Error Handling Strategies in Multiphase Inverse Modeling

Stefan Finsterle[*] *and Yingqi Zhang*

Lawrence Berkeley National Laboratory, Earth Sciences Division

One Cyclotron Road, MS 90-1116, Berkeley, California, SAFinsterle@lbl.gov

Abstract

Parameter estimation by inverse modeling involves the repeated evaluation of a function of residuals. These residuals represent both errors in the model and errors in the data. In practical applications of inverse modeling of multiphase flow and transport, the error structure of the final residuals often significantly deviates from the statistical assumptions that underlie standard maximum likelihood estimation using the least-squares method. Large random or systematic errors are likely to lead to convergence problems, biased parameter estimates, misleading uncertainty measures, or poor predictive capabilities of the calibrated model. The multiphase inverse modeling code iTOUGH2 supports strategies that identify and mitigate the impact of systematic or non-normal error structures. We discuss these approaches and provide an overview of the error handling features implemented in iTOUGH2.

Keywords

multiphase flow; inverse modeling; residual analysis; robust estimation; iTOUGH2

[*] Corresponding author; SAFinsterle@lbl.gov, Earth Sciences Division, 1 Cyclotron Road, MS 90-1116, Berkeley, CA 94720; phone: (510) 486-5205; fax: (510) 486-5686

1. Introduction

Numerical modeling is currently applied to characterize, predict, and optimize subsurface systems of increasing complexity. This increase in complexity arises foremost from the challenges that need to be addressed to ensure the sustainability of water, energy, and environmental systems. To be able to reliably predict the response of these systems to natural or man-made changes in the forcing terms, it is essential to consider coupled processes and to include many intricate hydrogeologic features. Advances in both process understanding and computational methods have enabled us to simulate subsurface systems with a higher degree of realism. However, the number of parameter values to be determined has also increased, so has the amount and variety of data that need to be collected for the calibration of a site-specific model. For example, while it may be sufficient to use basic geologic information along with water table measurements and a simple flow model to estimate groundwater flow in a confined aquifer, predicting the migration of contaminants in the vadose zone may require the development of a complex multiphase reactive transport model, the collection of hydrologic and geochemical field data, and the measurement of two-phase hydraulic properties on core samples.

Inverse modeling provides a framework to quantitatively integrate information about hydrogeological processes and the structure of the subsurface into a site-specific prediction model. Parameter estimation by inverse modeling involves minimizing an objective function that measures the misfit between the observed and calculated system state for all observation times. Many minimization algorithms as well as the interpretation of the inversion results are based on assumptions about the structure of the

residuals, i.e., the differences between the data and the corresponding model output once the objective function has been minimized. For example, these residuals are commonly expected to be random, uncorrelated, relatively small, and Gaussian. In practical applications, however, the residuals often contain significant systematic deviations or relatively large random errors that occur much more frequently than predicted by a Gaussian distribution. Such errors in the assumptions are likely to lead to convergence problems, biased parameters, misleading uncertainty estimates, or poor predictive capabilities of the calibrated model. An excellent discussion of structural errors can be found in Doherty and Welter (2010).

In this paper, we discuss the source and structure of such discrepancies and present strategies to mitigate their impact on inverse modeling results. We limit the discussion to approaches that are implemented in the multiphase inverse modeling code iTOUGH2 (Finsterle, 2004).

2. Theory

2.1 Residuals, measurement and modeling errors

The vector of residuals **r** is defined as the difference between the observed (indicated by an asterisk) and calculated (indicated by a hat) system responses at *m* discrete points in space and time where available measurements are considered suitable for model calibration:

$$r_i = (z^* - \hat{z})_i = (z + e^*)_i - (z + \hat{e})_i = (e^* - \hat{e}) = (b^* + \varepsilon^*) - (\hat{b} + \hat{\varepsilon}) \qquad i = 1, \ldots, m \quad (1)$$

The measurement z^* is the sum of the true (but unknown) system state z and the (true) measurement error e^*. Similarly, the model output is the sum of the true system state and the (true) modeling error $\hat{e}$. The error terms can be further described as containing a systematic component (b) and random component (ε), that is, $e^* = b^* + \varepsilon^*$ and $\hat{e} = \hat{b} + \hat{\varepsilon}$. These errors are termed "true" errors, as they refer to the true system response z. However, because z is unknown, we have to describe these errors using statistical terms.

While our ultimate goal is to identify the model structure and model-related parameters that best explain (and ultimately predict) the true system response z within acceptable uncertainty, this determination is based on an evaluation of residuals that are defined as the sum of two unavoidable error terms. In essence, the purpose of a physically based forward model is to capture the explainable part of the observable system response (b). In this sense, the model is deterministic and a function of the adjustable parameters. The unexplainable, random part of the residuals (ε), on the other hand, needs to be described by a stochastic model. Parameter estimation based on this formulation is thus inherently uncertain, and a careful examination of the measurement and modeling errors is essential to avoid a misinterpretation of its results. In this paper, we focus on estimation uncertainties that result from contamination by measurement noise, not from lack of sensitivity or from strong parameters correlations.

