WATER RESOURCES RESEARCH, VOL. 34, NO. 11, PAGES 2939–2947, NOVEMBER 1998

Robust estimation of hydrogeologic model parameters

Stefan Finsterle and Julie Najita

Earth Sciences Division, Lawrence Berkeley National Laboratory, University of California, Berkeley

Abstract. Inverse modeling has become a standard technique for estimating hydrogeologic parameters. These parameters are usually inferred by minimizing the sum of the squared differences between the observed system state and the one calculated by a mathematical model. The robustness of the least squares criterion, however, has to be questioned because of the tendency of outliers in the measurements to strongly influence the outcome of the inversion. We have examined alternative approaches to the standard least squares formulation. The robustness of these estimators has been tested by means of Monte Carlo simulations of a synthetic experiment, in which both non-Gaussian random errors and systematic modeling errors have been introduced. The approach was then applied to data from an actual gas-pressure-pulse-decay experiment. The study demonstrates that robust estimators have the potential to reduce estimation bias in the presence of noisy data and minor systematic errors, which may be a significant advantage over the standard least squares method.

1. Introduction

Inverse modeling has become a standard technique for estimating hydrogeologic parameters. The predictions calculated with a mathematical model are matched to a set of observations by adjusting parameters that are considered unknown or uncertain. The parameters that best reproduce the observed data are believed to be the most likely ones. This intuitive approach can be formalized by making certain assumptions about the distribution of the measurement errors, yielding a likelihood function that has to be maximized to obtain the best estimate parameter set. For example, if each data point has a measurement error that is random and normally distributed around the true system state, it can be shown that minimizing the sum of the squared weighted residuals leads to a maximum likelihood estimator of the unknown parameters [see, e.g., *Bickel and Doksum*, 1991; *Carrera and Neuman*, 1986].

Reviews of parameter estimation procedures in groundwater hydrology by *Yeh* [1986], *Kool et al*. [1987], *Carrera* [1987], *Sun* [1994], and *McLaughlin and Townley* [1996] reveal that the weighted least squares criterion is almost exclusively used as the performance measure to be minimized. The normality assumption of the measurement error is often justified by referring to the central limit theorem, which states that the distribution of a large number of small random measurement errors converges to a normal distribution. Furthermore, the normality assumption allows one to calculate confidence intervals of the estimated parameters and perform tests for significance. Finally, a large number of algorithms have been developed specifically for the minimization of sums of squares [see, e.g., *Gill et al*., 1981].

It is interesting to recall that C. F. Gauss introduced least squares during the last decade of the eighteenth century without giving a probabilistic justification of the method. In 1809 he showed that if the errors are normal, then least squares gives maximum likelihood estimates. His reasons for assuming normality, however, were tenuous, and he rejected the approach again in 1821. Instead, he cites practical reasons for choosing the square as a "measure of loss" and freely admits that the choice is quite arbitrary [see *Gauss*, 1821, p. 8]. In the same publication he proves the minimum variance theorem for linear models, which does not depend on the distribution of the errors.

Despite its popularity, an estimate based on least squares has the drawback of being significantly affected by violations of the underlying distributional assumptions. In particular, the presence of outliers in the data may lead to poor matches of the "good" data, which induces a bias of the estimated model parameters. Given the fact that field measurements show many more outlier points than one would expect from the tail of the normal distribution, their potential impact on inverse modeling results should be carefully assessed. This brings us to the issue of robustness. An estimator that is insensitive to small departures from the underlying assumptions is considered robust. While the problem of robustness is well recognized in statistics [*Andrews et al*., 1972; *Huber*, 1981], only a few alternative approaches to least squares have been proposed for applications in the earth sciences. *Claerbout and Muir* [1973], *Vasco* [1991], *Rosa and Horne* [1991], and *Xiang et al*. [1993] used the L_1 norm, i.e., the sum of the absolute residuals, as the criterion for minimization, explicitly referring to the issue of robustness. *Vasco et al*. [1994] used the L_p norm for the inversion of seismic travel times with non-Gaussian errors, where they determined the optimal value of p from the kurtosis of the data error distribution.

In this paper we discuss the impact of errors on the solution of inverse problems in multiphase flow modeling and explore the performance of robust estimators in comparison to the standard least squares method. We are interested in estimating hydrogeologic parameters of two-phase flow systems. The governing equations describing nonisothermal flow of gases and liquids in porous media are highly nonlinear in the parameters, which makes the corresponding optimization problem nonlinear as well. Standard algorithms developed for nonlinear least squares optimization [see *Gill et al*., 1981] seem capable of minimizing the modified objective functions. We therefore focus on the issues of estimation bias, outlier identification, and

Paper number 98WR02174.
0043-1397/98/98WR-02174$09.00

a posteriori error analysis rather than the minimization algorithm.

We first give a summary description of the forward problem and the numerical model used to simulate multiphase flow experiments. We then review the basic concepts of robust parameter estimation. Finally, we discuss inversions of synthetically generated data with various error distributions using the proposed robust estimators.

2. The Forward Problem

One of the most important elements in parameter estimation by data inversion is the forward operator, i.e., the mathematical model that relates the parameters to the observables. This model can be a simple regression equation, a closed-form analytical solution, or a sophisticated numerical simulator. The purpose of the forward model is to explain the systematic part of the observed system state; that is, it must be able to accurately describe the physical behavior of the system under the conditions prevailing during data collection. This requirement makes the development of the forward model the most crucial step in parameter estimation. Any systematic error in the calculated system state immediately leads to a bias in the estimated parameters that may be much larger than any uncertainty from random errors in the data.

