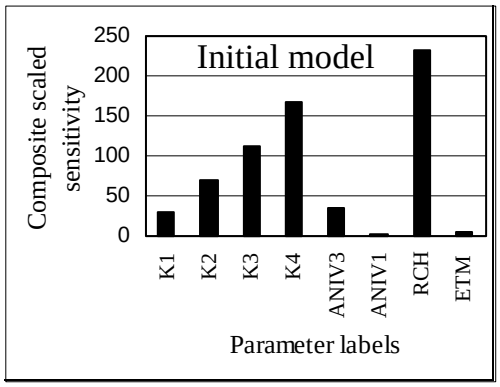
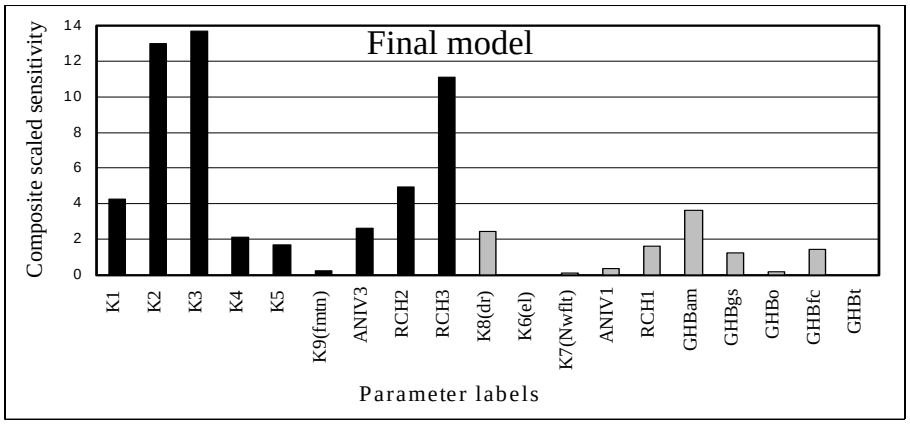
METHODS AND GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

U.S. GEOLOGICAL SURVEY
WATER-RESOURCES INVESTIGATIONS REPORT 98-4005

With application to:
UCODE, a computer code for universal inverse modeling, and
MODFLOWP, a computer code for inverse modeling with MODFLOW





METHODS AND GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

by Mary C. Hill

U.S. GEOLOGICAL SURVEY
WATER-RESOURCES INVESTIGATIONS REPORT 98-4005

With application to:
UCODE, a computer code for universal inverse modeling, and
MODFLOWP, a computer code for inverse modeling with MODFLOW

Denver, Colorado
1998

U.S. DEPARTMENT OF THE INTERIOR
BRUCE BABBITT, Secretary
U.S. GEOLOGICAL SURVEY
Thomas J. Casadevall, Acting Director

For additional information
write to:

Regional Research Hydrologist
U.S. Geological Survey
Water Resources Division
Box 25046, Mail Stop 413
Denver Federal Center
Denver, CO 50225-0046

Copies of this report can be purchased from:

U.S. Geological Survey
Branch of Information Services
Box 25286
Denver Federal Center
Denver, CO 80225-0425

PREFACE

The methods and guidelines described in this report are designed to promote accuracy when simulating complex systems with mathematical models that need to be calibrated, and in which the calibration is accomplished using inverse modeling. This report focuses on the implementation of the described methods in the computer codes UCODE (Poeter and Hill, 1998) and MODFLOWP (Hill, 1992), which perform inverse modeling using nonlinear regression, but the methods have been implemented in other codes. The guidelines as presented depend on statistics described in this work, but other statistics could be used. Many aspects of the approach are applicable to any model calibration effort, even those conducted without inverse modeling. The methods and guidelines presented have been tested in a variety of ground-water modeling applications, many of which are cited in this report, and are described in the context of ground-water modeling concepts. They are, however, applicable to a much wider range of problems.

CONTENTS

FIGURES

TABLES

METHODS AND GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

By Mary C. Hill

ABSTRACT

This report documents methods and guidelines for model calibration using inverse modeling. The inverse modeling and statistical methods discussed are broadly applicable, but are presented as implemented in the computer programs UCODE, a universal inverse code that can be used with any application model, and MODFLOWP, an inverse code limited to one application model. UCODE and MODFLOWP perform inverse modeling, posed as a parameter-estimation problem, by calculating parameter values that minimize a weighted least-squares objective function using nonlinear regression. Minimization is accomplished using a modified Gauss-Newton method, and prior, or direct, information on estimated parameters can be included in the regression. Inverse modeling in many fields is plagued by problems of instability and nonuniqueness, and obtaining useful results depends on (1) defining a tractable inverse problem using simplifications appropriate to the system under investigation and (2) wise use of statistics generated using calculated sensitivities and the match between observed and simulated values, and associated graphical analyses. Fourteen guidelines presented in this work suggest ways of constructing and calibrating models of complex systems such that the resulting model is as accurate and useful as possible.

INTRODUCTION

Problem

In many fields of science and engineering, mathematical models are used to represent complex processes. Commonly, quantities simulated by the mathematical model are more readily measured than are model input values, and model calibration is used to construct a model and estimate model input values. In model calibration, various parts of the model, including the value of model input values, are changed so that the measured values (often called observations) are matched by equivalent simulated values, and, hopefully, the resulting model accurately represents important aspects of the actual system.

The model inputs that need to be estimated are often distributed spatially and(or) temporally, so that the number of parameter values could be infinite. The number of observations, however,

generally is limited and able to support the estimation of relatively few model input values. Addressing this discrepancy is one of the greatest challenges faced by modelers in many fields. Generally a set of assumptions are introduced that allows a limited number of values to be estimated, and these values are used to define selected model inputs throughout the spatial domain or time of interest. In this work, the term "parameter" is reserved for the values used to characterize the model input. Alternatively, some methods, such as those described by Tikhonov (1977) typically allow more parameters to be estimated, but these methods are not stressed in the present work.

Not surprisingly, formal methods have been developed that attempt to estimate parameter values given some mathematically described process and a set of relevant observations. These methods are called inverse models, and they generally are limited to the estimation of parameters as defined above. Thus, the terms "inverse modeling" and "parameter estimation" commonly are synonymous, as in this report.

For some processes, the inverse problem is linear, in that the observed quantities are linear functions of the parameters. In many circumstances of practical interest, however, the inverse problem is nonlinear, and solution is much less straightforward than for linear problems. This work discusses methods for nonlinear inverse problems.

Despite their apparent utility, inverse models are used much less than would be expected, with trial-and-error calibration being much more commonly used in practice. This is partly because of difficulties inherent in inverse modeling technology. Because of the complexity of many real systems and the sparsity of available data sets, inverse modeling is often plagued by problems of insensitivity, nonuniqueness, and instability. Insensitivity occurs when the observations do not contain enough information to support estimation of the parameters. Nonuniqueness occurs when different combinations of parameter values match the observations equally well. Instability occurs when slight changes in, for example, parameter values or observations, radically change inverse model results. All these problems are exacerbated when the inverse problem is nonlinear.

Though the difficulties make inverse models imperfect tools, recent work has clearly demonstrated that inverse modeling provides capabilities that help modelers take greater advantage of their models and data, even when the systems simulated are very complex. The benefits of inverse modeling include (1) clear determination of parameter values that produce the best possible fit to the available observations; (2) diagnostic statistics that quantify (a) quality of calibration, (b) data shortcomings and needs, (3) inferential statistics that quantify reliability of parameter estimates and predictions; and (4) identification of issues that are easily overlooked during non-automated calibration. Quantifying the quality of calibration, data shortcomings and needs, and confidence in parameter estimates and predictions are important to communicating the results of modeling studies to managers, regulators, lawyers, and concerned citizens, as well to the modelers themselves.

Purpose and Scope

This report describes the theory behind inverse modeling and guidelines for its effective application. It is anticipated that the methods discussed will be useful in many fields of the earth sciences, as well as in other disciplines. The expertise of the author is in the simulation of groundwater systems, so the examples presented in this report all come from this field, which is characterized by three-dimensional, temporally varying systems with a high degree of spatial variability and sparse data sets.

For convenience, the methods and guidelines are presented in the context of the capabilities of specific inverse models. The models chosen are UCODE (Poeter and Hill, 1998) and MODFLOWP (Hill, 1992). These models were chosen because they were designed using the methods and guidelines described in this report, and because UCODE is a universal inverse code with broad applicability, and MODFLOWP is an inverse code programmed using the most accurate methods available for calculation of sensitivities.

The report is dominated by sections on methods and guidelines of inverse modeling using nonlinear regression. Because computer execution time is nearly always of concern in inverse modeling, a section is dedicated to issues related to this problem. There have been a number of field applications using the methods and guidelines presented in this report, and these are listed. Finally, a section is devoted to the use of the guidelines with inverse models with capabilities that differ from those of UCODE and MODFLOWP.

Previous Work

The methods presented are largely derived from Hill (1992) and Cooley and Naff (1990) and references cited therein. Various aspects of the suggested guidelines have a long history, and relevant references are cited when the guidelines are presented. To the author's knowledge, no similar set of guidelines that provide as comprehensive a foundation as those presented here have been presented elsewhere.

Acknowledgments

The author would like to acknowledge the following colleagues and students for insightful discussions and fruitful collaborations: Richard L. Cooley, Richard M. Yager, Claire Tiedeman, Frank D'Agnese, and Ned Banta of the U.S. Geological Survey, Eileen P. Poeter of the Colorado School of Mines, Evan R. Anderman of ERA Ground-Water Modeling, LLC, Heidi Christiansen Barlebo of the Geological Survey of Denmark and Greenland, and Steen Christensen of Aarhus University, Denmark. In addition, thought-provoking questions from students and MODFLOWP users throughout the years have been invaluable.

METHODS OF INVERSE MODELING USING NONLINEAR REGRESSION

Nonlinear regression, instead of the easier to use linear regression, is needed when simulated values are nonlinear with respect to parameters being estimated. This is common in groundwater problems, as discussed by Hill (1992) and Sun (1994), among others, and in other systems. Model nonlinearity produces important complications to regression and has been the topic of considerable investigation in several fields. Seber and Wild (1989) is an excellent upper-level text on nonlinear regression.

Weighted Least-Squares and Maximum-Likelihood Objective Functions

The objective function is a measure of the fit between simulated values and the observations that are being matched by the regression. The purpose of regression is to calculate values of defined parameters that minimize the objective function; the resulting values are said to be "optimal," "optimized," or "estimated by regression." The weighted least-squares objective function S(b), used in UCODE and MODFLOWP can be expressed as:

$$S(\underline{b}) = \sum_{i=1}^{ND} \omega_i [y_i - y'_i(\underline{b})]^2 + \sum_{p=1}^{NPR} \omega_p [P_p - P'_p(\underline{b})]^2 \qquad (1)$$

where,

b is a vector containing values of each of the NP parameters being estimated;

ND is the number of observations (called N-OBSERVATIONS in the UCODE documentation);

NPR is the number of prior information values (called NPRIOR in the UCODE documentation);

NP is the number of estimated parameters (called N-PARAMETERS in the UCODE documentation);

y_i is the ith observation being matched by the regression;

$y'_i(\underline{b})$ is the simulated value which corresponds to the ith observation (a function of b);

P_p is the pth prior estimate included in the regression;

$P'_p(\underline{b})$ is the pth simulated value (restricted to linear functions of b in UCODE and MODFLOWP);

ω_i is the weight for the ith observation;

ω_p is the weight for the pth prior estimate.

The simulated values related to the observations are of the form $y'_i(\underline{b}) = f(\underline{b},\underline{\xi}_i)$, where $\underline{\xi}_i$ are independent variables such as location and time, and the function may be nonlinear in b and $\underline{\xi}_i$. Commonly, complex problems require numerical solution, and the function is actually a numerical model.

The simulated values related to the prior information are restricted in this work to be of the form $P'_p(\underline{b}) = \Sigma a_{pj} b_j$, which are linear functions of $\underline{b}$. Most prior information equations have only one term with a coefficient equal to 1.0, so the contribution to the objective function is simply the prior information value of a parameter minus its estimated value. Additional terms are needed when the prior information relates to a linear function that includes more than one parameter value. For example, additional terms are included in a ground-water inverse model to account for the following circumstances: seasonal recharge rates are estimated and measurements of annual recharge are available, so that the P_p value equals the seasonal recharge rate and the summation includes terms for the seasonal recharge rates; or storage coefficients in two model layers are estimated and an aquifer test was conducted that measured the combined storage coefficient, so that the Pp value equals the storage coefficient from the aquifer test, and the summation includes terms for the layer storage coefficients.

A simple problem and its weighted least-squares objective function surface are shown in figure 1. The figure was constructed by calculating equation 1 for this problem using different sets of parameter values T1 and T2. The log of the resulting numbers were contoured to produce the contour map of figure 1B. For a linear problem, the objective function surface would be a smooth bowl, and the contours would be concentric ellipses or parallel straight lines symmetrically spaced about a trough. The nonlinearity of Darcy's Law with respect to hydraulic conductivity results in the much different shape shown in figure 1B.

The differences $[y_i - y'_i(\underline{b})]$ and $[P_p - P'_p(\underline{b})]$ are called residuals, and represent the match of the simulated values to the observations. Weighted residuals are calculated as $\omega_i^{1/2}[y_i - y'_i(\underline{b})]$ and $\omega_p^{1/2}[P_p - P'_p(\underline{b})]$ and represent the fit of the regression in the context of how the residuals are weighted.

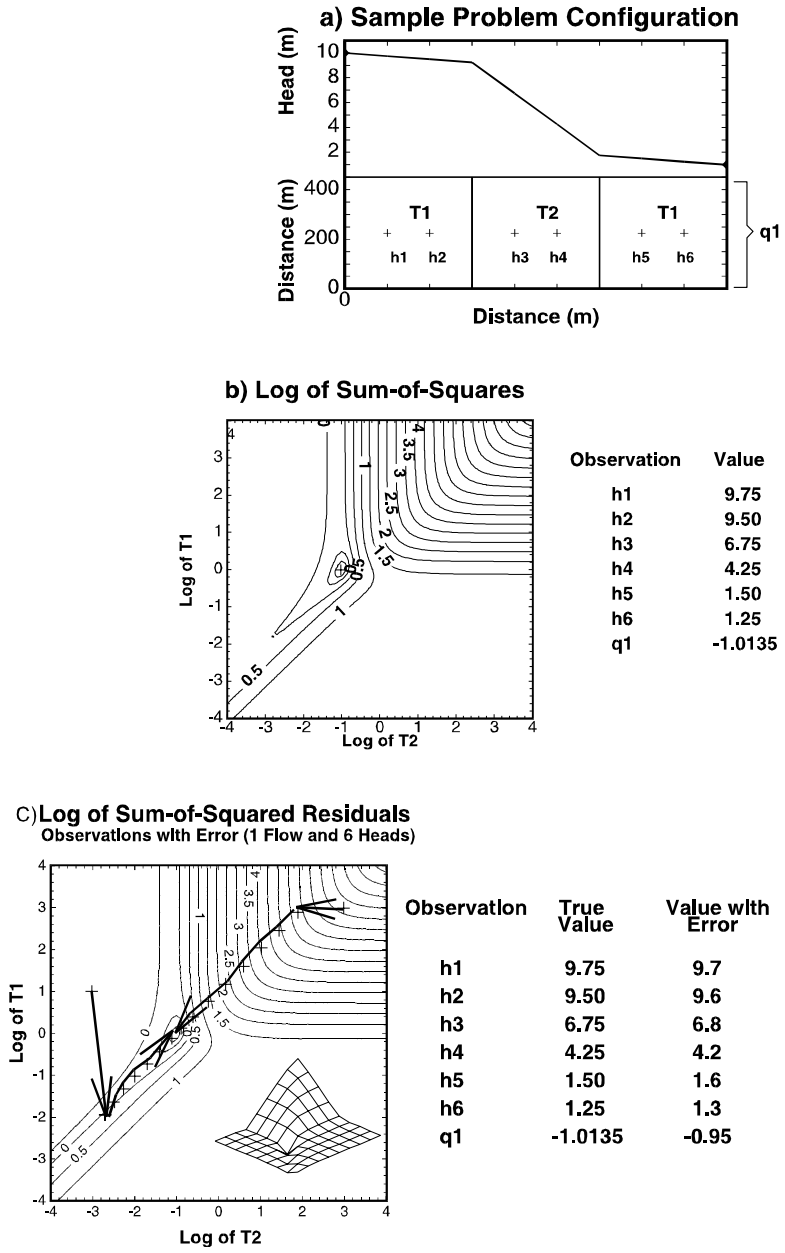


Figure 1: Objective function surfaces for a simple example. (a) The sample problem is a one-dimensional porous media flow field bounded by constant heads and consisting of three transmissivity zones and two transmissivity values. (b) Log of the weighted least-squares objective function that includes observations of hydraulic heads h1 through h6, in meters, and flow q1, in cubic meters per second. The observed values contain no error. (c) Log of the weighted least-squares objective function using observations with error, and a three-dimensional protrayal of the objective function surface. Sets of parameter values produced by modified Gauss-Newton iterations are identified (+), starting from two sets of starting values and progressing as shown by the arrows. (from Poeter and Hill, 1997)

In equation 1, a simple diagonal weight matrix was used to allow the equation to be written using summations instead of matrix notation. More generally, the weighting requires a full weight matrix, and equation 1 is written as:

$$S(\underline{b}) = [\underline{y} - \underline{y}'(\underline{b})]^T \underline{\omega}[\underline{y} - \underline{y}'(\underline{b})] = \underline{e}^T \underline{\omega} \underline{e} \qquad (2)$$

where the weight matrix, $\underline{\omega}$ and the vectors of observations and simulated values, $\underline{y}$ and $\underline{y}'(\underline{b})$ include terms for both the observations and the prior information, as displayed in Appendix A, and $\underline{e}$ is a vector of residuals. Full weight matrices are supported for most types of observations and prior information in MODFLOWP. With a full weight matrix, MODFLOWP calculates weighted residuals as $\underline{\omega}^{1/2}[\underline{y} - \underline{y}'(\underline{b})]$, where the square-root of the weight matrix is calculated such that $\underline{\omega}^{1/2}$ is symmetric.

An alternative derivation of the least-squares objective function involves the maximum-likelihood objective function. In practical application, the maximum-likelihood objective function reduces to the least-squares objective function (as shown in Appendix A), but the maximum-likelihood objective function is presented here and its value is calculated and printed by UCODE and MODFLOWP because it can be used as a measure of model fit (Carrera and Neuman, 1986; Loaiciga and Marino, 1986). The value of the maximum-likelihood objective function is calculated as:

$$S'(\underline{b}) = (ND+NPR) \ln 2\pi - \ln |\underline{\omega}| + (\underline{y} - \underline{y}')^T \underline{\omega}(\underline{y} - \underline{y}') \qquad (3)$$

where $|\underline{\omega}|$ is the determinant of the weight matrix, and it is assumed that the common error variance mentioned in Appebdix A and C equals 1.

Modified Gauss-Newton Optimization

The Gauss-Newton optimization method is an iterative form of standard linear regression, and works well only if modified by the addition of, for example, a damping parameter and a Marquardt parameter, as described below. The modified Gauss-Newton method presented here closely follows that of Cooley and Naff (1990, ch. 3), which is similar to methods presented by Seber and Wild (1989), and other texts on nonlinear regression.

Normal Equations and the Marquardt Parameter

Parameter values that minimize the objective function are calculated using normal equations. One of the differences between linear regression and nonlinear regression is that in linear regression parameter values are estimated by solving the normal equations once, while nonlinear regression is iterative in that a sequence of parameter updates is calculated, solving linearized nor-

mal equations once for each update. Thus, in nonlinear regression there are parameter-estimation iterations. The normal equations and the iterative process for the modified Gauss-Newton optimization method used in UCODE and MODFLOWP can be expressed as:

$$(\underline{C}^T\underline{X}^T{}_r\,\underline{\omega}\,\underline{X}_r\,\underline{C} + \underline{I}m_r)\underline{C}^{-1}\underline{d}_r = \underline{C}^T\,\underline{X}^T{}_r\,\underline{\omega}\,(\underline{y} - \underline{y}'(\underline{b}_r)) \qquad (4a)$$

$$\underline{b}_{r+1} = \rho_r\underline{d}_r + \underline{b}_r \qquad (4b)$$

where

r is the parameter-estimation iteration number;

$\underline{X}_r$ is the sensitivity matrix evaluated at parameter estimates $\underline{b}_r$, with elements equal to $\frac{\partial y'_i}{\partial b_j}$ (calculated by the sensitivity equation method in MODFLOWP and using forward or central differences in UCODE);

$\underline{\omega}$ is the weight matrix (can be a full matrix in MODFLOWP);

$(\underline{X}^T\underline{\omega}\,\underline{X})$ is a symmetric, square matrix of dimension NP by NP that is an estimate of the Fisher information matrix, and which is used to calculate statistics described in the section "Parameter Statistics";

$\underline{C}$ is a diagonal scaling matrix with element c_{jj} equal to $[(\underline{X}^T\underline{\omega}\,\underline{X})_{jj}]^{-1/2}$, which produces a scaled matrix with the smallest possible condition number (Forsythe and Strauss, 1955; Hill, 1990);

$\underline{d}_r$ is a vector with the number of elements equal to the number of estimated parameters. It is used in eq. 4b to update the parameter estimates;

$\underline{I}$ is an NP dimensional identity matrix;

m_r is the Marquardt parameter (Marquardt, 1963); and

ρ_r is a damping parameter.

Figure 1C shows the paths that this modified Gauss-Newton method followed from two sets of starting parameter values to the minimum of the objective-function surface of the simple example problem.

A quasi-Newton term can be added to the matrix on the left-hand side of equation 4a, as described in Appendix B, to aid convergence of the modified Gauss-Newton equations in some circumstances. The modified Gauss-Newton method used in this work also could be termed a Levenberg-Marquardt method.

The Marquardt parameter is used to improve regression performance for ill-posed problems (Theil, 1963; Seber and Wild, 1989). Initially m_r=0 for each parameter-estimation iteration r. For

iterations in which the vector $\underline{d}$ defines parameter changes that are unlikely to reduce the value of the objective function (as determined using the condition described by Cooley and Naff, 1990, p. 71-72), m_r is increased according to $m_r^{new} = 1.5\ m_r^{old} + 0.001$ until the condition is no longer met.

The damping parameter, ρ_r, can vary in value from 0.0 to 1.0 and modifies all values in the parameter change vector $\underline{d}_r$ by the same factor. Thus, in vector terminology, the direction of $\underline{d}_r$ is preserved. The damping parameter is used for two reasons: (1) to ensure that the absolute values of fractional parameter value changes are all less than a value specified by the user (MAX-CHANGE of UCODE; DMAX of MODFLOWP), and (2) to damp oscillations that occur when elements in $\underline{d}_r$ and $\underline{d}_{r-1}$ define opposite directions (Cooley, 1993), implemented as described in Appendix B. Fractional parameter value changes are calculated for each parameter as

$$(b_j^{r+1}-b_j^r) / |b_j^r| \qquad j=1,NP \qquad (5)$$

where b_j^r is the jth element of vector $\underline{b}_r$, that is, the value of the jth parameter at parameter estimation iteration r. If the largest absolute value of the NP values of equation 5 is greater than MAX-CHANGE (or DMAX for MODFLOWP), ρ_r is calculated in many circumstances as

As discussed by Cooley and Naff (1990, p.70), modified Gauss-Newton optimization typically converges within "a number of iterations equal to five or twice the number of parameters, whichever is greater." Convergence will tend to occur sooner for well-conditioned problems, and later for poorly conditioned problems. It is rarely fruitful to increase the number of iterations to more than twice the number of parameters, which can take large amounts of computer time. It generally is more productive to consider alternative models (See the guidelines discussed later in this report).

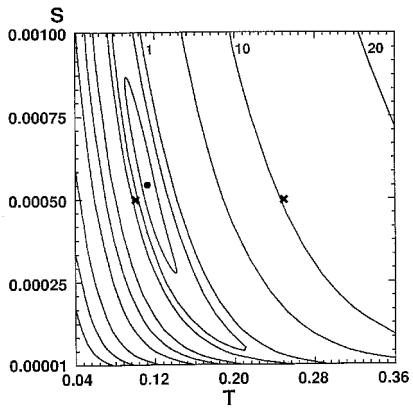
The performance of the modified Gauss-Newton method can be descibed using figure 2 which shows the effects of the linearization that occurs at each iteration of the modified Gauss-Newton method. The data shown in figure 2A represent ground-water level drawdown over time caused by pumpage from a single well. The model used is the Theis equation, which is a nonlinear functionof tranmissivity and the storage coefficient. In this problem, the nonlinear model $f(\underline{b},\underline{\xi})$, which was presented after equation1, is the Theis equation, the observations are the drawdowns listed in figure 2A, and the parameters to be estimated are the transmissivity and the storage coefficient.

(A)

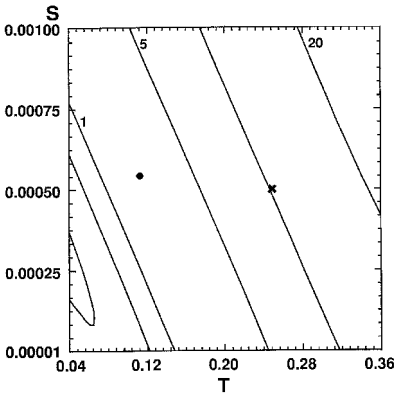
Time, in seconds	Drawdown, in feet
480	1.71
1020	2.23
1500	2.54
2040	2.77
2700	3.04
3720	3.25
4920	3.56

Pumpage = 1.16 ft^3/s
Distance from pumping to observation well = 175 ft

(B)



(C)



(D)

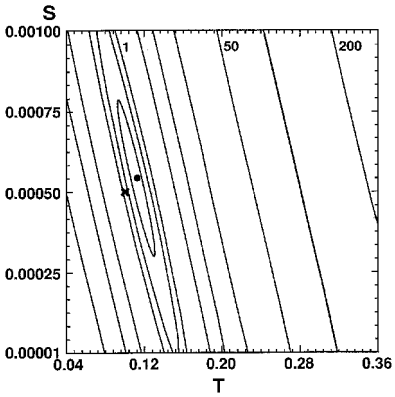


Figure 2: Objective-function surfaces for a Theis equation model. The system characteristics and ten observed drawdowns as reported by Cooley and Naff (1990, p.66) are shown in (A). The resulting nonlinear objective-function surface is shown in (B), with the minimum designated using a large dot. The same dot appears in (C) and (D). Objective-function surfaces for the same range of parameter values linearized using the Gauss-Newton approximation about the parameter values identified by the X's are shown in (C) and (D).

