claims that "there is no clear evidence that [nonlinear regression] methods [using simple parameterizations] actually work under field conditions." Indeed, Beven and Binley (1992) even suggest that for some problems it may be best to abandon the concept of parameterizations simple enough to produce an optimal set of parameter values. A concept as useful as parsimony should not be given up lightly, yet there has been no conclusive evaluation of how complex parameterizations need to be to produce useful results.

This study originally had two purposes: (1) to present an approach that makes nonlinear regression methods more useful for the types of problems typical in ground water; and (2) to use a synthetic test case to evaluate the method and some general issues of model calibration. Because several articles describing and applying the approach have recently been published or are in review (Anderman et al. 1996; Barlebo et al. 1996; Poeter and Hill 1996, 1997; Hill 1998; D'Agnese et al. 1998, in press; Barlebo et al. in press), this paper will focus less on presentation of the approach and more on its evaluation using the synthetic test case. Issues of concern are whether the approach can be used as a scientific hypothesis-testing and data analysis tool that is likely to yield substantial insight into, and accurate models of, complex ground water systems, and whether problems simple enough to produce a well-posed nonlinear regression are useful in terms of model calibration and accurate predictions.

Methods and Previous Works

The synthetic test case is set in the framework of numerical ground water flow model calibration and prediction. In this work, model calibration is divided into model construction and parameter evaluation and estimation (as in, for example, Gupta and Sorooshian 1985; Sun 1994). Model construction includes: (1) Choosing a physical equation and numerical methods and developing or choosing a computer program; (2) defining system discretization, representation of boundary conditions, and so on; and (3) selecting what aspects of the physical system to represent with parameters. Using this terminology, the same system characteristics may be classified differently in different applications. For example, when using zonation, defining the zones generally is classified under model construction as defined above. If the locations of the zone boundaries are represented by parameters, these would become part of the parameter evaluation and estimation (for example, Hyndman and Gorelick 1996). Given a model structure, model parameters can be evaluated for their importance under calibration and prediction conditions, and are estimated to achieve a model that in some way reproduces measured values. The resulting match is used to reevaluate model structure. Commonly, calibration is achieved using only trial and error. Problems with using trial and error alone have been discussed by many authors, including Carrera and Neuman (1986), Cooley and Naff (1990) and Hill (1992, p. 3).

Model calibration can be addressed more effectively by replacing trial and error with inverse modeling as much as possible, where inverse modeling refers to using formal optimization and stochastic methods to evaluate and estimate parameter values. In this work, inverse modeling is accomplished using nonlinear regression (Bard 1974; Cooley 1977, 1979, 1982; Sun 1994); associated advantages are discussed by, for example, Poeter and Hill (1997) and Hill (1998).

In this work, nonlinear regression is accomplished using the inverse ground water flow model MODFLOWP (Hill 1992). This model allows a wide variety of system characteristics to be calculated with defined parameters and allows for general definition of parameters so that spatially distributed quantities such as hydraulic conductivity and areal recharge can be defined using zones of constant value, interpolation methods, and some stochastic methods.

To conduct a definitive study of regression methods, it is necessary to use a complex synthetic test case in which all aspects are known. This is only possible if the synthetic test case is a numerical model. The true system in this work is a steady-state, three-dimensional numerical ground water flow model with five layers and five times smaller grid spacing than the calibrated model. The test case is characterized by aquifer heterogeneity, a confining unit, areal recharge, and ground water interaction with a lake and a stream. The turning bands stochastic method (Mantoglou and Wilson 1982), as implemented by Wilson (1989), was used to produce hydraulic-conductivity and areal-recharge distributions, and the areal extent of a confining unit. Some aspects of this test case were used by Eppstein and Dougherty (1996). Calibration is accomplished using nonlinear regression to estimate parameter values that represent aquifer and confining unit hydraulic conductivities, lakebed and streambed conductances, and areal recharge. This work is distinguished by the wide range of parameter types estimated in the regression; most studies only estimate parameters related to the hydraulic-conductivity distribution (for example, RamaRao et al. 1995). Model calibration was conducted by two of the authors of this report who knew only the information presented in the section "The Synthetic Test Case," except that they did not know the true head distribution.

As implemented, this is believed to be the most complicated test case that has been used in the evaluation of ground water inverse modeling. Other complex test cases include the following: Chu et al. (1987) estimated transmissivity and dispersivity of a two-dimensional synthetic test case. Gomez-Hernandez and Gorelick (1989) used a two-dimensional synthetic test case to investigate effective hydraulic-conductivity values, but did not investigate many of the model calibration issues studied in the present work. Poeter and McKenna (1995) present an innovative method of evaluating the hydraulic-conductivity distribution in detail using a three-dimensional test case, but do not consider other aspects of model construction or data availability.

In this report, the nonlinear regression method is presented briefly, data on the true system available for model calibration is presented, model construction and calibration using nonlinear regression are described, and the calibrated models are compared with the

true system characteristics. Predictions from the true and calibrated models are presented and compared with management criteria. Finally, the results are evaluated to determine the strengths and weaknesses of the regression and parameterization methods used.

Nonlinear Regression

This section briefly describes the regression methods used. Aspects of the approach are discussed further by Hill (1992), Anderman et al. (1996), Poeter and Hill (1997), D'Agnese et al. (1998, in press), and Barlebo et al. (in press); the most complete description is by Hill (1998). Nonlinear regression was used to find parameter values that minimize the weighted sum of squares objective function, S ($\underline{b}$), calculated as (Seber and Wild 1989, p. 27):

$$S(\underline{b}) = (\underline{y} - \hat{\underline{y}})^T \underline{\omega} (\underline{y} - \hat{\underline{y}}) \quad (1)$$

where

$\underline{b}$ = an np × 1 vector containing values of the estimated parameters
np = the number of estimated parameters
$\underline{y}$ = an n × 1 vector of observed hydraulic heads, flows, and prior information
n = the number of observations of hydraulic head, flows, and items of prior information used in the regression
$\hat{\underline{y}}$ = an n × 1 vector of simulated (using $\underline{b}$) hydraulic heads, flows, and prior information
$(\underline{y} - \hat{\underline{y}})$ = an n × 1 vector of residuals (observed minus simulated values)
$\underline{\omega}$ = an n × n weight matrix.

Weighted residuals are important indicators of model fit, and are calculated as $\underline{\omega}^2 (\underline{y} - \hat{\underline{y}})$. The objective function is minimized with respect to the parameter values using a modified Gauss-Newton method.

It is desirable to estimate parameters with the smallest possible variance to achieve estimated values that are most likely to be close to the true values. To do this using Equation 1, linear theory indicates that three conditions need to be satisfied (Bard 1974; Tarantola 1987):

1. The model needs to be correct.
2. The weight matrix needs to be proportional to the inverse of the variance-covariance matrix of the measurement errors of the observed hydraulic heads, flows, and prior parameter information.
3. The measurement errors need to be random.

In addition, if Equation 1 is derived by classical Gauss-Markov arguments, the errors need not be normally distributed (Helsel and Hirsch 1992); if it is derived using maximum-likelihood arguments, normality is needed (Carrera and Neuman 1986).

Two aspects of nonlinear regression as implemented in the present work are discussed in more detail — weighting and diagnostic statistics.

Weighting

To assign the weighting needed in Equation 1, it was assumed that measurement errors were uncorrelated, producing a diagonal weight matrix with nonzero elements proportional to one divided by the variance of the measurement errors. The actual lack of measurement error in the synthetically produced observation was unknown by the modelers, and weights were assigned based on expected measurement error (Cooley et al. 1986; Hill 1992, p. 48), as represented by standard deviations and coefficients of variation used to calculate the variances. Because the guidance provided by condition (2) allows room for adjustment, the weights are said to be subjectively determined. In practice, the determination of weights is always somewhat subjective except when they are automatically updated as part of the regression (Huber 1981; Barlebo et al. in press). Although useful for problems with large data sets (as in Neele et al. 1993), automatic updating can obscure the use of model fit in discovering erroneous data and model error when data sets are sparse, as is typical in ground water problems. In the approach presented in this work, weighting is not automatically adjusted. It is sometimes adjusted based on regression results after careful consideration.

The weighting procedure used in this work is a variation of common methods described by Theil (1963), Carrera and Neuman (1986), Cooley and Naff (1990) and Hill (1992), but eliminates use of the common error variance. Here, the weight matrix is assumed to equal (instead of being proportional to) the inverse of the measurement-error variance-covariance matrix and the flexibility previously assigned to the common error variance is now used to allow the calculated error variance to differ from its expected value of 1.0. The change reduces confusion for problems with more than one kind of observation — so-called coupled (Sun and Yeh 1990; Sun 1994), multiresponse (Seber and Wild 1989), or joint (Neele et al. 1993) problems. The convenience of the method results from (1) added clarity about the meaning of the weights and, therefore, the ability to compare any weighting used against expected values, and (2) the ability to use the standard error of the regression to infer possible dominance of measurement error versus model error.

Conditions 1 and 3 above are satisfied only if the weighted residuals from all types of observations and prior information appear to be statistically consistent with each other or if any statistical inconsistency can be explained by the correlation of the weighted residuals expected through the regression (Cooley and Naff 1990, p. 167-172; Hill 1992, p. 66-69). With this requirement, the method described here is consistent with methods using the common error variance.

With the weight matrix defined as being equal to the inverse of the variance-covariance matrix of the measurement errors, the objective function (Equation 1) is dimensionless. The regression standard error is commonly used to evaluate model fit, and is calculated as:

$$s = \left(\frac{S(\underline{b})}{n - np} \right)^{1/2} \quad (2)$$

and also is dimensionless. Using s directly as a measure of model fit is sometimes unsatisfactory because it cannot be used to compare models with different weighting and because it has little intuitive appeal. To obtain values that more effectively reflect model fit in this work, s is multiplied by the standard deviations or coefficients of variation used to calculate the weights for the head observations. The resulting statistic is defined here as the fitted standard deviation or the fitted coefficient of variation.

Diagnostic Statistics

During calibration, many statistics can be used to diagnose problems with the calibration and the regression in addition to the

standard error and fitted statistics described above. The statistics described here have proved to be extremely effective.

Composite scaled sensitivities and parameter correlation coefficients are used to measure the information available from the data to estimate parameters or, equivalently, to determine which parameters could likely be estimated uniquely with the available data. Composite scaled sensitivities are dimensionless quantities calculated as:

$$css_i = \left[(1/n)\sum_{j=1}^{n}\omega_{jj}\left(b_i\frac{\partial y_j}{\partial b_i}\right)^2\right]^{1/2} \quad (3)$$

These quantities were derived from similar statistics used by Cooley et al. (1986), and were first presented by Hill (1992). They resemble the CTB statistics derived independently by Sun and Yeh (1990). Each composite scaled sensitivity is the square root of a diagonal element of the Fisher information matrix (Roa 1973, p. 33; Carrera and Neuman 1986) scaled by the parameter value and n. The authors' experience indicates that if composite scaled sensitivities range over more than about two orders of magnitude, the regression is commonly unstable.