We use the TOUGH2 code [*Pruess*, 1991] as the forward model to simulate multiphase fluid flow in porous media. We consider flow of two components κ (water and air) in two phases β (liquid and gas). The mass balance equations for an arbitrary subdomain V_n bounded by the surface Γ_n can be written in the following integral form:

$$\frac{d}{dt}\int_{V_n} M\, dV = \int_{\Gamma_n} \mathbf{F}\cdot\mathbf{n}\, d\Gamma + \int_{V_n} q\, dV \tag{1}$$

The accumulation term M represents mass of component κ (κ = w: water; κ = a: air) per unit volume:

$$M^{\kappa} = \phi \sum_{\beta=l,g} S_{\beta}\rho_{\beta}X_{\beta}^{\kappa} \tag{2}$$

Here ϕ is porosity, S_β and ρ_β are the saturation and density of phase β, respectively, and X_β^κ is the mass fraction of component κ in phase β. The mass flux term consists of contributions from the liquid ($\beta = l$) and gaseous ($\beta = g$) phase:

$$\mathbf{F}^{\kappa} = \sum_{\beta=l,g} X_{\beta}^{\kappa}\mathbf{F}_{\beta} \tag{3}$$

where the following multiphase version of Darcy's law governs the phase fluxes:

$$\mathbf{F}_{\beta} = -k\frac{k_{r\beta}}{\mu_{\beta}}\rho_{\beta}(\nabla p_{\beta} - \rho_{\beta}\mathbf{g}) \tag{4}$$

Here k denotes the absolute permeability, $k_{r\beta}$ is relative permeability, μ_β is dynamic viscosity, p_β is the pressure of phase β, and $\mathbf{g}$ is the acceleration of gravity vector. In (1), $\mathbf{n}$ is the inward unit normal vector, and q represents sinks and sources.

The continuum equations (1) are discretized in space based on an integral finite difference formulation. Time is discretized fully implicitly as a first-order finite difference. Discretization results in a set of nonlinear coupled algebraic equations solved simultaneously by means of Newton-Raphson iterations. A conjugate gradient method is used to solve the linear equations arising at each iteration. For more details, see *Pruess* [1991].

The governing multiphase flow equations described above contain a large number of parameters that may be subjected to estimation by inverse modeling. These parameters include hydrogeologic properties such as the absolute permeability, the porosity, and the rock compressibility. The constitutive relations describing capillary pressure and relative permeability as a function of saturation are parameterized models that include fitting parameters such as the residual liquid saturation, the pore-size distribution index, and the gas entry pressure. Other types of parameters to be estimated are the initial and boundary conditions. For example, the initial gas saturation or formation pressure may be estimated by inverse modeling. Boundary conditions or unknown sinks and sources can also be considered parameters to be estimated. Even geometric features such as fracture spacing or the skin zone radius around a well are potential candidates. In general, any parameterized input to the numerical simulator is eligible for estimation by inverse modeling. The question whether unique and reasonable estimates for all these parameters can be obtained, however, depends solely on the available data, their sensitivity with respect to each of the parameters, and their correlation structure and quality. This aspect of inverse modeling is discussed in more detail by *Finsterle and Persoff* [1997].

3. Robust Estimators

When performing inversions based on noisy data, we have to be concerned about the distributional robustness of the estimator. An estimator is considered robust if it is relatively insensitive to small deviations of the underlying distribution. Data can be represented by a statistical model of the form

$$z_i^* = z_i(\mathbf{p}) + (b_m + \varepsilon_m)_i + (b_d + \varepsilon_d)_i \tag{5}$$

Here z_i^* is the observed value at a calibration point i, and z_i is the corresponding modeling result, which is a function of the unknown parameter vector $\mathbf{p}$. The residual, i.e., the difference between the measured and calculated value, is the sum of the error in the model, $e_m = \tilde{z} - z(\mathbf{p})$, and the error in the data, $e_d = z^* - \tilde{z}$, where $\tilde{z}$ is the true value. Both modeling error and data error have a systematic component b and a random component ε. For the discussion that follows, we prefer to distinguish between systematic and random errors, regardless of their source, because it is usually not relevant or possible to identify whether a deviation between the model prediction and the data is attributable to an error in the data or the model. The systematic error in the residuals is denoted by $b_r \equiv b_d + b_m$, and the random part is termed $\varepsilon_r \equiv \varepsilon_d + \varepsilon_m$. In most cases, systematic errors from an incomplete model description outweigh the systematic measurement errors, i.e., $|b_m| \gg |b_d|$, whereas random modeling errors such as round-off errors or numerical oscillations are usually smaller compared with the random errors in the data, i.e., $|\varepsilon_m| \ll |\varepsilon_d|$.

Given these definitions, the classical assumption can be described as follows: (1) ε_r are independent, (2) ε_r are normally distributed with mean zero and variance σ_r^2, and (3) there are no systematic errors, i.e., $b_r = 0$.

We will discuss two types of violations of the standard assumption. The first considers random errors that do not follow a Gaussian distribution. This might occur if the error distribution is contaminated by a few large outliers. Since the number

of data points used in an inversion is finite, even a small number of deviate points cause the least squares fit to be distorted, leading to parameter estimates with low precision. A similar effect occurs if the error distribution is heavy-tailed, for example, if a Gaussian distribution is contaminated by a large number of relatively small outliers.

The second type of violation occurs in the presence of systematic errors that usually yield an asymmetric distribution of the residuals. If certain portions of the data exhibit a systematic error, the corresponding residuals are likely to become deviate points. If a certain systematic error affects a single point used for calibration, it cannot be determined whether the large residual stems from a systematic error or is an outlier as a result of a random process; such a distinction is also insignificant. If multiple calibration points are affected by the same systematic error source, the corresponding residuals are strongly correlated and tend to have the same sign over a certain interval in space and time. The ensemble of residuals contaminated with systematic errors, however, can be viewed as one or several outlier points. The interpretation of systematic errors as equivalent, usually large outliers is the main reasoning for subjecting them to robust estimation methods. However, it is obvious that if the entire data set or model is flawed, such errors cannot be mitigated by using robust estimators.

Systematic errors may be local both in time and space. For example, inconsistent initial or boundary conditions often result in systematic deviations between the data and the model prediction at early or late times during a transient experiment, leading to errors in a specific time segment. Similarly, a data set from a sensor that is either defective or placed in a unit that is poorly represented in the model leads to erroneous residuals at a specific point in space, again corrupting the inversion. Note that these types of systematic errors may not appear as obvious outliers and are therefore difficult to identify.