The actual, nonlinear, objective-function surface is shown in figure 2B. Approximations of the objective function surface produced by linearizing the model, here the Theis equation, about the parameter values marked by the x's are shown in figures 2C and 2D. The problem is linearized by replacing the model (here the Theis equation) with the first two terms of a Taylor series expansion, and using the linearized model to replace y'_i in equation 1. The mathematical form of the lin-

earized model is presented in Appendix C. Not surprizingly, the linearized surfaces approximate the nonlinear surface well near the parameter values for which the linearization occurs, and less well further away.

For each iteration of the modified Gauss-Newton method, the model is linearized either about the starting parameter values or the parameter values estimated at the last parameter-estimation iteration. Then, equation 4a is solved to produce a vector, $\underline{d}_r$,which generally extends from the set of parameter values about which the linearizaion occurs to the minimum of the linearized objective-function surface.

Stated anthropogenically, at the current set of parameter values, the regression "sees" a linearized objective-function surface and tries to change the parameter values to reach the minimum of that linearized surface. Figure 2C shows a linearized objective-function surface obtained by using a Taylor series expansion about a set of parameter values far from the minimum. The parameter values which minimize the linearized surface are far from those that minimize the nonlinear surface, so that proceeding all the way to the linearized minimum is likely to hamper attempts to find the minimum of the nonlinear surface. Proceeding part way to the linearized surface, however, could be advantageous. In figure 2C, moving all the way to the minimum of the linearized objective-function surface would produce a negative value of transmissivity, and the fractional change in the parameter value would exceed 1. In this circumstance, the damping parameter of the modified Gauss-Newton method, ρ_r in equation 4b, could be used to limit the change in the transmissivity value, or the transmissivity parameter could be log-transformed to ensure positive values, as discussed below.

Figure 2D shows an objective function surface obtained by linearizing about a point near the minimum and shows that a linearized model closely replicates the objective-function surface near the mimimum. This has consequences for the applicability of the inferential statistics, such as confidence intervals, discussed later in this report, and these consequences are briefly outlined here. If the designated significance level is large enough, the inferential statistics calculated using linear theory are likely to be accurate if the other required assumptions hold. As the significance level declines, a broader range of parameter values needs to be included in calculating the inferential statistics, and the more nonlinear parts of the objective-function surface become important. In that circumstance, the stated significance level of the linear inferential statistics becomes less reliable. Thus, a 90-percent confidence interval (10-percent significance level) might be well estimated using linear theory, while a 99-percent confidence interval (1-percent significance level) might not.

Convergence Criteria

Convergence criteria are needed for the modified Gauss-Newton iterative process. In

UCODE and MODFLOWP, parameter estimation converges if either one of two convergence criteria are satisfied. First, convergence is achieved when the largest absolute value of d_j^r/b_j^r, j=1,NP, is less than a user-defined convergence criterion (TOL of UCODE and MODFLOWP). That is,

$$|d_j^r / b_j^r| < TOL \quad \text{for all } j=1,NP \qquad (7)$$

where d_j^r is the jth element of $\underline{d}_r$, the parameter change vector of equation 4; b_j^r is the ith element of $\underline{b}_r$, the vector of parameter values being changed in equation 4; and NP is the number of estimated parameters. If b_j^r equals 0.0, 1.0 is used in the denominator. Preferably, this convergence criterion is satisfied by the final calibrated model with TOL assigned a value no larger that 0.01.

Second, the nonlinear regression converges if the sum of squared objective function (eq. 1 or 2) changes less than a user-defined amount (SOSR of UCODE and MODFLOWP) for three sequential iterations. This convergence criteria often is useful early in the calibration process to avoid lengthy simulations that fail to improve model fit.

Log-Transformed Parameters

The parameters in vector $\underline{b}$ of equation 1 can either be the native values directly relevant to the system being considered, or the log-transform of the native values. Log-transforming parameters can produce an inverse problem that converges more easily, and prevents the actual parameter values from becoming negative (Carrera and Neuman, 1986). In UCODE and MODFLOWP, the log-transform is implemented using the natural logarithm, but the input and output include base 10 logarithms because these are easier for most modelers to use (MODFLOWP was converted to the base 10 user interface in version 3.3).

UCODE and MODFLOWP are designed so that even when there are log-transformed parameter values, the user generally sees the more readily understood native values. Thus, for example, even when parameters are log-transformed, the starting parameter values specified by the user are native values. There are, however, three situations in which either model input or output are affected by a parameter being log-transformed.

The one model input situation occurs when there is prior information on the log-transformed parameter value, in which case there can only be one parameter included in the prior information (one term in the summation presented after eq. 1), and the specified statistic needs to be related to the base 10 log of the parameter. The statistic can be calculated using methods described under Guideline 6, described later in this report.

The first model output situation is fairly subtle and will not be noticed by most users. It involves calculation of the damping parameter and the convergence criteria of equation 4, which are calculated to control or measure the change in the native parameter values. The printed damping parameter value, therefore, can not always be derived easily by the user. Calculation of the damping parameter is described in Appendix B.

The second model output situation is that log-transformed parameter estimates, standard deviations, coefficients of variation, and confidence interval limits appear in the output file along with the exponential of these values. In most circumstances, the log-transformed values are ignored by the user and the native values are used instead. Related issues are discussed in the section "Standard Deviations, Linear Confidence Intervals, and Coefficients of Variation" later in this report.

Lack of Limits on Estimated Parameter Values

Upper and lower limits on parameters that constrain possible estimated values are commonly available in inverse models (for example, PEST, Doherty, 1994) and are suggested by, for example, Sun (1994, p. 35). While such limiting constraints on parameter values may, at first, appear to be necessary given the unrealistic parameter values that can be estimated through inverse modeling, Hill and others (1998), using a complex synthetic test case, demonstrate that this practice can disguise more fundamental modeling errors. Poeter and Hill (1996) use a simple synthetic test case to further demonstrate the concept, and in Anderman and others (1996), unrealistic optimized values of recharge in a field problem revealed important model construction inaccuracies. As discussed in the section "Guideline 5: Use Prior Information Carefully," unrealistic estimated parameter values are likely to indicate either (1) that the data do not contain enough information to estimate the parameters, or (2) there is a more fundamental model error. In the first circumstance, the best response is to use prior information on the parameter value, which will tend to produce an estimate that is close to the most likely value, instead of at the less likely value that generally constitutes the imposed upper or lower limit. In the second circumstance, the best response is to find and resolve the error. UCODE, like MODFLOWP, does not support constraining limits on parameter values because a circumstance in which the use of such limits is the best way to proceed has not been identified.

Weights for Observations and Prior Information

Observations and prior information typically need to be weighted in the regression. In most circumstances, diagonal weight matrices are used, and it is useful to introduce weighting in this simpler context. Hill (1992) presents a detailed discussion of the assumptions implied by using a diagonal weight matrix.

Weighting performs two related functions. The most fundamental function is that the weighting needs to produce weighted residuals that have the same units so that they can be squared

and summed using equation 1 or 2. Obviously, summing numbers with different units produces a nonsensical result. The second function of the weighting is to reduce the influence of observations that are less accurate and increase the influence of observations that are more accurate.

These two functions are directly related to the theoretical requirement of the weighting, as derived in Appendix C. This requirement is that the weight matrix be proportional to the inverse of the variance-covariance matrix of the observation errors. For a diagonal weight matrix, this requirement means that the weights of equation 1 need to be proportional to 1 divided by the variance of the measurement error. More detail on how to determine values for weights and interpret regression results relative to the weighting is presented in the section "Guideline 6: Assign weights which reflect measurement errors."

Diagnostic and Inferential Statistics

A powerful aspect of using nonlinear regression is the useful statistics that can be generated. The statistics presented here can be used diagnostically to measure the amount of information provided by the data and to identify model error (bias), or to infer the uncertainty with which values are calculated. The statistics also can be used to determine what aspects of the model are important to predictions of interest. Difficulties common to nonlinear regression make diagnostic statistics invaluable to its success, and the diagnostic use of statistics is stressed in this work.

The sections below show how the statistics are calculated. Use of the statistics during regression is discussed in the following section "Guidelines for Effective Model Calibration," and example figures are provided there as listed in the following section “Example Figures.”

Statistics for Sensitivity Analysis

Dimesionless Scaled Sensitivities and Composite Scaled Sensitivities

When a diagonal weight matrix is used, the scaled sensitivities, ss_{ij} are calculated as in Hill (1992):

$$ss_{ij} = \left(\frac{\partial y'_i}{\partial b_j}\right) b_j \omega_{ii}^{1/2} \qquad (8)$$

where

y'_i is the simulated value associated with the ith observation;

b_j is the jth estimated parameter;

$\frac{\partial y'_i}{\partial b_j}$ is the sensitivity of the simulated value associated with the ith observation with respect to the

jth parameter, and is evaluated at $\underline{b}$;

$\underline{b}$ is a vector which contains the parameter values at which the sensitivities are evaluated. Because the problem is nonlinear with respect to many parameters of interest, the value of a sensitivity will be different for different values in $\underline{b}$; and

ω_{ii} is the weight of the ith observation.

Similar scaling was employed by Harvey and others (1996). These scaled sensitivities are dimensionless quantities that can be used to compare the importance of different observations to the estimation of a single parameter or the importance of different parameters to the calculation of a simulated value. In both cases, greater absolute values are associated with greater importance.

For the full weight matrix that can be used in MODFLOWP, the two indices on the weight matrix need to be different, and scaled sensitivities are calculated as:

$$ss_{ij} = \sum_{k=1}^{ND} \left[\left(\frac{\partial y'_k}{\partial b_j} \right) b_j \omega_{ik}^{1/2} \right] \qquad (9)$$

Composite scaled sensitivities are calculated for each parameter using the scaled sensitivities for all observations, and indicate the total amount of information provided by the observations for the estimation of one parameter. The composite scaled sensitivity for the jth parameter, css_j, is calculated as (Hill, 1992; Anderman and others, 1996; Hill and others, 1998):

$$css_j = \left[\sum_{i=1}^{ND} (ss_{ij})^2 \Big|_{\underline{b}} / ND \right]^{\frac{1}{2}} \qquad (10)$$

where ND is the number of observations being used in the regression and the quantity in parentheses equals the scaled sensitivities of equation 8 or 9. The composite scaled sensitivity was derived from a similar statistic used by R.L. Cooley (U.S. Geological Survey, written commun., 1988), equals a scaled version of the square root of the diagonal of the Fisher information matrix ($\underline{X}^T \underline{\omega}\, \underline{X}$), and is similar in form and function to the CTB statistic of Sun and Yeh (1990), but is scaled differently. The composite scaled sensitivity is independent of the observed values and, therefore, model fit.

One-percent Scaled Sensitivities

While dimensionless sensitivities are needed to compare the importance of different types of observations to the estimation of parameter values, for other purposes it is useful to have dimensional quantities. One-percent scaled sensitivities are calculated as

$$dss_{ij} = \frac{\partial y'_i}{\partial b_j} \frac{b_j}{100} \qquad (11)$$

and approximately equal the amount that the simulated value would change if the parameter value increased by one percent.

The one-percent scaled sensitivities cannot be used to form a composite statistic because they generally have different units. Also, the omission of the weighting from equation 11 means that the one-percent scaled sensitivities do not reflect the influence of the observations on the regression as well as the dimensionless scaled sensitivities. The omission of the weighting has an advantage, however, in that one-percent scaled sensitivities can be calculated easily for any simulated quantity without having to assign the weighting. In MODFLOWP, sensitivities for hydraulic heads are calculated for the entire grid because sensitivities are calculated using the sensitivity equation sensitivities. These sensitivities are used to calculate one-percent scaled sensitivities for hydraulic heads that can be contoured just like hydraulic heads can be contoured. The resulting one-percent scaled sensitivity maps can be used to identify where additional observations of hydraulic head would be most important to the estimation of different parameters and to compare the sensitivity of hydraulic heads throughout the model to different paramters. Similar maps can be produced by UCODE by defining every point (or many points) in the model grid as an observation. For analytical models, they would be defined at enough points to create an accurate map.

Prediction Scaled Sensitivities

Sensitivities can be calculated for simulated predictions as dz'_l/db_j, where z'_l is the simulated prediction. These sensitivities indicate the importance of the parameter values to these predictions, and can be scaled to produce statistics by which to compare the relative importance of different parameters. One useful scaling results in a statistic that indicates the percentage change in the prediction produced by a one-percent change in the parameter value. This is defined in this work as a prediction scaled sensitivitiy (pss_j), and is calculated as:

$$pss_j = (dz'_l/db_j)\ (bj/100)\ (100/z'_l) \qquad (12)$$

The prediction scaled sensitivity is the one-percent sensitivity of equation 11 calculated for predictions, and multiplied by the last term of equation 12.

A different scaling is needed when $z'_l = 0.0$, or if the change in the predictive quantity relative to, perhaps, a regulatory limit, is of interest. In such circumstances, the predictive scaled senstivity can be calculated as:

$$pss_j = (dz'_l/db_j)\ (bj/100)\ (100/a'_l) \qquad (13)$$

The resulting statistic is the change in the prediction caused by a one-percent change in the parameter value, expressed as a percentage of a_l. The value used for a_l could be the regulatory limit or another number that was relevant to a given situation.

Prediction scaled sensitivities are not calculated by UCODE or MODFLOWP, but can easily be calculated from one-percent sensitivities calculated for predictions. For UCODE, this is accomplished using PHASE=44; for MODFLOWP this is achieved by following the directions for the post-processing program YCINT (Hill, 1994).

In some circumstances the prediction on interest is the difference between two simulations. For example, in ground-water problems, the prediction of interest is often the drawdown or change in flow to a stream caused by pumpage. These were termed differences by Hill (1994), and UCODE and MODFLOWP are designed to calculate sensitivities related to differences. For UCODE, this is accomplished using PHASE=45; for MODFLOWP, this is achieved as described for YCINT in Hill (1994).

There are two points that need to be considered when calculating prediction scaled senstivities. First, generally it is important to calculate prediction scaled sensitivities for the parameters which were not estimated by regression during model calibration well as the paramters that were estimated by regression. Parameters that could not be estimated by regression because of insensitivity, as indicated by composite scaled senstivities, can be important to preditions, and prediction scaled senstivities are likely to display that importance.

Second, often it is important to calculate the prediction scaled sensitivities for other sets of parameter values besides the optimal parameter values. This tests the robustness of the conclusions drawn from the prediction scaled sensitivities with respect to model nonlinearity.

Statistical Measures of Overall Model Fit

Model fit is evaluated by considering the magnitude of the weighted and unweighted residuals (defined after eq. 1) and their distribution both statistically and relative to independent variable values such as location and time. The first step generally is searching the table of residuals and weighted residuals printed by UCODE or MODFLOWP for the largest (in absolute value) residuals and weighted residuals. In initial model runs, these largest residuals and weighted residuals can indicate gross errors in the model, the data, or how the observed quantity is simulated, and(or) the weighting. In subsequent model runs, after the gross errors have been corrected, the following statistics become increasingly important.

Objective-Function Values

The value of the weighted least-squares objective function often is used informally to indi-

cate model fit. It is rarely used for more formal comparisons because its value nearly always decreases as more parameters are added, and the negative aspect of adding parameters is not reflected. The negative aspect of adding parameter values is that as the data available for the estimation get spread over more and more parameter values the certainty with which the parameter values are estimated decreases. The measures presented below more effectively account for this circumstance.

Calculated Error Variance and Standard Error

A commonly used indicator of the overall magnitude of the weighted residuals is the calculated error variance, s^2, which equals:

$$s^2 = \frac{S(\underline{b})}{(ND + NPR - NP)} \qquad (14)$$

where $S(\underline{b})$ is the weighted least-squares objective function value of equation 1 or 2 and the other variables are defined after equation 1. The square root of the calculated error variance, s, is called the standard error of the regression and also is used to indicate model fit. Smaller values of both the calculated error variance and the standard error indicate a closer fit to the observations, and smaller values are preferred as long as the weighted residuals do not indicate model error (see below).

If the fit achieved by regression is consistent with the data accuracy as reflected in the weighting, the expected value of both the calculated error variance and the standard error is 1.0. This can be proven by substituting equation 2 into equation 4 and taking the expected value. For non-statisticians, it may be more convincing to perform a similar calculation using generated random numbers instead of residuals. Assuming a diagonal weight matrix, this can be accomplished using any software package that can generate random numbers and perform basic calculations. Simply do the following: (1) Generate n random numbers using any distribution (such as normal, uniform, and so on). These are equivalent to the residuals of equation 1 or 2. (2) Square each random number. (3) Divide each squared number by the variance of the distribution used. If weights are defined to be one divided by the variances, these numbers are equivalent to squared weighted residuals. (4) Sum the numbers from (3) and divide by n. (5) Compare this value to 1.0. As n increases, the value should approach 1.0.

Significant deviations of the calculated error variance or the standard error from 1.0 indicate that the fit is inconsistent with the weighting. For the calculated error variance, significant deviations from 1.0 are indicated if the value 1.0 falls outside a confidence interval constructed using the calculated variance. The confidence interval limits can be calculated as (Ott, 1993, p.332):

$$\frac{ns^2}{\chi_u^2} ; \frac{ns^2}{\chi_L^2} \qquad (15)$$

where,

n is the degrees of freedom, here equal to ND+NPR-NP (See equation 1 for definitions);

χ_u^2 is the upper tail value of a chi-square distribution with n degrees of freedom, with the area to the right equal to one-half the significance level of the confidence interval (the significance level is 0.05 for a 95-percent interval);

χ_L^2 is the lower tail value of a chi-square distribution with n degrees of freedom with the area to the left equal to one-half the significance level.

The calculated standard error can be evaluated similarly by taking the square root of the limits of equation 5. Equivalently, the test can be conducted using a χ^2 test statistic, as presented by Ott (1993, p.234).

Values of the calculated error variance and the standard error are typically greater than 1.0 in practice, reflecting the presence of model error as well as the measurement error typically represented in the weighting, or larger than expected measurement error (see Guideline 8).

When the weight matrix is defined as suggested in Guideline 4, the calculated error variance and standard error are dimensionless.The dimensionless standard error is not a very intuitively informative measure of goodness of fit. A more intuitive measure is the product of the calculated standard error and the statistics used to calculate the weights (generally standard deviations and coefficients of variation; see the discussion for guideline 4). Such products are called fitted standard deviations and fitted coefficients of variation by Hill and others (1998) and in general can be called fitted error statistics. These statistics clearly represent model fit both to modelers and resource managers. For example, if a standard deviation of 0.3 m is used to calculate the weights for most of the hydraulic-head observations and the calculated standard error is 3.0, the fitted standard error of 0.9 m accurately represents the overall fit achieved for these hydraulic heads. If a coefficient of variation of 0.25 (25 percent) is used to calculate weights for a set of springflow observations and the calculated standard error is 2.0, the fitted coefficient of variation of 0.50 (50 percent) accurately represents the overall fit achieved to these springflows. Generally this approach applies only if the fitted error statistic summarizes the fit to a fairly large number of observations. Application to a single observation can produce misleading results.

The AIC and BIC Statistics

The calculated error variance and standard error are sometimes criticized for not sufficiently representing the drawbacks associated with increasing the number of estimated parameters. The AIC and BIC statistics were developed in the time-series literature to address this criticism (Brock-

well & Davis, 1989). To reflect the fact that adding too many parameters produces unreliable parameter estimates, the following two statistics equal the sum of the maximum-likelihood objective function (eq. 3) evaluated at the optimal parameter values, $S'(\underline{b}')$, and terms that become large as more parameters are added. Although these statistics were developed for time-series problems, Carrera and Neuman (1986) successfully used them to discriminate between different parameterizations of a test case of ground-water flow. The statistics are stated below; see the cited references for their derivations and additional discussion.

The statistic AIC was developed by Akaike (1974) and equals:

$$AIC(\underline{b}') = S'(\underline{b}') + 2 \times NP. \quad (16)$$

The statistic BIC also was developed by Akaike (1978) as a response to concern that AIC sometimes promoted use of more parameters than was required. The version of this statistic used by Carrera and Neuman (1986) is:

$$BIC(\underline{b}') = S'(\underline{b}') + NP \times \ln(ND+NPR). \quad (17)$$

For both statistics, smaller values indicate a more accurate model. If the statistics for a model with fewer parameters are only slightly larger than the statistics of another model, however, it may be better to select the model with fewer parameters, unless the investigator has other information indicating the validity of the more complicated model.

Graphical Analysis of Model Fit and Related Statistics

The assumed model can be analyzed to determine if the simulated dependent-variable values indicate a valid regression using the methods described below. The methods were suggested for ground-water inverse modeling by Cooley and Naff (1990), using the work of Draper and Smith (1981), and are discussed in Hill (1992, 1994). Data files to support these analyses are produced by UCODE and MODFLOWP. Examples of many of the graphs described are presented later in this report in the context of the guideline it is likely to support, as listed in the section "Example Figures."

Weighted Residuals Versus Weighted Simulated Values and Minimum, Maximum, and Average Weighted Residuals

It can be shown (Draper and Smith, 1981) that, in most situations, weighted residuals and weighted simulated values should be independent, so it is informative to consider graphs that assess the independence of these two variables. Ideally, weighted residuals are scattered evenly about 0.0, and their size is not related to the simulated values. In ground-water problems, for example, ideally the weighted residuals are not consistently larger in areas of high hydraulic head than in areas of low hydraulic head. Examples of such graphs, a discussion of the theory behind them, and some situations in which adjustments are needed because the two data sets are not independent are

presented in Hill (1994).

Statistics printed by UCODE and MODFLOWP that summarize the distribution of the weighted residuals are the minimum, maximum, and average weighted residuals, and the observations for which the minimum and maximum weighted residuals occur. The minimum and maximum weighted residuals display the range of weighted residuals at a glance. In practice, especially in the initial stages of calibration, the minimum and maximum weighted residuals often identify observations that are misrepresented in the simulation, suffer from incorrect data interpretation, or simply have not been entered incorrectly. The average weighted residual is a simple arithmetic average of the weighted residuals and ideally equals zero. In linear regression the average always equals zero for the optimized parameter values; in nonlinear regression the value of the average weighted residual generally approaches zero as calibration proceeds.

Weighted Observations Versus Weighted Simulated Values and Correlation Coefficient R

Ideally, simulated values are close to observations, so that weighted simulated values are close to weighted observations. When weighted observations are plotted against weighted simulated values, the hope is that the points fall close to a line with slope equal to 1.0 and an intercept of zero. A summary statistic that reflects how well this is accomplished is the correlation coefficient between the weighted observations and the weighted simulated values. The correlation coefficient, R, is calculated as (Cooley and Naff, 1990, p. l66):

$$R = \frac{(\underline{\omega}^{1/2}\ \underline{y} - \underline{m}_y)^T(\underline{\omega}^{1/2}\ \underline{y}' - \underline{m}_{y'})}{[(\underline{\omega}^{1/2}\ \underline{y} - \underline{m}_y)^T(\underline{\omega}^{1/2}\ \underline{y} - \underline{m}_y)(\underline{\omega}^{1/2}\ \underline{y} - \underline{m}_{y'})^T(\underline{\omega}^{1/2}\ \underline{y} - \underline{m}_{y'})]^{1/2}} \quad (18)$$

where $\underline{y}$, $\underline{y}'$, and $\underline{\omega}$ were defined for equation 2. $\underline{m}_y$ and $\underline{m}_{y'}$ are vectors with all ND elements equal to:

$$m_{y_i} = \sum_{q=1}^{ND} (\underline{\omega}^{1/2}\underline{y})_q / ND \quad (19)$$

$$m_{y'_i} = \sum_{q=1}^{ND} (\underline{\omega}^{1/2}\underline{y}')_q / ND \quad (20)$$

Thus, m_{y_i} is simply a vector with each component equal to the average of the weighted dependent-

variable observations, and $m_{y'_i}$ is an analogous vector using the weighted simulated values. Generally, R needs to be greater than 0.90. When there is prior information, R also is calculated with $\underline{y}$, $\underline{y}'$, and $\underline{\omega}$ augmented with prior information as in Appendix A, in which case ND+NPR replaces ND when calculating m_{y_i} and $m_{y'_i}$.

Graphs Using Independent Variables and the Runs Statistic

Evaluating weighted and unweighted residuals and weighted and unweighted observations and simulated values as related to the independent variables of a problem, such as space and time, is crucial. Ideally, weighted residuals plotted on maps or time graphs such as hydrographs show no discernible patterns and appear random. Distinct patterns can indicate significant model error that may make simulated predictions incorrect and misleading. Distinct patterns often are present, however, especially in time graphs.