Correlation coefficients are calculated from the elements of the variance-covariance matrix on the parameters, which are calculated as:

$$v_{ij} = s^2\left[\left(\frac{\partial \underline{y}}{\partial \underline{b}}\,\omega\,\frac{\partial \underline{y}}{\partial \underline{b}}\right)^{-1}\right]_{ij} \quad (4)$$

The correlation between parameters i and j is calculated as $v_{ij}(v_{ii}v_{jj})^{-1/2}$. In general, correlations exceeding 0.95 indicate that all the parameter values may not be estimated uniquely.

One of the most obvious features of using nonlinear regression is that, for the well-posed problems developed in this work, optimal parameter values are obtained. These values and their individual 95% linear confidence intervals (Seber and Wild 1989; Hill 1994, p. 26-38) are used diagnostically in two ways. First, they are used to indicate whether unrealistic optimal parameter estimates suggest the presence of model error or not. If the model is correct and sufficient data are used and are being simulated correctly, optimized parameter values are expected to be reasonable. Unreasonable optimal parameter values often indicate problems with the model, the data, or the way the data are being related to the model. Linear confidence intervals on unrealistic optimized parameter values that include or nearly include realistic values suggest that the data are insufficient for conclusive evaluation, and the problem is less likely to be model error.

Confidence intervals on optimal parameter values are also used to indicate possible model simplifications. If the confidence intervals of two optimized recharge parameters, for example, largely overlap, it is likely that they could be represented as one recharge parameter.

The Synthetic Test Case

Synthetic Valley is an undeveloped alluvial valley surrounded by low permeability bedrock (Figure 1a). Surface water features are Blue Lake and the Straight River.

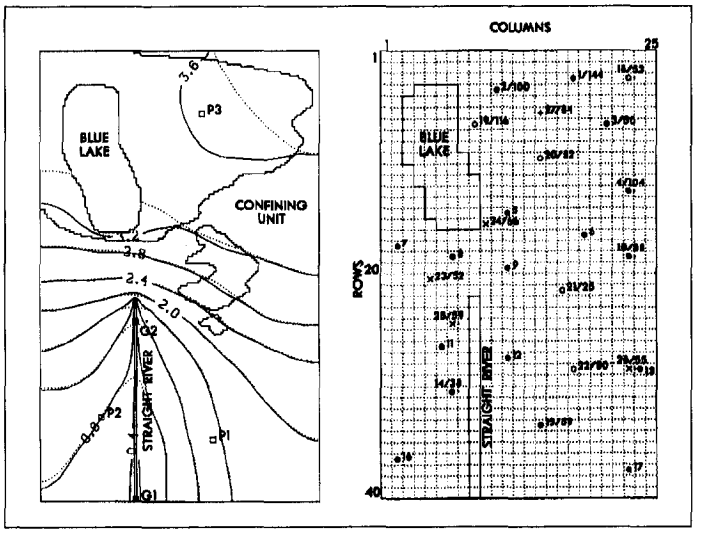

Figure 1a. (left) The 6096-m by 3810-m areal extent of the true and calibrated models used to represent Synthetic Valley, the true areal extent of Blue Lake and the confining unit, the location of stream gauges G1 and G2, the proposed locations of production wells P1, P2, and P3 (Development scenario A includes pumpage at P1 and P3; scenario B at P2 and P3), the true hydraulic heads for the water table (solid contours) and the basal part of the ground water system (dotted contours).

Figure 1b. (right) 152.4-m finite-difference grid spacing used in the calibrated models, finite-difference cells used to simulate the Straight River and Blue Lake, well numbers and locations, and, following the well numbers, hydraulic conductivities (in meters per day) measured using slug tests and, for well 27, an aquifer test. Figure 1b shows the location of wells available from the beginning of the study (●), wells drilled in field season 1 (o), wells drilled in field season 2 (x), and the well drilled in field season 3 (+).

Development Scenarios and Management Criteria

The design of the model and data collection efforts depend on the proposed development, so it is described first. Proposed development includes: (1) a 7600 m^3/d (cubic meters per day) well in the southern part of the valley at either well location P1 or P2 (Figure 1a), and (2) a 1900 m^3/d well at location P3 (Figure 1a). Development scenario A includes the pumpage at P1 and P3; development scenario B includes the pumpage at P2 and P3. Management criteria are: (1) Drawdown of the water table cannot exceed 0.6 m anywhere in the northern half of the valley; and (2) streamflow at gauge site G2 (Figure 1a) cannot be decreased by more than 20%. Because the development is characterized by constant rates of pumpage and the management criteria involve long-term effects of pumpage, predictive simulations are steady state.

Model Discretization and Boundary Conditions

The finite-difference cells used to represent Blue Lake and the Straight River in the true model are outlined in Figure 1a. The Straight River is a head-dependent boundary; Blue Lake is represented as a volume of very high hydraulic conductivity. The top of the model is simulated as a water table free-surface boundary subjected to areal recharge, which is represented as a defined flux. The four sides and the bottom of the model are no-flow boundaries. Five model layers were used in the true model.

Data

All data used for model construction and calibration are either calculated using the true system or derived from true system characteristics; they were not corrupted by added random noise. Thus, the data are more accurate than the data available for most model calibrations. This lack of what is commonly called measurement error allows for a definitive evaluation of model error, which plagues all model calibrations, but which generally cannot be characterized.

Data were collected at 17 existing wells and 10 additional wells drilled in subsequent field seasons (Figure 1b), and included drillers' logs, water table levels, and water levels at the screened intervals for all wells, and, for three wells, depth to bedrock. The valley sediments are mostly medium to coarse sands, some gravels, and a thin layer of lake clay in the north (confining unit of Figure 1a). Data from similar valleys indicate that horizontal hydraulic conductivity of the non-clay deposits ranges between 3 and 150 m/d (meters per day), which spans one and a half orders of magnitude. Remarkable vertical homogeneity displayed in the well logs was used in model calibration to justify using aquifer horizontal hydraulic conductivities that do not vary with depth, and are the same for all model layers representing aquifer material. In this way, the test case is simpler than most field sites, in which vertical inhomogeneity dominates, but the overall complexity of the test case is substantial despite this simplification.

Contour maps constructed using measured water table levels and hydraulic heads at the screened intervals are not shown in this report, but are similar to the true hydraulic-head maps shown in Figure 1a except that contours in the northeastern corner are smoother than the true contours. The true hydraulic-head maps were not seen by the authors calibrating the model until the end of the study. For all wells, water table levels are higher than the water levels at the screens. Where the confining unit occurs, the hydraulic head declines with depth by 0.031 to 0.183 m; where no confining unit occurs, it is as large as 0.03 m, but is less than 0.003 m for nine of the 13 wells.

Mean annual precipitation is 91 cm/y (centimeters per year), and there is no surface water flow into the valley. At the southern boundary, the stage in the Straight River is zero, the datum for all other elevations. The river gradient is 0.0002, so the stage at the head waters (2743 m upstream) is 0.55 m. Generally, the river ranges from 7.6 to 22.9 m wide and is 0.3 m or less deep. Gauged streamflows at G1and G2 of Figure 1a are 25,046 and 2730 m^3/d, respectively.

Blue Lake is a 5 m deep, sandy-bottom lake with no surface water inflow or outflow. A previous study of Blue Lake yielded the following: stage = 3.35 m; area of the lake = 1.510×10^6 m^2; lake evaporation = 69 cm/y. Because there are no surface water flows to or from Blue Lake, the difference between precipitation (91 cm/y) and evaporation from the lake (69 cm/y) recharges 910 m^3/d to the ground water system.

The ground water budget for Synthetic Valley can be expressed as: P – ET + 910 m^3/d = 25,046 m^3/d, where P is precipitation, ET is evapotranspiration from the land surface, 910 m^3/d is the net flow from the lake, and the right-hand side equals the measured flow at G1. Thus, P – ET = 24,136 m^3/d, which is equivalent to 0.00111 m/d, or 40.6 cm/y over Synthetic Valley minus the area of Blue Lake.

The long-term effect of local ground water pumpage on Blue Lake will be a decline in lake level. It will not be an increase in the net flow from the lake to the ground water system because Blue Lake is a closed lake with no surface water inflow or outflow, and precipitation on and evaporation from the lake surface generally are not affected by the lake-level changes.

Model Calibration

Calibration progressed through the development of seven models named CAL0 (the initial model), CAL0-G1 (tests the importance of the G1 flow measurement), CAL0+PR (tests using prior information to make one of the CAL0 estimated parameter values more reasonable), CAL1 (constructed after field season 1), CAL2 (constructed after field season 2), CAL3 (constructed after field season 3), and NO LAKE (tests the importance of representing the lake). The data available for the CAL0 models consisted of the data for wells 1 through 17, as described above, streamflows at G1 and G2, of which G2 and G1 minus G2 are used in the regression, and the net loss from the lake. Additional data were collected during the three field seasons.

Model parameters were defined to calculate the model characteristics listed in Table 1. All defined parameters were estimated by nonlinear regression, except as noted in Table 1.

Model Discretization and Boundary Conditions

The areal finite-difference grid and the cells used to represent Blue Lake and the Straight River in the calibrated models are shown in Figure 1b; the lake and river were head-dependent boundaries. To avoid adding nonlinearity that would promote longer, more unstable regression runs, during calibration the top layer of all calibrated models was represented as confined instead of unconfined. Resulting inaccuracies were found to be negligible for these steady-state models. The sides and bottom of the models were no-flow boundaries. Vertical discretization is shown in Figure 2. In the north, the bottom of layer 1 coincides with the bottom of the lake; the bottom of layer 2 coincides with the top of the clay confining unit (which is simulated as vertical leakance between layers 2 and 3). In the south, the bottom of layer 1 is deep enough to ensure that no cells go dry; the bottom of layer 2 was placed about 15 m lower for all models except CAL0, in which layer 3 is absent and layer 2 extends to the bottom of the model. The bottom of the model was derived from bedrock elevations measured at wells 1, 5, and 27 and some geophysical data.