The vast majority of parameter estimation problems in hydrology are solved by the classical least-squares method, i.e., by minimizing the variance of the residuals, which

may include prior information and other regularization terms. Moreover, it is often stated that this method leads to maximum likelihood estimates (see, e.g., Carrera and Neuman, 1987). The reasonableness of this approach is based on all or various combinations of the following assumptions: (1) The conceptual model (which includes all parameters that are fixed during the inversion) is capable of reproducing the true system state once the parameters are adjusted (Vrugt and Boutem, 2002); (2) the final residuals have a random structure, i.e., there are no systematic measurement or modeling errors; (3) the solution to the inverse problem is unique and stable; (4) a relatively large number of data points is available; (5) the residuals are statistically independent; and (6) there are no modeling errors, i.e., the residuals can be described by the statistics of the measurement errors. In some cases, additional assumptions are made (implicitly or explicitly). If one or several of the assumptions listed above are violated, the least-squares approach should be refined or replaced with a suitable alternative. If least-squares methods are used nonetheless, the results have to be interpreted with caution or discarded, or an effort has to be made to bring the residuals in better compliance with the underlying assumptions. The remainder of this paper explores some of these issues.

2.2 Residual analysis

Before we discuss ways to mitigate the impact of certain errors on inverse modeling results, we first address the non-trivial question of how the existence of such errors can be detected. With the exception of synthetic inversion studies in which the model structure is perfectly known, it is impossible to obtain complete confidence that the true system response is identified, and that the calibrated model is an accurate enough

representation of the true system for reliable predictions to be made (Oreskes et al., 1994). Even when we acknowledge that a model is by definition a simplification of the real system—and thus always contains errors—this does not address the question of whether these errors are acceptable, or whether a calibrated model stays within the error bounds that were originally deemed tolerable.

In this section, we focus on the residual analysis as a means to detect errors that may not have been properly accounted for during model setup and inversion. Focusing on residuals rather than on the uncertainty of the estimated parameters is justified by the fact that if the model is not able to reasonably reproduce the observed data, it is likely not a good representation of the true system, in which case an examination of the estimation uncertainties becomes meaningless.

When judging the goodness-of-fit and the structure of the residuals, it is necessary to define expectations about how well the calibrated model will match the data overall, and what deviations between the data and the model are considered acceptable. These expectations are generally reflected in the covariance matrix $\mathbf{C}_{zz}$, the inverse of which is used to weigh the residuals in the objective function. As discussed above, describing the expectations by a covariance matrix assumes that the final residuals will have a random structure, which in turn implies that the deterministic forward model captures the systematic, explainable part of the system response. This is why the covariance matrix is commonly thought of as describing the measurement errors only. In most applications, the covariance matrix is assumed to be diagonal. Its elements can be constant (at least for

each data type or sensor) or variable (e.g., expressed as a fraction of the measured value). If residuals are heteroscedastic, the Box-Cox transformation (Box and Cox, 1964) can be applied to the measured and simulated data:

$$\tilde{z}(z;\lambda) = \begin{cases} \dfrac{(z^{\lambda} - 1)}{\lambda \cdot g^{\lambda - 1}} & \text{if } \lambda \neq 0 \\ g \cdot \ln(z) & \text{if } \lambda = 0 \end{cases} \tag{2}$$

Here, g is the geometric mean of the data, and the Box-Cox parameter λ is either known or estimated along with the hydrogeological parameters. This rank-preserving power transformation can make the residuals more Gaussian-like.

While the covariance matrix can be multiplied by any scalar without impacting the parameter estimates, it is desirable to construct the covariance matrix to reflect expectations about the final residuals. Then, we can statistically test the goodness-of-fit by calculating the *a posteriori* error variance

$$s_0^2 = \frac{\mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r}}{m - n} \tag{3}$$

and test whether it significantly deviates from the *a priori* variance σ_0^2, which is by definition equal to 1.0 (in Eq. 3, n is the number of parameters). Failing the Fisher model test $F_{m-n,\infty,\alpha} \leq s_0^2 / \sigma_0^2 \leq F_{m-n,\infty,1-\alpha}$ indicates that there is an error in the functional or stochastic model. This formal analysis, however, has limited applicability in hydrogeology, mainly because our expectations about the final residuals are more difficult to formulate than in other fields (such as land surveying, where the functional model is well defined and the statistics of the measurement errors are accurately known). Nevertheless, the estimated error variance of Eq. (3) serves as an overall goodness-of-fit

criterion, which is also a key component in comparisons of model performance (see the discussion of model identification criteria in Carrera and Neuman (1986) for details).

Even if the overall fit is considered acceptable, a more detailed analysis is needed to reveal potential trends in the residuals, which indicate that there is a systematic error in the model or the data, or that the random components are correlated. In general, an inspection of the residuals can be used to help identify aspects of the model that need to be modified, or potential flaws in the measurement device. In addition, large residuals (outliers) may be detected by visual inspection of a scatter plot, or by use of a more rigorous approach based on mathematical statistics. Note that if the statistics of the residuals significantly deviate from normal, the least-squares estimates are likely to be biased, and the formal error analysis (which establishes quantitative relationships among the objective function, the covariance matrix, and the confidence level) is not valid.

