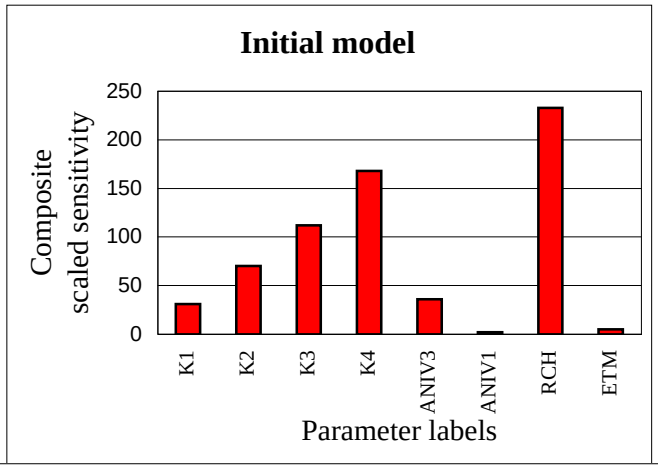
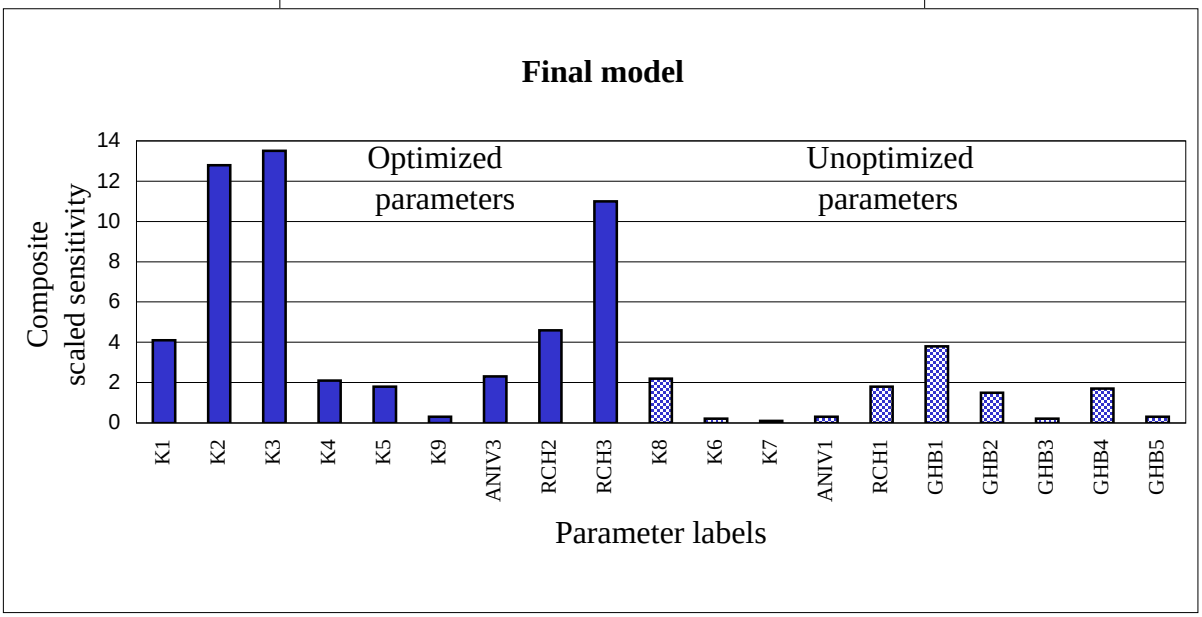
METHODS AND GUIDELINES FOR EFFECTIVE MODEL CALIBRATION

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.

Second printing, 2001
For the capabilities described in this report, MODFLOWP has been replaced by MODFLOW-2000. MODFLOW-2000 and UCODE are available from
http://water.usgs.gov/software/ground_water.html/
MODFLOW-2000 is documented by the following reports:

Harbaugh, A.W., Banta, E.R., Hill, M.C., and McDonald, M.G., 2000, MODFLOW-2000, the U.S. Geological Survey modular ground-water model, User's guide to the Modularization concepts and the Ground-Water Flow Process: U.S. Geological Survey Open-File Report 00-92, 121 p.

Hill, M.C., Banta, E.R., Harbaugh, A.W., and Anderman, E.R., 2000, Documentation of MODFLOW-2000, the U.S. Geological Survey modular ground-water model, User's guide to the Observation, Sensitivity, and Parameter-Estimation Process and three post-processing programs: U.S. Geological Survey Open-File Report 00-184, 209 p.