Before we introduce the robust estimators, we would like to emphasize that the main effort in estimating parameters by inverse modeling should be placed on avoiding systematic errors and minimizing random errors. The robust estimators presented here do not exempt the experimentalist and modeler from a comprehensive test design, careful execution of the experiment, accurate model development, and conscientious analyses of the inverse modeling results. However, systematic errors in the conceptual model and non-Gaussian random errors in the data are inherent in inverse modeling, and the problems associated with systematic errors seem to be accentuated rather than alleviated by the use of the standard least-squares estimator.

An overview of robust statistical procedures with mathematically rigorous definitions of their underlying concepts is given by *Huber* [1981, 1996]. In this paper we follow a more intuitive approach and introduce the robust estimators by discussing their common property of reducing the weight of deviant points. The performance of the robust estimators is illustrated and compared to the method of least squares using synthetically generated data. Finally, we will discuss an application of the method to previously analyzed data from a laboratory experiment.

Fitting a model to data for parameter estimation can be formulated as a minimization problem of the form

$$\min S = \sum_{i=1}^{m} \omega(y_i;\, \mathbf{p}) \tag{6}$$

Here ω is an arbitrary loss function, which is a function of the weighted residuals

$$y_i = \frac{z_i^* - z_i(\mathbf{p})}{\sigma_i} \tag{7}$$

where σ_i is the measurement error assumed to be independent and m is the total number of calibration points. At the minimum of S the derivatives of the objective function (6) with respect to the parameters p_j vanish

$$\sum_{i=1}^{m} \frac{1}{\sigma_i} \psi \frac{\partial y_i}{\partial p_j} = 0 \qquad j = 1, \cdots, n \tag{8}$$

where the function ψ is defined as the derivative of the loss function, $\psi \equiv \partial\omega/\partial y$.

It is important to realize that the loss function ω is arbitrary. Its choice can be based on probabilistic considerations, with ω being the negative logarithm of the joint probability density function. When adopting this viewpoint, the parameters $\mathbf{p} = \hat{\mathbf{p}}$ of a model $y(\mathbf{p})$ that minimize (1) are the maximum-likelihood estimates for $\mathbf{p}$. For example, if the errors are normally distributed, the loss function can be directly derived from the joint Gaussian distribution to be $\omega(y) = (1/2)y^2$ and $\psi(y) = y$, which yields the standard weighted least squares method [see, e.g., *Carrera and Neuman*, 1986]. Note that the ψ function serves as a weighting function in (8). It can be seen that least squares assigns greater weights to increasingly deviant points, reflecting the assumption that outliers are very unlikely according to the normal distribution. Consequently, if we suppose that the weighted residuals follow a distribution with a longer tail, that is with a somewhat larger probability of encountering points removed from the central region, we should choose a ψ function that yields decreasing relative weights for deviant points. It is expected that reducing the weight of outliers makes the estimator more robust.

Many functions with the desired properties have been proposed in the literature [see *Andrews et al*., 1972]. Some are maximum-likelihood estimators for known error distributions, whereas others do not correspond to a standard probability density function. We have selected five estimators for this study. They include (1) least squares (LS), (2) least absolute deviates (LAD) or L_1 estimator, (3) the maximum-likelihood estimator for measurement errors following a Cauchy distribution, (4) one of the robust estimators proposed by Huber, and (5) the Andrews estimator. Their functional forms are summarized in Table 1. The loss function $\omega(y_i)$ of the five estimators is shown in Figure 1, with the parameter $c = 1$ for the Huber and Andrews estimator.

Note that for the Andrews estimator, observations with weighted residuals larger than $c\pi$ are considered to be true outliers and are not counted at all in the estimation of the parameters. This property may lead to difficulties when using the Andrews estimator in a nonlinear optimization problem where the initial guess $\mathbf{p}_0$ is far away from the best estimate, in which case the initial residuals are too large. As a consequence, the gradient of the objective function becomes unstable, making it difficult for the minimization algorithm to converge. It is therefore suggested to first perform a standard least squares fit before switching to the Andrews estimator.

The robust estimators have been implemented into the ITOUGH2 code [*Finsterle*, 1997]. ITOUGH2 solves the inverse problem for TOUGH2 models. With ITOUGH2, any

Table 1. Estimator, Loss Function, and ψ Function

Estimator	Distribution	Loss Function ω	ψ Function
Least squares	Gaussian	$\omega = \frac{1}{2}y^2$	$\psi = y$
L_1 estimator	double exponential	$\omega = \lvert y \rvert$	$\psi = \begin{cases} 1 & y > 0 \\ -1 & y < 0 \end{cases}$
Cauchy	Cauchy	$\omega = \log(1 + \frac{1}{2}y^2)$	$\psi = \dfrac{y}{1 + \frac{1}{2}y^2}$
Huber	[a]	$\omega = \begin{cases} y^2/2 & \lvert y \rvert \le c \\ c\lvert y \rvert - c^2/2 & \lvert y \rvert > c \end{cases}$	$\psi = \begin{cases} -c & y < -c \\ y & \lvert y \rvert \le c \\ c & y > c \end{cases}$
Andrews	[a]	$\omega - \begin{cases} 1 - \cos(y/c) & \lvert y \rvert \le c\pi \\ 2 & \lvert y \rvert > c\pi \end{cases}$	$\psi = \begin{cases} \sin(y/c) & \lvert y \rvert \le c\pi \\ 0 & \lvert y \rvert > c\pi \end{cases}$

[a]No standard probability distribution available.

TOUGH2 input parameter can be estimated based on any type of observation for which a corresponding TOUGH2 output variable is calculated. Different algorithms are available to minimize the objective function.

4. Performance Comparison

In this section the performance of the robust estimators in comparison with the standard least squares method is demonstrated using synthetically generated data sets. The reason for performing synthetic inversions is that the conceptual model, as well as the error structure, is known and can be varied to test different hypotheses. An application to real data will be discussed in section 5.