If such discrepancies between the assumed and actual distributions of the residuals are not obvious, a simple moment analysis of the residuals can be performed, separate for each data set or each observation type, and for all appropriately scaled residuals. The mean of the residuals, $\bar{r}$, is expected to be close to zero, and the variance, s_r^2 should be consistent with that specified by the prior covariance matrix $\mathbf{C}_{zz}$. A large variance either indicates that the data were noisier than expected, or that there is a trend or outliers in the residuals. The third moment characterizes the degree of asymmetry of the distribution. A positive (negative) skewness signifies an asymmetric tail extending to more positive (negative) residuals. The fourth moment or kurtosis measures the peakedness or flatness

of the distribution relative to the Gaussian distribution. A distribution with positive (negative) kurtosis is relatively peaked (flat) in comparison with a normal distribution.

Since the calibrated model is expected to yield simulated values that are close to the observed data, points of model results versus data should be close to a line with an intercept of zero and a slope of one. The statistics of such a regression line, including its linear coefficient of determination, provide easy measures of how well these expectations are met. Autocorrelations among the residuals within a time series can be identified using an autoregressive model or the so-called "runs statistic" (Cooley, 1979). A run is defined as a series of residuals that are either positive or negative. The number of sequential residuals with the same sign is the length of the run. In a random data set, the probability that the sign of two adjoining residuals changes follows a binomial distribution, which forms the basis of the runs test. If runs are on average longer or shorter than expected, the residuals are most likely not the result of a random process.

To further analyze the residuals, it is necessary to estimate the uncertainty of the calculated system response. Under a linearity and normality assumption, the covariance matrix of the model prediction (also termed the data resolution matrix) is given by

$$\mathbf{C}_{\hat{z}\hat{z}} = s_o^2 \cdot \mathbf{J}\left(\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J}\right)^{-1} \mathbf{J}^T = \mathbf{J}\mathbf{C}_{pp}\mathbf{J}^T \tag{4}$$

The Jacobian matrix **J** holds the derivatives of the calculated system response at the calibration points with respect to the parameters of interest, i.e., $J_{ij} = \partial \hat{z}_i(\mathbf{p}) / \partial p_j$. The covariance matrix of the estimated parameters is given by $\mathbf{C}_{pp} = s_0^2 \left(\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J}\right)^{-1}$. A measure of uncertainty can be obtained by taking the square-root of the diagonal

elements of $\mathbf{C}_{\hat{z}\hat{z}}$. It represents a prediction of the mean, and its expected variability is less than that of an individual data point (i.e., $\sigma_{\hat{z}}^2 < \sigma_z^2$), provided that the Fisher model test is passed. The covariance matrix of the residuals is given by (Weisberg, 1980):

$$\mathbf{C}_{rr} = \mathbf{C}_{zz} - \mathbf{C}_{\hat{z}\hat{z}} \tag{5}$$

Note that the residuals are always correlated, even if the elements of $\mathbf{C}_{zz}$ are independent. The elements of $\mathbf{C}_{rr}$ depend on the number and location of the observation points and the sensitivities of the calculated system response and these points to the model parameters; they do not depend on the actually measured value. Next, we calculate a measure termed local reliability or partial redundancy:

$$y_i = \left(\frac{\sigma_r^2}{\sigma_z^2}\right)_i = \left(\frac{\sigma_z^2 - \sigma_{\hat{z}}^2}{\sigma_z^2}\right)_i = 1 - \left(\frac{\sigma_{\hat{z}}}{\sigma_z}\right)_i^2 \quad i = 1, \ldots, m \tag{6}$$

The local reliability is a measure of how much a data point is controlled by redundant observations. If y_i is close to zero, even a large error in the corresponding data point z_i^* cannot be detected. A y_i value close to one indicates a well-controlled observation. Adding more observation points in the vicinity of this measurement may improve the accuracy, but does not necessarily improve the reliability of the inverse modeling system. Observations with relatively small y_i values are poorly controlled, whereas relatively large y_i values indicate a high degree of redundancy. For a poorly controlled observation, the size of the actual error can be significantly larger than the residual. Note that y_i can be evaluated *a priori* and can therefore be used to improve the design of an experiment. Experience suggests that good experimental designs yield y_i values in the range $0.25 < y_i < 0.75$, signifying a good balance between control and data efficiency.

The normalized or studentized residual

$$w_i = \left(\frac{r}{\sigma_r} \right)_i \quad i = 1, \ldots, m \tag{7}$$

is a normally distributed random variable with zero mean and a variance of one. Hence, the size of a residual can be statistically tested to see whether it is acceptable or a potential outlier. If $|w| > u_{1-\alpha}$, where $u_{1-\alpha}$ is the quantile of the standard normal distribution on the $1-\alpha$ confidence level, then the corresponding residual is likely to be an outlier under the normality assumption; the risk of discarding an acceptable data point is α. Outliers should be discarded, or—if many outliers are detected—the normality assumption has to be questioned, robust estimators have to be used, or systematic errors removed. Note that the *w*-test checks each residual individually, whereas the *F*-test examines the ensemble of all residuals. Since the model is nonlinear, all the measures discussed above are only approximately correct. Sampling methods need to be employed to test the appropriateness of the linearity assumption.