A summary statistic that checks for the randomness of weighted residuals is the runs test (Cooley, 1979; Draper and Smith, 1981, p. 157-162). A sequence of residuals of the same sign is called a run, and the number of runs is counted and the value assigned to the variable u. For example, for the sequence of numbers -5, -2, 4, 3, 6, -4, 2, -3, -9, has the five runs (-5,-2), (4,3,6), (-4), (2), (-3, -9), so that u=5. Using the total number of positive residuals (n_1), and the total number of negative residuals (n_2), u can be defined as a random variable that depends on the order of the negative and positive values. It can be shown that if n1>10 and n2>10, u is normally distributed with mean, μ, and variance, σ^2, equal to:

$$\mu = [2n_1n_2/(n_1+n_2)]+1.0, \quad (21)$$

$$\sigma^2 = [2n_1\, n_2\, (2n_1n_2-n_1-n_2)]/[(n_1+n_2)^2(n_1+n_2-1)]. \quad (22)$$

The actual number of runs in a data set is compared with the expected value using test statistics. The test statistic for too few runs equals

$$z_f = (u-\mu+0.5)/\sigma; \quad (23)$$

the test statistic for too many runs equals

$$z_m = (u-\mu-0.5)/\sigma. \quad (24)$$

Critical values for z_f and z_m are printed by UCODE and MODFLOWP. Otherwise, critical values can be determined from a normal probability table available in most statistics texts.

In UCODE and MODFLOWP, the weighted residuals are analyzed using the sequence in which the observations are listed in the input file. The runs test is included because it takes the order of the residuals into account, which is ignored in all the other summary statistics. If observations are grouped by location in transient simulations, too few runs commonly indicate positive serial correlation between residuals at individual locations.

Normal Probability Graphs and Correlation Coefficient R^2_N

For a valid regression, the errors in the observations and the prior information used in the regression need to be random and the weighted errors need to be uncorrelated (Draper and Smith, 1981). In addition, inferential statistics such as confidence intervals generally require that the observation errors be normally distributed (Helsel and Hirsch, 1992). The actual errors are unknown, so the weighted residuals are analyzed. If the model accurately represents the actual system and the observation errors are random and the weighted errors are independent, the weighted residuals are expected to either be random, independent, and normally distributed, or have predictable correlations. The first step is to determine whether the weighted residuals are independent and normally distributed. If they are not, further analysis is needed to determine if the violations are consistent with the expected correlations.

The test for independent, normal weighted residuals is conducted using normal probability graphs of weighted residuals. Such graphs can be constructed as discussed by Hill (1994), using files created by UCODE or MODFLOWP. The files are designed so that the graphs can be constructed using commonly available x-y plotting software using arithmetic axes. If the weighted residuals are independent and normally distributed, they will fall on an approximately straight line in the normal probability graph. The associated summary statistic is R^2_N the correlation coefficient between the weighted residuals ordered from smallest to largest and the order statistics from a N(0,1) probability distribution function (Brockwell and Davis, 1987, p. 304). This statistic tests for independent, normally distributed weighted residuals and was chosen instead of other statistics, such as chi-squared and Kolomogorov-Smirnov, because it is more powerful for commonly used sample sizes (Shapiro and Francia, 1972). The correlation coefficient is calculated as:

$$R^2_N = \frac{[(\underline{e}_o - \underline{m})^T \underline{\tau}]^2}{[(\underline{e}_o - \underline{m})^T (\underline{e}_o - \underline{m})](\underline{\tau}^T \underline{\tau})} , \qquad (25)$$

where all vectors are of length ND for R^2_N evaluated only for the observation weighted residuals, and length ND+NPR for R^2_N evaluated for the observation and prior information weighted residuals; $\underline{m}$ is a vector with all components equal to the average of the weighted residuals, $\underline{e}_o$ is a vector of weighted residuals ordered from smallest to largest, and $\underline{\tau}$ is a vector with the ith element equal to the ordinate value of a N(0,1) probability distribution function for a cumulative probability equal to $u_i = (i-0.5)/ND$. A normal probability table (as in Cooley and Naff, 1990, p. 44, or any standard statistics text) can be used to determine that, for example, if u = 0.853l, then $\tau_i = 1.05$. UCODE and MODFLOWP print the ordered weighted residuals of $\underline{e}_o$ and R^2_N.

If R^2_N is too much less than its ideal value of 1.0, the weighted residuals are not likely to

be independent and normally distributed. The critical values for R^2_N for significance levels 0.05 and 0.10 are shown in Appendix D and the relevant critical values are printed by MODFLOWP and UCODE with R^2_N.

Determining Acceptable Deviations from Independent Normal Weighted Residuals

Weighted residuals may fail the tests described above because there are too few weighted residuals or because of the fitting process of the regression, and not because the model is inadequate. The fitting process can produce correlations between residuals that, for example, result in a normal probability graph of weighted residuals that is not linear. This can be tested by generating values that conform to the expected violations, as described by Cooley and Naff (1990, p. 176). The steps involved in the test are as follows.

1. Sets of normally distributed random numbers are generated with and without the regression-induced correlations.
2. Normal probability graphs of the weighted residuals are compared with graphs of the independent normally distributed random numbers (called d's by Cooley and Naff, 1990). If similar deviations from a straight line are apparent in these graphs, it can be concluded that the nonlinear shape of the weighted residuals graphs could result from too few weighted residuals.
3. Normal probability graphs of the weighted residuals are compared with graphs of the correlated normally distributed numbers (called g's by Cooley and Naff, 1990). If similar deviations from a straight line are apparent in these graphs, it can be concluded that the nonlinear shape of the weighted residuals graphs could result from the regression-induced correlation.

A computer program for generating the random numbers was presented by Cooley and Naff (1990), and was slightly modified for use with MODFLOWP, as discussed by Hill (1992); the computer program is called RESANP. An output file is produced by MODFLOWP that can be used as the input file to RESANP; UCODE includes RESANP as a subroutine, and files with the random numbers are produced when input variables are set appropriately.

Parameter Statistics

Although composite scaled sensitivities are good measures of the information the data contain for a single parameter, they do not reflect that there are many parameters being estimated simultaneously, and they do not reflect the actual precision of the parameter estimates. The following statistics fill these roles. In addition, one of them, the correlation coefficient, is independent of model fit, an attribute it shares with the composite scaled sensitivities. This attribute is used extensively in Guideline 3 later in this report.

Variances and Covariances

The reliability and correlation of parameter estimates can be analyzed by using the variance-covariance matrix, $\underline{V}(\underline{b}')$, for the final estimated parameters, $\underline{b}'$ (Bard, 1974, p. 59), calculated as:

$$\underline{V}(\underline{b}') = s^2(\underline{X}^T \underline{\omega}\, \underline{X})^{-1} \qquad (26)$$

where $\underline{V}(\underline{b}')$ is an NP by NP matrix; s^2, the calculated error variance, is calculated using equation 14, and $\underline{X}$ (which is calculated using the optimal parameters $\underline{b}'$) and $\underline{\omega}$ are augmented to include sensitivities and weights for prior information on the parameters (Appendix A). The diagonal elements of matrix $\underline{V}(\underline{b}')$ equal the parameter variances; the off-diagonal elements equal the parameter covariances. For a problem with three estimated parameters, the matrix would appear as:

$$\begin{matrix} \text{var}(1) & \text{cov}(1,2) & \text{cov}(1,3) \\ \text{cov}(2,1) & \text{var}(2) & \text{cov}(2,3) \\ \text{cov}(3,1) & \text{cov}(3,2) & \text{var}(3) \end{matrix} \qquad (27)$$

where var(1) is the variance of parameter 1, cov(1,2) is the covariance between parameters 1 and 2, and so on. The variance-covariance matrix is always symmetric, so that cov(1,2)=cov(2,1), and so on. The utility of equation 26 depends on the model being nearly linear in the vicinity of $\underline{b}'$ and on the appropriate definition of the weight matrix. The source of these restrictions is presented in the proofs of Appendix C.

While equation 26 equals the variance-covariance of the parameter estimates only if evaluated for the optimal parameter values, the calculation can be done for any set of parameter values, and some of the statistics calculated using this matrix are very useful for diagnosing problems with the regression (Anderman and others, 1996; Poeter and Hill, 1997; and Hill and others, 1998). To be concise in the present work, the matrix of equation 26 and statistics derived from it will be referred to by the same names used when evaluated at the optimal parameter values. In practice, it is important to indicate whether the parameter values used for the calculation are optimal or not. Statistics derived from the variance-covariance matrix on the parameters that are printed by UCODE and MODFLOWP are described in the following sections.

Two variations on the variance-covariance matrix of equation 26 are important. First, equation 26 usually is evaluated using the parameters estimated by regression, and the resulting parameter variance-covariance matrix is the one printed at the end of the regression. In many situations, however, some parameters are excluded from the regression because of insensitivity and(or) nonuniqueness, as determined using the sensitivity measures discussed above and the correlation coefficients presented below. These parameters are, therefore, excluded from calculation of the parameter variance-covariance matrix. It is important, however, to periodically calculate sensitivities and the variance-covariance matrix for all parameters to reevaluate insensitivity and nonuniqueness, and to evaluate the parameter from the perspective of predictions. This can be accomplished easily using UCODE and MODFLOWP by activating unestimated parameters and adding prior information on these parameters if available. Then, sensitivities can be calculated once, the sensitivity matrix ($\underline{X}$) augmented to include senstivities for the unestimated parameters, and equation 26 calculated using the augmented sensitivity matrix. This is accomplished by replacing the starting parameter values with the final parameter values for both UCODE and MOD-

FLOWP, and by using PHASE 22 of UCODE or by setting IPAR=1, TOL=$1x10^{6}$ and DMAX=$1x10^{-6}$ in MODFLOWP. In this work, this variation is called the parameter variance-covariance matrix for all parameters.

A second variation of the variance-covariance matrix of equation 26 can be used to determine if parameters that are highly correlated given the observations used in the regression are also highly correlated relative to the predictions of interest. This is important to determining whether parameters are estimated adequately given the desired predictions, as discussed in Guideline 14. This variation of equation 26 requires that the sensitivity and weight matrices be augmented to include predictions. This change can be implemented easily when using UCODE or MODFLOWP by adding the predictions to the list of observations using the method described above for the first variation and the suggestions discussed in the following paragraph. In this work, this second variation is called the parameter variance-covariance matrix with predictions.

The value specified for the prediction as the 'observed value' does not affect the calculated prediction correlation coefficients, but the weight does. It is possible to establish a value for the weight based on three logical arguments. First, the weight can be estimated based on expected measurement error, as was done for observations (see guideline 4). Second, the weight can be estimated using a statistic that reflects an acceptable range of uncertainty in the prediction (This is more consistent with the scaling of the CTB statistic of Sun and Yeh, 1990). Third, it may be useful to decrease the value of the statistic specified for the weight so that the value of the weight and the dominance of the predictive quantity is increased. The third option ensures that the predictions are not overwhelmed by the other data. To ensure that the result is correct for all of the predictions, predictions can be included individually or in groups, depending on the problem.

Standard Deviations, Linear Confidence Intervals, and Coefficients of Variation

Parameter standard deviations equal the square root of the parameter variances. Parameter standard deviations are perhaps most useful when used to calculate two other statistics: confidence intervals for parameter values and coefficients of variation. Linear confidence intervals calculated as described by Hill (1994) and references cited therein require trivial amounts of execution time and are calculated and printed by UCODE and MODFLOWP. The more accurate nonlinear confidence intervals of Vecchia and Cooley (1987) and Christensen and Cooley (1996) discussed below in section "Nonlinear Confidence and Prediction Intervals" require substantial execution time and are not calculated by the current versions of UCODE or MODFLOWP.

A linear confidence interval for each parameter β_j is calculated as

$$b_j \pm t\left(n, 1.0 - \frac{\alpha}{2}\right) s_{b_j} \qquad (28)$$

where

$t\left(n, 1.0 - \frac{\alpha}{2}\right)$ is the Student-t statistic for n degrees of freedom and a significance level of α; and

s_{b_j} is the standard deviation of the jth parameter.

Confidence intervals are referred to in a way that can be confusing, and that is derived from their definition. Technically, a confidence interval is a range that has a stated probability of containing the true value. As such, confidence intervals are referred to using the true, unknown value that is being estimated. Thus, equation 28 is said to be the confidence interval for β_j, the true, unknown jth parameter value, and the width of the confidence interval can be thought of as a measure of the likely precision of the estimate. Narrow intervals indicate greater precision. If the model correctly represents the system, the interval also can be thought of as a measure of the likely accuracy of the estimate. This was discussed in more detail by Hill (1994).

The derivation of equation 28 requires an assumption that is not needed to perform the regression -- that is, the assumption that the true errors and, therefore, for a linear problem, the parameter estimates, be normally distributed. For further discussion, see the section "Normal Probability Graphs and the Correlation Coefficient R^2_N."

When plotted on graphs with the related estimated values, linear confidence intervals provide a vivid graphical image of the precision with which parameters are estimated using the data included as observations in the regression, given the constructed model.

The coefficient of variation for each parameter equals the standard deviation divided by the parameter value and provides a dimensionless number with which the relative accuracy of different parameter estimates can be compared.

For log-transformed parameters, confidence intervals and coefficients of variation of the transformed parameters can be difficult to interpret. In UCODE and MODFLOWP the confidence intervals are untransformed by taking the exponential of the confidence interval limits, and these are printed. The coefficients of variation are untransformed by untransforming the parameter variance, $(s_{\log b})^2$ as:

$$s_b^{\,2} = \exp\left[2.3\left(s_{\log b}\right)^2 + 2\log b\right]\left[\exp\left(2.3\left(s_{\log b}\right)^2\right) - 1.\right] \qquad (29)$$

where the exponentials and logs are in base 10, *b* is the native parameter, and log*b* is the estimated log-transformed parameter. The coefficient of variation of the native parameter is calculated by di-

viding the square root of its variance by *b*. It should be noted that the linear confidence intervals for the true, unknown native parameters are symmetric when plotted on a log scale, but are not symmetric when plotted on an arithmetic scale.

Correlation Coefficients

Correlation coefficients are calculated as the covariance between two parameters divided by the product of their standard deviations. Using the notation of equation (27), the correlation between the ith and jth parameter is calculated as:

$$cor(i,j) = \frac{cov(i,j)}{var(i)^{1/2} var(j)^{1/2}} \qquad (30)$$

Correlation coefficients range in value from -1.0 to 1.0, with values close to -1.0 and 1.0 indicative of parameter values that cannot be uniquely estimated with the observations used in the regression. Poeter and Hill (1997) provide a good description of correlation coefficients calculated for a simple test case; Anderman and others (1996) show how they can be used to evaluate the worth of different kinds of observations. Correlation coefficients are typically displayed as a matrix, such as:

$$\begin{matrix} 1.0 & 0.96 & 0.05 \\ 0.96 & 1.0 & 0.46 \\ 0.05 & 0.46 & 1.0 \end{matrix} \qquad (31)$$

Correlation coefficient matrices are always symmetric and the diagonal elements always equal 1.0. Correlation coefficients can be calculated using any of the variations of the parameter variance-covariance matrix discussed above. Correlation coefficients calculated using the parameter variance-covariance matrix with all parameters are called correlation coefficients for all parameters; correlation coefficients calculated using the parameter varience-covariance matrix with predictions are called prediction correlation coefficients.

Influence Statistics

While dimensionless scaled sensitivities indicate the importance of an observation to the estimation of a parameter, the actual effect of the observation in the regression also depends on calculated residuals. The Cook's D and DFBETA influence statistics incorporate this effect. The Cook's D statistics can be calculated for each observation as described by Cook and Weisberg (1982) and Helsel and Hirsch (1992). DFBETAs are calculated for each parameter j and each observation i .

Anderman (1996) and Yager (in press) show how the DFBETA influence statistic can be

used to identify the interaction between the observations and estimated parameter values in ground-water problems.

Prediction Uncertainty

The uncertainty with which predictions are simulated can be approximated using confidence and prediction intervals. Confidence intervals are for the true, unknown predictions, which are not random variables, and result from the uncertainty with which the parameters are estimated, as represented by the variance-covariance matrix on the parameters (eq. 26). Prediction intervals also account for random measurement error in the quantity for which the interval is constructed, and are needed to construct an interval for an anticipated measurement of the prediction. Confidence and prediction intervals are discussed in many texts, such as Seber and Wild (1989) and Helsel and Hirsch (1992), as well as by Cooley and Naff (1990) and Hill (1994).

Linear Confidence and Prediction Intervals

Approximate linear confidence and prediction intervals for predictions can be calculated using output files produced by MODFLOWP and computer program YCINT of Hill (1994), or by setting the input variables appropriately for UCODE, in which YCINT has been converted to a subroutine. Linear confidence intervals are calculated as:

$$z'_l \pm t_s\left(n,\, 1.0 - \frac{\alpha}{2}\right) s_{z'_l} \qquad (32)$$

where z'_l is the lth simulated value;

$t_s\,(n,\, 1.0\text{-}\alpha/2)$ is the critical value, and equals the value for which there is an $\alpha/2$ probability that a student-t distributed random variable would be larger;

n is the degrees of freedom, here equal to ND+MPR-NP;

α is the significance level and is commonly 0.05 or 0.10 (5 and 10 percent), and

$s_{z'_l}$ is the standard deviation of the prediction, calculated as

$$s_{z'_l} = \left[\sum_{i=1}^{NP} \sum_{j=1}^{NP} \frac{\partial z'_l}{\partial b_j} \underline{\underline{V}}(\underline{b}') \frac{\partial z'_l}{\partial b_i} \right]^{\frac{1}{2}} . \qquad (33)$$

The calculated confidence interval is said to have a (1-α) probability of including the true value of the predicted quantity. Corresponding to the values noted above, 90- and 95-percent confidence intervals are the most common.

Approximate linear prediction intervals are calculated as:

$$z'_l \pm t\left(n, 1.0 - \frac{\alpha}{2}\right)\left(s_{z'_l} + s_a\right) \qquad (34)$$

where s_a is the standard error of the regression adjusted for the expected measurement error of the prediction (see Hill, 1994, p. 32; Miller, 1981).

Individual confidence intervals calculated using equation 32 are exact for linear models with normally distributed residuals, assuming that the model is correct. As these conditions are violated to a greater degree, the calculated intervals become progressively less accurate, so that the actual significance level of the interval can be substantially different than intended. This is of serious concern for the nonlinear problems considered in this work, as discussed by Donaldson and Schnabel (1987). Recent publications in the ground-water literature, however, indicate that in many ground-water flow problems linear intervals are accurate enough to be useful (Christensen and Cooley, in press). Other types of ground-water problems have not been evaluated.

The calculation of confidence and prediction intervals can (and often needs to) include more parameters than were included in the regressions performed for model calibration, as discussed above in the section 'Prediction Scaled Sensitivities' and under guideline 13.

The individual intervals defined above apply when the uncertainty of only one quantitiy is of interest. When more than one quantity is of interest, different intervals are needed, and these are called simultaneous intervals. Simultaneous intervals calculated using linear theory are always of equal size or larger than equivalent linear individual intervals, reflecting the greater uncertainty involved in trying to define intervals which are likely to include the true value of two or more predictions at the same time. As more intervals are considered, the intervals tend to become wider. The largest intervals are calculated when the number of predictions equals the number of parameters included in the uncertainty analysis. Additional predictions do not increase the size of the simultaneous intervals.

Simultaneous intervals are difficult to calculate exactly, but can be approximated using the equations listed in Hill (1994), as discussed by Miller (1981). The equations for simultaneous confidence and prediction intervals are of the same form as equations 32 and 34, respectively, and differ only in the critical values used. If the number of intervals considered is represented by k, the interval limits can be calculated using critical values from a Bonferroni-t distribution or from an F distribution. The Bonferroni critical value is

$$t_B(n,1.0-\alpha/2k). \qquad (35a)$$

The F distribution critical value is

$$[d \times F_\alpha(d,n)]^{1/2}, \qquad (35b)$$

where d equals k or the number of parameters (NP), whichever is less. Intervals calculated with the

F distribution critical value are called Scheffe intervals. Scheffe intervals are labeled either as Scheffe d=k or Scheffe d=NP.

Both Bonferroni and Scheffe intervals are approximate, and tend to be large. Thus, for any finite value of k, the smaller interval should be used.

In some cases k is not finite. For example, if a prediction of interest is the largest simulated value over a defined area, the predicted quantity can not be specified exactly before the simulation, and the number of predictions considered simultaneously needs to be thought of as infinite. In this circumstance, the only applicable approximate simultaneous interval is the Scheffe d=NP.

As discussed in the section "Prediction Scaled Sensitivities", in some circumstances the prediction on interest is the difference between two simulations. Both UCODE and MODFLOWP can calculate linear confidence and prediction intervals on differences, as discussed by Hill (1994).

Calculation of linear confidence intervals requires only the sensitivities calculated for the optimized parameter values and, therefore, takes very little computer execution time.

Nonlinear Confidence and Prediction Intervals

Accurate evaluation of parameter and prediction uncertainty for nonlinear models requires nonlinear confidence and prediction intervals, as discussed by Veccia and Cooley (1987), Cooley (1997) and Christensen and Cooley (in press). Calculation of nonlinear confidence intervals requires the equivalent of a full regression for each limit of each interval, so can entail substantial additional computer execution time. For many nonlinear problems, a practical approach is to calculate linear confidence and(or) prediction intervals, and then to calculate nonlinear intervals for selected predictions. Unfortunately, nonlinear intervals are not calculated with the present versions of UCODE and MODFLOWP.

Testing for Linearity

The methods presented in this section are only applicable if the model is sufficiently linear. Although the modified Gauss-Newton optimization method and many of the statistical methods discussed are useful even for problems which are quite nonlinear, more stringent requirements on linearity are needed for the linear confidence and prediction intervals to adequately represent parameter and prediction uncertainty. The assumption of linearity upon which the linear confidence intervals are based can be tested using the modified Beale's measure (also called Linssen's measure) described by Cooley and Naff (1990) and Hill (1994). Although the modified Beale's measure indicates nonlinearity of the confidence region of the parameters, and does not directly measure nonlinearity of confidence intervals, no better indicator of nonlinearity is available. Increasingly problematic situations occur as predictive quantities or situations differ more from cal-

ibration observations and situations.

The modified Beale's measure can be calculated using MODFLOWP and the computer program BEALEP of Hill (1994); a slightly modified version of BEALEP can be executed by UCODE using PHASE=33. Many practical problems are nonlinear, in which case the linear intervals are inaccurate.

Example Figures

Table 1 lists most of the statistics and graphical analyses discussed in this section of the report, and the figures and guidelines in which they are presented and discussed in the next section. Note that the use of these statistics and graphs is not restricted to the suggested application.

Table 1: Statistics and graphical analyses, related figures, and the guidelines in which the figures are presented. [1]

Statistic or graph (ordered by function)	Figure[2]	Guideline
Sensitivity Analysis Statistics		
Dimensionless scaled sensitivities	13, Table 3	11
Composite scaled sensitivity	3, 4, 16, 17, Table 3	3, 11, 14
One-percent scaled sensitivity map	none [3]	11
Parameter correlation coefficients[4]	5, 6, 16	3, 14
Linear confidence intervals on parameters	10, 14, 16	9, 13, 14
Model Fit Statistics and Graphical Analysis		
Fitted error statistics	11	10
Graph of weighted residuals versus weighted simulated values	12	10
Graphs using independent variables	7, 8	8
Runs test	9	8
Normal probability graphs	15	13
Evaluate Estimated Parameter Values		
Compare estimated parameter values with reasonable ranges	10	9
Linear confidence intervals on parameters	10, 14, 16	9, 13, 14
Parameter correlation coefficients	5, 6, 16	3, 14
Influence statistics	none [5]	
Evaluate Predictions		
Prediction scaled sensitivities	16, 17	14
Linear and nonlinear confidence intervals on predictions	none [6]	

1. The statistics and graphs often are useful for other guidelines as well. See Table 2.
2. Unless otherwise indicated.
3. No example is provided in this report.
4. Repeated because of their frequent application for two purposes.
5. No example is provided in this report. See Anderman (1996) and Yager (in press).
6. No example is provided in this report. An example is shown by Christensen and Cooley (in press).

GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

A clear, thorough discussion of an entire modeling protocol is presented by Anderson and Woessner (1992, p. 4-9). The guidelines presented here fit into that protocol, enhancing the calibration, prediction, and uncertainty analysis phases, and emphasizing the testing of different conceptual models. Preliminary steps of the protocol include identifying the purpose of the model and selecting or programming a model with the appropriate capabilities, and the guidelines presented in this work assume these have been accomplished.

Ideally, the model is constructed and the data are collected with the purpose of the model in mind, with the evolving model used to guide data collection efforts. Formally using the model in these effort is complicated because, as noted by Sun (1994, p. 210), there is an inherent difficulty associated with the optimal design of experiments for nonlinear problems, i.e., the solution of optimal design depends on the values of the unknown parameters. In addition, in the three-dimensional, transient problems common to many fields, evolution of the conceptual models may be significant, and new data may challenge previous conceptual models, as well as change the optimized parameter values. Sun (1994) presents some elegant methods of addressing this problem; those presented here tend to be simpler, and, in some circumstances, may serve as preliminary steps to a more sophisticated evaluation.

To ensure that a reasonably accurate model is used to guide data collection, the guidelines presented in this work do not suggest using the model to evaluate potential new data or to formally consider the desired prediction until Guidelines 12 and 14, respectively. This is not intended to diminish the importance of considering these issues throughout data collection and model development, but to provide steps by which the available data can be used to develop a model that is as accurate as possible for each phase of the analysis. Once a reasonable model is developed, it may be used to visit previously considered guidelines. Thus, the guidelines are not intended to be followed sequentially once, but may be repeated many times during model calibration.

The guidelines are summarized in table 1 and are explained further in the text. The guidelines are presented in the context of ground-water model calibration, but are applicable to other fields. Many aspects of the approach have had a long history in a variety of fields. The idea of starting simple and building complexity, emphasized in guideline 1, is discussed by Parker (1994), among others. The principle of parsimony and some of the other characteristics have been discussed or used by Cooley and others (1986), Constable and others (1987), Cooley and Naff (1990) and Parker (1994). Most of the graphical analyses of Guideline 8 were suggested for application to ground-water problems by Cooley and Naff (1990), as derived from Draper and Smith (1981). The approach developed by Hill and others (1998) is close to the approach presented here, and they test the approach using a complex synthetic test case. Simple synthetic test cases are used to demonstrate many aspects of the approach in Poeter and Hill (1996, 1997).