Anderman, E.R. and Hill, M.C., 2000, Documentation of MODFLOW-2000, the U.S. Geological Survey modular ground-water model, the Hydrogeologic Unit Flow (HUF) Package: U.S. Geological Survey Open-File Report 00-342, 89p.

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.

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: Guidelines for effective model calibration

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: Guidelines for effective model calibration

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.

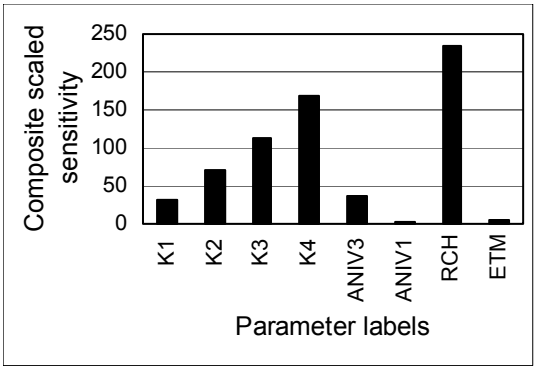


Figure 3: Composite scaled sensitivities for parameters of the initial Death Valley regional ground-water flow system model of D'Agnese and others (1998, in press). K* are hydraulic-conductivity parameters, ANIV* are vertical anisotropy parameters, RCH is an areal recharge parameter, and ETM is a maximum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

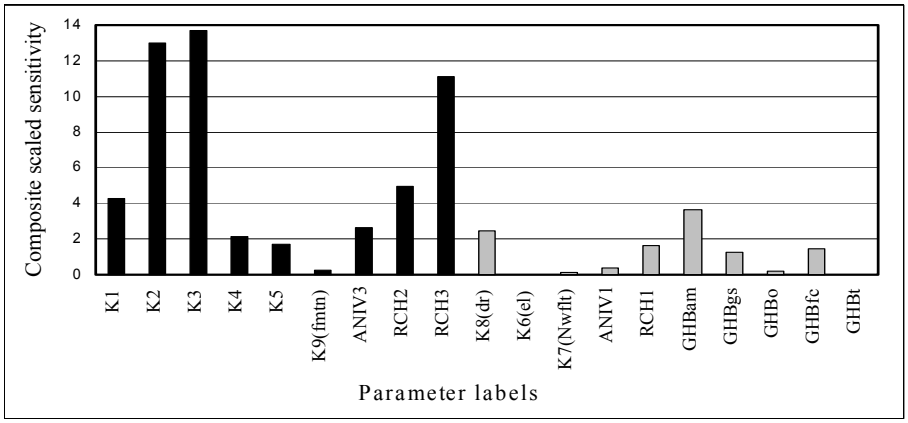


Figure 4: Composite scaled sensitivities for the parameters of the final calibrated Death Valley regional ground-water system model of D'Agnese and others (in press). K* are hydraulic-conductivity parameters, ANIV* are vertical anisotropy parameters, RCH is an areal recharge parameter, ETM is a maximum 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.

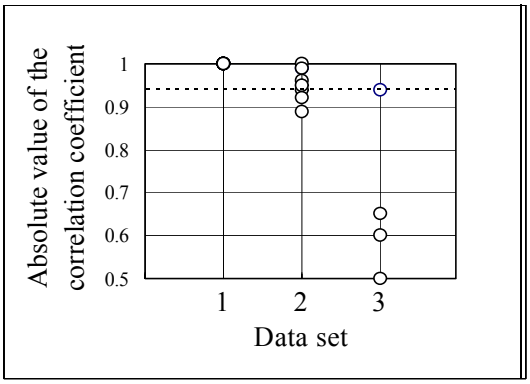


Figure 5: Parameter correlation coefficients for the same five parameters for three data sets from the Cape Cod sewage plume model of Anderman and others (1996), evaluated for the initial parameter values. Data set 1 includes only hydraulic heads, and all parameters are extremely correlated (the absolute 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

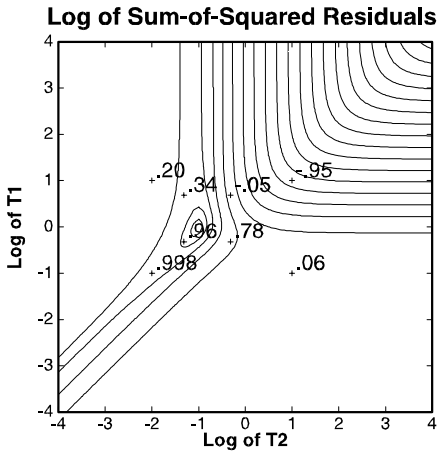


Figure 6: Correlation of parameters T1 and T2 of figure 1 at specified parameter values, plotted on a $\log_{10}$ weighted least-squares objective function surface. T1 and T2 are in square meters per day. (from Poeter and Hill, 1997)

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.