Finally, the relative contribution of each data point, each data set, and each observation type to the objective function may be used to identify errors in portions of the data or the model. Since the objective function is built using the weighted residuals, an imbalance in these contributions may also signify an error in the stochastic model. This concludes the discussion of the residual analysis as implemented in iTOUGH2.

2.3 Correlated and non-normal residuals

In most hydrogeological inverse modeling applications, errors are assumed to be (1) independent, (2) normally distributed, and (3) devoid of systematic errors. In this section, we discuss steps that could be taken to address violations of these assumptions.

First we note that Eqs. (3) and (4) as well as related equations (e.g., those describing the parameter update in the Gauss-Newton or Levenberg-Marquardt minimization algorithms) include a covariance matrix **C** that may contain off-diagonal terms. This means that if correlations are known to exist, they can be specified and used directly in the matrix formulations of these equations. The more difficult question is how statistical correlations can be determined *a priori*. A simple approach to partially account for temporal correlations among the residuals in a data set is to apply a first-order autogregressive time series model (AR1):

$$\tilde{r}_i = \begin{cases} r_1\sqrt{(1-\rho^2)} & \text{if } i = 1 \\ r_i - \rho \cdot r_{i-1} & \text{if } i = 2,\ldots,m \end{cases} \tag{8}$$

If the process is indeed autoregressive, the corrected residuals $\tilde{r}$ become uncorrelated for the appropriate value of the autoregressive coefficient ρ, which can be calculated iteratively or estimated along with the hydrogeological parameters. If autocorrelation arises from an incorrect functional model rather than from the random part of the error term, it may be more appropriate to refine the model, estimate parameters of a trend model, or represent the physics of the measurement condition in parameterized form, as will be discussed below.

The covariance matrix $\mathbf{C}_{zz}$ serves multiple purposes within a formal inversion framework. It (1) scales data of different quality, (2) scales observations of different types, so the weighted residuals are dimensionless and can be combined within a single objective function, (3) weighs the contribution of a residual to the overall misfit measure, and (4) is the stochastic model for maximum-likelihood estimation assuming normally distributed residuals. The last point is almost universally invoked if the generalized least-squares approach is used, and the normality assumption is justified by referring to the central limit theorem. However, in practical applications, large residuals occur more frequently than predicted by the tails of the normal distribution; they thus receive an undue weight in the inversion, and may bias the estimation results. In addition to the standard least-squares objective function, a number of alternative loss functions are available in iTOUGH2 to mitigate the impact of large residuals. The performance of these robust estimators has been examined using synthetic and real data (Finsterle and Najita, 1998). In the illustration below, we will make use of the Andrews estimator (Andrews et al., 1972):

$$\hat{r} = \begin{cases} 1 - \cos(\tilde{r} / c) & \text{if } |\tilde{r}| \le c\pi \\ 2 & \text{if } |\tilde{r}| > c\pi \end{cases} \tag{9}$$

Here, c is a parameter, $\tilde{r} = r / \sigma_z$ is the weighted (potentially transformed) residual, and $\hat{r}$ is the loss function that enters the objective function $S = \sum \hat{r}$ to be minimized.

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. For practical applications, a c value of 0.5 approximately represents Gaussian residuals with the tails of the distribution truncated.

2.4 Systematic errors

The distinction between systematic and random errors, and that between modeling and measurement errors, is somewhat arbitrary or difficult to establish in practice. Correlations among random errors may suggest the presence of a systematic process that is not explicitly modeled. Moreover, a systematic difference between measurements and model results can be attributed to an error in the measured data or an inappropriate representation of the measurement process in the forward model. Such errors can be viewed as an inconsistency between the model and the real system, and this discrepancy can be reduced by either adjusting the experiment or correcting measured data, or by refining the model so it properly captures the conditions that prevailed during data collection.

Analyzing systematic deviations between the model and the data is the means to gain insight into the explainable and predictable parts of the system behavior, and to extract information from the data. However, if the cause of the deviation is not related to the parameters of interest, the discrepancy has to be removed prior to or during the inversion to avoid an estimation bias. If sufficient complementary data are available, potential systematic errors can be included in the evolution or observation models using an appropriate parameterization, and these parameters can be estimated concurrently. This approach has been demonstrated in Finsterle and Persoff (1997) and will be further discussed in the example below.

3. Example

Using a synthetic laboratory experiment involving two-phase flow processes, we examine the ability of the diagnostic features and mitigation measures implemented in iTOUGH2 to detect and partly remove the impact of random and systematic errors in both the model and data. In the synthetic experiment (which was simulated by TOUGH2), water is injected at a constant pressure into a 0.5 m long horizontal column filled with uniform, partially saturated soil. The flow rate at the outlet and the pressure at the center of the column are used to estimate soil parameters. Various errors are introduced into the model and the synthetic data (see Table 1), purposely violating some of the assumptions underlying the standard least-squares approach for maximum-likelihood estimation. In our example, a small gap between the soil and the core holder leads to leakage. While outflow rate exhibits rate-dependent random fluctuations, with an outlier present, a homoscedastic measurement error with a standard deviation of σ_f = 5 ml/min is assumed for the initial inversion. Noise in the pressure data is assumed to have a standard deviation of σ_p = 100 Pa. However, the manometer has a memory effect that leads to temporally correlated pressure signals (generated using a Gaussian noise with an autocorrelation coefficient of 0.5), and the head data are erroneously converted to absolute pressure using a reference pressure of one bar instead of one atmosphere. Moreover, the precise location of the pressure sensor within the sample is not known.