Calibration Results

Some of the questions posed and answered at major steps of the calibration are presented in Table 2 to display the hypothesis-testing framework and to demonstrate how nonlinear regression and the diagnostic statistics were used. There were no problems with uniqueness in any of the regressions — optimal parameter values were readily identified for each model and parameter correlation was low; the largest correlation of 0.95 was calculated for CAL0-G1, and even then starting the regression at several sets of initial values indicated that the optimal parameter values were unique. Using the diagnostic statistics to evaluate the importance of different parameters to the predictions, as suggested by Hill (1998), might have been useful, but was not considered in this study.

The hydraulic conductivity distribution of the CAL0 models is defined using the zones shown in Figure 3a, which produced the best overall results of the many zone configurations considered. Interpolations based on linear triangular finite elements are used to define the hydraulic conductivity distributions of the other models; Figure 3b shows the finite elements used for CAL0. For the interpolations, most of the estimated values were at nodes of the finite-

Table 1
Parameters of the Calibrated Models

[Parameter labels: KRB^0, the conductance of the streambed at the head waters of the Straight River; K_1^0, K_2^0, and K_3^0, zonal horizontal hydraulic-conductivity values (Figure 3a); KRB_1, KRB_2, and KRB_3, zonal streambed-conductance values (Figure 7); KLB, leakance (vertical hydraulic conductivity divided by thickness) of the lakebed; RCH, areal recharge rate; KV, leakance of the confining unit (Figure 4); ANIV, vertical anisotropy of aquifer material; w/prior, parameters had prior information.]

	Model Characteristic Calculated Using the Parameter					
Model	Streambed Conductance	Horizontal Hydraulic Conductivity	Lakebed Conductance	Areal Recharge	Confining Unit Leakance	Vertical Anisotropy
CAL0 CAL0-G1 CAL0+PR	KRB^0	K_1^0 K_2^0 K_3^0	KLB	RCH	[1]KV	[1]ANIV
CAL1	KRB_1 KRB_2 KRB_3	[2]12 w/prior	do.	do.	KV	do.
CAL2	do.	[2]16 w/prior [3]+3	do.	do.	do.	do.
CAL3	do.	do.	[4]none	do.	do.	do.
NO LAKE	do.	do.	[4]none	do.	do.	do.

[1]Not estimated by nonlinear regression because of insensitivity, as indicated by small composite scaled sensitivities.

[2]The parameter labels are K*, where * is a well number from Figure 1b. These parameters are used in the interpolation scheme shown in Figure 3b with prior information used in the regression equal to the slug-test value shown in Figure 1b; for example, for well 1, it is 144 m/d.

[3]The parameter labels are K*, where * identifies location A, B, or C of Figure 3b. These parameters have no prior information.

[4]For CAL3 and NO LAKE, KLB is calculated as the hydraulic conductivity of the underlying cell, divided by the product of ANIV and the vertical distance to the center of the underlying cell (6.9 m in all models). No separate KLB parameter was defined.

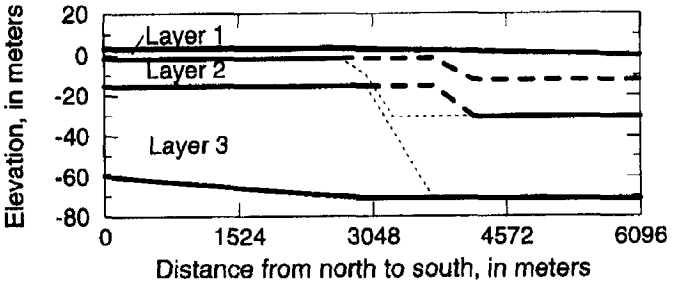

Figure 2. North-south cross-section showing the layers used in the calibrated models, including divisions used in all models (—), top is the water table; divisions used for all CAL0 models (···); and divisions used for all models except the CAL0 models (---). For any model, all north-south cross-sections are the same.

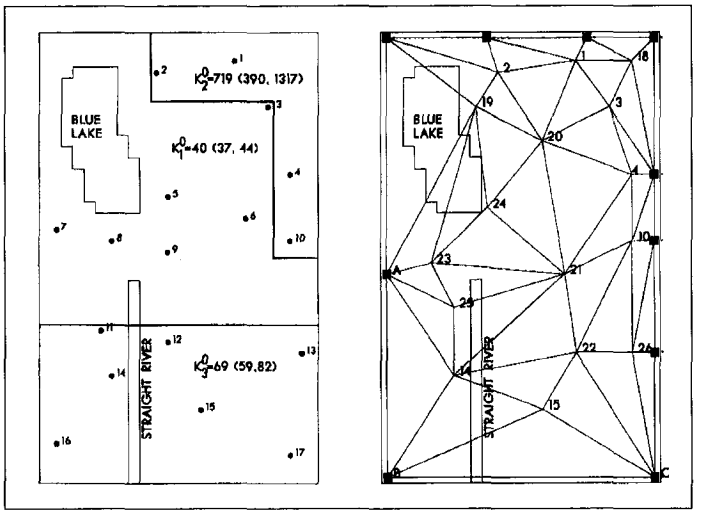

Figure 3a. (left) Zones used to parameterize the horizontal hydraulic-conductivity distribution for all model layers for the CAL0 models, with the wells used to construct the zones, the parameter labels of Table 1, and the estimated values and the 95% linear individual confidence intervals, in meters per day.

Figure 3b. (right) Triangular finite elements are used to parameterize the horizontal hydraulic-conductivity distribution for all model layers for the CAL3 model. The hydraulic conductivity is changed by the regression at the numbered apexes of the triangles (the numbers are well numbers) and at the squares labeled A, B, and C.

element grid located where slug tests had been conducted. The slug-test values are used as prior information for these parameters, with weights on the prior information just large enough (or, equivalently, coefficients of variation just small enough) to achieve convergence of the regression. The final coefficients of variation are about 20%, which is somewhat smaller than the 30% that would be consistent with the authors' prior beliefs about the accuracy of the prior information. Thus, this application needs to be regarded as a regularization procedure instead of Bayesian prior (Backus 1988).

The weights used for observations in the regression were calculated using the standard deviations and coefficients of variation of the measurement errors in hydraulic heads and flows, respectively, shown in Table 3, which were based on typical measurement error for these data types. The covariance between flows G2 and G1-G2 (Hill 1992, p. 43), was not included in the weighting, but its omission is not expected to significantly affect the results. The statistics used to calculate the weights were modified within reasonable limits during calibration to achieve statistically consistent weighted residuals. The standard errors, s, in Table 3 are all less than 1.0, indicating that the model fit is better than would be consistent with the assigned weighting. This produces the small fitted statistics of Table 3 and Figure 5, and is because, unbeknownst to the modelers, the data have no noise added to them.

Graphs of weighted residuals against weighted simulated values for CAL0 and CAL3 are shown in Figure 6. Graphs for CAL0-G1 and CAL0+PR were similar to the CAL0 graph; graphs for the other models were similar to the CAL3 graph. For CAL0 (Figure 6a), six out of the seven weighted residuals between weighted simulated values of 15 and 30 are positive, suggesting they may be nonrandom. For CAL3, there is no indication of model bias related to the head and flow data, but estimated parameter values tend to be slightly smaller, on average, than the measured slug-test values, as indicated by predominantly positive weighted residuals.

Observed streamflow gains and those simulated using the CAL0 and CAL3 models are shown in Figure 7. For the CAL0 simulations, there were streamflow gain observations at G2 and between G1 and G2; for the CAL3 simulations there were streamflow gain observations every 152.4 m. In general, streamflows are matched within 150 m^3/d. The 910 m^3/d lake budget, which is smaller than any of the per cell streamflow gains, was matched within 2% by all models.

Table 2
Short Description of Results from the Calibrated Models

Model Label; Number of Regression Data and Parameters	Questions Addressed	Answers Derived from Regression Results for the Listed Model * Conclusion contradicts the true model. Confidence intervals are 95%, linear intervals. Diagnostic statistics are *italicized.*
CAL0 34 heads 3 flows 6 parameters	1. What are the basic relations between the parameters and the head and flow data? 2. Can zones of constant hydraulic conductivity be used to produce a good model? 3. Does the areal recharge vary spatially? 4. Does streambed hydraulic conductivity vary spatially? 5. Does the confining unit extend all the way under the lake?	1. Head data alone: most ***correlation coefficients*** are 1.0, indicating that all parameters except ANIV are completely correlated. Adding lake budget data reduces correlation slightly. Adding the G2 and (G1-G2) flow data reduces correlation substantially, but ***composite scaled sensitivities*** suggest that these data are insufficient to estimate KV and ANIV. 2. No. The zones shown in Figure 3a produced a good fit to heads and flows (Figures 5 and 7), but K_2^0 (719 m/d) ***exceeds the expected maximum*** (152 m/d), and its ***confidence interval*** (390;1317) excludes reasonable values. This indicates that the model is significantly biased. 3. No. Estimated recharge rates applied to distinct zones fell within each others' ***confidence intervals,*** indicating that one areal recharge rate was sufficient. 4. Yes(*). The best ***model fit*** to the data resulted when the simulated streambed hydraulic conductivity was increased by a factor of three from the head waters to the southern end. 5. No(*; see 2b for CAL3). Having the confining unit extend under the lake such that it covered both the dotted and striped areas of Figure 4a resulted in poor ***model fit.***
CAL0-G1 34 heads 2 flows 6 parameters	1. Does model accuracy depend on the unusual situation found in CAL0, in which 100% of the flow exiting the system was measured at G1?	1. Yes. Omitting the G1 flow, only 11% of the flow exiting the system (the G2 flow) is included in the regression. The G2 flow was matched closely because it alone reduced what would otherwise be extreme ***correlation*** of the estimated parameters. Overall ***model fit*** was closer than for CAL0 (see s, Table 3, Figure 5), but CAL0-G1 predictions were less accurate than CAL0 predictions (Figure 10).
CAL0+PR 34 heads 3 flows 1 prior 6 parameters	1. Can prior information be used to make K_2^0 more reasonable? 2. Does this produce a more accurate model?	1. Yes. Imposing a prior information value of 122 m/d, with a confidence interval of (74;200) (equivalent to a standard deviation of 0.25 for the log-transformed prior), resulted in a high, but more reasonable, estimate of 273 m/d, with a confidence interval of (177;421). 2. No. CAL0+PR predictions are less accurate than CAL0 predictions (Figure 10).
CAL1 44 heads 19 flows 12 prior 18 parameters	1. Can the slug-test data be used to realistically represent spatial heterogeneity of the hydraulic conductivity? 2. Is it most likely that the spatial variation of flow into the Straight River is caused by variations in streambed hydraulic conductance or subsurface hydraulic conductivity?	1. Yes. The slug-test data were used as prior information to estimate a smoothly varying hydraulic conductivity field (Figure 9c). This distribution appears to represent the true distribution with sufficient accuracy in that all ***estimated parameter values*** are reasonable and ***weighted residuals*** are generally random. 2. (*) The ***model fit*** the measured flows into the Straight River more closely using the streambed conductances of Table 1 than by variations in subsurface hydraulic conductivity represented with four finite element nodes along the Straight River. It is likely that variations of either hydraulic property cause similar variations in the measured and predicted heads and flows. The streambed hydraulic conductivity is actually constant.
CAL2 52 heads 19 flows 16 prior 25 parameters	1. Do the data provide enough information to estimate the hydraulic conductivity at any of the boundary nodes?	1. Yes. ***Composite scaled sensitivities*** indicated that hydraulic-conductivity parameters probably could be estimated by the regression at boundary points A, B, and C (Figure 3b) and the resulting regression was, indeed, successful. No new prior information was used.
CAL3 54 heads 19 flows 16 prior 24 parameters	1. How do values from the aquifer test at well P3 (Figure 1a) in field season 3 compare to simulated values? 2. Evaluate two assumptions used so far: a. The leakance of the lakebed is constant over the area of the lake. b. The confining unit does not extend westward under the lake.	1. The ***simulated and aquifer-test horizontal hydraulic conductivities were comparable;*** aquifer-test derived vertical leakance of the confining unit of 0.00337/d was ***significantly lower than the optimized value*** (outside the ***confidence interval*** of Figure 8b), indicating possible model error. 2a. As an alternative consistent with available data, the separate lakebed was omitted from the model so that water flowing to and from the lake simply flowed through the aquifer material beneath the lake. This produced a similar ***model fit,*** and was used in the final CAL3 model. 2b. With the lakebed represented as in 2a, good model fit was achieved with the confining unit extending under the lake (Figure 4b). This was used in the final CAL3 model.
NO LAKE 54 heads 18 flows 16 prior 24 parameters	1. Is including the lake in this simulation important given the predictive quantities of interest?	1. No. When the lake was omitted from the simulation, estimated parameter values were adjusted by the regression so that results were similar to CAL3. The estimated value of K19, the hydraulic conductivity parameter closest to the center of the lake, was unrealistically high, but its ***confidence interval*** included realistic values (Figure 8d).