We consider a simulated laboratory experiment in which water is injected at a constant pressure into a one-dimensional, horizontal column filled with uniform, partially saturated sand (see Figure 2). The synthetic data include flow rate measurements at the inlet and pressures observed at the center of the column. The duration of the experiment is 10 min, with a sampling interval of 30 s. In order to simulate measurement errors, the calculated flow rates and pressures are perturbed by a random value drawn from a prescribed probability distribution.

It is important to realize that no systematic errors have been introduced so far. In this synthetic experiment, the conceptual model, the governing equations, the functional form of the relative permeability and capillary pressure curves, etc. are known and free of errors. The only unknowns are two parameters selected for estimation, namely, the porosity ϕ and the initial gas saturation in the sample, S_{gi}. The true values used for generating the data are $\phi = 0.35$ and $S_{gi} = 0.30$.

It is the purpose of this study to examine the performance of various estimators in the presence of errors $e = z^* - z$ that are not normally distributed. A distribution with pronounced tails, which will be used to model non-Gaussian errors, is the Cauchy distribution given by

$$f(e) = \frac{1}{\sigma\pi[1 + \frac{1}{2}(e/\sigma)^2]} \tag{9}$$

Since the forward model is a strongly nonlinear function of the parameters and since we consider a non-Gaussian error distribution, no analytical solution can be provided for the

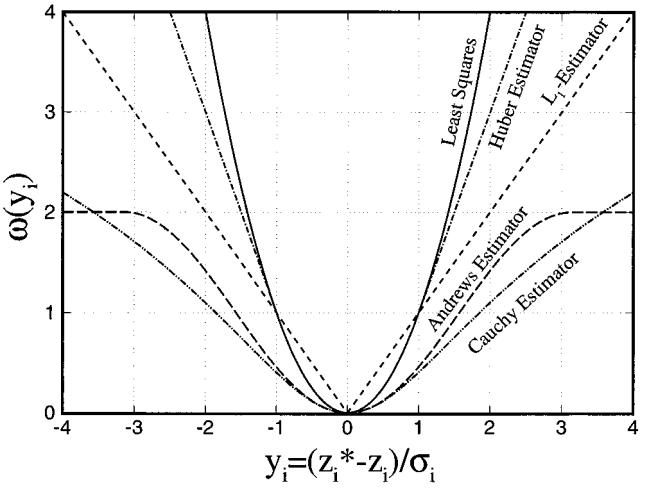

Figure 1. Loss function ω of five estimators as a function of the weighted residual.

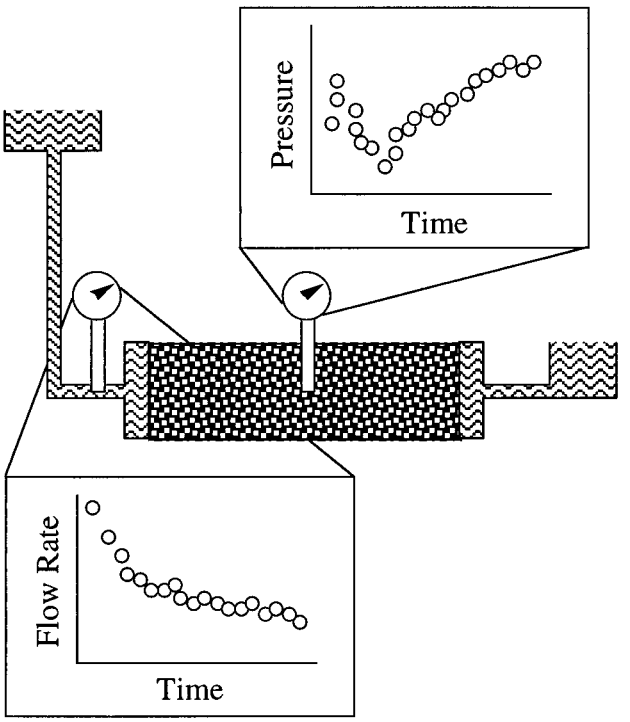

Figure 2. Schematic of synthetic laboratory experiment.

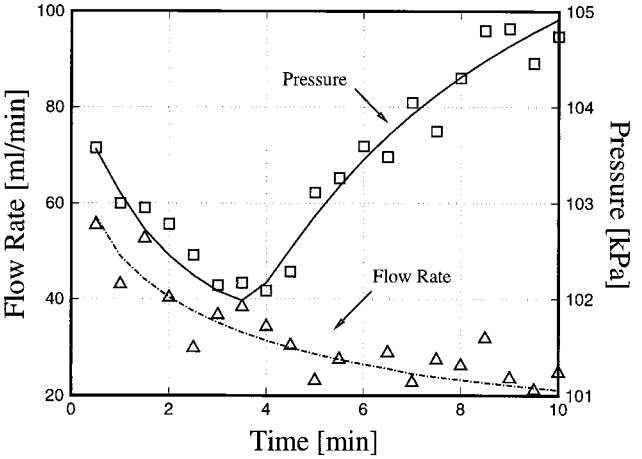

Figure 3. True system response (lines) and one set of perturbed data (symbols) used for parameter estimation.

expected distribution of the estimates. Monte Carlo simulations are performed instead. We generated 200 realizations for each data point, $z_k^* = z + e_k$, $k = 1, \cdots, 200$, where the scaling factors for the flow and pressure measurements are $\sigma_q = 5$ mL/min and $\sigma_p = 200$ Pa, respectively. Figure 3 shows the true, simulated flow rate and pressure response, and one of the 200 hypothetical data sets that will be used for estimation by inverse modeling. Note that no obvious outliers are present that could easily be removed by screening the data.

Initial guesses different from the true values were assigned to the two unknown parameters ϕ and S_{gi}, and the Levenberg-Marquardt algorithm was used to minimize the objective function, yielding different parameter estimates for each given data set. The procedure was repeated 200 times for each realization of a synthetic data set, and the solutions were plotted in the two-dimensional parameter space.