The impact of these errors on the estimation of soil properties is examined by performing three calibrations. In the first calibration, two hydrogeologic properties—the logarithm of permeability k and porosity ϕ—are estimated along with the initial gas saturation S_{gi},

which is recognized as uncertain and thus included in the inversion. In the second calibration, potential systematic errors (leakage permeability k_{leak}, data shift A_p, and wrong sensor location X_p) are parameterized and jointly estimated. Finally, the third calibration addresses autocorrelation ρ, heteroscedasticity λ, and the presence of outliers. For the first two inversions, the standard weighted least-squares objective function was used:

$$S = \sum_{i=1}^{m} \left(\frac{z^* - z(\mathbf{p})}{\sigma_z} \right)_i^2 \tag{10}$$

For the third inversion, the Andrews estimator (see discussion of Eq. 9) was used, where the observations and corresponding model output were Box-Cox transformed (Eq. 2) and the transformed residuals corrected using the autoregressive AR1 model (Eq. 8). In all inversions, the objective function was minimized using the Levenberg-Marquardt algorithm with an eigenvalue-based Tikhonov matrix (Finsterle and Kowalsky, this issue). Porosity and initial gas saturation were constrained to their physical ranges; all other parameters were unconstrained. For each of these calibrations, we discuss the estimation bias and whether insights can be gained from a detailed residual analysis.

Fig. 1 shows the synthetically generated pressure and flow rate data, which are taken every 10 seconds. They are derived from the true system response by adding the random and systematic errors summarized in Table 1. Both the autocorrelation in the pressure noise and the heteroscedasticity in the flow rate data are visible. The calculated system response with the initial parameter set and no error handling deviates significantly from the data, as expected.

Fig. 2 shows the measured against calculated pressures and flow rates, scaled by the respective standard deviations σ_p and σ_f. The deviations from the unit-slope line are the residuals. Table 2 lists the parameter estimates, and Table 3 the results of the residual analysis.

While adjusting permeability, porosity, and initial gas saturation during the first inversion leads to a substantial reduction in the objective function, a simple visual inspection of the residuals (Fig. 2b) immediately reveals the systematic deviations caused by the conversion error and a potential leak. Moreover, the estimates for porosity and initial gas saturation end up at their upper and lower bounds, respectively, which often indicates that the model is an unlikely representation of the true system.

Next, the potential leak along the sample edge (simulated as a cylindrical opening of unknown permeability), the data shift (an unknown constant added to the measured pressures), and the location of the pressure sensor (i.e., the coordinate where the calculated pressure is extracted from the mode) are parameterized and estimated along with the hydrologic properties, significantly improving the fit. Nevertheless, relatively many large residuals are still identified using the criterion of Eq. (7), and the Fisher model test fails on a significance level of 5%. Moreover, the distributions of residuals are skewed and flatter than the normal distribution, suggesting that the normality assumption is violated. Even though the estimated parameters are physically reasonable, they are

biased. The residual analysis indicates that these parameters should not be taken as the final result, but that the analysis needs to be refined.

Finally, estimating the autoregressive coefficient and Box-Cox parameter, and using a robust estimator rather than least squares leads to an acceptable model fit and parameter estimates that are consistent with the true values given their uncertainties (see Table 2). The sole outlier in the flow rate was identified by the Andrews estimator and its effect automatically removed from the inversion.

The residual analysis further indicates that the local reliability measures for all residuals are above a value of 0.25, i.e., all calibration points are sufficiently controlled by neighboring, partly redundant measurements. The relatively large intercept combined with the unit slope and high Pearson's coefficient clearly point to the presence of a constant shift in the pressure data. The runs statistic appears to be misleading for the first inversion, where too few runs during the early stages of the experiment are compensated by many runs at the later stage, leading to the conclusion that the flow-rate residuals are random. For the final inversion, the pressure residuals are classified as non-random, which is due to the autocorrelation that results in too few runs.

4. Concluding Remarks

Parameter estimation by automatic model calibration generally relies on simplifying assumptions about the structure of the residuals in order to be able to use the powerful tools of Gaussian statistics. However, systematic and non-Gaussian errors in both the

model and the data are common in the study of complex multiphase flow systems. The analytics implemented in iTOUGH2 provide some means to identify and mitigate such errors. In this paper, we demonstrated that systematic errors in the model or the data can often be parameterized and then subjected to the estimation process. While this may lead to higher dimensional inverse problems with related stability, non-uniqueness, and performance issues, we consider it essential to pay special attention to these types of errors as they may severely bias the estimates. Moreover, in hydrogeology, the statistical properties of measurement errors are often not well known or poorly represented by the common distributional assumptions. The effects of wrong distributional assumptions about the errors can be mitigated by providing a more flexible stochastic model and estimating its parameters, concurrently with the hydrogeological properties of interest and the correction factors for systematic errors. Finally, we also demonstrated the effectiveness of a robust estimator to reduce or eliminate the impact of outliers.