Selected estimated parameter values and their 95% linear, individual confidence intervals are shown in Figure 8. These figures show how the estimated values and their precision changed for the different models. Notice that parameters with prior information (Figure 8d) never have confidence intervals greater than the confidence interval on the prior information (also noted by Carrera and Neuman 1986). Also, confidence intervals on hydraulic-conductivity values without prior information tend to increase as the estimated value increases, which is consistent with the smaller sensitivities commonly calculated for larger hydraulic conductivities.

The results presented in Table 2 demonstrate a number of situations that are likely to be common in practice. In CAL0, correlation coefficients and composite scaled sensitivities are used to detect data that provide insufficient or marginally sufficient information for the estimation of two of the defined parameters. The CAL0-G1 model demonstrates that with one outflow measurement including only 11% of the flow leaving the system, correlations are as high as 0.95, but a better-fitting model results (the standard error of the regression was less than for the CAL0 model; Table 3). The better fit apparently resulted from there only being one correlation-reducing observation (the G2 flow); the simulated and observed G2 flow were nearly identical, which is typical of single correlation-reducing observations. The consequence is that any error in a single correlation-reducing observation or in the way it is simulated will be directly transmitted to the estimated parameters, so that such an observation can act like an influential outlier on the regression. As discussed below, less accurate predictions were obtained with CAL0-G1 than with CAL0.

CAL0+PR shows prior information being used in a way that diminishes model accuracy, while CAL1 through CAL3 show prior information being used in a way that improves model accuracy. Thus, for this problem, using prior information to force optimal parameter values to be realistic reduced model accuracy; using prior information to include complexity not directly supportable by the other data improved model accuracy.

Table 3

Statistics for Weighting and of the Final Models

[s, standard error of the regression (Equation 2), a value of 1.0 indicates that the model, on average, fits the hydraulic head, streamflow gain, and lake loss data as closely as is consistent with the statistics used to calculate the weights, smaller values indicate a better fit; R_N^2, correlation coefficient for the weighted residuals and expected normal values, with values ranging from 0.0 to 1.0, values close to 1.0 indicate that the weighted residuals are independent, random, and normally distributed.]

	Model						
Statistic	**CAL0**	**CAL0-G1**	**CAL0+PR**	**CAL1**	**CAL2**	**CAL3**	**NO LAKE**
Statistics used to calculate the weights for the observations:							
Preliminary standard deviation for heads (m)	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Preliminary coefficient of variation for:							
Streamflow gains	0.063	0.063	0.063	0.12	0.17	0.17	0.17
Lake loss	0.13	0.13	0.13	0.23	0.33	0.33	0.33
Statistics of the final models:							
s	0.78	0.60	0.99	0.60	0.45	0.33	0.42
[1] R_N^2	0.960	0.939	0.879	0.962	0.980	0.981	0.976
	(0.943)	(0.943)	(0.943)	(0.969)	(0.972)	(0.972)	(0.972)
Fitted statistics of the final models (s times the statistic used to calculate the weights):							
Fitted standard deviaion for heads (m)	0.078	0.060	0.099	0.066	0.045	0.033	0.042
Fitted coefficient of variation for:							
Streamflow gains	0.049	0.038	0.063	0.070	0.075	0.055	0.070
Lake loss	0.099	0.076	0.12	0.14	0.15	0.11	0.14

[1] R_N^2, evaluated for the observations and the prior information. Critical values are in parentheses: if the residuals are independent and normally distributed, then there is only a 5% chance that R_N^2 is less than this value.

Comparison with the True Hydrogeology

The true system has a 30.4-m grid spacing, compared to the 152.4-m spacing used in the calibrated models, and has five model layers instead of two or three. The area of Blue Lake and the location of the Straight River are the same in the true and calibrated models. The side no-flow boundaries are the same in all models. In the true model, the elevation of the impermeable bottom of the ground water flow system was not as smooth as in the calibrated models (Figure 2), and varied from about –46 m on the east to about –91 m on the west.

The confining unit in the final calibrated models (Figure 4b) extends over more of the modeled area than the true confining unit (Figure 1a) because (1) the confining unit present at well 21 was assumed erroneously to be continuous with the larger confining unit to the north; and (2) none of the available data indicates that the confining unit is actually absent along the northeastern model boundary. Neural network analysis of the confining unit data (Risso 1993) correctly represented the break in the southern part of the confining unit, but did not resolve the second problem. No simulations were done with the neural network results as part of the present study.

True parameter values are plotted with the optimized parameter values and their confidence intervals in Figure 8. The true recharge rate plotted in Figure 8a is the average of a stochastically generated distribution with a small variance (5%), so that the true recharge is essentially constant, as concluded in the calibration. Only the CAL0-G1 and CAL0+PR models excluded the true value from the confidence interval. This reflects the importance of the G1 flow in the accurate estimation of recharge and the ability of prior information inappropriately applied to one parameter to affect other estimated parameters. For the vertical leakance of the confining unit (Figure 8b), the confidence intervals only include values that are greater than the actual values. This probably results from the excessive areal extent of the confining unit in the calibrated models discussed previously.

The value of hydraulic conductance of the streambed used in the true model is 244 m^2/d per meter of stream along the length of the Straight River instead of being variable, as simulated in all calibrated models. For all simulations, Figure 8c shows that the smallest estimated conductances are simulated for KRB_1 in the northern reach of the river, where the true underlying hydraulic conductivity is less than the calibrated value (Figure 9; see Figure 1a for stream location). The analysis in Appendix A indicates that little grid-size effect would be expected.

The lakebed was represented in the true system as in CAL3 and NO LAKE, so that there was no distinct lakebed. There is, therefore, no true value of lakebed leakance to compare with the estimates and their confidence intervals, and these values are not presented in this report.

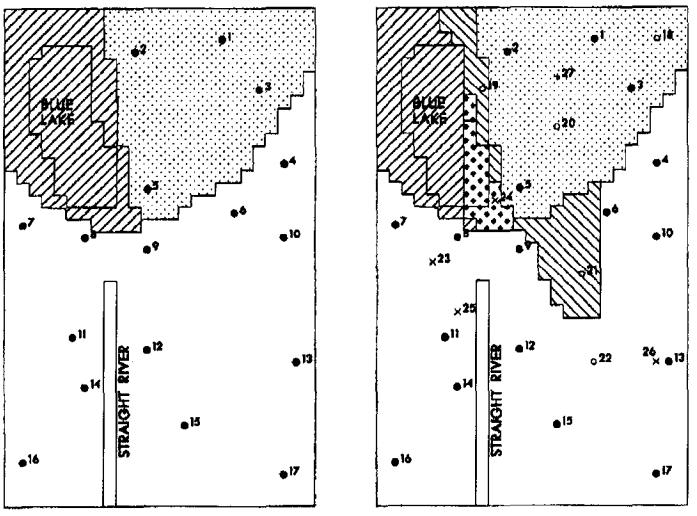

Figure 4a. (left) The simulated extent of the confining unit in the CAL0 calibration. Simulations included a confining unit covering both the dotted and striped areas; in the final CAL0 model the confining unit extended only over the dotted area.

Figure 4b. (right) The simulated extent of the confining unit in the CAL3 and NO LAKE models, with the areas added for the CAL1(///), CAL2 (♦), and CAL3(\\\) models. The locations of wells available at the beginning of the study (●), drilled in field season 1 (o), drilled in field season 2 (x), and the well drilled in field season 3 (+) are shown.

Vertical anisotropy equals 10 in most of the true system instead of 2.5, as used in the final calibrated models; the inaccurate estimate is consistent with the small composite scaled sensitivities of this parameter.

Aquifer horizontal hydraulic conductivity for the true system is shown in Figure 9a and is essentially vertically homogeneous, as assumed in model calibration. The calibrated hydraulic-conductivity distribution for the CAL0 model, with its three zones of constant value, is shown in Figure 3a. It is clear why zonation could not successfully represent this true hydraulic conductivity distribution. The triangular finite-element grid used for interpolation for the CAL3 model is shown in Figure 3b; the same grid is used for the NO LAKE model, and the grid used for the CAL1 and CAL2 models is similar but has fewer nodes.