.

Table 1:

Guideline	Description
1. Apply the principle of parsimony	Start simple and add complexity as warranted by the hydrogeology and the inability of the model to reproduce observations.
2. Use a broad range of information to constrain the problem	For example, in ground-water model calibration, use hydrology and hydrogeology to identify likely spatial and temporal structure in, for example, areal recharge and hydraulic conductivity, and use this structure to limit the number of parameters needed to represent the system. Do not add features to the model to attain model fit if they contradict other information about the system.
3. Maintain a well-posed, comprehensive regression problem	a) Define parameters based upon their need to represent the system, within the constraint that the regression remains well-posed. Accomplish this using composite scaled sensitivities (css_j) and parameter correlation coefficients. b) Maintain a comprehensive model in which as many aspects of the system as possible are represented by parameters, and as many parameters as possible are estimated simultaneously by regression.
4. Include many kinds of data as observations in the regression	Adding different kinds of data generally provides more information about the system. In ground-water flow model calibration, it is especially important to provide information about flows. Hydraulic heads simply do not contain enough information in many circumstances, as indicated by the frequency with which extreme values of parameter correlation coefficients occur when using only hydraulic heads.
5. Use prior information carefully	a) Begin with no prior information to determine the information content of the observations. b) Insensitive parameters (parameters with small composite scaled sensitivities) can be included in regression using prior information to maintain a well-posed problem, but during calibration it often is advantageous to exclude them from the regression to reduce execution time. Include these parameters for Guidelines 13 and 14. c) For sensitive parameters, do not use prior information to make unrealistic optimized parameter values realistic.
6. Assign weights which reflect measurement errors	Initially assign weights to equal $1/\sigma_i^2$, where σ_i^2 is the best available approximation of the variance of the error of the ith measurement (This is for a diagonal weight matrix; see text for full weight matrix.)
7. Encourage convergence by making the model more accurate	Even when composite scaled sensitivities and correlation coefficients indicate that the data provide sufficient information to estimate the defined parameters, nonlinear regression may not converge. Working to make the model represent the system more accurately obviously is beneficial to model development, and generally results in convergence of the nonlinear regression. Use model fit and the sensitivities to determine what to change.

Table 1:

Guideline	Description
8. Evaluate model fit	Use the methods discussed in the sections "Statistical Measures of Model Fit" and "Graphical Analysis of Model Fit and Related Statistics".
9. Evaluate optimized parameter values	a) Unreasonable estimated parameter values could indicate model error. b) Identify parameter values that are mostly determined based on one or a few observations using dimensionless scaled sensitivities and influence statistics. c) Identify highly correlated parameters.
10. Test alternative models	Better models have three attributes: better fit, weighted residuals that are more randomly distributed, and more realistic optimal parameter values.
11. Evaluate potential new data	Use dimensionless scaled sensitivities, composite scaled sensitivities, parameter correlation coefficients, and one-percent scaled sensitivities. These statistics do not depend on model fit or, therefore, the possible new observed values.
12. Evaluate the potential for additional estimated parameters	Use composite scaled sensitivities and parameter correlation coefficients to identify system characteristics for which the observations contain substantial information. These system characteristics probably can be represented in more detail using additional estimated parameters.
13. Use confidence and prediction intervals to indicate parameter and prediction uncertainty.	a) Calculated intervals generally indicate the minimum likely uncertainty. b) Include insensitive and correlated parameters, perhaps using prior information, or test the effect of excluding them. c) Start by using the linear confidence intervals, which can be calculated easily. d) Test model linearity to determine how accurate these intervals are likely to be. e) If needed and as possible, calculate nonlinear intervals (This is not supported in the present versions of UCODE and MODFLOWP). f) Calculate prediction intervals to compare measured values to simulated results. g) Calculate simultaneous intervals if multiple values are considered or the value is not completely specified before simulation.
14. Formally reconsider the model calibration from the perspective of the desired predictions	Evaluate all parameters and alternative models relative to the desired predictions using prediction scaled sensitivities (pss_j), confidence intervals, composite scaled sensitivities, and parameter correlation coefficients.

From the perspective of stochastic inverse methods, the approach presented here can be thought of as a strategy to approximate the mean, or effective, values. Stochastic methods generally require that the mean of any spatially distributed quantity, such as hydraulic conductivity, be constant or a simple function. Unfortunately, geologic media often defy these limitations. The method presented here can be used to test whether the mean is constant, and, if not, to provide an estimate of what could be a very complex spatial distribution, often with sharp contrasts. Once these large-

scale variations are established, it may be useful to use stochastic methods to assess the influence of smaller scale variations. To date, methods of determining large-scale variations, such as those described in this work, and methods of charaterizing small-scale variations, such as stochastic methods, have been integrated very little, and this is an area for future research.

Guideline 1: Apply the principle of parsimony

Using the principle of parsimony, the model is kept as simple as possible while still accounting for the system processes and characteristics evident in the observations and while respecting other information about the system. In many fields, including ground-water hydrology, the known complexities of the systems being simulated often seem overwhelming, and being parsimonious in model development can require substantial restraint.

It is important to apply the principle of parsimony to various aspects of model construction and calibration. For example, it is important to use a mathematical model that is only as complex as is needed for the system being considered, or which is designed such that unneeded capabilities do not add complexity. It also is important to investigate the processes and characteristics that are likely to be most dominant first and add processes or complexity gradually, always testing the importance of the added complexity to the observations available for model calibration and the predictions of interest. For inverse modeling, it is important to begin calibration estimating very few parameters that together represent most of the features of interest and to increase the complexity of the parameterization slowly. The remaining guidelines suggest methods for accomplishing this.

Guideline 2: Use a broad range of information to constrain the problem

In most fields, there is information about the modeled system that cannot, given present methods, be directly included as observations in the regression. Effective use of this information can mean the difference between a parsimonious model that represents the system well and a parsimonious model that produces nonsense.

For example, if a ground-water model is to have any credibility, it must respect what is known about the hydrology and hydrogeology. Using hydrogeologic data to constrain model calibration is practical in many cases. Most ground-water problems consider relatively shallow geologic systems, and there is often substantial geologic data. This is in contrast to many fields of geophysics and other Earth sciences in which the depth of the region of interest precludes being able to constrain the calibration much with the known geology. Often, it is geologic data that allows useful well-posed ground-water inverse models to be developed, as suggested in guideline 3. Hydrogeologic data often indicate that sharp contrasts probably occur in the hydraulic-conductivity distribution, which need to be represented to simulate the ground-water system and which cannot usually be represented well by, for example, most geostatistical methods. A good example of using hydrologic and hydrogeologic data in ground-water flow model development of an incredibly complex system using geoscientific information systems (GSIS) is described by D'Agnese and others (1996, 1998, and in press). The GSIS approach can be described as a fully three-dimensional GIS

that is able to represent common geologic relationships such as faults and sequential layering. Other approaches have been suggested by Poeter and McKenna (1995), McKenna and Poeter (1995) and Eppstein and Dougherty (1996). This is an area ripe for further development.

There will inevitably be some overlap in the information used to constrain a problem as described in this guideline, and information used as prior information on parameters as discussed in Guideline 5. For example, the results of hydraulic tests may be used to determine that two hydrogeologic units have similar hydraulic-conductivity values and probably can be combined to form one parameter in the regression, producing what may be an important constraint on the problem. Later, the same results may be used to determine a prior information value for the combined or individual hydrogeologic units.

Guideline 3: Maintain a well-posed, comprehensive regression problem

A well-posed regression problem is one that will converge to an optimal set of parameter values given reasonable starting parameter values. Given commonly available data, the requirement of maintaining a well-posed regression produces rather simple models with relatively few estimated parameters. Often, however, it is this simple level of model complexity that can be supported by the data based on regression methods. Thus, determining the greatest possible level of model complexity while maintaining a well-posed regression can be thought of as an objective analysis of the information provided by the data. Prior information can be used to support additional complexity (See Guideline 5). Developing simplifications that produce a meaningful model is difficult and requires the constraints discussed in Guideline 2.

Hydrologic and hydrogeologic information, and composite scaled sensitivities and parameter correlation coefficients, can be used to define parameters and to decide which parameters to estimate using regression. Composite scaled sensitivities and parameter correlation coefficients are well-suited for this purpose because they depend only on the sensitivities and are independent of the actual values observed. Evaluated for the starting parameter values, they can be used to determine what sets of parameters are likely to be estimated given a model and a set of observations (Anderman and others, 1996), as described in the following paragraphs.

If some parameters have composite scaled sensitivities that are less than about 0.01 times the largest composite scaled sensitivity, it is likely that the regression will have trouble converging. Often, it is useful to plot the composite scaled sensitivities as a bar chart, as in D'Agnese and others (1996,1998, in press) and Barlebo and others (1996; in press). The bar chart for starting parameter values used by D'Agnese and others (1998) shown in figure 3 indicates that the K4 and RCH parameters are likely to be easy to estimate by regression with this model, while the ANIV1 and ETM parameters are not. In general, it appears that the available observations contain substantial information about K (hydraulic conductivity) and RCH (areal recharge) parameters, and less information about ANIV (vertical anisotropy) and ETM (maximum evapotranspiration) parameters.

Composite scaled sensitivities were calculated often during model clibration and were used to determine what new parameters to introduce, and whether previously excluded parameters should be included. The composite-scaled sensitivities for the final model are shown in figure 4. Note that there are more K (hydraulic conductivity) and RCH (recharge) parameters, and that most of these were estimated by regression. This is consistent with the initial evaluation that the data contained substantial information for these types of parameters. There is one new type of parameter: GHB, which represents the hydraulic conductivity of the head-dependent boundary conditions being used to represent ground-water supported springs. None of the GHB parameters were estimated in the regression in the final model because they tended to produce a good match solely to the flow of the spring or set of springs at which they were applied, and any error in the spring flow measurement would be fit by the model through adjustment of the GHB parameters. Instead, their values were determined based primarily on hydrogeologic arguments.

Parameter correlation coefficients indicate whether the estimated parameter values are likely to be unique. For the parameters of figures 3 and 4, all correlation coefficients were less than 0.95, suggesting that uniqueness was not a problem. A situation in which uniqueness was a problem is presented by Anderman and others (1996), as displayed in figure 5. Figure 5 shows correlation coefficients calculated for initial parameter values for the same five parameters of the same model for three sets of observation data: (1) hydraulic heads only, (2) hydraulic heads and a lake seepage value, and (3) hydraulic heads, lake seepage, and an advective-travel observation. Figure 5 clearly shows that with only hydraulic heads (data set 1), all parameters are completely correlated (the absolute values of all correlation coefficients equal 1.0), so that any parameter estimates found by the regression are not unique. Adding one lake seepage measurement (data set 2) reduced correlations some, but only the data set including the advective-travel observation (data set 3) was sufficient to uniquely estimate all of the parameters.

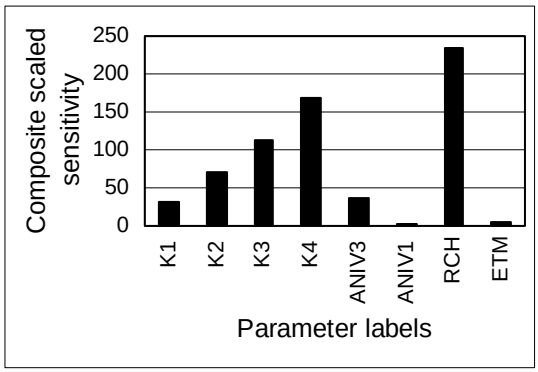


Figure 3: Composite scaled sensitivities for parameters of the initial Death Valley regional ground-water flow system model of D'Agnese and others (1998, in press). K* are hydraulic-conductivity parameters, ANIV* are vertical anisotropy parameters, RCH is an areal recharge parameter, and ETM is a maximum evapotranspiration parameter.

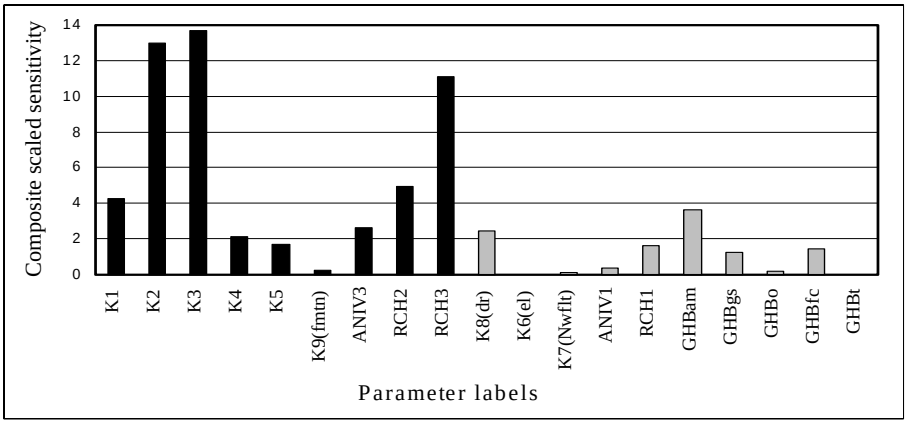


Figure 4: Composite scaled sensitivities for the parameters of the final calibrated Death Valley regional ground-water system model of D'Agnese and others (in press). K* are hydraulic-conductivity parameters, ANIV* are vertical anisotropy parameters, RCH is an areal recharge parameter, ETM is a maximum evapotranspiration parameter, and GHB* are parameters related to the conductance of head-dependent boundaries used to represent springs. Parameters estimated by regression have black bars; parameters defined but not estimated by regression have grey bars.

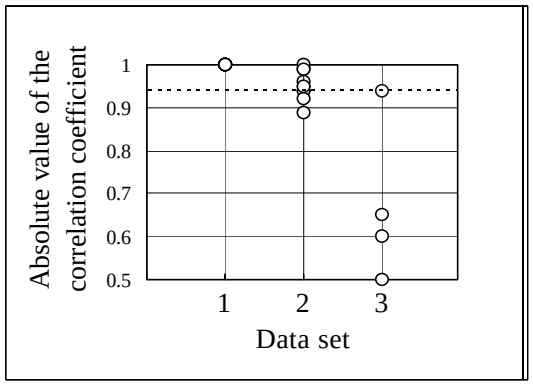


Figure 5: Parameter correlation coefficients for the same five parameters for three data sets from the Cape Cod sewage plume model of Anderman and others (1996), evaluated for the initial parameter values. Data set 1 includes only hydraulic heads, and all parameters are extremely correlated (the absolute value of all correlation coefficients equals 1.0). Data set 2 includes hydraulic heads and one flow observation, and many parameter pairs are still extremely correlated; data set 3 also contains an advective-travel observation, which reduced correlation considerably.

T1	T2	Correlation
10.0	0.01	0.20
10.0	10.00	-0.95
5.0	0.05	0.34
5.0	0.50	-0.05
*1.0	0.10	0.86
0.5	0.50	0.78
0.5	0.05	0.96
0.1	10.00	0.06
0.1	0.01	0.998

* true values

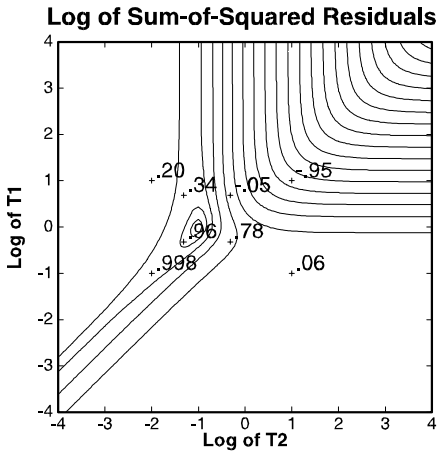


Figure 6: Correlation of parameters T1 and T2 of figure 1 at specified parameter values, plotted on a $\log_{10}$ weighted least-squares objective function surface. T1 and T2 are in square meters per day. (from Poeter and Hill, 1997)

Two concerns about using calculated correlation coefficients exist: the effects of model nonlinearity and inaccurate calculated senstivities. The first of these also affects composite scaled sensitivities.

The nonlinearity of inverse problems can make composite scaled sensitivities and correlation coefficients quite different for different sets of parameter values. Figure 6 demonstrates this for correlation coefficients calculated for the simple test case from figure 1. This figure shows that though there is a distinct minimum to this objective function surface, so that the parameters can

clearly be estimated uniquely, correlation coefficients close to 1.0 are calculated for some sets of parameter values. For most sets of parameter values, however, the values are significantly less that 1.0, correctly indicating that unique parameter values can be estimated. Thus, in this problem, the misleading results can be detected by calculating correlation coefficients for several sets of parameter values.

The effects of both nonlinearity and scaling by the parameter value also make composite scaled sensitivities different for different sets of parameter values. If the differences that occur for a reasonable range of parameter values are too extreme, composite scaled sensitivities are inadequate for the purposes they serve in the guidelines. Their utility can be tested by calculating values for several sets of parameter values. They have been useful in many ground-water flow and transport problems (Christiansen and others, 1995, Anderman and others, 1996; D'Agnese and other, 1996, 1988; Barlebo and others, 1996; Poeter and Hill, 1997; Hill and others, 1998).

The second concern about calculated correlation coefficients is that they can be substantially affected by sensitivities that are accurate to less than about four or five significant digits (O. Osterby, Aarhus University, Denmark, written commun., 1997). This is a more serious issue for UCODE, in which the sensitivities are calculated by less accurate difference methods, and can occur even when the more accurate central difference method is used to calculate sensitivities. It is important, therefore, to follow the suggestions provided in the UCODE documentation (Poeter and Hill,1998) to enhance sensitivity accuracy. Inaccurate sensitivities are less of a problem for MODFLOWP, which uses the sensitivity-equation method to calculate sensitivities.

UCODE and MODFLOWP calculate and print correlation coefficients and composite scaled sensitivities for the final parameter values of any run, whether the regression converges or not. Composite scaled sensitivities also can be printed at initial and intermediate parameter-estimation iterations.

Guideline 4: Include many kinds of data as observations in the regression

Guideline 4 stresses the importance of using as many kinds of observations as possible. For example, in ground-water flow problems, it is important to augment commonly available hydraulic-head observations with flow observations. The latter serve to constrain solutions much more than the relatively easy to fit hydraulic heads and, therefore, using observations that reflect the rate and(or) direction of ground-water flow tends to promote the development of more accurate models. MODFLOWP supports many types of observations relevant to ground-water flow problems, such as hydraulic heads, temporal changes in hydraulic head, streamflow gains and losses, and advective travel (Hill, 1992; Anderman and Hill, 1997). An advantage of UCODE is that it allows any quantity to be used as an observation for which a simulated equivalent value is printed in any application model output file, or for which a simulated equivalent value can be calculated from the values printed in any application model output file. A detailed analysis of the importance of different types of observations and how to conduct such an analysis is presented by Anderman and others (1996).

In some circumstances, it may appear that guideline 4 could be addressed by using contoured values to increase the number of observations. In a ground-water example, Neuman (1982), Clifton and Neuman (1982), Neuman and Jacobson (1984), and Carrera and Neuman (1986) used kriging to interpolate hydraulic-head measurements to generate hydraulic heads used in the regression. When kriging is used, the associated kriging variances and variogram can be used to calculate the variance-covariance matrix on hydraulic-head observation errors needed to calculate the weighting. The advantage of interpolation methods is that more hydraulic-head values are available for the regression. As shown by Cooley and Sinclair (1976) and noted by Hill (1992), the disadvantage of interpolation methods is that the interpolated hydraulic heads are not based on the physics of ground-water flow, so that interpolated values generally do not respect the underlying processes represented in the model. This problem can be severe if aquifer properties change rapidly because the interpolation method would tend to make the ‘observed’ hydraulic-head distribution unrealistically smooth. Use of interpolated values in the regression procedure produces correlation between the errors, so use of a full weight matrix may be important. These problems are avoided if the observations are used directly in the regression.

Guideline 5: Use prior information carefully

Using prior information allows direct measurements of model input values to be included in the regression. Prior information is treated differently than observations in this work because relevant observations generally can be measured more accurately than model-input values. Indeed, that is the most fundamental characteristic of the problems considered in this work. If the measurements of the model input values were accurate and applicable to the scale of the model, model calibration would be unnecessary or less important. Thus, it is suggested that the generally more accurate observations be emphasized more than the relatively less accurate prior information. Prior information takes on an important, but less central role in the suggested methodology. For problems with more accurate prior information, the prior information might be treated more like the ob-

servation data are treated here.

Initially omitting prior information on parameters from the regression encourages understanding of the information directly available from the observations. Two reasons generally would motivate the use of prior information. First, if the sensitivity for a parameter is low, as indicated by a small composite scaled sensitivity, regression including the parameter often will not converge. Two possibilities generally exist: specify prior information on the parameter or set the parameter value so that it is not changed during the regression (which is roughly equivalent to prior information with a very large weight). Specifying prior information usually will result in a parameter estimate that is close to the value specified in the prior information, so that the estimate will be equal or close to the prior value regardless of which option is chosen. Execution time is less when the parameter value is set because this eliminates the need to calculate sensitivities for the parameter, so it is suggested that this option be followed for model calibration. This will continue to be the best option as long as the parameter remains insensitive, which can be checked during calibration by occasionally calculating composite scaled sensitivities for the estimated parameters and the parameter in question. An exception to this guideline occurs when the user purposely defines more parameters than can be directly supported by the data to represent suspected system complexity, and this generally requires substantial use of prior information to obtain a well-posed regression problem. An example of this use of prior information and its effect on model accuracy is presented in a synthetic test case by Hill and others (1998).

The other common reason for using prior information on parameters is when the parameter value estimated by the regression is unreasonable. This problem is discussed in the previous section of this report titled "Lack of Limits on Estimated Parameter Values." As noted there, the most productive response to this problem depends on the amount of information the observations provide on the parameter in question. If little information is provided, the problem falls into the category of insensitive parameters, and the guidelines discussed in the paragraph above apply. If substantial information is provided, the unrealistic estimated parameter value is likely to indicate problems with the model or the data, as discussed by Anderman and others (1996) and Poeter and Hill (1996). To determine whether enough information is provided by the observations such that the unrealistic estimated parameter value indicates a problem with the model or the observations, the linear confidence interval on the parameter can be considered. If the confidence interval includes no realistic parameter values, the unrealistic estimate is likely to indicate problems with the model or the observations. If the confidence interval includes realistic parameter values, it is not clear whether there is a problem with the model or the data. Examples of the first circumstance are described by Anderman and others (1996), Poeter and Hill (1996), and Hill and others (1998). An example of the latter circumstance is described by Christiansen and others (1995) and Barlebo and others (in press) for a problem in which only hydraulic-head observations are used. In that application, addition of concentration observations produced more realistic parameter values, indicating that the problem was primarily due to inadequate data. UCODE and MODFLOWP prints linear

confidence intervals on the parameter values (eq. 28).

Guideline 6: Assign weights that reflect measurement errors

The weights are an important part of the regression, and assigning appropriate values can be confusing. The guideline presented here has a solid statistical basis and provides substantial guidance in most circumstances. For regression methods to produce parameter estimates with the smallest possible variance, the weighting needs to be proportional to the inverse of the variance-covariance matrix of the measurement errors (Appendix C). For a diagonal weight matrix, this means that the weights need to be proportional to one divided by the variance of the measurement errors. This definition of the weights results in two consequences that have substantial intuitive appeal: (a) Relatively accurate measurements are weighted more heavily than relatively inaccurate measurements, and (b) although different observations may have different units, weighted quantities have the same units and can, therefore, be summed in equation 1 or 2. Based on this guideline, information independent of the model is used to determine the weights, so that issues related to the weights are less likely to obscure model error or problems related to the data.

For problems with observations of a simgle type and measured with apparently equal error, on average, it generally is easiest to set all weights equal to 1.0, as was done for the Theis problem of figure 2. In this situation, the calculated error variance has the units of the observations.

For problems with more than one kind of observation, as well as prior information on the parameters, it is more convenient to define the weighting to equal the inverse of the variance-covariance matrix of the measurement errors instead of being proportional to it (Hill and others, 1998). This guideline encourages the user to compare the weights used to what the weights should be theoretically. If it is suspected that another weighting is needed to achieve, for example, randomly weighted residuals at optimal parameter values, this can be tested and placed in context relative to the assumed measurement error statistics. In addition, the assumed statistics of the measurement errors can be compared with the fit to the data achieved by the regression to provide a check on the weights used, as discussed under guideline 8.

UCODE and MODFLOWP read statistics from which the variances of the observation errors and then the weights are calculated. The statistics can equal the variance, standard deviation, or coefficient of variation of the measurement error of the observations or prior information. Values for these statistics rarely are known in practice. Although assignment of values for the statistics, therefore, is subjective, in most circumstances the estimated parameter values and calculated statistics are not very sensitive to moderate changes in the weights used. Several examples of using commonly available data to determine weights are described in the following paragraphs. MODFLOWP also allows a full weight matrix, with covariances as well as variances, to be used. The following examplesfocus primarily on determining the more commonly used diagonal weighting,

but one example of determining covariances is presented.

The statistics used to calculate the weights often can be determined using readily available information and a simple statistical framework. For example, in a ground-water problem, consider an observation well for which the elevation was determined by an altimeter and is considered to be accurate to within 3 ft. To estimate the variance of the measurement error, this statement needs to be quantified to, for example, the probability is 95 percent that the true elevation is within 3 ft of the measured elevation. If the measurement errors are assumed to be normally distributed, a table of the cumulative distribution of a standardized normal distribution (Cooley and Naff, 1990, p. 44, or any basic statistical text, such as Davis, 1986) can be used to determine the desired statistics as follows.