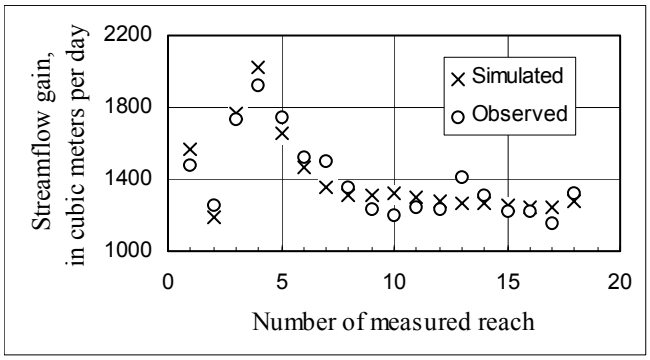


Figure 7: Observed and simulated streamflow gains for model CAL3 of Hill and others (1998).

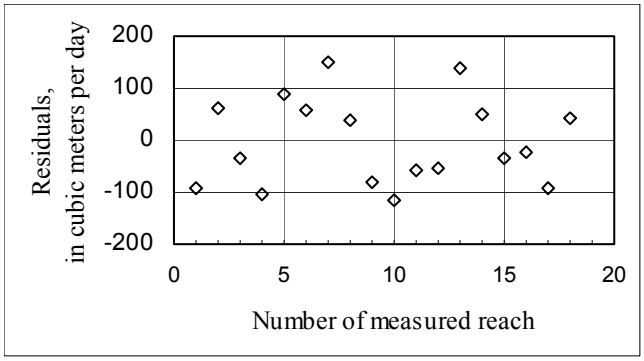


Figure 8: Residuals equal to the observed minus the simulated streamflow gains of figure 7.

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.