While iTOUGH2 provides considerable flexibility in estimating a variety of parameters of different types for different purposes, the results of the inversion must first be examined using a detailed residual analysis, before the estimated parameter values can be further assessed. A number of statistical measures and graphical analysis tools were presented; they point towards aspects of the functional and stochastic models that may need to be refined. In addition to these statistical analyses, a sound understanding of the physics of multiphase flow, of the structure of geologic media, and of measurement processes is essential to avoid biased, misleading, or erroneous estimates.

The error handling capabilities of iTOUGH2 are currently enhanced in three ways. First, joint hydrogeophysical data inversion approaches are investigated (Finsterle and Kowalsky, 2008), testing their capability to identify the structure of the subsurface along with the parameters that govern multiphase flow and transport through these structures. Reducing structural error is expected to also significantly reduce estimation errors. Second, the approximation error theory developed by Kaipio and Somersalo (2004) and demonstrated for a hydrogeophysical application by Lehikoinen et al. (2010) will be implemented to address structural errors when highly simplified models are used during the inversion. Third, it is recognized (e.g., Moore and Doherty, 2006) that restricting the solution space by the zonation approach and other parameterization schemes may lead to calibrated models that have limited predictive power. We therefore examine alternative ways to parameterize heterogeneity and other features of the system, and to increase the robustness of inversions of highly parameterized models. The ultimate goal is to provide a theoretically sound and practically useful framework for integrating experimental design, inverse modeling, and uncertainty quantification for the study of multiphase flow systems.

Acknowledgment

We would like to thank Eric Laloy and an anonymous reviewer for their very constructive comments, as well as Mike Kowalsky (LBNL) for his careful review of an earlier version of the manuscript. This work was supported by the U.S. Dept. of Energy under Contract No. DE-AC02-05CH11231.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J. W., 1972. Robust Estimates of Location: Survey and Advances, Princeton Univ. Press, Princeton, N. J., pp. 374.

Box, G, Cox, D., 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B, 26(2), 211–252.

Carrera, J., Neuman, S.P., 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. Water Resources Research 22, 199–210.

Cooley, R.L., 1979. A method of estimating parameters and \assessing reliability for models of steady state groundwater flow: 2. Application of statistical analysis. Water Resources Research 15(3), 603–617.

Doherty, J., and Welter, D., 2010. A short exploration of structural noise. Water Resources Research 46, W05525, doi:10.1029/2009WR008377.

Finsterle, S., 2004. Multiphase inverse modeling: Review and iTOUGH2 applications. Vadose Zone Journal 3, 747–762.

Finsterle, S., Kowalsky, M.B., 2008. Joint hydrological-geophysical inversion for soil structure identification. Vadose Zone Journal 7:287–293, doi:10.2136/vzj2006.0078, 2008.

Finsterle, S., Kowalsky, M.B., 2010. A truncated Levenberg-Marquardt algorithm for the calibration of highly parameterized nonlinear models, *Computers and Geosciences,* 2010.

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

Finsterle, S., Najita, J., 1998. Robust estimation of hydrogeologic model parameters. Water Resources Research 34(11), 2939–2947.

iTOUGH2, 2010. Lawrence Berkeley National Laboratory, Berkeley, CA, http://esd.lbl.gov/iTOUGH2, [Accessed October 1, 2010].

Kaipio, J., Somersalo, E. 2004. Statistical and Computational Inverse Problems. Applied Mathematical Sciences 160, Springer-Verlag, Berlin, Germany, pp. 344.

Lehikoinen, A., Huttunen, J.M.J., Finsterle, S., Kowalsky, M.B., Kaipio, J.P., 2010. Dynamic inversion for hydrological process monitoring under model uncertainty, Water Resources Research, W04513, doi:10.1029/2009WR008470.

Moore, C., Doherty, J., 2006. The cost of uniqueness in groundwater model calibration. Advances in Water Resources 29 (4), 605–623.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 264, 641–646.

Vrugt, J., Boutem, W., 2002. Validity of first-order approximations to describe parameter uncertainty in soil hydraulic models. Soil Sciences Society of America Journal 66, 1740–1751.

Weisberg, S., 1980. Applied Linear Regression, John Wiley & Sons, New York, NY, pp. 283.

Figure Captions

Fig. 1. Synthetic (a) pressures and (b) flow rates; calibration data (symbols) are derived from true system behavior (solid lines); initial parameter set leads to significantly different response (dash-dotted line).

Fig. 2. Measured versus calculated pressures and flow rates for (a) initial parameter set and after (b) the first, (c) second, and (d) final inversion.