Performing Monte Carlo simulations in this fashion is a means to map out the full probability distribution of the parameter estimates in n dimensions. Figure 4 shows the family of solutions for the least squares fit. The sample mean is identical with the true parameter values, confirming that the

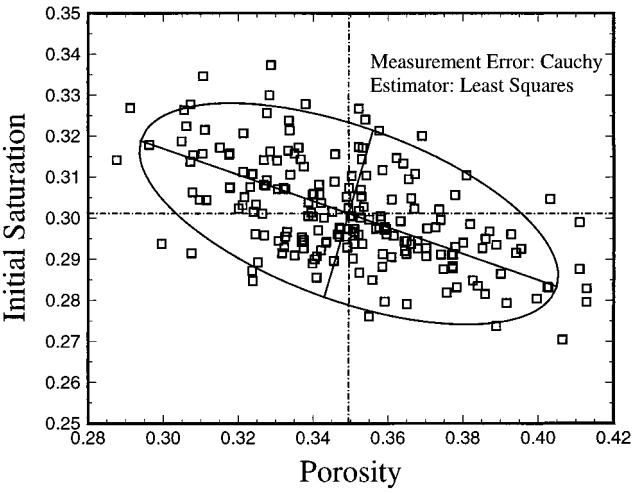

Figure 4. Solutions from 200 least squares fits to hypothetical data sets with measurement errors following a Cauchy distribution.

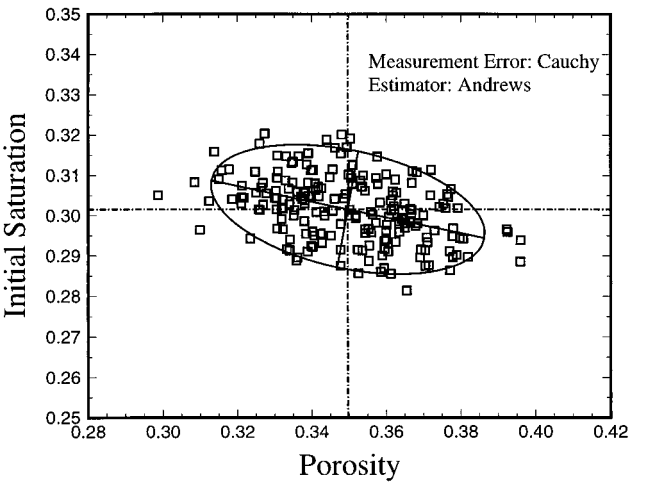

Figure 5. Solutions from 200 fits to hypothetical data sets with measurement errors following a Cauchy distribution using the Andrews robust estimator.

least squares estimator is unbiased. The sample covariance matrix is visualized as an ellipse; its size is a measure of estimation uncertainty. Because the estimates are not normally distributed, the ellipse does not correspond to a joint confidence region on a given confidence level that can be easily calculated from the sensitivity matrix. The same restriction applies to the robust estimators discussed below.

Figure 5 shows the set of solutions for the Andrews robust estimator. The solutions are more tightly clustered about the true parameter set, which is correctly identified. Recall that in almost all applications, only one realization of the data is available for parameter estimation. Under the presumption that deviate points are more frequently encountered than predicted by the normal distribution, using the robust estimator increases the chance of obtaining estimates that are close to the true parameter set. The results also obtained with the other robust estimators are summarized in Table 2, showing that all estimators are unbiased and that smaller uncertainties are obtained with estimators that more strongly reject deviate points.

The reduction in variability obtained with the robust estimators as compared to the least squares method is relatively minor in this first case, where no obvious outliers or systematic errors are present. Recall that the purpose was to demonstrate that the robust estimators are unbiased and that they are less affected by deviate points than the standard method. The reduction in estimation uncertainty (compare Figures 4 and 5) can be deemed satisfying, considering that the distribution of the synthetic data is symmetric and close to normal.

Next, we examine the performance of the robust estimators

Table 2. Sample Statistics of Inverse Modeling Results for Five Estimators

	Least Squares	L_1 Estimator	Huber Estimator[a]	Cauchy Estimator	Andrews Estimator[b]
Mean ϕ	0.349	0.352	0.352	0.352	0.350
Mean S_{gi}	0.301	0.301	0.301	0.301	0.301
s. d. ϕ	0.026	0.023	0.021	0.020	0.018
s. d. S_{gi}	0.013	0.011	0.010	0.010	0.008

[a]Parameter $c = 1.0$ for Huber estimator.
[b]Parameter $c = 0.5$ for Andrews estimator.

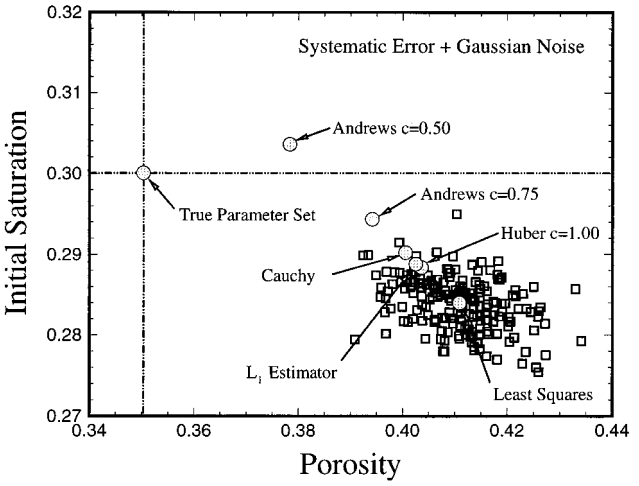

Figure 6. Mean estimated parameter sets from 200 inversions of hypothetical data sets with systematic errors using five different estimators. The individual solutions from least squares fits are shown as squares.

in the case where a systematic error is present. Systematic errors in either the data or the model almost always lead to an error in the estimated parameters. It is likely that part of the modeling errors can be compensated for by moving the parameters away from their true values. For example, it is possible to counteract phase dispersion effects by reducing the parameter controlling capillary strength of the soil [*Pruess*, 1996]. Predictions of saturation distributions may be more reliable when using a biased parameter value obtained from inverse modeling rather than the true value, provided that the latter is known at all. While the parameters estimated by data inversion can be considered optimal for the given model and are thus the preferred ones for predictions based on a similar conceptualization, these model-related parameters may not be adequate when used in an application with a different model structure. It is imperative that any systematic errors be eliminated as completely as possible. However, a systematic component will almost always remain in the final residuals due to an incomplete representation of all aspects of a hydrological system in a numerical model. For these cases we test the performance of the robust estimators in comparison with least squares.