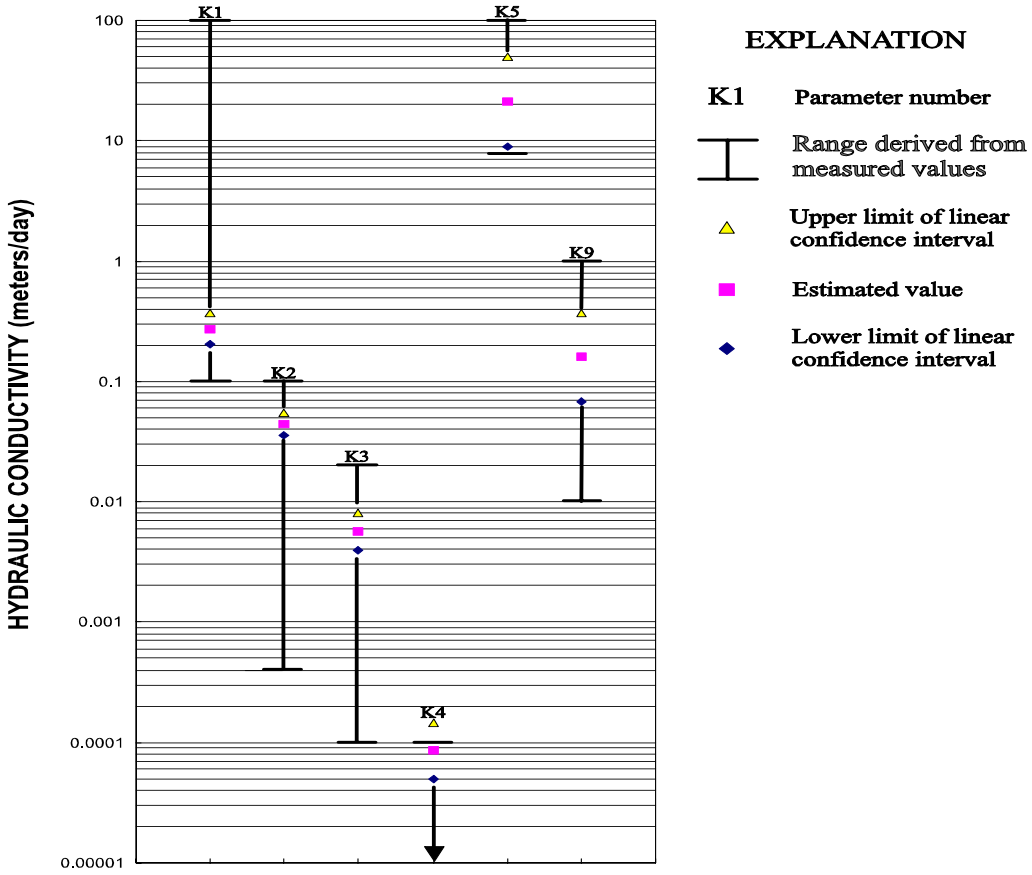


Figure 10: Optimized hydraulic-conductivity values, 95-percent linear confidence intervals, and the range of hydraulic-conductivity values derived from field and laboratory data. (from D'Agnese and others, 1998)

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.

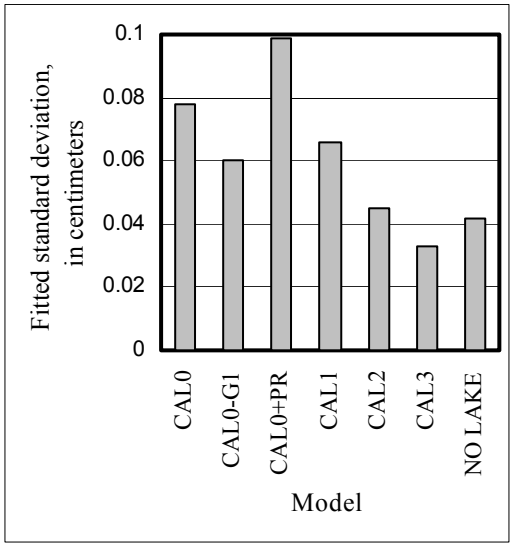


Figure 11: Fitted standard deviations for hydraulic heads for seven models from a controlled experiment in model calibration. (from Hill and others, 1998)

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.

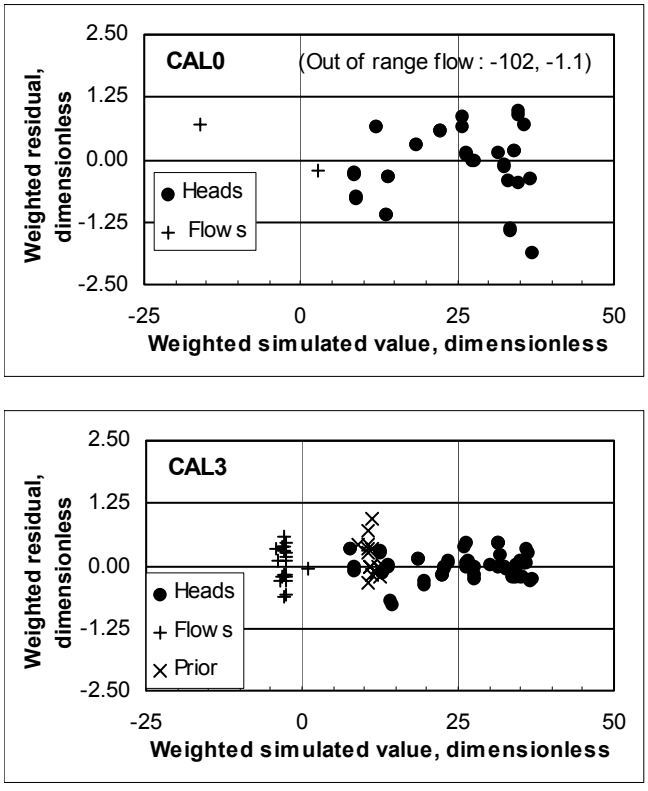


Figure 12: Weighted residuals versus weighted simulated values for models CAL0 and CAL3 of Hill and others (1998).

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".