The calibrated hydraulic-conductivity distribution for model CAL3 is shown in Figure 9b. Most of the major highs and lows of the true hydraulic conductivity are represented. Processes that depend on the smaller scale variations of hydraulic conductivity, such as flow into the Straight River, can only be simulated well if other aspects of the model can make up for the overly smooth estimated hydraulic-conductivity distribution. As mentioned before, low estimated streambed conductance at the headwaters of the Straight River is probably making up for the hydraulic conductivity being too high. The hydraulic heads simulated using CAL3 are not shown, but are nearly identical to the true hydraulic heads (Figure 1a) except in the northeastern corner, where the confining unit was simulated differently and the CAL3 hydraulic-head contours are smoother.

Comparison with the True Predictions

The simulated predictions relevant to management criteria 1 and 2 are (1) the maximum drawdown anywhere in the northern part of the study area (not supposed to exceed 0.6 m), and (2) the percent change in streamflow at gauging station G2 (not supposed to exceed 20%), calculated as 100 times the simulated change divided by 2730 m^3/d, the observed flow under unstressed conditions. As described previously, development scenario A includes pumpage at P1 and P3, scenario B includes pumpage at P2 and P3. Blue Lake is a closed lake so that pumpage was expected to affect the lake level. This was simulated for each prediction by reducing the lake level so that the contribution to the ground water system was the same with pumpage as it had been without pumpage. In the true system, this was accomplished by assigning the lake cells very large hydraulic conductivities.

Figure 10 shows that prediction accuracy for the best-fitting CAL2, CAL3, and NO LAKE models is extremely good; the CAL0-G1 model had the least accurate predictions.

The maximum drawdown may be located anywhere in the northern part of the model. The true maximum drawdown is located along the eastern boundary for scenario A, and along the northern boundary for scenario B. For the three CAL0 models, in which the vertical leakance of the confining unit is very small, the maximum simulated drawdown for both the A and B development scenarios is located along the eastern model boundary at or near row 5, column 25 (Figure 1b). For the other calibrated models, it was simulated at or adjacent to proposed well location P3 (Figure 1a).

From a management perspective, the predictions made by the different models are similar. For development scenario A, all models indicate that management criteria 1 and 2 would be violated, which is correct. For development scenario B, the predictions for the maximum drawdown are close to the management criterion (which is correct), with the greatest discrepancy simulated by the CAL0-G1 model; the predictions for the percent streamflow change are closer to the management criterion for the CAL1, CAL2, CAL3, and NO LAKE models than truly occurs, and the CAL0, CAL0-G1, and CAL0+PR predictions exceed the management criterion more than truly occurs.

The less accurate predictions of the CAL0+PR model relative to the CAL0 model show that, in this circumstance, when the linear confidence interval on the unrealistic parameter value included no reasonable values, the model with more accurate predictions was

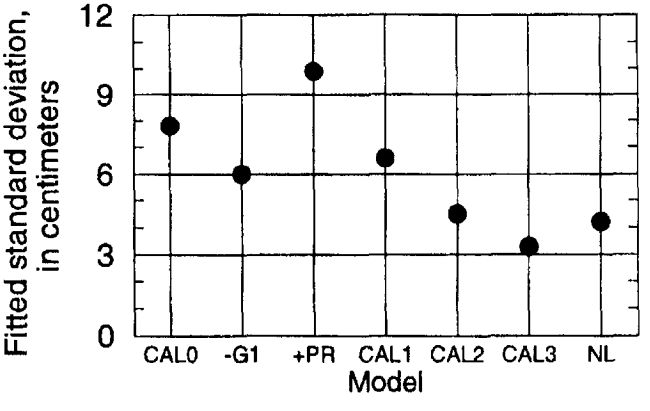

Figure 5. Fitted standard deviations of hydraulic heads for the final models. (NL is the NO LAKE model)

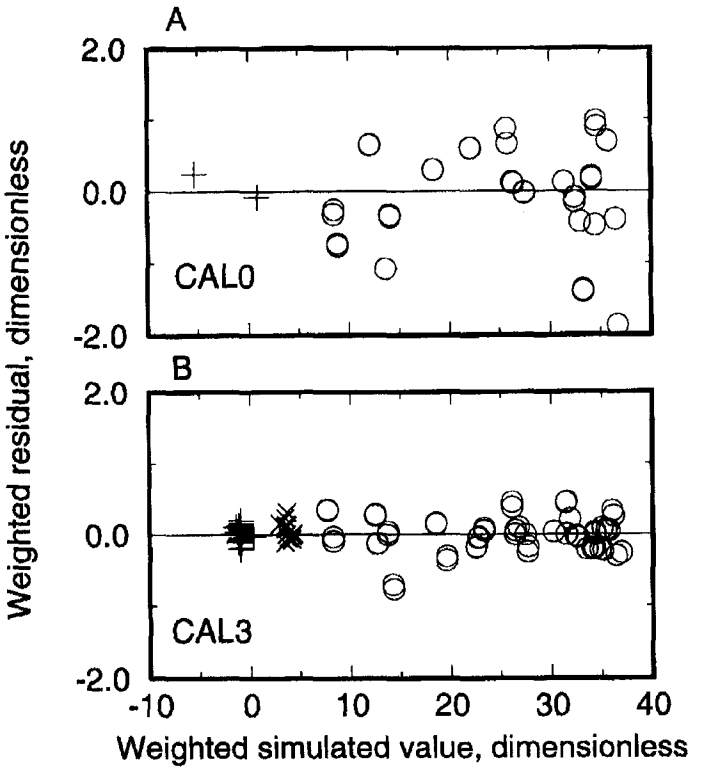

Figure 6. Weighted residuals versus weighted simulated values for the (a) CAL0 and (b) CAL3 models, for hydraulic heads (o), flows (+), and prior information from slug tests (x). In (a), one out of range flow at (-102, -1.2) is not shown.

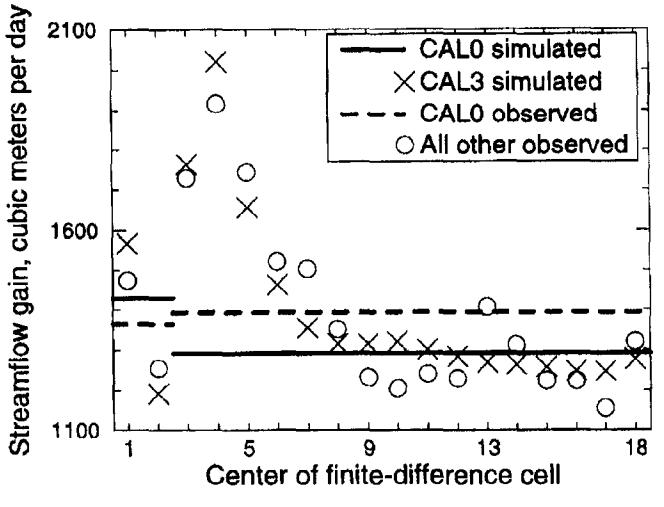

Figure 7. Measured and simulated streamflow gains along the Straight River for the CAL0 and CAL3 models. Values are per finite-difference cell, and finite-difference cell 1 is at the upstream end of the river. For the CAL3 model, parameter KRB_1 applies to cells 1 and 2, KRB_2 applies to cell 3, and KRB_3 applies to cells 4 through 18. Gauge G1 is located between cells 2 and 3; Gauge G2 is located at the downstream end of cell 18.

the model that fit the data better, even though one of the estimated parameter values was unreasonable. This is consistent with Troutman's (1983, p. 801) results for rainfall-runoff modeling, and suggests that when an unreasonable value is well supported by observation data (as indicated by large composite scaled sensitivities and a confidence interval that excludes reasonable parameter values), use of prior information often is not a productive mechanism with which to resolve the problem. More accurate predictions were obtained in this study with a more complex parameterization; prior information was used to achieve a stable regression with the additional parameters.

The results show that, for this problem, the most obvious indicator of a model likely to produce inaccurate predictions was an unreasonable estimated parameter value. In addition, weighted residuals were slightly nonrandom for the less accurate models. Models with a closer fit to the data (Figure 5 and s of Table 1) generally, but not always, produced more accurate predictions.

The predictions shown in Figure 10 are, of course, uncertain, which could be important to their use for management decisions. Analysis of measures of prediction uncertainty is beyond the scope of this report.

Discussion

The power of this work comes from the test case being complex enough to test how the methods are likely to work in actual application. Thus, test case realism is the first issue discussed in this section. Subsequent sections focus on three issues crucial to the use of calibrated models to evaluate ground water systems: Model nonuniqueness and its practical consequences, appropriate representation of small- and large-scale heterogeneities of the hydraulic-conductivity distribution, and the effect of model error.

Test Case Realism

The test case is more realistically complex than any other test case used for this purpose in that (1) the flow system is fully three-dimensional and has commonly encountered types of boundary conditions; (2) the areal hydraulic-conductivity distribution, areal recharge, and areal extent of the confining unit are reasonably complicated, except as noted below; and (3) the problem is posed in terms of a management problem. The test case, however, is simpler than most field problems in that (1) the hydraulic conductivity of the aquifer material did not vary with depth much, a condition rarely, if ever, encountered in natural environments (the flow system, however, was still three-dimensional because of the boundary conditions and the presence of a confining unit); (2) both the true and calibrated systems were truly at steady state, thus avoiding transient effects; (3) the location and impermeability of the lateral and bottom boundaries were better known than in most field problems, and (4) the range of hydraulic conductivities was somewhat narrower than in many field problems.