In order to introduce a systematic error into our synthetic data, we rotated the column by 90° in the simulation and allowed the water to redistribute under gravity for 5 min. The column is brought back into horizontal position, before synthetic data were generated as described above (see discussion of Figure 2). While the average saturation in the column is unchanged by the sample manipulation, there are now slightly higher gas saturations near the inlet and slightly lower saturations near the outlet of the column. The synthetic data are then inverted with a model that assumes uniform initial gas saturation. This type of error (nonuniform initial saturation) is likely to occur even under well-controlled laboratory conditions. Note that the error could be attributed to the data (the sample was not handled properly prior to testing), or it could be seen as a modeling error; that is, the simplifying assumption of uniform initial conditions is not adequate and should be replaced by a refined model with varying saturations along the column. Because it is the residuals that are minimized during the inversion, the distinction between measurement error and modeling error is irrelevant, and it is a matter of convenience whether more effort should be placed on achieving well-controlled experimental conditions, or whether a more sophisticated model should be developed to capture potential flaws in the experiment.

Gaussian noise was added to the synthetic data to simulate random measurement errors, and 200 inversions were performed with each of the five estimators. The results are visualized in Figure 6. The true parameter set is indicated at $\phi = 0.35$ and $S_{gi} = 0.30$. Note that S_{gi} is here interpreted as the true average saturation at the beginning of the experiment.

Using the method of least squares results in a porosity estimate of $\phi = 0.411$ and an initial gas saturation of $S_{gi} = 0.283$. The porosity estimate is significantly biased due to the systematic error. The higher gas saturation near the inlet of the column is partly compensated for by an increase in porosity. The impact of higher gas saturation near the inlet is restricted to early-time pressure and flow rate data and could also be accounted for by an increase in the estimate of S_{gi}. However, the data are much more sensitive to initial gas saturation, which affects both relative permeability and storativity of the sample. Overall, the two parameters are negatively correlated, as can be seen from the orientation of the cloud of inverse modeling results (see Figures 4, 5, and 6). Physically, the negative correlation is a result of the fact that the pressure at the observation point responds according to the sample's diffusivity. Diffusivity decreases with an increase in porosity. It also decreases with an increase in initial gas saturation due to both the strong reduction in relative liquid permeability and the increase in storativity. In other words, if one parameter is increased, the other has to be decreased in order to yield a similar average system behavior. As a result of this correlation structure and the relative sensitivity of the two parameters, the systematic error in the early-time residuals reduces the initial gas saturation estimate by a relatively small amount.

The robust estimators result in mean estimates that exhibit a smaller bias compared to the least squares solution. Note that one cannot expect to identify the true parameter set because a systematic error is indeed present and will affect the estimates, regardless of the estimator being used. It is only a question of how strongly the deviate points at early times deflect the estimates from the expected parameter set. Furthermore, the estimated uniform initial saturation is conceptually different from the true initial average saturation of 0.30.

The solutions indicated in Figure 6 seem to be aligned along the direction in the parameter space that is least constrained, following the correlation structure discussed above. The results of the robust estimators are further away from the least squares solution and closer to the true parameter set according to the amount of weight given to deviate points. Consequently, the Huber estimator lies in between the L_1 estimator and least squares, and the Cauchy estimator with a decreasing weight with increasing deviation performs even better. The Andrews estimator, which cuts off all points with a residual greater than c times the prior standard deviation, moves closer to the true parameter set with decreasing c value. However, the smaller c means that fewer data points are actually used in the inversion; hence trade-off for this reduction in bias is an increase in estimation uncertainty.

It is important to realize that the robust estimators perform favorably only for a certain type of systematic error. The systematic error has to be constrained to a relatively small subset of the available data. This subset may consist of early-time or

late-time data showing effects from inappropriate initial and boundary conditions or of data from a single faulty sensor. If the majority of the data is corrupted, however, the robust estimators are not expected to perform better than least squares and may in fact bias the solution toward the wrong parameter set by discarding the good data. Nevertheless, we believe that the robust estimators have a potentially significant advantage over the standard least squares method in many applications, as will be discussed in section 5.

5. Application

The robust estimators are applied to data from laboratory experiments. Data from three gas-pressure-pulse-decay (GPPD) experiments performed at three different pressure levels were inverted simultaneously to estimate the absolute permeability $\log(k)$, the Klinkenberg slip factor $\log(b)$, and the porosity ϕ of a very tight graywacke core plug. The experiments and the analysis procedure are described in detail by *Finsterle and Persoff* [1997]. Here we focus on the performance of the least squares and robust estimators, taking advantage of the earlier findings regarding the uniqueness of the solution, the sensitivity of the data, and the potential sources of errors.