Table 1

Error strategy employed to address deviation between test case and standard assumptions

#	Error Description	Error Type	Error Handling Strategy
A	Gas leak	Systematic, solution-dependent, nonlinear experimental or modeling error	Include into model as parameterized process; estimate leakage parameters
B	Wrong initial saturation	Systematic, nonlinear modeling error	Parameterize and estimate initial condition
C	Uncertain location of pressure sensor	Systematic experimental or modeling error	Parameterize and estimate sensor location
D	Rate-dependent measurement errors	Error in stochastic model; heteroscedastic error structure	Estimate Box-Cox transformation parameter
E	Autocorrelation in pressure data	Error in stochastic model; correlation	Estimate autoregressive coefficient
F	Offset in pressure data	Systematic data error	Estimate offset in data
G	Outlier in flow-rate data	Error in stochastic model	Use robust estimator

Table 2

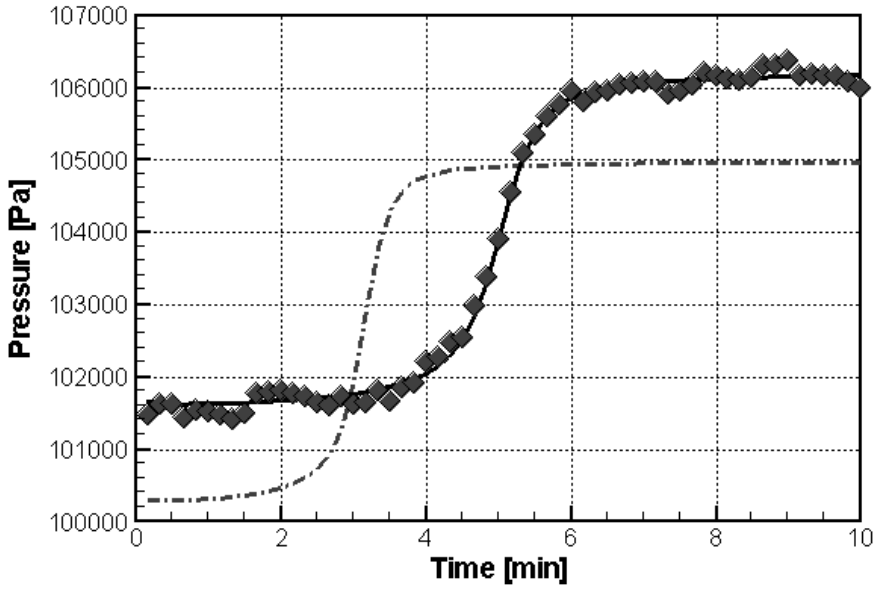
Estimated parameter sets for inversions with different error handling

Parameter	**True**	**Initial**	**Inversion 1**	**Inversion 2**	**Inversion 3**	
					Value	σ_p
$\log(k\ [m^2])$	-11.70	-11.50	-11.71	-11.71	-11.70	0.06
Porosity ϕ [–]	0.30	0.25	0.60	0.27	0.31	0.04
Init. sat. S_{gi} [–]	0.35	0.30	0.00	0.29	0.34	0.03
$\log(k_{leak}\ [m^2])$	-11.0	-15.0	n/a	-15.0	-11.0	0.37
Shift A_p [Pa]	1325	0	n/a	1260	1300	30
Location X_p [m]	0.23	0.25	n/a	0.23	0.23	0.02
AR1 ρ [–]	0.5	0.0	n/a	n/a	0.28	0.15
Box-Cox λ [–]	0.5	1.0	n/a	n/a	0.41	0.11