The regression data used to calibrate the seven models developed in this work were hydraulic heads, lake seepage, streamflow gains, and hydraulic-conductivity values produced by slug tests, all of which were derived from the synthetic test case. The data were generally typical of the type of data available in most field studies, except as follows: (1) The "observed" values of hydraulic head and flow were derived directly from the true models, with no random errors added; (2) the hydraulic-head data were more evenly distributed in space than commonly occurs; (3) the streamflow-gain measurements included 100% of the outflow from the system (except for the CAL0-G1 model) and, for some of the models, were of higher resolution (every 152.4 m) than streamflow gains generally can be determined from measured streamflows given common measurement errors; and (4) the slug-test data were equal to the hydraulic conductivity of a 30.6-m square finite-difference cell,

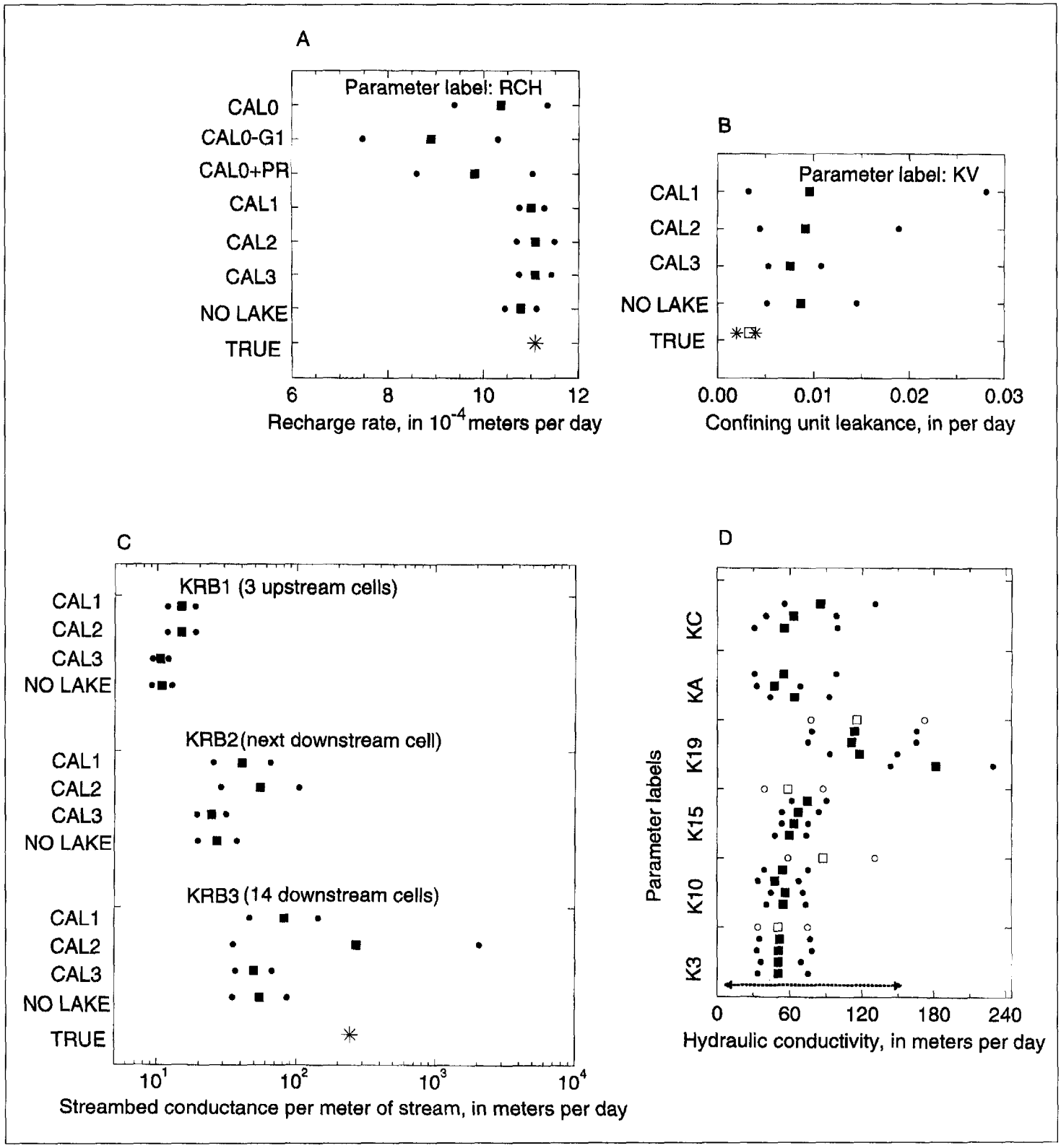

Figure 8. (a–d). Estimated values of selected parameters used to calculate the following model characteristics: (a) recharge rate; (b) vertical leakance of the confining unit; (c) streambed conductance; and (d) horizontal hydraulic conductivity. The graphs show the estimated values (■); their 95% linear, individual confidence intervals (●); true values (*) or, in (Figure 8b), the range of true values (between the *s). Figure 8d includes measured values (□); for measured values used as prior information in the regression (Figure 8d), the 95% linear, individual confidence intervals (o), calculated using the statistics specified for the prior information in the regression; the expected range of values is plotted using a line with arrows. Prior information, if used, is plotted as the first row for each parameter; subsequent rows are results from the CAL1, CAL2, CAL3, and NO LAKE models. See Table 1 for definition of parameter labels.

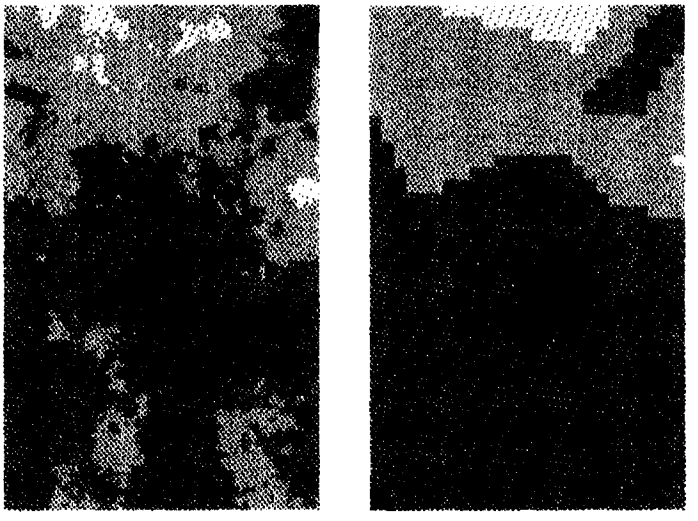

Figure 9a. (left) Map showing the true hydraulic-conductivity distribution.

Figure 9b. (right) Map showing the optimized distribution of hydraulic conductivity for model CAL3.

Gray shade	Hydraulic conductivity, in m/d
(darkest)	0-36
	36-72
	72-108
	108-144
(lightest)	144-180 [occurs only in one cell of (A)]

which is a larger amount of material than slug tests usually sample. Repeating some of the regressions using noisy data would be an important contribution, but this was considered to be beyond the scope of this report.

Despite the simplifications, the test case appears to have been complex enough to serve the objectives of this work.

Model Uniqueness and Its Practical Consequences

Given the advantages inherent in the synthetic test case, one might have expected to find one model that was clearly the best and thus be able to avoid the problem of nonuniqueness normally found in ground water flow problems. Nonetheless, nonuniqueness proved to be a problem. Models CAL2, CAL3, and NO LAKE yielded a remarkably similar quality of calibration (Table 3, Figure 5) and prediction (Figure 10), despite different assumptions regarding the hydraulic-conductivity distribution, areal extent of the confining unit, vertical anisotropy, and representation of the lakebed and the lake. Even CAL1, for which the calibration was not as good, produced accurate predictions. This is important because it indicates that the lack of uniqueness pervasive in ground water models does not necessarily indicate that the models produce inaccurate predictions and, therefore, are useless. Instead of uniqueness, prediction accuracy appears to depend on the type and accuracy of the available data and the calibration methodology. Basically, if the calibration is sufficiently constrained, different calibrated models are likely to produce results of similar accuracy.

Appropriate Representation of Small- and Large-Scale Heterogeneity

Representing the hydraulic-conductivity field using three zones in the CAL0 models had the advantage of allowing a clear evaluation of the relation between the data and the parameters. In this test case, use of head data alone or in combination with the lake-budget data resulted in extremely correlated parameters, which would have prohibited estimation of individual parameter values. Such an evaluation generally is not possible with more complex parameterizations in which some parameters are not estimated and(or) prior information is used, because both affect the statistics used to identify parameter correlation.

Representing the hydraulic-conductivity field using three zones had the disadvantage of being too unrealistic in this test case, as indicated by one unrealistic hydraulic-conductivity value estimated by the regression, somewhat nonrandom weighted residuals, and inaccurate predictions.

Representing the hydraulic-conductivity field using a simple interpolation method based on linear triangular finite elements produced good fitting models capable of accurate predictions of drawdown and changes in streamflow gain. Use of other interpolation methods could have some advantages; for example, kriging allows the distance of influence of each interpolation point to be easily adjusted. We do not expect that using another interpolation method would significantly change the results, but this was not tested in the present study because such a good fit was achieved with the simpler methods. Methods that allow smaller scale variation, such as grid-scale parameterizations, also were not tested for the same reason.

Using the slug-test data as prior information on estimated parameters located at interpolation points of the finite-element grid allowed the regression to adjust the estimated parameter values so that the simulated hydraulic conductivities were more representative of the hydraulic conductivities at the larger scale. This is in contrast to the pilot-points method (Marsily et al. 1984; Certes and Marsily 1991; RamaRao et al. 1995), in which the hydraulic conductivity at the slug-test measurement points would have been set (or nearly set) to the slug-test value and parameters located at additional points would be estimated. It is also in contrast to other geostatistical methods (Kitanidis 1995; Yeh et al. 1995), in which the point values are used with little or no change and the majority of the effort is spent modeling the variogram. Estimated values differed from prior estimates by as much as 57%, and 26% for the best-fitting model and it is thought that much of the difference is related to scale. Certes and Marsily (1991) suggest that the scale effect can be neglected, but that didn't appear to be the case in the present study.

Accurate predictions were obtained despite the lack of pumping in calibration conditions, and its presence in prediction conditions. Although changes in the flow field can change effective values of hydraulic conductivity, this problem did not appear to be significant in the present test case.