The previous analyses revealed that the late-time data from two of the three experiments exhibit a systematic error that can be attributed to gas leakage from the apparatus. The residual plot after concurrent least squares fitting of all three experiments is reproduced in Figure 7. The corresponding histogram of the residuals (Figure 8) reveals the non-Gaussian error distribution. The long tail of the distribution is a result of the trend in the late time data. Since least squares minimizes the variance of the residuals, it also introduces a trend in the residuals of the presumably good data from experiment 1. Using least squares, the resulting parameter set is likely to be biased (see Table 3). In particular, the porosity estimate turns out to be too high, because increasing the pore space is the only way to account for the volume of gas that in fact leaked to the laboratory environment. The robust estimators seem to be able to identify the late-time data from experiments 2 and 3 as being unreasonably large. Instead of minimizing the variance of all residuals, they preferentially match the less affected early-time data as well as the good data from experiment 1. By doing so, the systematically negative late time residuals from experiments 2 and 3 are actually increased, since less weight is assigned to these deviate points. Figure 9 shows the residual plot after matching the data with the Cauchy estimator. Similar results were obtained with the other robust estimators considered in this study. The parameter estimates and the standard deviations of the final residuals are summarized in Table 3. Note the significantly lower porosity estimates for the robust estimator and the corresponding increase in the standard deviation of the final residuals, indicating that the late-time data are less honored. While the true porosity is not known, a value near 1% is believed to be reasonable. A low value is also obtained by an inversion in which the leakage rate is estimated along with the hydrogeologic parameters. This approach of incorporating, in a parameterized form, the process that presumably led to the systematic error, is described in detail by *Finsterle and Persoff* [1997]. If the process is correctly identified, the additional parameterization effectively eliminates the impact from systematic errors, as is evident from the purely random structure of the final residuals shown in Figure 10. The fact the large residuals disappeared once a deterministic correction term was introduced confirms the systematic nature of the errors in retrospect.

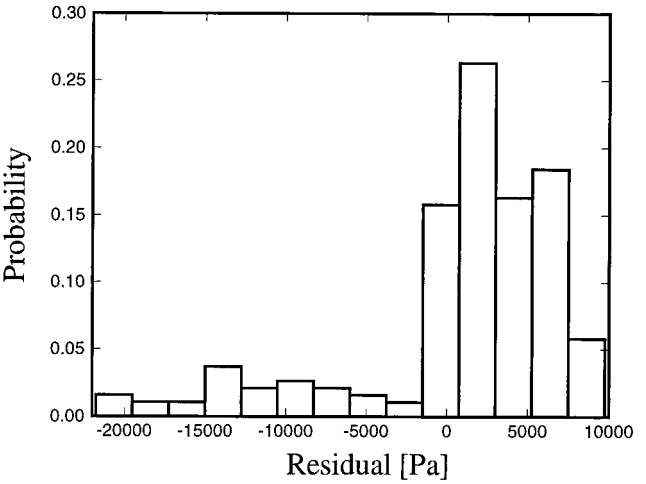

Figure 8. Histogram of residuals shown in Figure 7, revealing non-Gaussian error distribution.

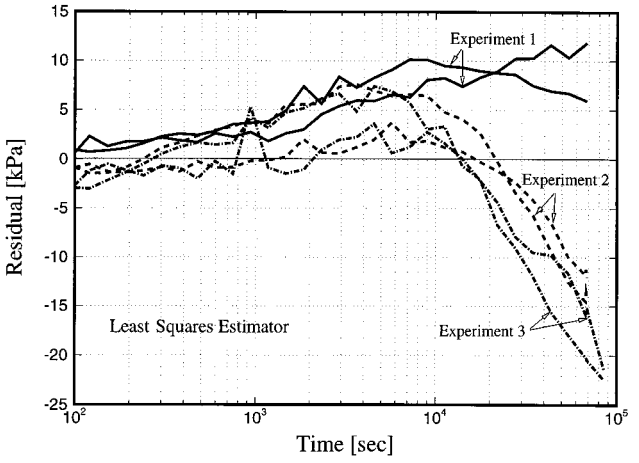

Figure 7. Residuals as a function of time after matching with the least squares estimator [after *Finsterle and Persoff*, 1997].

Table 3. Estimated Parameter Sets From Least Squares and Robust Estimators and Standard Deviation of Final Residuals

Estimator	Parameters $\log k$, m^2	$\log b$, Pa	Porosity ϕ, %	s. d. of Residuals, Pa
Least squares[a]	−20.68	7.31	1.81	4480
L_1 estimator	−20.71	7.36	1.20	5100
Huber	−20.71	7.36	1.19	7370
Cauchy[b]	−20.75	7.39	1.04	8200
Andrews	−20.70	7.34	1.09	7340
No systematic error[c]	−20.67	7.31	1.05	1100

[a]Residual plot shown in Figure 7.
[b]Residual plot shown in Figure 9.
[c]Systematic errors eliminated by parameterization; residual plot shown in Figure 10.

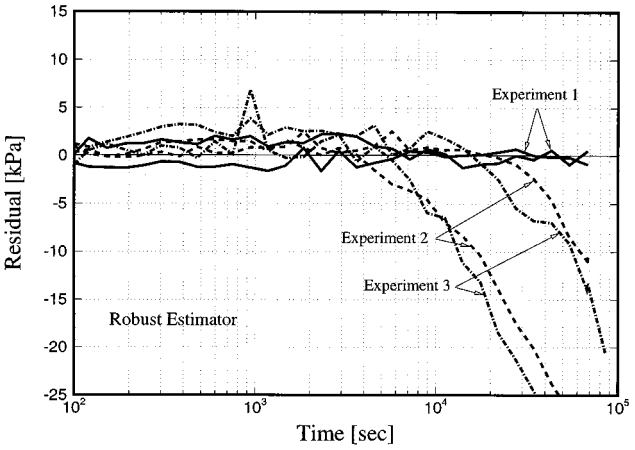

Figure 9. Residuals as a function of time after matching with the Cauchy estimator.

6. Summary and Conclusions

This study was motivated by the general observation that errors in either the data or the model used for inverse modeling usually exhibit a non-Gaussian distribution. The impact of outliers and systematic errors on the estimated parameter set was examined for both the standard least squares method as well as four alternative objective functions, which were termed robust estimators. While the standard error of the residuals is by definition smallest when using least squares, it was found that the robust estimators are less affected by the presence of random errors following a heavy tailed distribution, leading to more consistent estimates. These findings are in accordance with results from a study of location estimates [*Andrews et al*., 1972] and the various discussions of that topic by *Press et al*. [1992].