1. Use the table to determine that a 95-percent confidence interval for a normally distributed variable is constructed as the measured value plus and minus 1.96 times the standard deviation of the value.
2. As applied to the situation here, the 95-percent confidence interval is thought to be plus and minus 3 ft, so that 1.96 x s_{y_i} = 3.0 ft, or s_{y_i} = 1.53, where s_{y_i} is the estimated standard deviation.

In UCODE and MODFLOWP, the standard deviation (1.53 ft) can be specified and the variance will be calculated, or the variance (2.34 ft^2) can be specified. If elevations of wells are obtained from U.S. Geological Survey (USGS) topographic maps, the accuracy standards of the USGS can be used to quantify errors in elevation. The USGS (1980, p. 6) states that on their topographic maps, "...not more than ten percent of the elevations tested shall be in error more than one-half the contour interval." If this were thought to be the dominant measurement error, a 90-percent confidence interval would be plus and minus one-half the contour interval. Assuming that the error is normally distributed, a 90-percent interval is constructed by adding and subtracting 1.65 times the standard deviation of the measurement error. Thus, the standard deviation of the measurement error can be calculated as one-half the contour interval divided by 1.65, or (contour interval)/(2 x 1.65). The value of 1.65 was obtained from a normal probability table.

A similar procedure can be used for observations that are a sum or difference between measured values. For example, consider streamflow measurements between two gaging stations. In ground-water modeling, often it is the difference between the two flow measurements that is used as an observation in the regression, and these are called streamflow gain or loss observations. Consider a situation in which the upstream and downstream streamflow measurements are 3.0 ft^3/s and 2.5 ft^3/s, so that there is a 0.5 ft^3/s loss in streamflow between the two measurement sites. Also assume that the measurements are each thought to be accurate to within 5 percent (using, for example, Carter and Anderson, 1963), and the errors in the two measurements are considered to be independent. Stated quantitatively, perhaps the hydrologist is 90 percent certain that the first measurement is within 0.15 ft^3/s (5 percent) of the true value, and 95 percent certain that the second

measurement is within 0.125 ft^3/s (5 percent) of the true value. Assuming that the errors are independent and normally distributed, the standard deviation of the first measurement is calculated using the method described above from $1.65\, s_{q_1} = 0.15$ ft^3/s, so $s_{q_1} = 0.091$. The standard deviation of the second measurement is calculated from $1.96\, s_{q_2} = 0.125$ ft^3/s, so $s_{q_2} = 0.064$. The uncertainty of the difference between the two flows needs to be calculated using their variances, which can be calculated by squaring the standard deviations to produce $s^2_{q_1} = 0.0083$ $(ft^3/s)^2$ and $s^2_{q_2} = 0.0041$ $(ft^3/s)^2$. The variance of the loss of 0.5 ft^3/s equals $s^2_{q_1} + s^2_{q_2} = 0.0124$ $(ft^3/s)^2$. The coefficient of variation (standard deviation, $0.0124^{1/2}$, divided by the loss, 0.5 ft^3/s) for the loss in streamflow is, therefore, 0.22, or 22 percent. In UCODE and MODFLOWP, the variance, standard deviation, or coefficient of variation could be specified by the user. The choice generally is based on convenience.

In some circumstances there is a series of measurements from which differences are calculated. For example, there may be three streamflow measurements, q1, q2, and q3, along the length of a stream with gains or losses produced by subtracting each measurement from the next downstream measurement, resulting in two gain/loss observations, q2-q1 and q3-q2. The errors in the two differences are not statistically independent because the error in q2 is included in both differences. Hill (1992) shows that in this circumstance the covariance between the two differences equals the negative of the variance of the q2 measurement. This covariance cannot be included in UCODE, which is restricted to a diagonal weight matrix that includes only the variances of the measurement errors. Christensen and others (in press) extended the results of Hill (1992, p. 43) to measurements along branching streams, and S. Christensen extended MODFLOWP to include full weight matrices. It was found, however, that inclusion of the off-diagonal covariance terms in the weight matrix had negligible effect on the regression or statistical analysis in the problem considered (S. Christensen, 1997, oral commun.). Ignoring the covariances as is required in UCODE, and as is often done in applications of MODFLOWP, is not expected to effect results substantially in many circumstances.

The methods presented above also can be used to determine weighting for prior information, but there are two additional issues to consider. First, if the weighting is determined using the arguments presented above, the prior information fits into the framework of either classical or Bayesian statistics, the later being the framework from which the term prior information originates. Sometimes, however, larger weights (smaller statistics) are assigned to the prior information to achieve a stable regression, in which case the term regularization needs to be used instead of prior information (Hill and others, 1998; Backus, 1988). Setting parameter values to constants that are not changed by the regression can be thought of as an extreme case of regularization. When regu-

larization is used, confidence intervals on parameters and predictions may not represent model uncertainty accurately. Thus, classifying what is called prior information throughout this work as either prior information or regularization is very important.

The second issue unique to prior information occurs when the associated parameter is log-transformed. In this situation, the statistic on the prior information needs to relate to the log of the parameter value. The methods discussed above are directly applicable, but an extra step is needed because it is easier to establish a range of plausible values for native than for transformed values. Thus, if the prior estimate for a hydraulic conductivity is 100 m/d, and the true value is expected to fall within a range of 80 to 120 m/d with a certainty of about 95 percent, a 95-percent confidence interval for the native value has approximate limits of 80 and 120. Taking the log (base 10) of these values produces limits of 1.90 and 2.08, about a prior estimate of 2.0. If it is assumed that the uncertainty in the hydraulic conductivity can be approximated by a log-normal distribution, the log-transformed value is normally distributed. Changing the limits 1.90 and 2.08 slightly to form a symmetric interval with limits 1.91 and 2.09, the methods described above can be used to determine that the standard deviation relevant to the log-transformed parameter equals 0.045, and this value would need to be used as the statistic.

It generally is impossible to identify all measurment errors that contribute to an observation or prior information value, and the variances, standard deviation, and coefficients of variation calculated by using the methods discussed in this section are clearly approximate. Indeed, a problem related to Guideline 6 as described above is what should be included in the so-called "measurement errors". While this point can be argued extensively, a definition that has proven to be useful for the purpose of determining weighting is that measurement error is error related to any aspect of the measurement not accounted for by the model considered. Unambiguous types of measurement errors are errors in the measuring device and the location of the measurement in three-dimensional space. Ambiguous contributions include, for example, heads measured in wells that only partially penetrate the numerical layer to which they are assigned. This is more ambiguous because the model could be refined to accommodate this, and it could be debated whether this is model error or measurement error. Despite such ambiguities, the above definition for measurement error works relatively well in practice, partly because the regression often is not very sensitive to the weighting used, and the definition is sufficient to produce weighting based on common sense that is at least approximately correct.

A final useful aspect of defining the weighting as described here was discussed previously in the section “Calculated Error Variance and Standard Error.” Stated briefly, if the model fit is consistent with the assigned weighting, the calculated error variance and the standard error are close to 1.0. Larger values, which are common in practice, indicate that the model fits the data less well than would be accounted for by expected measurement error. Thus, if the standard error is 5.0, it can be said that the model fit was, on average, five times worse than was consistent with the pre-

liminary analysis of measurement error. Possible sources of the additional error are neglected measurement error, which would change the weighting, or model error. Hill and others (1998) show that some types of model error contribute to the calculated error variance but do not necessarily result in an inaccurate model.

Guideline 7: Encourage convergence by making the model more accurate

Nonlinear regression models of complex systems often do not converge. In general, convergence is improved as the model becomes a better representation of the system that produced the observations being matched by the regression, so that the goal of achieving convergence and a valid regression and the goal of model calibration generally are identical. Substantial insight about the model can be obtained by using the information available from unconverged regressions, such as dimensionless and one-percent scaled sensitivities, composite scaled sensitivities, parameter correlation coefficients, weighted and unweighted residuals, and parameter updates calculated by the regression. This information can be used to evaluate the parameters, observations, and fit of the existing model, and to detect inaccuracies in model construction.

Possible model modifications resulting from this analysis include estimating fewer parameters, modifying the defined parameters, modifying other aspects of model construction, including additional data as observations in the regression, and, rarely, changing the weighting used.

Guideline 8: Evaluate model fit

The most basic attribute of nonlinear regression methods is that, given a well-posed problem, parameter values are calculated that produce the best fit between simulated and observed values. The model can then be evaluated without wondering whether a different set of parameter values would be better.

Two common problems are strong indicators of model error: (1) the model does a poor job of matching observations, and (2) the optimized parameter values are unrealistic and confidence intervals on the optimized values do not include reasonable values. The first is discussed here under Guideline 8; the second indicator is discussed under Guideline 9.

The match to observations achieved through the regression can be evaluated using the methods described in the sections "Statistical Measures of Model Fit" and "Graphical Analysis of Model Fit and Related Statistics." Evaluations using these methods have been presented in a number of publications, including Cooley and others (1986), Yager (1991, 1993), D’Agnese and others (1998), and Hill and others (1998), and example graphs of weighted residuals can be found there.

Weighted residuals are indicative of model fit but, being dimensionless, can be confusing to interpret. Technically, they equal the ratio between the unweighted residual and the statistic used to define the weight. So, if the statistic was a standard deviation and the unweighted residual is

twice as large as the standard deviation, the value of the weighted residual is 2.0. To more clearly present model fit, often it is useful also to include maps of unweighted residuals in reports, as was done by D'Agnese and others (1998). Then very large residuals can be pointed out and discussed.

Two example graphs are presented here. Figure 7 shows observed and simulated streamflow gains along the length of a river. Figure 8 shows the related residuals, which are a good indication of model fit if the observed gains are all about equally reliable, as is the case in this example, but could be misleading if some of the measurements were known to be less accurate.

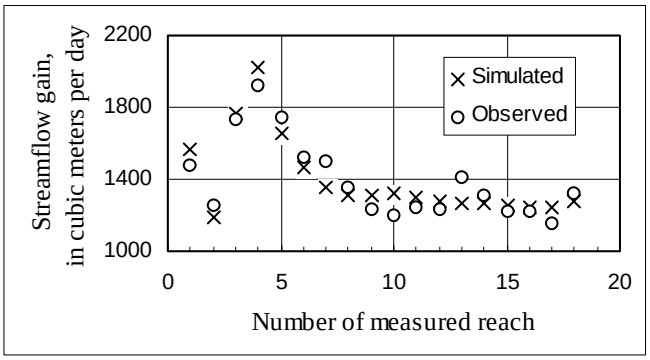


Figure 7: Observed and simulated streamflow gains for model CAL3 of Hill and others (1998).

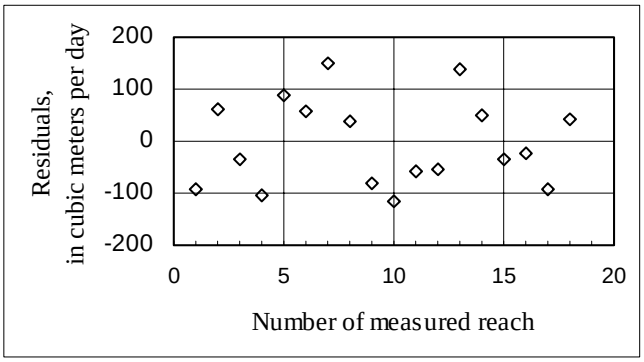


Figure 8: Residuals equal to the observed minus the simulated streamflow gains of figure 7.

Trying to identify trends (lack of nonrandomness) by visual inspection is not always reliable. Often it is useful to evaluate randomness using formal methods to avoid false identification of trends and to avoid missing trends that exist. One such method is the runs tests, as discussed in the section "Graphs using independent variables and the runs test". For example, Cooley and others (1986), use runs tests to evaluate spatially distributed weighted residuals. UCODE and MODFLOWP perform a runs test on the weighted residuals using the sequence in which the observations are listed in the input file. Figure 9 displays the runs statistic information printed by MODFLOWP.

```
STATISTICS FOR ALL RESIDUALS :
 AVERAGE WEIGHTED RESIDUAL  :  .100E+00
 # RESIDUALS >= 0. :     18
 # RESIDUALS < 0.  :     17
 NUMBER OF RUNS  :   17  IN   35 OBSERVATIONS

 INTERPRETTING THE CALCULATED RUNS STATISTIC VALUE OF     -.339
 NOTE: THE FOLLOWING APPLIES ONLY IF
       # RESIDUALS >= 0 . IS GREATER THAN 10 AND
       # RESIDUALS < 0.   IS GREATER THAN 10
 THE NEGATIVE VALUE MAY INDICATE TOO FEW RUNS:
    IF THE VALUE IS LESS THAN -1.28, THERE IS LESS THAN A 10 PERCENT
                                      CHANCE THE VALUES ARE RANDOM,
    IF THE VALUE IS LESS THAN -1.645, THERE IS LESS THAN A 5 PERCENT
                                      CHANCE THE VALUES ARE RANDOM,
    IF THE VALUE IS LESS THAN -1.96, THERE IS LESS THAN A 2.5 PERCENT
                                      CHANCE THE VALUES ARE RANDOM.
```

Figure 9: Runs test output from MODFLOWP for test case 1 of Hill (1992).

If the model fit is unsatisfactory, three possible problems need to be considered. Listed in order of the frequency with which they occur, the three problems are: (1) model error, including how parameters are defined; (2) data errors such as data entry errors or mistakes in the definition of associated simulated values; and (3) errors in the weighting of the observations or prior information. It is often difficult to identify the cause of a problem. In some circumstances, influence statistics, such as DFBETAs (Cook and Weisberg, 1982) that indicate the importance of each observation to the estimation of each parameter can be useful (Anderman and others, 1996; Yager, in press). Additional methods described in guideline 10 also can be useful to evaluate individual models.

As discussed in the section “Calculated Error Variance and Standard Error” and under Guideline 6, if the weights reflect the measurement errors as suggested in this work, weighted residuals that are, on average, larger than 1.0 indicate that the model is worse than would be expected given anticipated measurement error, and values smaller than 1.0 indicate that the model fits better than expected given anticipated measurement error.

If the model fit is unsatisfactory, the situation can be addressed as described at the end of Guideline 7.

Guideline 9: Evaluate optimized parameter values

Evaluate optimized parameter values by comparing the optimized values and their confidence intervals with independent information about the parameter values. The independent infor-

mation may include ranges of expected values, and (or) a relative ordering of values. This simple test can be an unexpectedly powerful indicator of model error, as shown by Poeter and McKenna (1995), Poeter and Hill (1996), Anderman and others (1996), and Hill and others (1998).

Using independent information on the parameters as suggested here is an alternative to using the information in the context of prior information values, and is discussed in this report in section "Lack of Limits on Estimated Parameter Values" and under Guideline 5. As noted there, unreasonable optimized parameter values can be disconcerting to modelers, but provide important indicators of problems with model construction, the observations, or both. An example of a graphical comparison of estimated hydraulic conductivities and ranges of expected values is shown in figure 10. In this example, the reasonable ranges are broad, but a number of conceptual models were rejected because optimized parameter values were outside these ranges. Thus, even in this circumstance, requiring reasonable optimized parameter values produced an important constraint to model development.

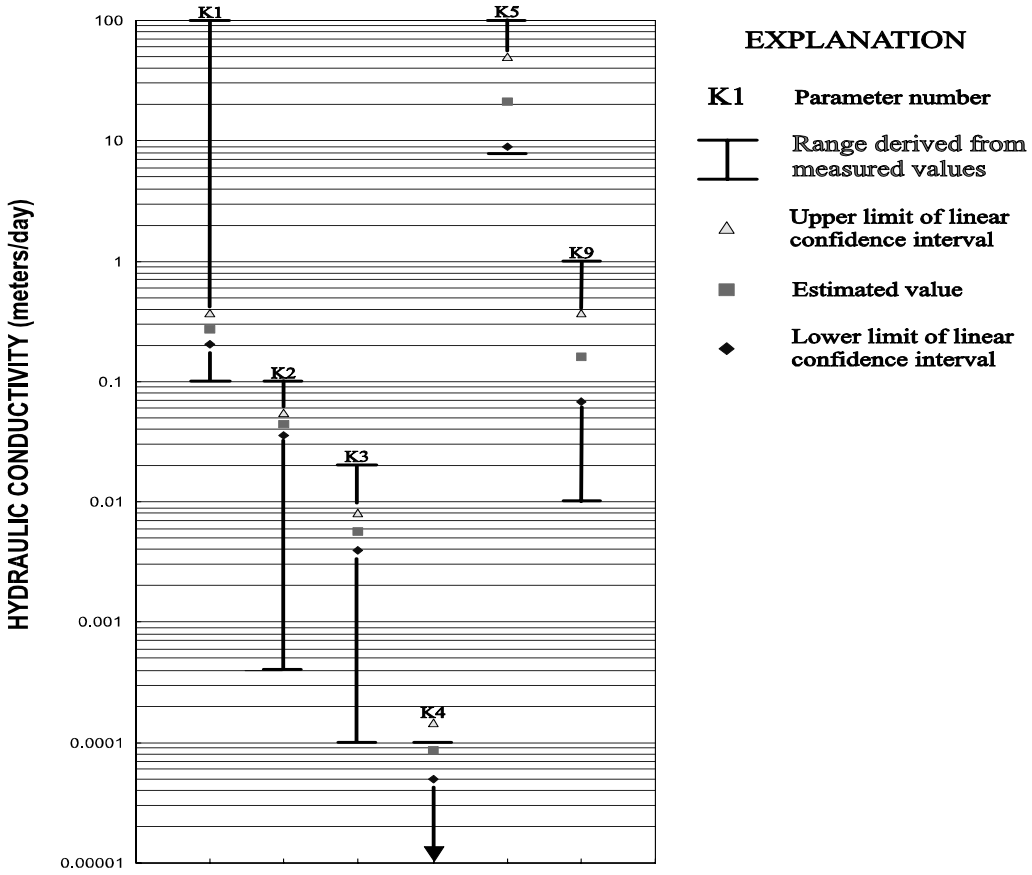


Figure 10: Optimized hydraulic-conductivity values, 95-percent linear confidence intervals, and the range of hydraulic-conductivity values derived from field and laboratory data. (from D'Agnese and others, 1998)

Consideration of confidence intervals on the optimized parameter values is needed to avoid concluding that there is a problem with the model when the real problem is insufficient data with which to estimate the defined parameters. 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 producing the unrealistic values is less likely to be model error. An example of this circumstance is discussed by Barlebo and others (in press). Confidence intervals are discussed further in Guideline 9.

Guideline 10: Test alternative models

In most problems, there is more than one possible representation of the system involved, and this guideline encourages testing all alternative models. Such testing is a viable alternative when inverse modeling is used. Models that are more likely to be accurate tend to have three attributes: better fit, weighted residuals that are more randomly distributed, and more realistic optimal parameter values. These attributes are discussed in the following paragraphs.

The first attribute is a better match to observed data, as indicated by smaller values of the calculated error variance (eq. 14), the standard error of the regression (the square-root of eq. 14), fitted error statistics, AIC and BIC statistics (eq. 16 and 17), or the maximum likelihood criteria (eq. 3), all of which are printed by UCODE and MODFLOWP. Other statistics, such as Kashyap's measure (Medina and Carrera, 1996), also can be used, and generally can be easily calculated using the printed statistics. A graph of fitted standard deviations for hydraulic heads from seven models of Hill and others (1998) is shown in figure 11.

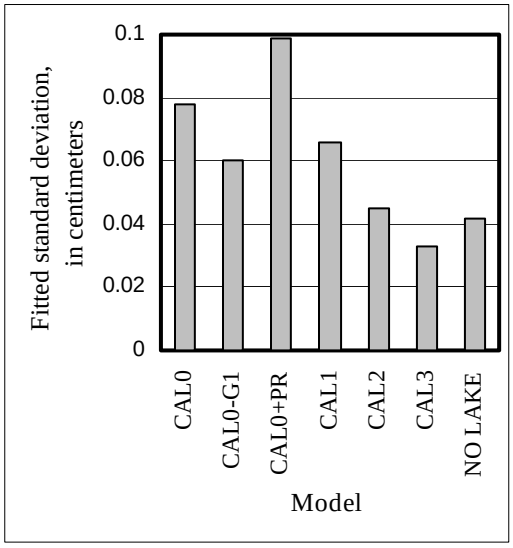


Figure 11: Fitted standard deviations for hydraulic heads for seven models from a controlled experiment in model calibration. (from Hill and others, 1998)

Besides summary statistics, it is important to consider graphs of the observations, simulated values, residuals, and weighted residuals, as discussed in Guideline 8.

The second attribute of better models is that weighted residuals (defined after eq. 1and 2) are more randomly distributed. This generally is determined using the graphs and related statistics discussed in the section "Graphical Analysis of Model Fit and Related Statistics." Graphs of weighted residuals against weighted simulated values, adjusted to account for using coefficients of variation calculated using the observed values in the weighing as discussed by Hill (1994), are shown for two models in figure 12. The weighted residuals from model CAL0 tend to be larger than those of CAL3, as indicated by the greater spread about the 0.0 weighted residual line. In this example, the weighting changed somewhat, so the spread does not necessarily indicate a closer fit between simulated and observed values. Figure 11, however, shows that the CAL3 model does fit the hydraulic-head data better than the CAL0 model. The two sets of weighted residuals of figure 12 are both reasonably random, although the grouping of positive CAL0 residuals in figure 12A for weighted simulated values between 15 and 30and the predominantly positive prior information weighted residuals for CAL3 may be of concern.

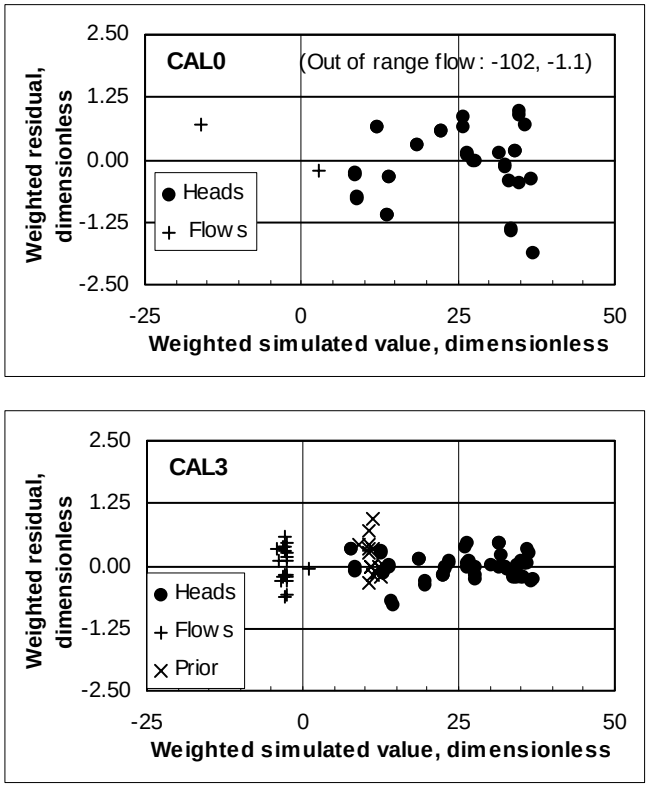


Figure 12: Weighted residuals versus weighted simulated values for models CAL0 and CAL3 of Hill and others (1998).

The third attribute of better models is that optimum parameter values will tend to be more reasonable, both in terms of the estimated values and their values relative to one another. Unrealistic optimized parameter values often are disconcerting to users, as mentioned in the section "Lack of Limits on Estimated Parameter Values" and under Guideline 9.

In some cases the model evaluation may indicate that the data are insufficient to identify a best model from several possible alternatives, in which case any predictions of interest need to be simulated using all reasonable models.

Poeter and McKenna (1995) present an innovative method of using indicator kriging to generate possible models that differed in the zonation used for the hydraulic-conductivity field. They then estimated hydraulic conductivities using MODFLOWP. The synthetic test case used allowed them to show that the additional analysis provided by nonlinear regression tended to produce more accurate transport predictions than could be attained without the use of regression. The additional analysis included determining the best-fit parameters for each model through regression, and then omitting models for which at least one of the following conditions occurred: (1) the best-fit parameter values were unrealistic in that obviously coarser deposits had lower hydraulic conductivities than finer grained deposits, (2) the best-fit parameter values were substantially different than expected, (3) the model fit was significantly worse than for other models, or (4) the regression did not converge. The dramatic improvement in the predictions produced by models screened using these criteria indicated that their application is likely to be useful for identifying more accurate models.

Guideline11: Evaluate potential new data

Potential new data can be evaluated in a number of ways using the methods discussed in this work. Here, dimensionless and one-percent scaled sensitivities and one-percent sensitivity maps are discussed as tools for evaluating potential new data. These statistics depend only on sensitivities and not on measured values. Thus, the type, location, and weighting of potential new data are evaluated.

The analysis is conducted by adding the potential data to the observation data sets of UCODE or MODFLOWP as if the data had already been collected. Specification of the statistic for the weighting can be used to represent the anticipated accuracy of the measurement. Any number can be specified for the observations because they do not affect the statistics being considered.

Anderman and others (1996) use composite scaled sensitivities and correlation coefficients (see figure 5 of this report) calculated for initial parameter values to evaluate the contribution to a ground-water flow model calibration of three types of data: hydraulic heads, an estimate of lake-aquifer interaction, and subsurface transport as represented by advective travel derived from concentration measurements. Although, in this case, the data had already been collected, it is proposed

both here and by Anderman and others (1996) that such an analysis is useful before data collection.

The example of Anderman and others (1996) demonstrates how model nonlinearity can produce misleading results. For the initial parameter values, the advective-transport path enters a lake near the source instead continuing on in the ground-water system, as is more probable given the concentration data. The short advective-travel path results in an underestimate of the importance of these data when evaluated using the composite scaled sensitivities and correlation coefficients calculated for the initial parameter values. Such model nonlinearity is common, and often it is useful to calculate the statistics for several combinations of parameter values to evaluate possible future data collection activities.