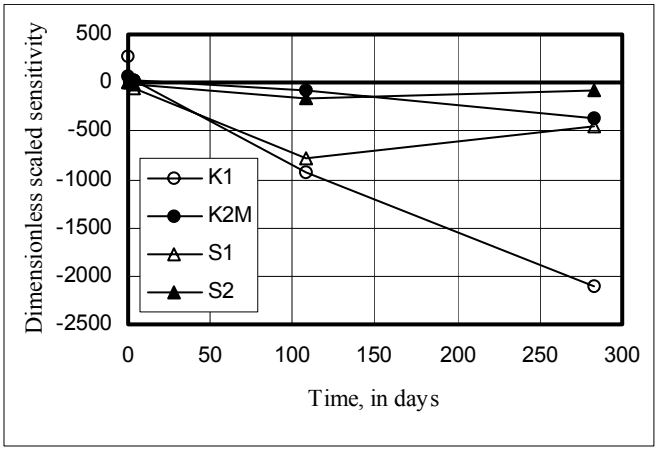


Figure 13: Dimensionless scaled sensitivites plotted against time. The values are from well 2 of test case 1 of Hill (1992). Time zero has no pumpage; at subsequent times constant pumpage is applied. The K1 parameter represents the hydraulic conductivity in the top of two layers. The K2M parameter represents 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

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

and their associated confidence intervals, calculated as:

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

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.

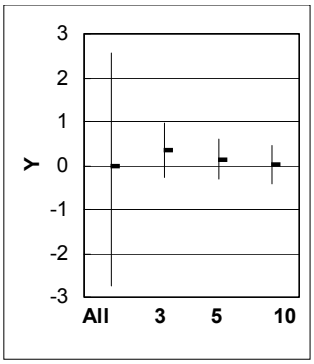


Figure 14: Confidence intervals for a population mean given different sample sizes. The population is composed of 300 random normally distributed numbers with a range noted by the bar labeled "All" and a mean noted by the mark in the center of that bar. The other bars are labeled with the sample size 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.

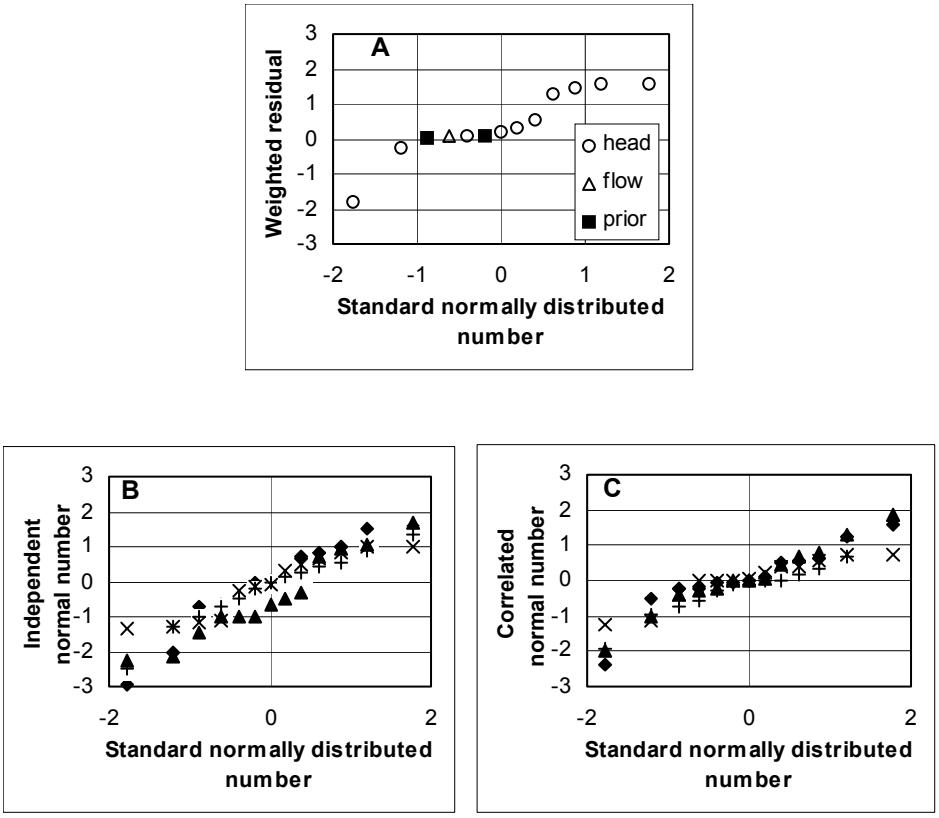


Figure 15: Normal probability graphs for the steady-state version of test case 1 of Hill (1992), including (A) weighted residuals, (B) normally distributed, uncorrelated random numbers, and (C) normally distributed random numbers correlated as expected given the fitting of the regression. In B and C,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