We also looked at the performance of all five estimators in the case where a subset of the data is corrupted by systematic errors. While the estimates from the least squares fit were more strongly biased, caution has to be exercised when using one of the robust estimators. The robust estimators only perform better for a specific type of systematic errors, and the error has to be contained within a limited portion of the data. Furthermore, systematic errors should be eliminated whenever possible, since they always affect the outcome of an inversion and thus reduce the reliability of subsequent prediction runs based on the estimated parameter set. On the other hand, it is recognized that measuring and modeling the state of a multiphase flow system is a difficult task, which almost always leads to some systematic errors in the residuals. While the major effort should be placed on obtaining accurate measurements and on careful model development, the use of robust estimators seems to be appropriate in many practical applications.

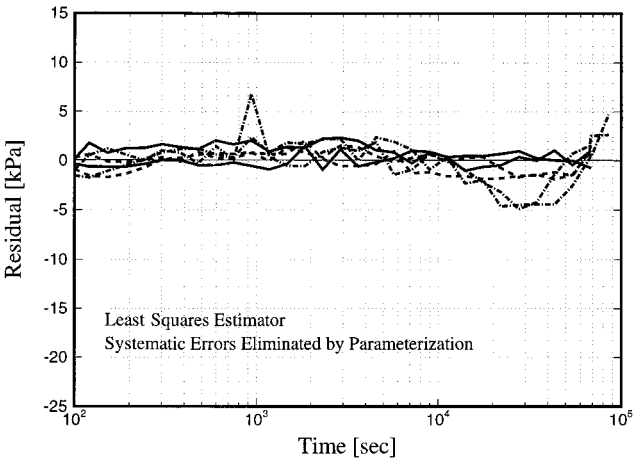

Figure 10. Residuals as a function of time after elimination of systematic error [after *Finsterle and Persoff*, 1997].

Acknowledgments. This work was supported, in part, by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Geothermal Technologies, of the U.S. Department of Energy, and by the Director, Office of Civilian Radioactive Waste Management, U.S. Department of Energy, through Memorandum Purchase Order EA9013MC5X between TRW Environmental Safety Systems, Inc. and the Ernest Orlando Lawrence Berkeley National Laboratory, under contract DE-AC03-76SF00098. We would like to thank K. Pruess, D. Vasco, and three anonymous reviewers for their comments and suggestions.

References

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton Univ. Press, Princeton, N. J., 1972.

Bickel, P. J., and K. A. Doksum, *Mathematical Statistics*, Holden-Day, Merrifield, Va., 1977.

Carrera, J., State of the art of the inverse problem applied to the flow and solute transport equations, in *Groundwater Flow and Quality Modelling*, *NATO ASI Ser*., *224*, 549–583, 1987.

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

Chavent, G., On the theory and practice of non-linear least-squares, *Adv. Water Resour*., *14*(2), 55–63, 1991.

Claerbout, J. F., and F. Muir, Robust modeling with erratic data, *Geophysics*, *38*, 826–844, 1973.

Finsterle, S., ITOUGH2 command reference, version 3.1, *Rep. LBNL-40041*, Lawrence Berkeley Natl. Lab., Berkeley, Calif., 1997.

Finsterle, S., and P. Persoff, Determining permeability of tight rock samples using inverse modeling, *Water Resour. Res*., *(33)*8, 1803–1811, 1997.

Gauss, C. F., *Theoria Combinatonis Observationum Erroribus Minimis Obnoxiae*, *Göttingische gelehrta Anzeign*, *33*, 321–327, 1821. (Translated by G. W. Stewart, Theory of the combination of observations least subject to errors, Soc. for Ind. and Appl. Math., Philadelphia, 1995.)

Gill, P. E., W. Murray, and M. H. Wrigth, *Practical Optimization*, Academic, San Diego, Calif., 1981.

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

Huber, P. J., *Robust Statistical Procedures*, 2nd ed., Soc. of Ind. and Appl. Math., Philadelphia, Pa., 1996.

Kool, J. B., J. C. Parker, and M. T. van Genuchten, Parameter estimation for unsaturated flow and transport models—A review, *J. Hydrol*., *91*, 255–293, 1987.

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

Neuman, S. P., Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res*., *9*(4), 1006–1021, 1973.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Rescipies in FORTRAN, the Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, New York, 1992.

Pruess, K., TOUGH2—A general-purpose numerical simulator for multiphase fluid and heat flow, *Rep. LBL-29400*, Lawrence Berkeley Natl. Lab., Berkeley, Calif., 1991.

Pruess, K., A Fickian diffusion model for the spreading of liquid plumes infiltrating in heterogeneous media, *Transp. Porous Media*, *24*, 1–33, 1996.

Rosa, A. J., and R. N. Horne, Automated well test analysis using robust (LAV) nonlinear parameter estimation, paper SPE 22679

presented at the 66th Annual Technical Conference and Exhibition of the Society of Petroleum Engineers, Dallas, Tex., Oct. 6–9, 1991.

Sun, N.-Z., *Inverse Problems in Groundwater Modeling*, Kluwer Acad., Norwell, Mass., 1994.

Vasco, D. W., Bounding seismic velocities using a tomographic method, *Geophysics*, *56*, 472–482, 1991.

Vasco, D. W., L. R. Johnson, R. J. Pulliam, and P. S. Earle, Robust inversion of IASP91 travel time residuals for mantle *P* and *S* velocity structure, earthquake mislocations, and station corrections, *J. Geophys. Res*., *99*, 13727–13755, 1994.

Xiang, Y., S. F. Sykes, and N. R. Thomson, A composite L_1 parameter estimator for model fitting in groundwater flow and solute transport simulations, *Water Resour. Res*., *29*(6), 1661–1673, 1993.

Yeh, W. G., Review of parameter estimation procedures in groundwater hydrology: The inverse problem, *Water Resour. Res*., *22*(2), 95–108, 1986.

S. Finsterle and J. Najita, Earth Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 90-1116, Berkeley, CA 94720. (e-mail: SAFinsterle@lbl.gov)

(Received August 22, 1997; revised June 12, 1998;
accepted June 24, 1998.)