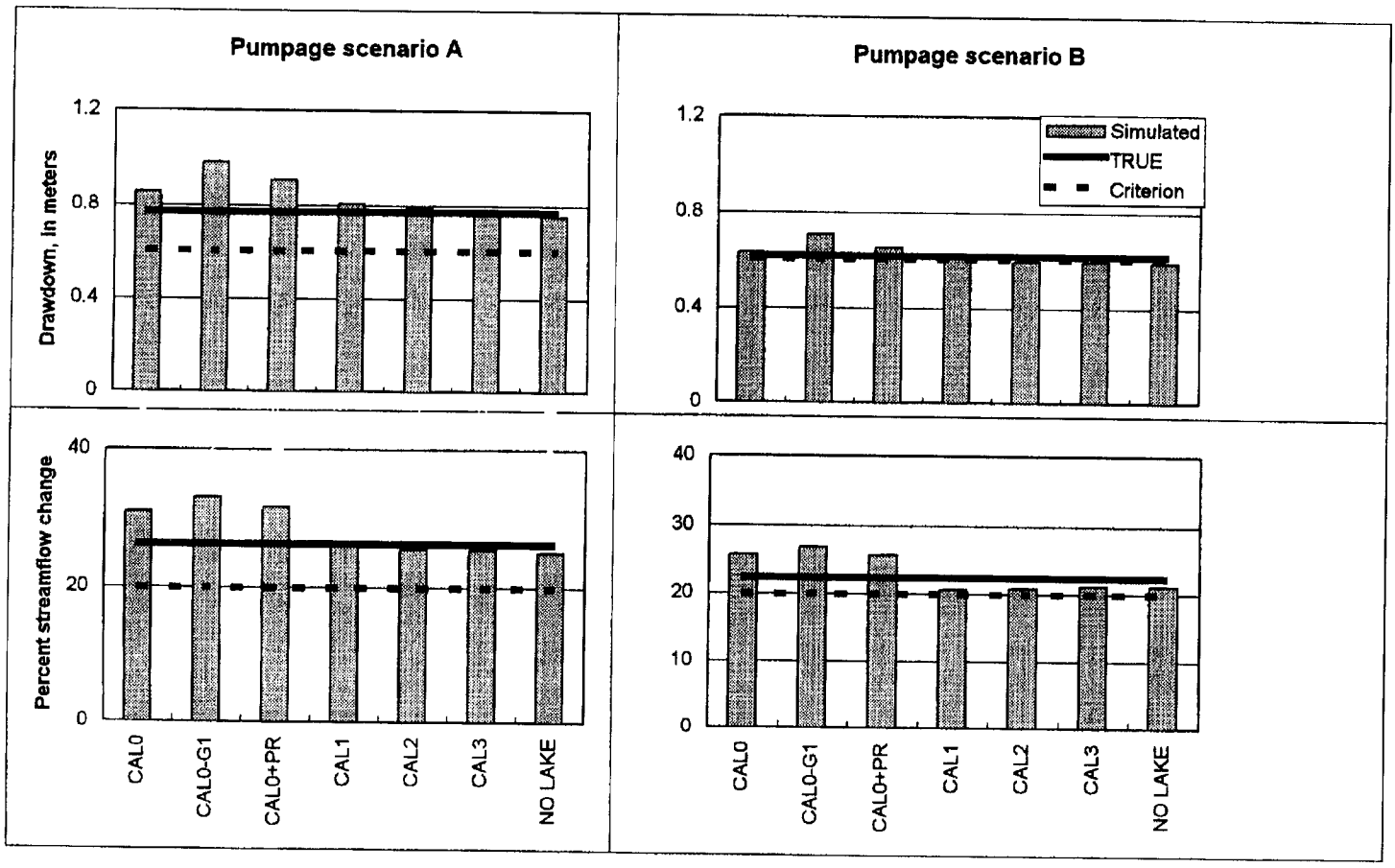

Figure 10. Graphs showing predictions for pumping scenarios A and B calculated using the calibrated and true models and the relevant management criteria.

Effects of Model Errors

In this test case, residuals result from model error because no random measurement errors were introduced into the simulated data. This provides a unique opportunity to investigate model error.

The R_N^2 values of Table 3 indicate that all sets of weighted residuals are, or are nearly, independent and normally distributed, and spatial plots of weighted residuals (not shown) generally revealed no spatial trends. Two aspects of this are important. The most basic is that the weighted residuals appear to be random. Randomness of residuals in regression analysis usually is thought to result from measurement errors (Draper and Smith 1981, p. 22-24). A possible explanation for the apparent randomness of the residuals in this synthetic test case is that each residual is the result of model errors from several sources, such as errors in parameterization, boundary conditions, and other aspects of model construction. Weighted residuals that are in effect random suggest that weighted true errors are also effectively random. This is significant because, if true, it suggests that certain types of model error might be accommodated by standard regression assumptions that require that the true errors be random (Seber and Wild 1989, p. 573; White 1981). In addition, the assumption of random true errors is required to assure asymptotic consistency and normality of estimated parameter values (Seber and Wild 1989, p. 563-572). This work indicates that the presence of model error may not violate these requirements unless the estimated parameter values, weighted residuals, or other measures indicate model bias.

In addition to being random, the weighted residuals were normally distributed. As with randomness, the normality of the weighted residuals is often thought to be a result of measurement error, which was not present in this synthetic problem. The normality of the weighted residuals suggests that despite the pervasive presence of model error in numerical ground water flow models, the assumption of normality often may be valid. This would result if a process similar to that described by the classical central limit theorem (Draper and Smith 1981, p. 24) were operating, and means that the formulation of Tarantola (1987, p. 58) would be valid. By the central limit theorem, contributions of error from multiple independent sources result in random, normally distributed values, regardless of the probability distribution of each of the individual errors. As stated above, indication of model bias, such as nonrandom weighted residuals or unrealistic optimal parameter values, could indicate that a limited number of errors are dominating so that the central limit theorem would not apply.

Conclusions

In this test case, properly used nonlinear regression either produced effective, though nonunique, calibrated models capable of accurately predicting two quantities important to resource management, or provided clear evidence of model or data inadequacy. The conclusions listed here relate both to effective use of the nonlinear regression methods (conclusions 1–4) and to challenging some common practices and commonly held beliefs in model calibration (conclusions 5–7).

1. The method of determining the weights for the weighted least-squares objective function tested in this study was used to correctly detect much smaller measurement error than would normally occur.

2. The most conclusive indicator of model bias was unrealistic optimal parameter estimates that also had confidence intervals that excluded reasonable values. Nonrandom weighted residuals were a less conclusive indicator of model bias in the present study, but this may not always be the case, especially when models are more biased than those considered here.
3. Including prior information in the regression diminished model accuracy when used to force optimal parameter values to be reasonable, but improved model accuracy when used to represent the hydraulic-conductivity distribution with more complexity than was supportable with the head and flow observations alone. Excluding all prior information initially allowed for clear evaluation of the contributions of different types of data.
4. The importance of flow data was clearly demonstrated in this study. This is important because few field studies have measurements representing all of the flow leaving the system, so that problems of completely correlated parameters and the effects of a single correlation-reducing observation, as documented for the CAL0-G1 model, are probably common. These problems can be clearly characterized and understood by first representing ground water systems very simply, and building complexity as warranted by the data and modeling objectives.
5. The results of this study indicate that, given present technology, hydraulic-conductivity values measured in the field often are not as directly applicable to a numerical model of the system as would be consistent with how these data are sometimes used in model calibration. Two aspects of the controlled experiment presented in this paper support this conclusion.
   (a) In four of the models, the hydraulic-conductivity distribution was adequately represented (as evidenced by accurate simulated predictions) by an interpolation scheme in which values that were held constant were limited to the model boundaries, while nearly all values within the modeled area were estimated. Unusually accurate slug-test values were used as prior information in the regression; corresponding estimated aquifer hydraulic-conductivities differed from the slug-test values by as much as 57%, suggesting that direct imposition of the slug-test values, as is done in some geostatistical methods, would probably produce a less accurate model. This situation largely reflects problems of scale. Here, the numerical grid spacing was five times larger than in the true system; in field applications the scale problem is likely to be more severe.
   (b) Streambed hydraulic-conductance estimates were affected by underlying subsurface heterogeneities that were not well represented by the simulated aquifer hydraulic-conductivity distribution; if field work had determined that the streambed conductance was constant along the river (which it was) and this had been imposed, the calibrated models probably would have produced less accurate predictions. This situation does not represent a scale problem as much as error in representing one part of the system affecting the parameters representing another part of the system.
6. For three of the calibrated models, the fit to the regression data was nearly equally good and there was no evidence of model bias. This lack of uniqueness is probably unavoidable in complex ground water problems, but the results of this work indicate that such nonuniqueness is not necessarily a debilitating problem. In the synthetic test case, all three models produced similar accurate predictions, probably because the data used in model construction and in the regression sufficiently constrained the solutions.
7. Weighted residuals (observed minus simulated hydraulic heads, flows, and prior information) resulting from the regression were, in general, random and normally distributed, which is surprising because no errors had been added to the synthetically generated observations. Thus, all error was model error. It is generally thought that if model error dominates measurement error, the regression results are invalid, but the results of this work imply that the dominance of model error does not necessarily produce an inaccurate model if there is no obvious indication of model bias.

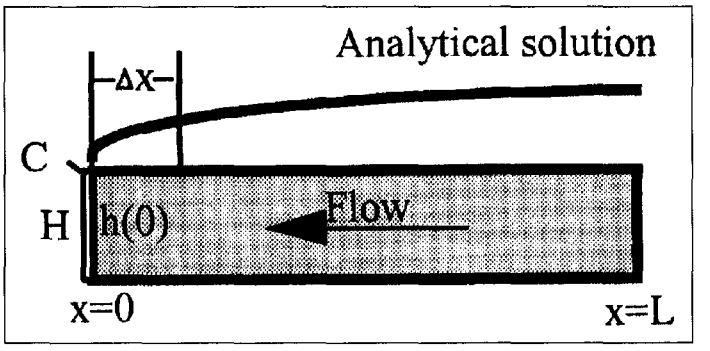

Figure A1. Simple system of ground water flow toward a fully penetrating stream with a streambed of different hydraulic properties. C is the conductance of the streambed for the analytical problem, h(0) is the hydraulic head on the aquifer side of the streambed for the analytical problem.

Taken with conclusion 2, conclusion 5 produces a dilemma, because unrealistic optimal parameter values (as determined by comparing optimal and measured values) are said to indicate a less accurate model, but parameter values cannot be expected to equal measured values because parameter values are accommodating model error. A useful resolution is derived by noting that if model error is so large that best fit parameter values are far from measured values, the resulting model is likely to produce less accurate predictions than a model for which the best parameter values are close to measured values. Thus, models with more realistic best-fit parameter values are more likely to be accurate.

Conclusions 4 and 6 suggest that improved accuracy of ground water models is likely to be attained by using available data more effectively and by designing methods to collect new kinds of data, perhaps using regression methods as a guide.

Appendix A

Effects of Grid Size on Simulated Streambed Conductance

This appendix evaluates the effects of grid size on the simulated streambed conductance using a simple one-dimensional analytical equation. The results are expected to approximate the effects related to the three-dimensional system considered in this study.

Flow to a stream with a streambed having conductance C, through a homogeneous ground water system of hydraulic conductivity K, given a constant recharge rate of W, can be idealized as shown in Figure A1, where h(0) is the hydraulic head on the

ground water system side of the streambed. Using the variables shown in Figure A1, recharge rate W, and constant layer thickness b, the analytical solution for hydraulic head is

$$h(x) = \frac{W}{Kb}\left(Lx - \frac{x^2}{2}\right) + h(0)$$

When a finite-difference grid is imposed, a linear gradient is enforced between finite-difference cell centers, over distance Δx in Figure A1. Note that the system can be approximated this way, regardless of whether the stream abuts the grid on the side, as in Figure A1, or on the top, as in most models. Farther from the stream, it is assumed that the finite-difference grid closely simulates the analytical head solution. Thus, the following equations apply:

$$h(x) = \frac{W}{Kb}\left(Lx - \frac{x^2}{2}\right) + h(0) \qquad x \geq \Delta x \qquad (A1)$$

$$h_{\Delta x}(x) = \frac{W(L - \Delta x/2)}{Kb}x + h(0) \qquad 0 < x < \Delta x \qquad (A2)$$

Subscript Δx indicates that the quantity is related to the linear gradient imposed near the stream. The flow through the finite-difference cell next to the stream is evaluated midway between cell centers, so equals $W(L - \Delta x/2)$. The condition of head continuity at $x = \Delta x$ was used to determine the constant of integration as h(0) in Equation A2.