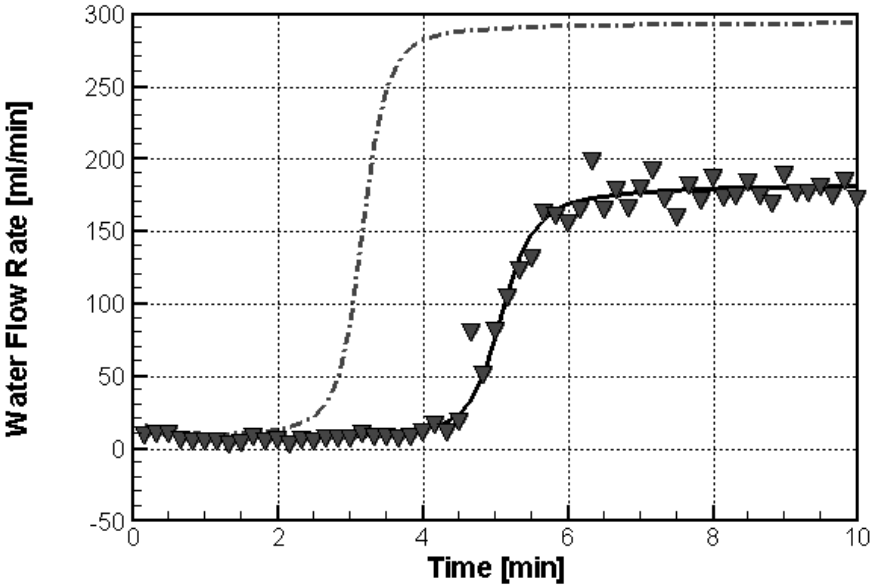
Table 3

Residual analysis for inversions with different error handling

Statistics		**Initial**	**Inversion 1**	**Inversion 2**	**Inversion 3**
Objective function [-]		58060	7384	441	109
Fisher model test		failed	failed	failed	passed
		Pressure Residuals			
Maximum studentized residual		-29.2	13.8	2.8	2.1
Number of large residuals		59	60	8	1
Mean		382	1050	0	-0
Standard deviation		1330	2010	122	95
Skewness		-1.39	-1.02	0.47	0.27
Kurtosis		0.30	1.19	-0.61	-0.57
Regression:	intercept	18500	3180	2100	298
	slope	0.83	0.98	0.98	1.0
	Pearson's *r*	0.79	1.00	1.00	1.0
Runs statistics:	Runs *R*	3	1	17	23
	E[*R*]	22	n/a	30	30
	Residuals are	not random	n/a	not random	not random
		Flow-Rate Residuals			
Maximum studentized residual		-57.2	7.4	9.4	7.94
Number of large residuals		48	26	10	3
Mean		112.8	8.8	1.1	1.4
Standard deviation		82.8	11.5	9.2	4.4
Skewness		0.31	-0.14	-2.28	-0.21
Kurtosis		-0.73	1.40	9.93	0.05
Regression:	intercept	8.16	10.2	-2.6	1.0
	slope	0.49	1.0	0.98	0.99
	Pearson's *r*	0.76	0.99	0.99	1.0
Runs statistics:	Runs *R*	1	21	39	31
	E[*R*]	n/a	18	28	28
	Residuals are	n/a	random	fluctuating	random



(a)



(b)

Fig. 1. Synthetic (a) pressures and (b) flow rates; calibration data (symbols) are derived from true system behavior (solid lines); initial parameter set leads to significantly different response (dash-dotted line).

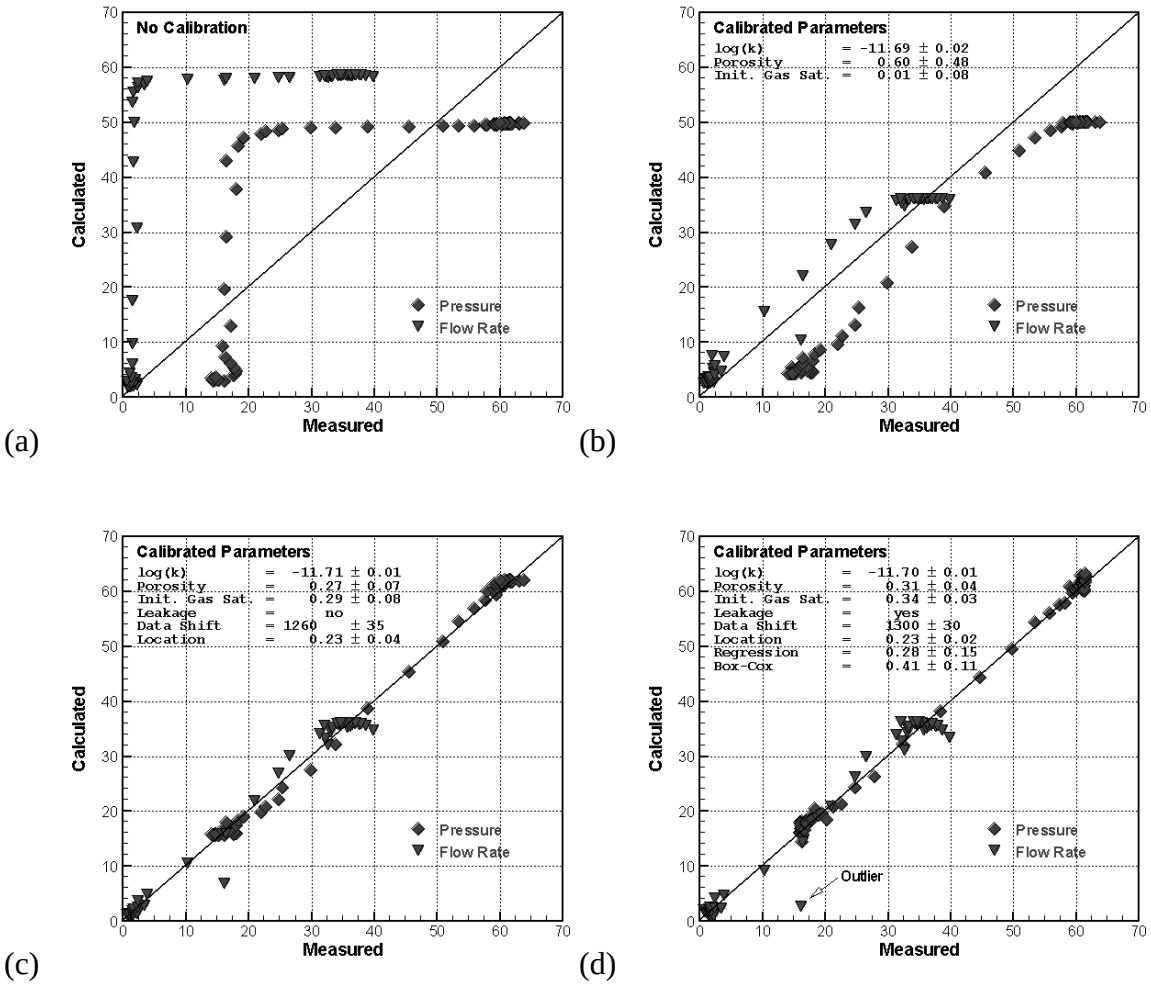


Fig. 2. Measured versus calculated pressures and flow rates for (a) initial parameter set and after (b) the first, (c) second, and (d) final inversion.

LEGAL DISCLAIMER