Dimensionless scaled sensitivities can be calculated for any potential observation, and they can be used to compare the likely importance of individual proposed data to the estimation of all of the parameters. Table 3 shows selected dimensionless scaled sensitivities from test case 1 of Hill (1992). Dimensionless scaled sensitivities that are larger in absolute value indicate greater likely importance. Here it can be seen that different observations are likely to be important to the estimation of different parameters. In the simple steady-state ground-water flow system for which these sensitivities are calculated, the dimensionless scaled sensitivities can be explained easily. For example, consider observation WELL1, which is a hydraulic head measured just beneath the river, which forms the only outflow boundary. Simulated hydraulic head at this location is dominated by the elevation of the water in the river, the characteristics of the riverbed, and the amount of water leaving the system. K1 and K2 are hydraulic conductivity parameters that apply along the entire length of the river and do not influence the spatial distribution of outflow to the river at steady-state, so that they do not affect simulated hydraulic head at WELL1. KRB is the hydraulic conducitivity of the riverbed, which does influence the simulated hydraulic head beneath the river, resulting in the reltivly large scaled sensitivity for observation WELL1. The composite scaled senstivities indicate that the four observations listed provide much more information for parameter K1 than for KRB, and an intermediate amount of information for K2.

Dimensionless scaled sensitivities also can be plotted against independent variables such as time and location. The graph of dimensionless scaled sensitivities plotted against time shown in figure 13 indicates the relative importance of hydraulichead measurements before and during pumpage. Additional uses of scaled sensitivities are discussed under Guideline 14 and in the section “Statistics for Sensitivity Analysis”.

Table 3: Selected dimensionless and composite scaled sensitivities from test case 1 of Hill (1992).

	Parameter name		
Observation name	K1	K2	KRB
WELL1	-0.652×10^{-4}	-0.289×10^{-4}	1.17
WELL2	180	34.5	1.17
WELL3	351	115	1.17
RIVER	0.399×10^{-2}	0.177×10^{-2}	0.109×10^{-4}
	Composite Scaled Sensitivities (css)		
	197	60.0	1.01

Dimensionless scaled sensitivities also can be plotted against independent variables such as time and location. The graph of dimensionless scaled sensitivities plotted against time shown in figure 13 indicates the relative importance of hydraulichead measurements before and during pumpage. Additional uses of scaled sensitivities are discussed under Guideline 14 and in the section “Statistics for Sensitivity Analysis”.

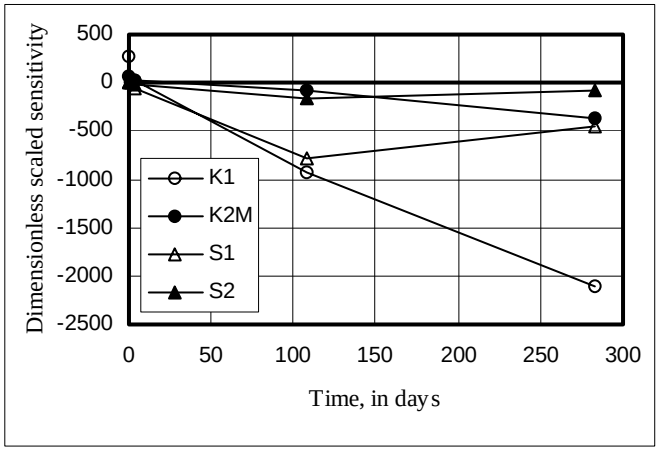


Figure 13: Dimensionless scaled sensitivites plotted against time. The values are from well 2 of test case 1 of Hill (1992). Time zero has no pumpage; at subsequent times constant pumpage is applied. The K1 parameter represents the hydraulic conductivity in the top of two layers. The K2M parameter represents a multiplicative parameter that, combined with an assumed linear trend, defines the hydraulic conductivity of the bottom layer. S1 and S2

are storage coefficients of the top and bottom layers, respectively.

Guideline 12: Evaluate the potential for additional estimated parameters

At any stage of model calibration, composite scaled sensitivities can be analyzed as described in Guideline 3 to determine if the available data are likely to support additional detail in representing the system characteristics associated with the defined parameters. Parameters with large composite-scaled sensitivities can be subdivided in ways that are consistent with other data, such as geologic and hydrogeologic data in ground-water problems. The new set of defined parameters can then be evaluated using the methods of Guideline 3, and regression pursued if warranted.

Guideline 13: Use confidence and prediction intervals to indicate parameter and prediction uncertainty

Confidence and prediction intervals can be constructed using the methods described in the sections "Parameter Statistics" and "Prediction Uncertainty" in the first part of this report. Thus, instead of reporting a single predicted value, a predicted value and a confidence or prediction interval are reported. For example, linear confidence intervals for a set of parameter values were shown in figure 10 in Guideline 9. Ideally, confidence intervals are intervals in which the true parameter value or true predictive quantity is likely to occur with some specified probability. Prediction intervals differ from confidence intervals in that they include the effect of measurement error (see eq. 34 and related text). Prediction intervals need to be used if the intervals are to be compared to measured values and are most commonly constructed for simulated predictions.

Confidence intervals are for the true average value (Ott, 1993, p.519). Confidence intervals on average values depend not only on the variance of the original population, but also on the sample size used to calculate the estimated average. This is confusing to many users, who are likely to look at, for example, the confidence intervals of figure 10 and conclude that they are too small. This judgment, however, needs to be made in the context of the confidence intervals being constructed for the average value. To demonstrate the significance of this, consider a simple example using a generated set of 300 normally distributed numbers. Figure 14 shows the range of the 300 numbers. Also included are estimated means calculated as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \qquad (36)$$

and their associated confidence intervals, calculated as:

$$\left(\bar{y} + \frac{2s}{\sqrt{n}}; \bar{y} - \frac{2s}{\sqrt{n}}\right) \qquad (37)$$

where s is the sample standard deviation and n is the sample size (300 for the example). From this simple example it can be seen how few samples are needed for the confidence interval for the average to be much smaller than the range of the population.

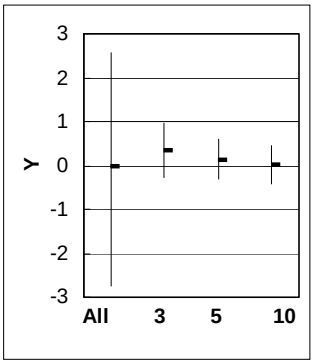


Figure 14: Confidence intervals for a population mean given different sample sizes. The population is composed of 300 random normally distributed numbers with a range noted by the bar labeled “All” and a mean noted by the mark in the center of that bar. The other bars are labeled with the sample size used (3, 5, and 10). The marks in the center of these bars are the sample means, and the lengths of the bars display the associated confidence interval.

In figure 10, the range of hydraulic conductivity within a selected volume is shown by the solid bars, as derived from measured values. This range is analogous to the entire range of the 300 generated random values in figure 14. The situation in figure 10 differs from the simple example of figure 14 in two important ways. First, and most fundamentally, the situation in figure 10 assumes that an effective hydraulic-conductivity value can be applied to a specified volume of subsurface material. The regression analysis is valid only in so far as this assumption is valid.

The second difference between the situations represented in figures 10 and 14 is that in figure 10 estimates are derived through regression. Thus, most of the data used to estimate the mean are measurements of other quantities--here, hydraulic heads and spring flows--which are used to estimate the effective hydraulic-conductivity value through nonlinear regression. In contrast, the data used in figure 14 are samples from the population for which the mean is being estimated.

Despite these differences, the discrepancy between the full range of values and the confidence intervals displayed both in figures 10 and 14 is important to remember when interpreting results such as these shown in figure 10.

As noted in the first part of this report, both linear and nonlinear confidence and prediction intervals can be calculated. Linear intervals take a minor computational effort; nonlinear intervals take substantial computational effort because each nonlinear confidence interval limit requires

computational effort equivalent to a full regression. The section “Testing for Linearity” discusses a test with which model nonlinearity can be evaluated.

Linear intervals use the assumption of normality of the parameter estimates in their construction. As discussed in the section “Normal Probability Graphs and Correlation Coefficient R_N^2,” the weighted residuals are the only quantities that can be readily tested for normality. A sample normal probability graph is shown in figure 15, along with graphs showing normally distributed random numbers generated with and without regression-induced correlations, as described in the section “Determining Acceptable Deviations from Independent Normal Weighted Residuals.” Figure 15 shows that most aspects of the nonlinear pattern evident in the weighted residuals can be explained by the regression-induced correlations.

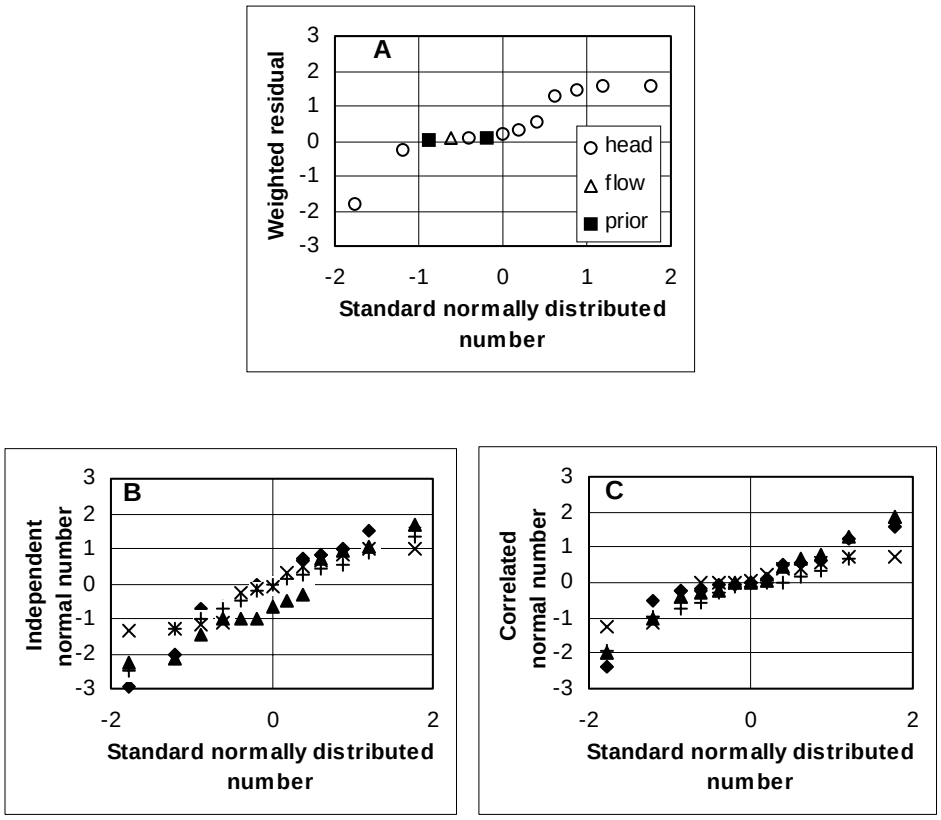


Figure 15: Normal probability graphs for the steady-state version of test case 1 of Hill (1992), including (A) weighted residuals, (B) normally distributed, uncorrelated random numbers, and (C) normally distributed random numbers correlated as expected given the fitting of the regression. In B and C, four sets of generated numbers are shown, each with a different symbol.

Christensen and Cooley (1996; in press) show that in nonlinear problems, nonlinear confidence intervals can be very different than linear intervals for some quantities, while they can be very close for others. It appears that linear confidence intervals are useful as a general indication of uncertainty in many circumstances, but, if at all possible given computer resources, some nonlinear intervals need to be calculated if the model is nonlinear.

Linear and nonlinear confidence intervals, along with any other method of uncertainty analysis, such as Monte Carlo methods and the methods presented by Sun (1994), are based on the assumption that the model accurately represents the real system. In truth, all models are simplifications of real systems, and the accuracy of the uncertainty analysis is in question. Accuracy of uncertainty analyses is very difficult to evaluate definitively. Steen Christensen and R.L. Cooley (written commun., 1997) compared nonlinear prediction intervals to measured heads and flows indicating good correspondence between the expected and realized significance level of the intervals. If model fit to data indicates model bias, the theory suggests the calculated intervals do

not reflect all aspects of system uncertainty, and, conservatively, they might be best thought of as indicating the least amount of uncertainty. That is, actual uncertainty might be larger than indicated by the confidence intervals. If prediction intervals are dominated by the measurement error term, they are less likely to be prone to error. Unfortunately, in many circumstances the confidence intervals are of more interest because they reflect model uncertainty most clearly. Cooley (1997) provides additional analysis of nonlinear confidence intervals.

Guideline 14: Formally reconsider the model calibration from the perspective of the desired predictions

It is important to evaluate the model relative to the desired predictions throughout model calibration, as discussed in the beginning of the section “Guidelines for Effective Model Calibrations”. For reasonably accurate models, it also is useful to consider the predictions more formally, as described below. In this work it is suggested that formal analysis using uncalibrated models is likely to produce misleading results, given the nonlinearity of the models considered. It can be difficult to determine when a model is sufficiently accurate, but at the very least the obvious errors in system representation and the relation of the observations to simulated equivalents need to be resolved, and weighted residuals need to be approximately random. The analysis is divided into two approaches.

First, predictions and linear confidence intervals on the predictions can be calculated for all reasonably accurate models to evaluate how different sets of observations and conceptual models are likely to affect both the simulated predictions and their likely precision. Linear confidence intervals are suggested instead of nonlinear confidence intervals or either kind of prediction interval because linear confidence intervals can be calculated quickly and represent the prediction uncertainty contributed by the model and the parameter estimates.

Second, the model parameters and the simulated predictions can be evaluated to determine which parameters and what system features are likely to be most important to prediction accuracy. This is accomplished using sensitivities related to the regression observations and the predictions, and statistics calculated from these sensitivities, and can be used to guide subsequent field and model calibration efforts. The procedure for such an analysis is outlined in figure 16.

		A. Precision of the parameter estimate	
		Poor: Large parameter composite scaled sensitivity, coefficient of variation, or confidence interval	**Good:** Small parameter composite scaled sensitivity, coefficient of variation, or confidence interval
Importance of the parameter to predictions of interest	**Not important:** Small prediction scaled sensitivity	I. Acceptable[1]	II. Acceptable[1]
	Important: Large prediction scaled sensitivity	IV. Improve estimation of this parameter and associated system features.[2]	III. Acceptable[1]

		B. Uniqueness of the parameter estimate	
		Poor: The absolute value of some of this parameter's correlation coefficients are close to 1.0.[3]	**Good:** All of this parameter's correlation coefficients have absolute values less than about 0.95.[3]
Importance of the parameter to predictions of interest	**Not important**: The same parameter pairs are extremely correlated.[4]	I. Acceptable[1]	II. Acceptable[1]
	Important: Previously correlated parameter pairs are uncorrelated.[4]	IV. Improve estimation of this parameter and associated system features .[2]	III. Acceptable[1]

1. Acceptable means that this parameter is estimated well compared to other parameters, from the perspective of simulating predictions, or is unimportant to the predictions of interest. Improved estimation of this parameter and improved representation of the system features with which this paramter is associated are likely to be less important to improving prediction accuracy than for other parameters
2. Improved estimation of this parameter and improved representation of the system features with which it is associated probably are important to improved prediction accuracy.
3. The parameter correlation coefficients needed for this analysis are calculated using unestimated as well as estimated parameters, and include only the observations and prior information used in the calibration.
4. The prediction correlation coefficients needed for this analysis are as in 3, but include predictions as well as the observations and prior information used in the calibration.

Figure 16: Classification of the need for improved estimation of a parameter and, perhaps, associated system features. The classification is based on statistics which indicate the importance of parameters to predictions of interest and (A) the precision of parameter estimates or (B) the uniqueness with which parameters are estimated by the regression.

Parameter correlation coefficients are cited in figure 16 as measures both of the uniqueness of the parameter estimate and the importance of parameters to the predictions of interest. In both cases, the correlation coefficients are variations of the parameter correlation coefficients printed at the end of most regression runs, as discussed above in the sections “Variances and Covariances” and “Correlation Coefficients.” An example of the utility of such correlation coefficients can be found in the following ground-water modeling example. Consider a ground-water flow model calibrated with hydraulic-head and streamflow gain or loss observation data. The calibrated model is being developed to predict (a) hydraulic head at a location where no measurement can be obtained, and (b) advective transport from the site of a contaminant spill. Correlation coefficients for all parameters are obtained using the calibrated model using all defined parameters (see section “Variances and Covariances”); the prediction correlation coefficients are obtained by adding the prediction hydraulic-head location and advective transport as ‘observations’ in the input file and again calculating the correlation matrix for the same set of parameters. A similar calculation is reported by Anderman and others (1996), showing that advective-travel was affected by individual parameter values, while hydraulic heads were not. In this circumstance, prediction of hydraulic heads did not require uncorrelated parameter estimates while prediction of advective travel did.

An example analysis of predictions is presented in figure 17. Prediction scaled sensitivities calculated using equation 12 are compared to parameter composite scaled sensitivities of equation 10. In the example, the predictions of interest are the cartesian components of advective travel simulated by particle tracking using the ADV Package of Anderman and Hill (1997). The figure shows the range and mean of the prediction scaled sensitivities for eight transported particles. The prediction scaled sensitivity is defined to equal the percent change in the advective transport caused by a one-percent change in parameter value. The figure clearly shows that parameters T3 and T4 are most important to the determination of advective-transport distance in all three coordicate directions, and that the observations used in the regression provide more information for parameter T3 than for parameter T4. This type of information can be invaluable for understanding model strengths and weaknesses and for planning additional modeling and data collection efforts.

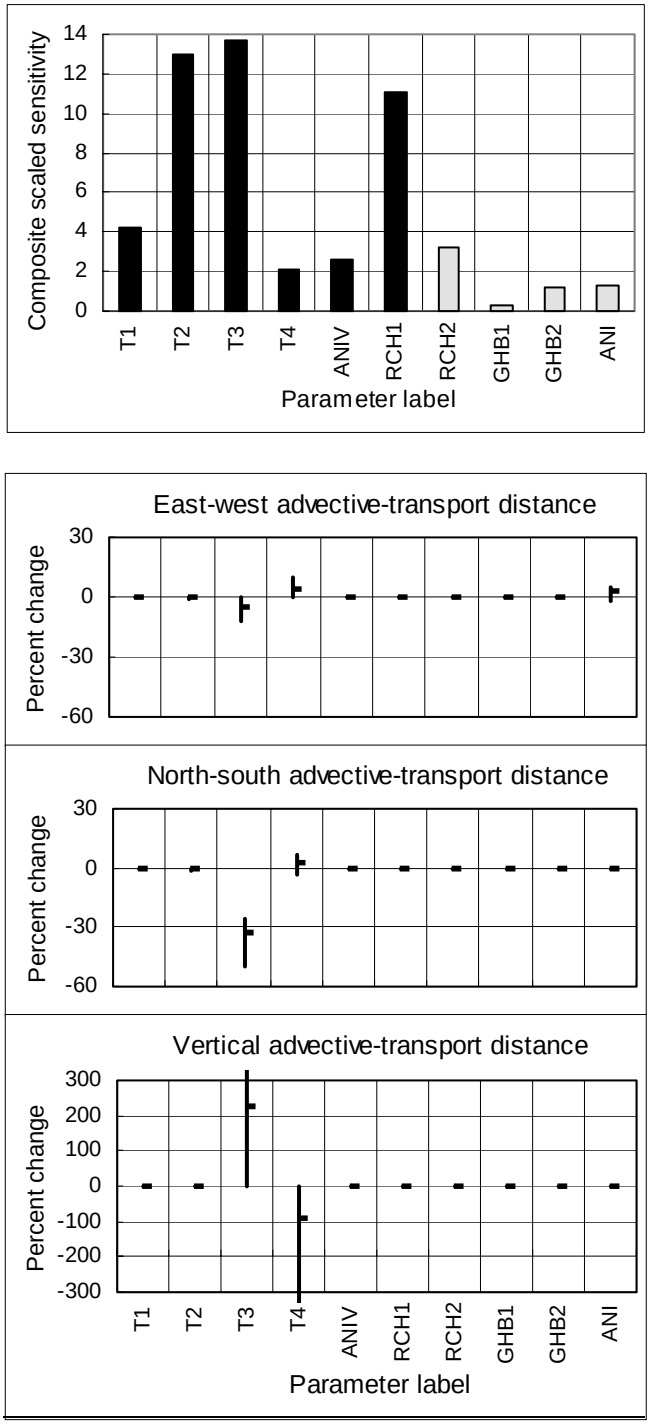


Figure 17: Composite scaled sensitivities for estimated parameters and prediction scaled sensitivities for the spatial components of predicted advective transport. The composite scaled sensitivites for parameters estimated in the regression are shown using black bars; those not estimated in the regression are shown using gray bars. The prediction scaled sensitivities are defined as the percent change in the prediction given a one-percent change in the parameter value, so 'Percent change' is used to label the vetical axes.

ISSUES OF COMPUTER EXECUTION TIME

Computer execution time is often a problem when using regression methods. Thus, a set of hints for effective use of regression also needs to include a few ideas about model construction as it affects execution time. The suggestion about starting with a relatively simple model of the ground-water system and building complexity as warranted by the system and by the available data, as discussed in guideline 1, also is relevant to the issue of minimizing execution time. Starting with a simple model often results in shorter execution times.

Execution times for regression, or inverse, simulations can be estimated using execution times for forward simulations (a simulation for hydraulic heads in a ground-water flow problem) as:

$$T_i = 2(NP)\, T_f\, (1+NP) \qquad (38)$$

where

T_i is the execution time for the regression (inverse) solution;

T_f is the execution time for the forward solution; and

NP is the number of parameters being estimated by regression.

This assumes that the number of parameter-estimation iterations approximately equals twice the number of parameters, that is, 2(NP), which is, on average, typical. The (1 + NP) term is for one forward simulation and one simulation to calculate sensitivities for each of the NP parameters. The NP sensitivity simulations solve a slight variation of the forward problem for the forward- or backward-difference sensitivities of UCODE, or sensitivity equations that result from taking the derivative of the forward equation with respect to the parameter in MODFLOWP. In both cases, each of the sensitivity simulations take, on average, the same amount of execution time as a forward simulation.

Experience indicates that inverse model execution times that exceed about 15 hours (an overnight simulation) commonly occur when the forward execution time exceeds 30 minutes. The number of grid rows, columns, and layers, and the number of time steps this execution time allows depends on the speed of the computer and the characteristics of the simulated system, including the contrasts present in the hydraulic-conductivity field.

Sometimes simple changes in the simulation can dramatically improve execution times. For example, the initial hydraulic conductivity structure of the model described by D'Agnese and others (1998, in press) was characterized by values in bordering finite-difference cells that differed by more than five orders of magnitude in many parts of the model. Introducing single cells of moderate hydraulic conductivity between the high and low valued cells in most of the model resulted in about a 6-fold decrease in execution time with little effect on simulated results.

Another simplification that can dramatically reduce execution time is to replace nonlinear forward problems with linear approximations as much as possible without substantial diminishment of accuracy. In ground-water flow simulations, for example, water-table and convertible layers (as they are called in MODFLOW and MODFLOWP) often can be replaced by confined layers with approximate thicknesses. This is nearly always good practice for steady-state simulations, but can be too inaccurate for transient simulations in which layers are substantially de-watered during the calibration period. The inaccuracy produced by this simplification can be evaluated by comparing forward simulations that include the water-table and convertible layers with those that include the approximate thicknesses.

EXAMPLE FIELD APPLICATIONS AND SYNTHETIC TEST CASES

The nonlinear regression methods, diagnostic and inferential statistics, and guidelines described above have been used successfully in a number of applications. Because nonlinear regression is a useful but imperfect tool for model calibration, its application is not always straightforward. Thus, it can be helpful to consider other applications when designing and reporting a modeling study using nonlinear regression. References describing applications using the methods described in this report include Cooley (1979, 1983a), Cooley and others (1986), Christiansen and others (1995), McKenna and Poeter (1995), Anderman and others (1996), Barlebo and others(1996), D'Agnese and others (1996, 1998, in press), Tiedeman and others (1997), and Eberts and others (in press). Similar approaches were used by Gailey and others (1991), Tiedeman and Gorelick (1993), Yager (1993), Kuiper(1994), Olsthoorn (1995), Christensen (1997), and Christensen and others (in press). In addition, Hill and others (1998) discuss the calibration of a complex synthetic test case using the methods discussed above, and Poeter and McKenna (1995) present a synthetic ground-water transport test case that evaluates stochastically generated zonations using nonlinear regression methods. The synthetic test cases provide the opportunity to conclusively evaluate the accuracy of the models calibrated using the methods described in this report, and in both studies better models determined as discussed in this report produced more accurate predictions. Poeter and Hill (1997) demonstrate many of the ideas presented in this report using simple examples.

USE OF THE GUIDELINES WITH DIFFERENT INVERSE MODELS

The methods and guidelines presented above were used to design UCODE and MODFLOWP. Many aspects of the methods and guidelines, however, are broadly applicable to other inverse models, with some modifications. Likely differences between UCODE and MODFLOWP and other inverse models are categorized below, with a few words about how the guidelines would be adapted.

Alternative Optimization Algorithm

Presently used alternative algorithms for the minimization of the least-squared objective function with respect to parameter values include adjoint-state methods (as used by, for example, Carrera and Neuman, 1986; Xiang and others, 1992; Tarantola, 1994), and global optimization methods such as simulated annealing, genetic algorithms (Wagner, 1995), and tabu search (Zheng and Wang, 1996). In adjoint-state methods, the derivative of the objective function with respect to the parameter values is calculated and can be used instead of the composite scaled sensitivities. There are no replacements for the one-percent and dimensionless scaled sensitivities and the parameter correlations in the adjoint-state method. It is not uncommon, however, for adjoint-state algorithms also to be programmed to calculate the sensitivities and variance-covariance matrix on the parameters as discussed above, in which case the guidelines apply directly.