Flow through the streambed at x = 0 can be described using Darcy's law as:

for the analytical system,

$$WL = C(h(0) - H), \text{ or } h(0) = H + \frac{WL}{C} \qquad (A3)$$

for the discretized system,

$$WL = C_{\Delta x}(h_{\Delta x}(0) - H), \text{ or } h_{\Delta x}(0) = H + \frac{WL}{C_{\Delta x}} \qquad (A4)$$

where C and $C_{\Delta x}$ are the streambed conductances of the analytical and discretized solutions, respectively, and h(0) and $h_{\Delta x}(0)$ are the hydraulic heads on the aquifer side of the streambed applicable to the analytical and discretized solutions, respectively. Flow out of the model is considered to be positive to coordinate with Equation (A1). Substituting Equations A3 and A4 into Equation A2 evaluated at x = 0 shows that $C_{\Delta x} = C$. Solve Equation A2 for x = 0 to yield $h_{\Delta x}(0) = h(0)$. Thus, grid size does not affect the variables related to simulation of the river.

This analysis does not completely represent the dynamics of flow to a stream in a three-dimensional, heterogenous system. The results, however, indicate that grid effects are not likely to be important in the calibration of streambed conductance in the present work.

References

Anderman, E.R., M.C. Hill, and E.P. Poeter. 1996. Two-dimensional advective transport in ground water flow parameter estimation. *Ground Water* 34, no. 6: 1001-1009.

Backus, G.E. 1988. Bayesian inference in geomagnetism. *Geophysical Journal,* 92, 125-142.

Bard, J. 1974. *Nonlinear Parameter Estimation.* New York: Academic Press.

Barlebo, H.C., M.C. Hill, and D. Rosbjerg. 1996. Identification of groundwater parameters at Columbus, Mississippi, using three-dimensional inverse flow and transport model. In Proceedings of the 1996 Model CARE Conference, Golden Colorado, September, ed. K. Kovar and P. van der Heidje.

Barlebo, H.C., M.C. Hill, D. Rosbjerg, and K.H. Jensen. In press. On concentration data and dimensionality in groundwater models. *Nordic Hydrology.*

Beven, K., and A. Binley. 1992. The future of distributed models, Model calibration and uncertainty prediction. *Hydrological Processes* 6, 279-298.

Box, G.E.P., and G.M. Jenkins. 1976. *Time Series Analysis, Forecasting and Control,* 2nd ed. San Francisco: Holden-Day.

Carrera, J., and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady-state conditions. *Water Resources Research* 22, no. 2: 199-242.

Certes, C., and de Marsily. 1991. Application of the pilot point method to the identification of aquifer transmissivities. *Advances in Water Resources* 14, no. 5: 284-300.

Chu, Wen-sen, E.W. Strecker, and D.P. Lettenmaier. 1987. An evaluation of data requirements for groundwater contaminant transport modeling. *Water Resources Research* 23, no. 3: 408-424.

Clifton, P.M., and S.P. Neuman. 1982. Effects of kriging and inverse modeling on conditional simulation of the Avra Valley aquifer in southern Arizona. *Water Resources Research* 18, no. 4: 1215-1234.

Cooley, R.L. 1977. A method of estimating parameters and assessing reliability for models of steady-state ground water flow, 1. Theory and numerical properties. *Water Resources Research* 13, no. 2: 318-324.

Cooley, R.L. 1979. A method of estimating parameters and assessing reliability for models of steady-state ground water flow, 2. Application of statistical analysis. *Water Resources Research* 135 no. 3: 603-617.

Cooley, R.L. 1982. Incorporation of prior information on parameters into nonlinear regression groundwater flow models—1. Theory. *Water Resources Research* 18, no. 4: 965-976.

Cooley, R.L., L.F., Konikow, and R.L. Naff. 1986. Nonlinear regression groundwater flow modeling of a deep regional aquifer system. *Water Resources Research* 22, no. 13: 1759-1778.

Cooley, R.L., and R.L. Naff. 1990. Regression modeling of ground-water flow: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 3, Chapter B4.

D'Agnese, F.A., C.C. Faunt, M.C. Hill, and A.K. Turner. In press. Death Valley regional ground water flow model calibration using optimal parameter estimation methods and geoscientific information systems. Advances in Water Resources.

D'Agnese, F.A. C.C. Faunt, A.K. Turner, and M.C. Hill. 1998. Hydrogeologic evaluation and numerical simulation of the Death Valley Regional ground water flow system, Nevada and California. U.S. Geological Survey Water-Resources Investigation Report 96-4300, 124 p.

Draper, N.R., and H. Smith. 1981. *Applied Regression Analysis,* 2nd ed. New York: John Wiley and Sons.

Eppstein, M.J., and D.E. Dougherty. 1996. Simultaneous estimation of transmissivity values and zonation. *Water Resources Research* 32, no. 11: 3321-3336.

Gelhar, L.W. 1993. *Stochastic Subsurface Hydrology.* Englewood Cliffs, New Jersey: Prentice-Hall.

Gomez-Hernandez, J.J., and S.M. Gorelick. 1989. Effective groundwater model parameter values, Influence of spatial variability of hydraulic conductivity, leakance, and recharge. *Water Resources Research* 25, no. 3: 405-419.

Gupta, V.K., and S. Sorooshian. 1985. The relationship between data and the precision of parameter estimates of hydrologic models. *Journal of Hydrology* 81: 57-77.

Helsel, D.R., and R.M. Hirsch. 1992. *Statistical Methods in Water Resources.* New York: Elsevier.

Hill, M.C., 1992, A computer program (MODFLOWP) for estimating parameters of a transient, three-dimensional, ground water flow model using nonlinear regression. U.S. Geological Survey Open-File Report 91-484.

Hill, M.C. 1994. Five computer programs for testing weighted residuals and calculating linear confidence and predictions intervals on results from the ground water parameter-estimation computer program MODFLOWP. U.S. Geological Survey Open-File Report 93-481.

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

Hoeksema, R.J., and P.K. Kitanidis. 1984. An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling. *Water Resources Research* 20, no. 7: 1003-1020.

Huber, P.J. 1981. *Robust Statistics.* New York: John Wiley and Sons.

Hyndman, D.W., and S.M. Gorelick. 1996. Estimating lithologic and transport properties in three dimensions using seismic and tracer data, The Kesterton aquifer. *Water Resources Research* 32, no. 9: 2659-2670.

Jacobson, E.A. 1985. A Statistical Parameter Estimation Method Using Singular Value Decomposition with Application to Avra Valley Aquifer in Southern Arizona, Ph.D. diss., University of Arizona, Department of Hydrology and Water Resources.

Kitanidis, P.K. 1995. Quasi-linear geostatistical theory for inversing. *Water Resources Research* 31, no. 10: 2411-2419.

Mantoglou, A., and J.L. Wilson. 1982. The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research* 18, no. 5: 1379-1394.

Marsily, G. de, G. Lavedon, M. Boucher, and G. Fasinino. 1984. Interpretation of inference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, In *Geostatistics for natural resources characterization,* vol. 122, ed. G. Verly, M. David, A.G. Journel, and A. Marechal, 831-849. New York: NATO ASI Series C.

McLaughlin, D., and L.R. Townle. 1996. A reassessment of the groundwater inverse problem. *Water Resources Research* 32, no. 5: 1131-1161.

Neele, F., J. VanDecar, and R. Snieder. 1993. The use of P wave amplitude data in a joint inversion with travel times for upper mantle velocity structure. *Journal of Geophysical Research* 98, no. B7: 12,033-12,054.

Parker, R.L. 1994. *Geopyhsical Inverse Theory.* Princeton, New Jersey: Princeton University Press.

Poeter, E.R., and M.C. Hill. 1996. Unrealistic parameter estimates in inverse modeling, A problem or a benefit for model calibrations? In Proceedings of the ModelCARE Conference, Golden Colorado, ed. K. Kovar and P. van der Heijde. IAHS Publ. no. 237: 227-285.

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

Poeter, E.R., and S.A. McKenna. 1995. Reducing uncertainty associated with ground water flow and transport predictions. *Ground Water* 33, no. 6: 899-904.

RamaRao, B.S., M.A. LaVenue, G. de Marsily, and M.G. Marietta. 1995. Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 1, Theory and computational experiments. *Water Resources Research* 31, no. 3: 475-493.

Reid, L.B. 1996. A functional inverse approach for three-dimensional characterization of subsurface contamination. Ph.D. diss., Department of Civil Engineering, Massachusetts Institute of Technology.

Risso, D. 1993. Written communication, University of Vermont, Burlington, Vermont.

Roa, C.R. 1973. *Linear Statistical Inference and Its Applications,* 2nd ed. New York: John Wiley.

Seber, G.A.F., and C.J. Wild. 1989. *Nonlinear Regression.* New York: Wiley.

Sun, N-Z. 1994. *Inverse Problems in Groundwater Flow and Transport.* Boston: Kluwer Academic Publishers.

Sun, N-Z., and W.W.-G. Yeh. 1985. Identification of parameter structure in groundwater inverse problem. *Water Resources Research* 21, no. 6: 869-883.

Sun, N-Z, and W. W.-G. Yeh. 1990. Coupled inverse problems in groundwater modeling. *Water Resources Research* 26, no. 10: 2507-2540.

Tarantola, A. 1987. *Inverse Problem Theory.* New York: Elsevier.

Theil, H. 1963. On the use of incomplete prior information in regression analysis. *American Statistical Association Journal* 58, no. 302: 401-414.

Tikhonov, A.N., and V.Y. Arsenin. 1977. *Solutions of Ill-Posed Problems.* New York: Winston and Sons.

Troutman, B.M. 1983. Runoff prediction errors and bias in parameter estimation induced by spatial variability of precipitation. *Water Resources Research* 19, no. 3: 791-810.

White, H. 1981. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76, 419-433.

Wilson, J. 1989. Written communication, New Mexico Institute of Technology, Socorro, New Mexico.

Yeh, T.-C. J., A.L. Gutjahr, and M. Jim. 1995. An iterative cokriging-like technique for ground water flow modeling. *Ground Water* 33, no. 1: 33-41.