Global optimization methods are most useful for problems with very irregular objective function surfaces that are not amenable to the much more numerically efficient gradient search methods, such as modified Gauss-Newton or adjoint states. For problems with such irregular objective functions, scaled sensitivities, composite scaled sensitivities, and parameter correlation coefficients are likely to change values so dramatically as parameter values change that they would be worthless. Other aspects of the guidelines, however, would still be applicable.

Alternative Objective Functions

The primary alternative to the least-squares objective function is the sum of the absolute values of weighted residuals. Minimizing this objective function requires methods that do not use sensitivities or derivatives of the objective function, so that there are no scaled sensitivities, composite scaled sensitivities, or parameter correlation coefficients. As with adjoint states, however, the algorithms developed for this objective function also have been programmed to calculate sensitivities, so that, again, the guidelines would apply directly.

Direct Instead of Indirect Inverse Models

The most dramatically different possibility that presently exists is direct inverse modeling (Yeh, 1986; Sun, 1994). In direct inverse modeling, values of the dependent variable (for example, hydraulic head for the ground-water flow equation or concentration for the transport equation) are determined using usually sparse field data and these values are used directly to calculate the model

input values. This is in contrast to indirect methods, such as the method discussed in this work. The direct methods have been considered longer than the indirect methods in inverse modeling, but have been shown consistently to be more unstable in the presence of typical measurement errors. The direct methods do not use sensitivities and rarely calculate them, so that there are no equivalents to the scaled sensitivities, composite scaled sensitivities, correlation coefficients, and confidence intervals. Other aspects of the guidelines would still apply.

Alternative Parameterization Approach

While this work emphasizes estimating relatively few parameter values, an alternative is to estimate many parameter values (often a unique value for each finite element or finite-difference cell for numerical models) for spatially distributed system characteristics, such as hydraulic conductivity in ground-water problems. Then smoothness criteria, or other types of regularization, are imposed to achieve a tractable regression problem. Such methods are presented by Tikkonov and Arsenin (1977), and tend to be most useful when very little is known about the distribution, or when the distribution is known to be smooth. For such a parameterization approach, the concept of starting simple and building complexity might be useful when designing the regularization method. Application of other aspects of the guidelines is unclear.

REFERENCES

Akaike, Hirotugu, 1974, A new look at statistical model identification: Institute of Electrical and Electronics Engineers Transactions on Automatic Control, v. AC-19, no. 6, p. 716-723.

_____ 1978, Time series analysis and control through parametric models, in Findley, D.F., ed., Applied time series analysis: New York, Academic Press, p. 1-25.

Anderman, E.R., 1996, The use of advective-transport observations to improve ground-water flow parameter estimation: Golden, Colorado, Colorado School of Mines, Ph.D. dissertation,124 p.

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

Backus, G.E., 1988, Baysian inference in geomagnitism: Geophysical Journal, v. 92, p.125-142.

Bard, Jonathon, 1974, Nonlinear parameter estimation: New York, Academic Press, 341p.

Barlebo, H.C., Hill, M.C., and Rosbjerg, Dan, 1996, Identification of groundwater parameters at Columbus, Mississippi, using a three-dimensional inverse flow and transport model, in van der Heidje, Paul and Kovar, Karel, eds., Calibration and reliability in groundwater modeling, Proceedings of the 1996 ModelCARE Conference, Golden, Colorado, September, 1996, International Association of Hydrologic Sciences, Publ. 237, p. 189-198.

Barlebo, H.C., Hill, M.C., Rosbjerg, Dan, and Jensen, K.H., in press, On concentration data and dimensionality in groundwater transport models: Nordic Hydrology.

Bentley, L.R., 1997, Influence of the regularization matrix on parameter estimates: Advances in Water Resources, v. 20, no. 4, p. 231-247.

Brockwell, P.J and Davis, R.A., 1989, Time series, Theory and methods: New York, Springer-Verlag, 519 p.

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

Carter, R.W. and Anderson, I.E., 1963, Accuracy of current meter measurements: American Society of Civil Engineers Journal, v. 89, no. HV4, p. 105-115.

Christensen, Steen, 1997, On the strategy of estimating regional-scale transmissivity fields: Ground Water, v. 35, no. 1, p. 131-139.

Christensen, Steen and Cooley, R.L., 1996, Simultaneous confidence intervals for a steady-state leaky aquifer groundwater flow model, in van der Heidje, Paul and Kovar, Karel, eds, Calibration and reliability in groundwater modeling, Proceedings of the 1996 ModelCARE Conference, Golden, Colorado, September, 1996: International Association of Hydrologic Sciences, Publ. 237, p. 189-198.

Christensen, Steen and Cooley, R.L, in press, Simultaneous confidence intervals for a steady-state leaky aquifer groundwater flow model: Advances in Water Resources Special Issue of Model Calibration and Reliability Evaluation.

Christensen, Steen, Rasmussen, K.R., Moeller, K., 1998, Prediction of regional ground-water flow to streams: Ground Water, v. 36, no. 2, p. 351-360.

Christiansen, Heidi, Hill, M. C., Rosbjerg, Dan, and Jensen, K.H., 1995, Three-dimensional inverse modeling using heads and concentrations at a Danish landfill, in Wagner, Brian and Illangesekare, Tissa, eds, Models for assessing and monitoring groundwater quality, International Association of Hydrologic Sciences, Publ. no. 227, p.167-175.

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

Constable, S.C., Parker, R.L., and Constable, C.G., 1987, Occam's inversion, A practical algorithm for generating smooth models from electromagnetic sounding data: Geophysics, v. 52, p.289-300.

Cook, R.D. and Weisberg, Sanford, 1982, Residuals and influence in regression: Chapman and Hall, New York, 230 p.

Cooley, R.L., 1979, A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 2, Application of statistical analysis: Water Resources Research, v. 15, no. 3, p. 603-617.

Cooley, R.L., 1983a, Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2, Applications: Water Resources Research, v. 19, no. 3, p. 662-676.

Cooley, R.L., 1983b, Some new procedures for numerical solution of variably saturated flow problems: Water Resources Research, 19(5):1271-1285.

Cooley, R.L., 1993, Regression modeling of ground-water flow, Supplement 1 -- Modifications to the computer code for nonlinear regression solution of steady-state ground-water flow problems: U.S Geological Survey Techniques of Water Resources Investigations, book 3, chapt. B4, supplement 1, 8p.

Cooley, R.L., 1997, Confidence intervals for ground-water models using linearization, likelihood, and bootstrap methods: Ground Water, v. 35, no. 5, p. 869-880.

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

Cooley, R.L., and R.L. Naff, 1990, Regression modeling of ground-water flow: U. S. Geological Survey Techniques in Water-Resources Investigations, book 3, chap, B4, 232 p.

Cooley, R.L. and Sinclair, 1976, Uniqueness of a model of steady-state ground-water flow: Journal of Hydrology, v. 31, p. 245-269.

D'Agnese, F.A. Faunt, C.C. Hill, M.C., and Turner, A.K., 1996, Death Valley regional ground-water flow model calibration using optimal parameter estimation methods and geoscientific information systems: in Kovar, Karel and van der Heidje, Paul, eds., Calibration and reliability in groundwater modeling, Proceedings of the 1996 Model CARE Conference, Golden, Colorado, September, 1996: International Association of Hydrologic Sciences, Publ. 237, p. 41-52.

D'Agnese, F.A. Faunt, C.C., Turner, A.K, and Hill, M.C., 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.

Davis, J.C., 1986, Statistic and data analysis in geology: New York, John Wiley, 646 p.

Doherty, J. 1994, PEST: Corinda, Australia, Watermark Computing, 122 p.

Donaldson, J.R. and Schnabel, R.B., 1987, Computational Experience with confidence regions and confidence intervals for nonlinear least squares: Technometrics, vol. 29, no. 1, p.67-87.

Draper, N.R., and Smith, H., 1981, Applied regression analysis (2nd ed.): New York, John Wiley & Sons, 709 p.

Eberts, S.M. and George, L.L., in press, Regional ground-water flow and geochemistry in the midwestern basin and arches aquifer system in parts of Indiana, Ohio, and Michigan: U.S. Geological Survey Professional Paper 1423-C.

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

Forsythe, G.E. and Strauss, E.G., 1955, On best conditioned matrices: American Mathematical Society proceedings, v. 10, no. 3, p. 340-345.

Gailey, R.M., Gorelick, S.M., and Crowe, A.S., 1991, Coupled process parameter estimation and prediction uncertainty using hydraulic head and concentration data: Advances in Water Resources, v. 14k no. 5, p. 301-314.

Harvey, J.W., Wagner, B.J., and Bencala, K.E., 1996, Evaluating the reliability of the stream tracer approach to characterize stream-subsurface water exchange: Water Resources Research, v. 32, no. 8, p. 2441-2451.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: Elsevier, 522 p.

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, 358 p.

Hill, M.C., 1994, Five computer programs for testing weighted residuals and calculating linear confidence and prediction intervals on results from the ground-water parameter estimation computer program MODFLOWP: U.S. Geological Survey Open-File Report 93-481, 81p.

Hill, M.C., Cooley, R.L., and Pollock, D.W., 1998, A controlled experiment in ground-water flow model calibration using nonlinear regression: Ground Water, v. 36 p. 520-535.

Knopman, D.S. and Voss, C.I., 1998, Further comments on sensitivities, parameter estimation and sampling design: Water Resources Research, v. 24, no. 2, p. 225-238.

Kuiper, L.K., 1994, Nonlinear-regression flow model of the Gulf Coast aquifer systems in the south-central United States: U.S. Geological Survey Water-Resources Investigations Report 93-4020, 171p.

Loaiciga, H.A. and Marino, M.A., 1986, Estimation and inference in the inverse problem: Proceeding of Water Forum ‘86, World Water Issues in Evolution, ASCE, Long Beach, CA, Aug. 4-6, p.973-980.

Marquardt, D.W., 1963, An algorithm for least-squares estimation of nonlinear parameters: Journal for the Society of Industrial and Applied Mathematics, v.11, no.2, p.431-441.

McDonald, M. G. and Harbaugh, A. W., 1988, A modular three-dimensional finite-difference ground-water flow model: U.S. Geological Survey Techniques of Water Resources Investigations, book 6, chapter A1, 586 p.

McKenna, S.A. and Poeter, E.P., 1995, Field example of data fusion in site characterization:Water Resources Research, v. 31, no. 12, p. 3229-3240.

Medina, A. and Carrera, J., 1996, Coupled estimation of flow and transport parameters: Water-Resources Research, v. 32, no. 10, p.3063-3076.

Miller, R.G., Jr., 1981, Simultaneous statistical inference: second edition, New York, Springer-Verlag, 299p.

Neuman, S.P., 1982, Statistical characterization of aquifer heterogeneities--An overview, in Narasimhan, T.N., ed., Recent trends in hydrology: Geological Society of America Special Paper 189, p. 81-102

Neuman, S.P. and Jacobson, E.A., 1984, Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels: Mathematical Geology, v. 16, no. 5, p. 499-521.

Olsthoorn, T.N., 1995, Effective parameter optimization for ground-water model calibration: Ground Water, v. 33, n. 1, p. 42-48.

Ott, Lyman, 1993, An introduction to statistical methods and data analysis: Boston, PWS-Kent Publishing Company, Fourth Edition, 1170p.

Parker, R.L.,1994, Geophysical inverse theory: Princeton University Press, Princeton, New Jersey, 386 p.

Poeter, E.P. and Hill, M.C., 1996, Unrealistic parameter estimates in inverse modeling: a problem or a benefit for model calibration?: Proceedings of the ModelCARE 96 Conference, Golden, CO, September 1996, International Association of Hydrological Sciences Publication no. 237, p. 277-285.

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

Poeter, E.P. and McKenna, S.A., 1995, Reducing uncertainty associated with groundwater flow and transport predictions: Ground Water, v. 33, no. 6, p.889-904.

Seber, G.A.F., and C.J. Wild 1989, Nonlinear Regression, John Wiley & Sons, NY, 768 p.

Sun, N.-Z., 1994, Inverse problems in ground-water modeling: Boston, Kluwer Academic Publishers, 337 p.

Sun, N.-Z. and Yeh, W., W.-G., 1990, Coupled inverse problems in groundwater modeling, 1, Sensitivity analysis and parameter identification: Water Resources Research, v. 26, no. 10, p. 2507-2525.

Tarantola, Albert, 1994, Inverse problem theory: New York, Elsevier, 613 p.

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

Tiedeman, C.R., Goode, D.J., and Hsieh, P.A., 1997, Numerical simulation of ground-water flow through glacial deposits and crystalline rock in the Mirror lake area, Grafton County, New Hampshire: U.S. Geological Survey Professional Paper 1572, 50 p.

Tiedeman, Claire and Gorelick, S.M., 1993, Analysis of uncertainty in optimal groundwater contaminant capture design: Water Resources Research, v. 29, no. 7, p. 2139-2153.

Tikhonov, A.N. and Arsenin, V.Y., 1977, Solution of ill-posed problems: New York, Winston and Sons.

U.S. Geological Survey, 1980, Accuracy specifications for topographic mapping, in Technical instructions of the National Mapping Division: Reston, Virginia, Chapter 1B4, p. 1-13.

Vecchia, A.V. and Cooley, R.L., 1987, Simultaneous confidence and prediction intervals for nonlinear regression models with application to a groundwater flow model: Water Resources Research, v. 22, no. 2, p. 95-108.

Wagner, B.J., 1995, Sampling design methods for groundwater modeling under uncertainty: Water Resources Research, v.31, no. 10, p. 2581-2591.

Xiang, Y., Sykes, J.F., and Thomson, N.R., 1992, A composite L1 parameter estimator for model fitting in groundwater flow and solute transport simulation: Water Resources Research, v. 29, no. 6, p.1661-1673.

Yager, R.M., 1991, Estimation of hydraulic conductivity of a riverbed and aquifer system of the Susquehanna River in Broome County, New York, U.S. Geological Survey Open-File Report 91-457, 54 p.

Yager R.M, 1993, Simulated three-dimensional ground-water flow in the Lockport Group, a fractured dolomite aquifer near Niagra Falls, New York: U.S. Geological Survey Water Resources Investigations Report 92-4189, 43 p.

Yager, R.M., in press, Detecting influential observations in nonlinear regression modeling of ground-water flow: Water Resources Research.

Yeh, W.W.-G., 1986, Review of parameter identification procedures in ground-water hydrology--The inverse problem: Water Resources Research, v. 22, no. 2, p. 95-108.

Zheng, Chunmiao and Wang, P. Patrick, 1996, Parameter structure identification using tabu search and simulated annealing: Advances in Water Resources, v. 19, no. 4, p. 215-224.

APPENDIX A: THE MAXIMUM-LIKELIHOOD AND LEAST-SQUARES OBJECTIVE FUNCTIONS

The maximum-likelihood objective function is developed by considering the random nature of $\underline{y}$, the observations. This random nature results from conceptualizing measurement error as random. If $\underline{Y}$ is the vector of jointly distributed random variables of which $\underline{y}$ is a realization, the joint probability distribution function (pdf), $f_{\underline{Y}}(\underline{y})$, depends on the true model and true parameter values. For the purpose of estimating parameters for a given assumed model, consider the joint pdf conditioned on a particular set of parameter values, $f_{\underline{Y}}(\underline{y}|\underline{b})$. This joint pdf can be thought of as the probability that different sets of possible observations would occur given the parameter values $\underline{b}$. In parameter estimation, the elements of $\underline{y}$ are known and we would like to estimate $\underline{b}$. A reasonable requirement of the estimates is that they maximize the probability of obtaining the observations, $\underline{y}$. This requirement is imposed by defining the objective function using the likelihood function, $l(\underline{b}|\underline{y})$, which is defined as:

$$l(\underline{b}|\underline{y}) = f_{\underline{Y}}(\underline{y}|\underline{b}). \tag{A1}$$

If the true errors are from a joint, normal distribution, the likelihood function equals (Brockwell and Davis, 1987, p. 247):

$$l(\underline{b}|\underline{y}) = \left(\frac{1}{2\pi}\right)^{ND/2} \left|\underline{V}(\underline{\varepsilon})\right|^{-1/2} \exp\left\{-\frac{1}{2}\underline{e}^T(\underline{V}(\underline{\varepsilon}))^{-1}\underline{e}\right\} \quad , \tag{A2}$$

where, as in equation 1 and 2,
$\underline{e} = \underline{y} - \underline{y}'$,
$\underline{y}'$ is a function of $\underline{b}$, and
ND is the number of observations.

Replacing $\underline{V}(\underline{\varepsilon})$ using equation C21 (see below), taking the natural log, and multiplying by -2 produces the maximum-likelihood objective function:

$$S'(\underline{b}) = -2\ln(l(\underline{b}|\underline{y})) = ND\ln 2\pi - \ln\left|\frac{1}{\sigma^2}\underline{\omega}\right| + \underline{e}^T\left(\frac{1}{\sigma^2}\underline{\omega}\right)\underline{\epsilon}. \tag{A3}$$

Because of the multiplication by a negative number, the maximization problem becomes a minimization problem, and the objective is to determine the parameter estimates that minimize equation A3. To include prior estimates of the parameters, $\underline{e}$ and $\underline{\omega}$ are augmented as described in Appendix B, ND is replaced by ND+NPR, and the determinant of A3 is expanded so that A3 can be expressed as:

$$S'(\underline{b}) = (ND+NPR)\ \ln 2\pi + (ND+NPR)\ \ln\sigma^2 - \ln|\underline{\omega}_d| - \ln|\underline{\omega}_p| + \underline{e}^T\left(\frac{1}{\sigma^2}\underline{\omega}\right)\underline{e}, \tag{A4}$$

where $\underline{\omega}_d$ and $\underline{\omega}_p$ are the sections of the weight matrix applicable to dependent variable observations and prior estimates of the parameters, respectively.

For any assumed model, set of observations, and defined weight matrix used in the parameter-estimation procedure, $\underline{\omega}$ is approximated and ND, σ^2, and $\underline{\omega}$ are constant. Eliminating terms of equation A4 that do not depend on $\underline{b}$ and multiplying by σ^2 yields:

$$S(\underline{b}) = \underline{e}^T \underline{\omega} \underline{e}. \qquad \text{(A5)}$$

Thus, for the optimization process, the maximum-likelihood objective function equals the sum-of-squares objective function (eq. 2).

The development of equation A5 from the maximum-likelihood objective function requires that the true errors be from a joint, normal distribution, a condition not required when the equation is derived in other ways.

References

Brockwell, P.J and Davis, R.A., 1989, Time series, Theory and methods: New York, Springer-Verlag, 519 p.

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

APPENDIX B: CALCULATION DETAILS

Three aspects of the calculations needed for the nonlinear regression methods described in this work require more detailed explanation. These include a more detailed description of the matrices and vectors of equations 2, 3, 4, and 26, discussing a possible addition to equation 4a, and calculation of the damping parameter and convergence of equation 4.

Vectors and Matrices for Observations and Prior Information

The primary vectors and matrices of concern in nonlinear regression are the measured values of vector $\underline{y}$, the simulated values of vector $\underline{y}'$, the sensitivities of matrix $\underline{X}$, the weights of matrix $\underline{\omega}$, the residuals of vector $\underline{e}$ (equal to $\underline{y} - \underline{y}'$) and the true errors of vector $\underline{\varepsilon}$. These vectors and matrices, including terms for both the observations and prior information used in the regression, are as follows. Except for $\underline{e}$ and $\underline{\varepsilon}$, these vectors and matrices are used in equations 2, 3 and 4a of this report. The vectors $\underline{e}$ and $\underline{\varepsilon}$ are included here because they appear frequently in regression literature. A few common relationships are displayed at the bottom of this section using vector notation.

$$\underline{y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{ND} \\ \hline P_1 \\ P_2 \\ \vdots \\ P_{NPR} \end{Bmatrix}, \qquad \underline{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,NP} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,NP} \\ & & \vdots & \\ \underline{x_{ND,1}} & \underline{x_{ND,2}} & \underline{\cdots} & \underline{x_{ND,NP}} \\ a_{1,1} & a_{1,2} & \cdots & a_{1,NP} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,NP} \\ a_{NPR,1} & a_{NPR,2} & \cdots & a_{NRP,NP} \end{bmatrix}, \qquad \underline{\omega} = \begin{bmatrix} \underline{V} & \underline{0} \\ \underline{0} & \underline{U} \end{bmatrix}$$

$\underline{V}$ is the weighting for the observations; $\underline{U}$ is the weight matrix for the prior information, and it is assumed that the true errors in the observations are independent of the true errors in the prior information.

$$\underline{y}' = \begin{Bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_{ND} \\ P'_1 \\ P'_2 \\ \vdots \\ P'_{NPR} \end{Bmatrix}, \qquad \underline{e} = \begin{Bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{ND} \\ u_1 \\ u_2 \\ \vdots \\ u_{NPR} \end{Bmatrix}, \text{ and} \qquad \underline{\varepsilon} = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{ND} \\ \upsilon_1 \\ \upsilon_2 \\ \vdots \\ \upsilon_{NPR} \end{Bmatrix}$$

These definitions are used in the following sections of this appendix.

Quasi-Newton Updating of the Normal Equations

For problems with large residuals and a large degree of nonlinearity, Dennis and others (1981) suggest substituting $\underline{X}_r^T \underline{\omega}\, \underline{X}_r + \underline{R}_r$ for $\underline{X}_r^T \underline{\omega}\, \underline{X}_r$ into equation 4a at selected iterations, where $\underline{R}_r$ is an estimate of the difference between $\underline{X}_r^T \underline{\omega}\, \underline{X}_r$ and the Hessian matrix $\left(\frac{\partial^2 S(\underline{b})}{\partial \underline{b}^2}\right)$, and is calculated by quasi-Newton updating as (Dennis and others, 1981):

$$\underline{R}_r = \underline{\mathbf{0}} \qquad \text{for } r = 0$$

$$\underline{R}_r = t\,\underline{R}_{r-1} + \frac{\underline{u}\Delta\underline{g}_r^T + \Delta\underline{g}_r \underline{u}^T}{\rho_{r-1}\underline{d}_{r-1}^T\ \Delta\underline{g}_r} - \frac{\rho_{r-1}\,\underline{d}_{r-1}^T\ \underline{u}\Delta\underline{g}_r\Delta\underline{g}_r^T}{(\rho_{r-1}\,\underline{d}_{r-1}^T\ \Delta\underline{g}_r)} \qquad \text{for } r > 0 \qquad \text{(B1)}$$

where

$$\Delta\underline{g}_r = \underline{g}_r - \underline{g}_{r-1}\ ; \qquad \underline{g}_r = {}^T\underline{X}_r\ \underline{\omega}\,\underline{e}_r\ ;$$

$$\underline{e}_r = [\underline{y} - \underline{y}'(\underline{b})]\ ; \qquad \underline{u} = (\underline{X}_r - \underline{X}_{r-1})^T\,\underline{\omega}\,\underline{e}_r - t\,\underline{R}_{r-1}\,\rho_{r-1}\,\underline{d}_{r-1}\ ;$$

$$t = min\left\{\left|\rho_{r-1}(\underline{d}_{r-1}^{T})(\underline{X}_r - \underline{X}_{r-1})^T\underline{\omega}\ \underline{e}_r / [\rho_{r-1}(\underline{d}_{r-1}^{T})\underline{R}_{r-1}\rho_{r-1}(\underline{d}_{r-1})]\right|;1.0\right\}$$

and all other variables are defined after equations 1 and 4. $\underline{R}_r$ is calculated starting at r = 1, but is only included in equation 4a in later iterations. Performance of the method depends on when $\underline{R}_r$ is included. Cooley and Hill (1992) found that it is most advantageous to include $\underline{R}_r$ after the sum of squared, weighted residuals no longer changes very much at each parameter-estimation iteration. In the UCODE and MODFLOWP, $\underline{R}_r$ is included for all iterations after the sum of squared, weighted residuals decreases by less than a user-defined percentage over two iterations; in UCODE the value is set to one percent. In MODFLOWP, $\underline{R}_r$ also can be included after a user-specified number of iterations. The more elaborate criteria for inclusion of $\underline{R}_r$ suggested by Dennis and others (1981) require additional model simulations. Considering the large problems that are expected to be simulated with UCODE and MODFLOWP and the modest expected benefit obtainable in most circumstances, the more elaborate criteria seemed impractical and were not included. When $\underline{R}_r$ is included in equation (4a), the elements of the diagonal scaling matrix, $\underline{C}$, are calculated as $[(\underline{X}_r{}^T\underline{\omega}\underline{X}_r + \underline{R}_r)_{ii}$.

Calculating the Damping Parameter and Testing for Convergence

For problems with one or more log-transformed parameters, requiring the absolute value of equation 5 to be less than DMAX (MAX-CHANGE for UCODE) for any parameter-estimation iteration, and requiring equation 7 to be satisfied to achieve convergence, produces inconsistent results. The following example illustrates the problem as manifested when applying DMAX.

If the estimated parameter is b_i = log K, where K is hydraulic conductivity, and DMAX=2.0, placing the restriction on log K requires that $(\log K)^{r+1}$, the estimate at the next parameter-estimation iteration, be between $(\log K)^r$-2.0$(\log K)^r$ and $(\log K)^r$+2.0$(\log K)^r$. If K at parameter-estimation iteration r is close to 1.0, say K=1.1, the restriction requires $(\log K)^{r+1}$ to be between -0.041 and 0.124, so that K^{r+1} is required to be within the narrow range 0.91 and 1.33. If K at parameter-estimation iteration r is far from 1.0, say $K=1\times10^{-4}$, the restriction requires that $(\log K)^{r+1}$ be between -12.00 and 4.00, so that K^{r+1} is allowed to vary within the very wide range of 1×10^{-12} and 1×10^{4}. More physically meaningful limitations are produced if the restriction is placed on the native parameter, which requires that K be between 0.0 and 3.3 in the first situation and between 0.0 and 3×10^{-4} in the second situation. In both situations, the lower limit of 0.0 is a result of estimating a log-transformed parameter and is always the lower limit for a log-transformed parameter when DMAX $\geq$ 1.0.

To address this problem, a number of quantities are calculated at each parameter-estimation iteration, as shown in table B1. The circumstances treated individually are: (1) parameters that are not log-transformed, (2) parameters that are log-transformed and the regression is trying to increase their value ($d^r_i<0$), and (3) parameters that are log-transformed and the regression is trying to increase their value ($d^r_i>0$). The objective that allows a single damping parameter to be chosen despite the individual circumstances is that the smallest of all value is needed, regardless of the of how it is calculated.

Table B1.-- Quantities used to test for convergence and to calculate damping parameter ρ_r for parameter-estimation iterations.

Parameter category	A. Convergence test on the fractional change in the native parameter value [1]	B. Equation for ρ_r if the absolute value of the quantity in column A is larger than DMAX [2]	C. Fractional parameter change used to adjust ρ_r for oscillation control [5]
Untransformed	d_i^r/b_i^r	[3] ρ_r=DMAX / ($\lvert d_i^r/b_i^r \rvert$)	$d_i^r/\lvert b_i^r \rvert$
Transformed, $d^r_i>0$	d_i^r-1	ρ_r=ln(DMAX+1)) / d_i^r	$d_i^r/\lvert b_i^r \rvert$
Transformed, $d^r_i<0$	d_i^r-1	[4] ρ_r=ln(DMAX-1)) / d_i^r	$d_i^r/\lvert b_i^r \rvert$

1. Largest absolute value needs to be less than TOL for convergence.
2. Otherwise ρ_r=1.0, except as needed for oscillation control. For each parameter-estimation iteration, the smallest of all ρ_r values is used and printed with the related parameter number in the output file.
3. To enable parameter values to increase more quickly after being assigned values near zero, b_i^0 replaces b_i^r if $\lvert b_i^r \rvert < \lvert b_i^0 \rvert /10^3$
4. Only use if DMAX<1.0; otherwise, ρ_r=1.0 except as determined for oscillation control.
5. Equation B5.

The equations are derived as follows. For untransformed parameters, the fractional change of the native parameter value simply equals the change calculated by solving equation 4a divided by the value of the parameter value, or d_i^r/b_i^r, where d_i^r is the ith element of vector $\underline{d}_r$ and b_i^r is the ith element of vector $\underline{b}_r$. For log-transformed parameters, the fractional change in the native value equals $(\exp(b_i^{r+1}) - \exp(b_i))/\exp(b_i^r)$, or, equivalently, $(\exp(b_i^{r+1})/\exp(b_i^r))$-1.0. Substituting $\exp(d_i^r) = \exp(b_i^{r+1})/\exp(b_i^r)$, which is derived from equation (4b) with ρ_r = 1.0, yields

$$\exp(d_i^r) - 1.0 \qquad \text{(B2)}$$

In column B of table B1, the equation for untransformed parameters is obvious, and the

equations for log-transformed parameters are derived using equation (B2). If $d_j^r > 0.0$, the DMAX restriction requires that $(\exp(b_j^{r+1})/\exp(b_j^r))-1.0 \leq$ DMAX, or, equivalently,

$$\rho_r d_i^r \leq \ln (\text{DMAX}+1.0), \qquad i=1,NP \qquad \text{(B3)}$$

If $d_j^r < 0.0$ and DMAX < 1.0, $(\exp(b_j^{r+1})/\exp(b_j^r))-1.0 > -$DMAX, or, equivalently,

$$\rho_r d_i^r > \ln (1.0-\text{DMAX}). \qquad \text{(B4)}$$

Dividing B3 and B4 by d_i^r and noting that division by a negative number transforms a "<" to a ">" gives the equations in table B1, column B.

An exception to equation (B4) is described in footnote 4 of table B1. This exception applies to log-transformed parameters if DMAX $\geq$ 1.0, because, as mentioned previously, the exponential of a log-transformed parameter is always greater than 0.0, and can never decrease enough to require ρ_r to be less than 1.0 if DMAX $\geq$ 1.0. Thus, if $d_i^r < 0$ for a log-transformed parameter and DMAX $\geq$ 1.0, parameter i is excluded from consideration when calculating ρ_r.

Oscillation control is achieved using a slightly modified version of the method described by Cooley (1983a, p. 1274; 1993). A preliminary damping parameter, ρ_r*, is calculated to minimize oscillations according to the following, where j_r is the parameter with the smallest ρ_r in iteration r.

$$\text{DMX}_r = d_i^r/|b_i^r|$$

$$\rho_r^* = 1 \qquad r = 0 \text{ or } j_r \neq j_{r-1} \qquad \text{(B5a)}$$

$$\left.\begin{array}{l} s = \text{DMX}_r/(\rho_{r-1}\,\text{DMX}_{r-1}) \\ \text{If } s \geq -1 \qquad \rho_r^* = \dfrac{3+s}{3+|s|} \\ \text{If } s < -1 \qquad \rho_r^* = 1/(2|s|) \end{array}\right\} \quad r>0 \text{ and } \quad j_j = j_{r-1} \qquad \text{(B5b)}$$

where the condition on j has been added to Cooley's method.

Typically, DMAX is larger than 1.0 and less than about 2.0. Use values less than 1.0 to reduce excessive parameter-value oscillations. Note that values less than 1.0 do not prohibit param-

eter values from changing sign because b_i^o replaces b_i^r when calculating ρ_r if $|b_i^r| < |b_i^o|/10^3$.

Solving the Normal Equations

Using double precision as suggested by Stewart (1972, p. 226-227), equation (4) has been solved accurately and efficiently in many applications using Cholesky LDL^T decomposition (Dennis and Schnabel, 1983, p. 50-51). Exceptions were plagued by strong correlations between parameters or insensitive parameters, and were resolved by reparameterization. Dennis and Schnabel (1983, p. 221) and Seber and Wild (1989, p. 621) suggest that solving the alternative formulation $\underline{X}\,\underline{d} = (\underline{y} - \underline{y})$ using QR or singular-value decomposition (Dennis and Schnabel, 1983, p. 49-51; Seber and Wild, 1989, p. 680-681; Press and others, 1989, p. 52-63) is more stable, but it is unclear whether or not they used the scaling and Marquardt parameter which adds stability to equation 4. Press and others (1989, p. 515-520) suggest using singular-value decomposition for linear regression, but use Gauss-Jordon elimination to solve a variation of equation 4 that includes similar scaling and implementation of the Marquardt parameter for nonlinear regression. Considering the success experienced using Cholesky decomposition, Cholesky decomposition is used in UCODE and MODFLOWP.

References

Cooley, R.L., 1983a, Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2, Applications: Water Resources Research, v. 19, no. 3, p. 662-676.

Cooley, R.L., 1993, Regression modeling of ground-water flow, Supplement 1 -- Modifications to the computer code for nonlinear regression solution of steady-state ground-water flow problems: U.S Geological Survey Techniques of Water Resources Investigations, book 3, chapt. B4, supplement 1, 8p.

Cooley, R.L. and Hill, M.C., 1992, A comparison of three Newton-like nonlinear least-squares methods for estimating parameters of ground-water flow models: in Russel, T.F., Ewing, R.E., Brebbia, C.A., Gray, W.G., and Pinder, G.F., eds., Proceeding, Computational methods in water resources IX: Denver, CO, v. 1, Numerical methods in Water Resources, ***publisher, p. 379-386.

Dennis, J.E. Gay, D.M. and Welsch, R.E., 1981, An adaptive nonlinear least-squares algorithm: ACM Transactions on Mathematical Software, v. 7, no. 3, p. 348-368.

Dennis, J.E., and Schnabel, R.B., 1983, Numerical methods for unconstrained optimization and nonlinear equations: Englewood Cliffs, New Jersey, Prentice-Hall, 378 p.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., 1992, Numerical recipes: Cambridge, Great Britain, Press Syndicate of the University of Cambridge, 2nd Edition, 963 p.

Seber, G.A.F., and C.J. Wild 1989, Nonlinear Regression, John Wiley & Sons, NY, 768 p.

Stewart, G.W., 1972, Introduction to matrix computations: New York, Academic Press, 423 p.

APPENDIX C: TWO IMPORTANT PROPERTIES OF REGRESSION

This appendix presents two basic properties of weighted linear regression, which are generally known as the Gauss-Markov theorem, in a manner which emphasizes the difficulties produced when the regression is nonlinear. More traditional derivations of the Gauss-Markov theorem can be found in Bard (1974) and Beck and Arnold (1977).

The two properties of concern are:

1. Parameters estimated by linear regression are unbiased.
2. The weight matrix needs to be defined a particular way for the parameter estimates to have the smallest variance, and for the parameter variance-covariance matrix to be calculated using equation 26.

Definitions and identities used in both proofs are presented first followed by the two proofs.

Identities

True linear model. The true model is unknown and correctly represents the system of concern. A true linear model can be represented as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_j X_j \ldots \beta_\nu X_\nu + \varepsilon \qquad E(\varepsilon)=0 \qquad (C1)$$

where,
y is a measurement of the dependent variable (here, hydraulic heads, flows, and so on);
β_j are the true (unknown) parameter values;
X_j are the independent variables (here, location, depth, time, etc.);
ν is the number of terms in the true model; and
ε is the true error, and needs to have a mean of zero, as shown, for the regression to be valid.

True nonlinear model. The true nonlinear model can not be represented as in C1, and requires the more general form presented after equation 1 -- that is, using vector notation, $y= F(\underline{\beta},\underline{\zeta}) + \varepsilon$., where F repesents the form of the unknown nonlinear function, $\underline{\zeta}$ represents the independent variables, and the other symbols are as defined for equation C1.

Linearized true nonlinear model. A linearized true nonlinear model is defined here for the purposes of this discussion. The model is linearized using a Taylor series expansion about the true parameter values and has the form of C1, within a constant additive vector, but the X_j are derivatives of the nonlinear model with respect to the parameters, evaluated at the true parameter values. Linearized models are further discussed below.

Approximate linear model. The approximate model is the model being developed to rep-

resent the system of concern, and is the model to be calibrated. A linear approximate model can be represented as:

$$y = b_0 + b_1X_1 + b_2X_2 + \ldots \beta_jX_j \ldots \beta_nX_n + e = y' + e \qquad (C2)$$

where
y is a measurement of the dependent variable (here, hydraulic heads, flows, and so on), as above;
b_j are the estimated parameter values;
X_j are the independent variables (here, location, depth, time, etc.);
n is the number of terms in the approximate model;
e is the true error; and
y' is the simulated equivalent of the measured dependent variable.

Approximate nonlinear model. As for the true model, the approximate nonlinear model can not be represented as in C1, and requires the more general form presented after equation 1 -- that is, using vector notation, $y= f(\underline{b},\underline{\xi}) + e$, where f repesents the form of the unknown nonlinear function, $\underline{\xi}$ represents the independent variables, and the other symbols are as defined for equation C2.

Linearized approximate nonlinear model. The linearized approximate nonlinear model is produced using a Taylor series expansion about a defined set of parameter values, $\underline{b}^*$. Within an additive vector that is constant for any b* (this vector is needed to derive the interative equation 4a, but is not important to the present discussion), the linearized approximate nonlinear model can be expressed in the form of equation C2. In this situation however, the X_j are no longer simply independent variables, but equal the derivatives of the approximate linear model with respect to the parameter values, evaluated at $\underline{b}^*$. These derivatives were defined for equation 8 and have the following characteristics:
1. Like the X_j for linear problems, the derivatives include the independent variables; but they also include the effects of other aspects of the nonlinear model.
2. Because of model nonlinearity, the values of the derivatives depend on the parameter values in $\underline{b}^*$.
3. The derivatives generally are called sensitivities because the represent the senstivity of the simulated value to a change in the parameter value.

Linearized models reproduce the same simulated value at $\underline{b}^*$ as the nonlinear model, by definition, and often closely mimick the nonlinear model for values of $\underline{b}$ near $\underline{b}^*$. As the linearized model is evaluated for values further from b*, simulated values will vary from those of the approximate nonlinear model depending on its degress of nonlinearity. This deviation is apparent in the sum-of-squared residuals surfaces of figure 2, which shows an objective-function surface calculated using the Theis equation as the approximate nonlinear model, and two objective-function sur-

faces calculated using a linearized approximate model. The linearized surfaces closely mimic the nonlinear surface near the b* values, marked by an x, and mimick it less well, and even poorly, for increasingly different sets of parameter values.

The importance of X and X. The different symbols, X and X, are used in C1 and C2 because they may be different. For linear problems they often are the same, but differences occur when the approximate model includes more or fewer terms than the true model (n does not equal ν). In addition, errors in measuring the independent variables could affect X_j and X_j, but this problem is not addressed in this report.

For a nonlinear model, an equation of the form C2 is used to represent the linearized approximate nonlinear model, as discussed above. For nonlinear problems, the X_j also vary depending of the set of parameter values about which model is linearized, and the difference between X_j and X_j, becomes greater as the optimized parameter values differ more from the true parameter values.

Functional form of observations. The ith observation used in the regression can be expressed in terms of the true linear model as:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_j X_{ji} \ldots \beta_\nu X_{\nu i} + \varepsilon_i \qquad \text{(C3)}$$

and in terms of the approximate linear or linearized model as:

$$y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots \beta_j X_j \ldots \beta_n X_{ni} + e_i = y'_i + e_i \qquad \text{(C4)}$$

All observations used in the regression together can be expressed in terms of the true linear model using matrix notation (vectors are underlined lower case or greek letter, matrices are underlined capital letters) as:

$$\underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon} \qquad \text{(C5)}$$

and in terms of the approximate linear or linearized model as:

$$\underline{y} = \underline{X}\,\underline{b} + \underline{e} \qquad \text{(C6)}$$

For the linearized approximate nonlinear model, each element of the $\underline{X}$ array is one of the derivatives, or sensitivities, discussed above. An expanded form was shown in appendix B.

Normal equations. To calculate parameter values that produce the closest match to the ob-

servations, the weighted least-squares objective function is minimized with respect to the parameter values. Using the approximate linear model, this produces what are called the normal equations, expressed in matrix notation as:

$$\underline{b} = (\underline{X}^T \underline{\omega}\, \underline{X})^{-1}\, \underline{X}^T \underline{\omega}\, \underline{y} \qquad \text{(C7)}$$

Despite some variation, the similarity between equations C7 and 4a is apparent, with the major difference being that C7 produces the actual optimal parameter values after being evaluated just once, while equation 4a produces a vector that is used to update the parameter values, and optimal parameter values are obtained only after a number of parameter-estimation iterations. Because, as noted above, the interative nature of the equations are not central to the issue addressed here, equation C7 is used.

Random variables. The primary random variable in the above equations are the true errors ε. Then, noting that functions of random variable are random, y is random from C1, e is random from C2, and b is random from C7. Because for any step of the analysis teh nonlinear model is linearized and X is evaluated for a defined set of parameters, X it is not thought of as being random.

Expected value. The expected value can be taken of any term, and is represented as E(.) or E[.], where the term appears within the parentheses or brackets. As noted above, ε has a mean of zero, so E(ε)=0.

Variance-covariance matrix of a vector. Proof 2 requires the evaluation of the variance-covariance matrix of the vector of estimated parameters and the true errors. The variance-covariance matrix of any vector $\underline{v}$ is calculated as $E[\ (\underline{v}-E(\underline{v}))\ (\underline{v}-E(\underline{v}))^T\ ]$.

Proof 1: **Parameters estimated by linear regression are unbiased.**

Take the expected value of the optimized parameters, as calculated using equation (C7):

$$E(\underline{b}') = (\underline{X}^T \underline{\omega}\, \underline{X})^{-1}\, \underline{X}^T \underline{\omega}\, E(\underline{y}) = (\underline{X}^T \underline{\omega}\, \underline{X})^{-1}\, \underline{X}^T \underline{\omega}\, \underline{X}\, \underline{\beta} \qquad \text{(C8)}$$

If $\underline{X} = \underline{X}$,

$$(\underline{X}^T \underline{\omega}\, \underline{X})^{-1}\, \underline{X}^T \underline{\omega}\, \underline{X} = \underline{I} \qquad \text{(C9)}$$

where I is an identity matrix. Substituting C9 into C8 yields:

$$E(\underline{b}) = \underline{\beta} \qquad \text{(C10)}$$

Thus, if $\underline{X} = \underline{X}$, the expected values of the estimates equal the true values, which means that the estimates are unbiased. In nonlinear models, the equality is unlikely to be true, so that unbiasedness is not guaranteed for nonlinear models, even if the mdoel if correct.

Proof 2: The weight matrix needs to be defined in a particular way for the parameter estimates to have the smallest variance.

It is desirable to estimate parameters with the smallest variance and, therefore, the greatest precision. The variance of the parameter estimates occur as the diagonal terms in the variance-covariance matrix of the parameters, which is calculated using the equation defined above as:

$$\underline{V}(\underline{b}) = E[(\underline{b}-E(\underline{b}))\,(\underline{b}-E(\underline{b}))^T]. \qquad \text{(C11)}$$

replacing $\underline{b}$ with equation (C7) and E(b) with equation (C10) yields:

$$\underline{V}(\underline{b}) = E[\,((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y} - \underline{\beta})\,((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y} - \underline{\beta})^T]. \qquad \text{(C12)}$$

Expanding the product on the right-hand side produces an equation with four terms:

$$\underline{V}(\underline{b}) = E[\,((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y})\,((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y})^T \; - \qquad \text{(C13)}$$
$$((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y})\,\underline{\beta}^T \; - \; \underline{\beta}\,((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y})^T \; + \; \underline{\beta}\,\underline{\beta}^T\,]$$

Use the matrix property $(\underline{AB})^T=\underline{B}^T\underline{A}^T$ to rearrange the first term as:

$$((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,\underline{y})\,((\,\underline{X}^T\underline{\omega}\,\underline{y})^{-1}\;\underline{X}^T\underline{\omega}\,\underline{y})^T = \qquad \text{(C14)}$$
$$(\,\underline{X}^T\underline{\omega}\,\underline{y})^{-1}\;\underline{X}^T\underline{\omega}\,\underline{y}\,\underline{y}^T\,\underline{X}(\underline{X}^T\underline{\omega}\,\underline{X})^{-1}$$

Take the expected value of each term and note that only $\underline{y}$ is stochastic to obtain:

$$\underline{V}(\underline{b}) \;=\; (\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,E[\underline{y}\,\underline{y}^T]\underline{\omega}\,\underline{X}(\,\underline{X}^T\underline{\omega}\,\underline{X})^{-1} - \qquad \text{(C15)}$$
$$(\,\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,E[\underline{y}])\;\underline{\beta}^T \; - \; \beta((\,\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,E[\underline{y}])^T \; + \; \underline{\beta}\,\underline{\beta}^T$$

In the first term, apply $\underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}$, so that:

$$E[\underline{y}\,\underline{y}^T] = E[(\underline{X}\,\underline{\beta} + \underline{\varepsilon})\,(\underline{X}\,\underline{\beta} + \underline{\varepsilon})^T] \qquad \text{(C16)}$$
$$= E[(\underline{X}\,\underline{\beta})(\underline{X}\,\underline{\beta})^T + (\underline{X}\,\underline{\beta})\,\underline{\varepsilon}^T + \underline{\varepsilon}\,\underline{X}\,\underline{\beta}^T + \underline{\varepsilon}\,\underline{\varepsilon}^T]$$

Taking the expected value of each term, and noting that only $\underline{\varepsilon}$ is stochastic and that the second and third terms of equation C16 equal zero because $E[\underline{\varepsilon}]=0$ produces:

$$E[\underline{y}\,\underline{y}^T] = (\underline{X}\,\underline{\beta})(\underline{X}\,\underline{\beta}^T) + E[\underline{\varepsilon}\,\underline{\varepsilon}^T] = \underline{X}\,\underline{\beta}\,\underline{\beta}^T\underline{X}^T + E[\underline{\varepsilon}\,\underline{\varepsilon}^T] \quad \text{(C17)}$$

Note that $E[\underline{\varepsilon}\,\underline{\varepsilon}^T] = \underline{V}(\underline{\varepsilon})$, the variance-covariance matrix of the true errors. This can be derived by applying the standard equation for calculating the variance-covariance matrix of a vector, so that $\underline{V}(\underline{\varepsilon}) = E[(\underline{\varepsilon}-E(\underline{\varepsilon}))\,(\underline{\varepsilon}-E(\underline{\varepsilon}))^T]$, and noting that $E(\underline{\varepsilon})=\underline{0}$.

Substituting these results into equation (C15) yields:

$$\begin{aligned}\underline{V}(\underline{b}) = {} & (\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,\underline{X}\ \ \underline{\beta}\,\underline{\beta}^T\ \ \underline{X}^T\underline{\omega}\,\underline{X}\ \ (\underline{X}^T\underline{\omega}\,\underline{X})^{-1} \\ & + (\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,E[\underline{\varepsilon}\,\underline{\varepsilon}^T]\,\omega\,\underline{X}\ \ (\underline{X}^T\underline{\omega}\,\underline{X})^{-1} \\ & - ((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,\underline{X}\ \ \underline{\beta})\,\underline{\beta}^T\ -\ \ \underline{\beta}((\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,\underline{X}\ \underline{\beta}\ )^T\ \ +\ \ \underline{\beta}\,\underline{\beta}^T\end{aligned} \quad \text{(C18)}$$

If $\underline{X} = \underline{X}$, then $(\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,\underline{X} = \underline{I}$, which gives the following:

$$\begin{aligned}\underline{V}(\underline{b}) = {} & \underline{\beta}\,\underline{\beta}^T + (\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\,\underline{X}^T\underline{\omega}\,E[\underline{\varepsilon}\,\underline{\varepsilon}^T]\,\underline{\omega}\,\underline{X}(\underline{X}^T\underline{\omega}\,\underline{X})^{-1} \\ & - \underline{\beta}\,\underline{\beta}^T - \underline{\beta}\,\underline{\beta}^T + \underline{\beta}\,\underline{\beta}^T\end{aligned} \quad \text{(C19)}$$

The $\underline{\beta}\,\underline{\beta}^T$ terms cancel, leaving:

$$\underline{V}(\underline{b}) = (\underline{X}^T\underline{\omega}\,\underline{X})^{-1}\ \underline{X}^T\underline{\omega}\,E[\underline{\varepsilon}\,\underline{\varepsilon}^T]\,\underline{\omega}\,\underline{X}\ \ (\underline{X}^T\underline{\omega}\,\underline{X})^{-1} \quad \text{(C20)}$$

If the weight matrix is defined such that

$$E[\underline{\varepsilon}\,\underline{\varepsilon}^T] = \underline{V}(\underline{\varepsilon}) = \sigma^2\,\underline{\omega}^{-1}, \quad \text{(C21)}$$

where σ^2 is the true common error variance, equation C20 reduces to:

$$\underline{V}(\underline{b}) = \sigma^2\,(\underline{X}^T\underline{\omega}\,\underline{X})^{-1} = s^2(\underline{X}^T\underline{\omega}\,\underline{X})^{-1} \quad \text{(C22)}$$

where the last equals sign is approximate and s2, the calculated error variance, approximates the unknown true common error variance. Equation C22 is the expression commonly used to calculate the variance-covariance matrix for the parameter values, but really only applies if $\underline{X} = \underline{X}$, and C21 applies.

If the equation for $\underline{V}(\underline{b})$ cannot be simplified to equation C22, equations of the form C18 or C20 should be used to calculate the variance-covariance matrix of the of the parameter estimates, although it is unclear how to evaluate C18 because $\underline{\beta}$ is unknown. For linear problems, equarion

C19 always produces a larger variance for the parameters and simulated predictions than is produced by otehr possible equations (Bard, 1974; Beck and Arnold, p. 232-234). Thus, the smallest variance parameter estimates are those for which equation C21 applies and, therefore, for which $\underline{X} = \underline{X}$ and the weighting is defined such that $\underline{\omega} = \underline{V(\varepsilon)}^{-1}$ (the weighting is closely related to the variance-covariance matrix of the true, unknown errors). Although not always valid, linear theory provides the only available guidance for defining the weight matrix for nonlinear problems.

Reference

Bard, Jonathon, 1974, Nonlinear parameter estimation: New York, Academic Press, 341p.

Beck, J.V. and Arnold, K.J., 1977, Parameter estimation in engineering and science: New York, John Wiley and Sons, 501p.

APPENDIX D: CRITICAL VALUES FOR THE CORRELATION COEFFICIENT FOR NORMAL PROBABILITY GRAPHS, R_N^2

Table D1: Critical values of R^2_N below which the hypothesis that the weighted residuals are independent and normally distributed is rejected at the stated significance level (from Shapiro and Francia, 1972; Brockwell and Davis, 1987, p.304)

[ND, the number of observations (N-OBSERVATIONS in the UCODE documentation);
NPR, the number of prior information values (NPRIOR in the UCODE documentation)]

ND or	Significance level		ND or	Significance level	
ND+NPR	0.05	0.10	ND+NPR	0.05	0.10
35	0.943	0.952	81	0.970	0.975
			83	0.971	0.976
50	0.953	0.963	85	0.972	0.977
51	0.954	0.964	87	0.972	0.977
53	0.957	0.964	89	0.972	0.977
55	0.958	0.965			
57	0.961	0.966	91	0.973	0.978
59	0.962	0.967	93	0.973	0.979
			95	0.974	0.979
61	0.963	0.968	97	0.975	0.979
63	0.964	0.970	99	0.976	0.980
65	0.965	0.971			
67	0.966	0.971	131	0.980	0.983
69	0.966	0.972	200	0.987	0.989
71	0.967	0.972			
73	0.968	0.973			
75	0.969	0.973			
77	0.969	0.974			
79	0.970	0.975			

References

Brockwell, P.J and Davis, R.A., 1989, Time series, Theory and methods: New York, Springer-Verlag, 519 p.

Shapiro, S.S., and Francia, R.S., 1972, An approximate analysis of variance test for normality: Journal of the American Statistical Association, v. 67, p. 215-216.