A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals

Steen Christensen

Department of Earth Sciences, University of Aarhus, Ny Munkegade b.520, DK-8000 Aarhus C, Denmark.
E-mail: *sc@geo.au.dk*

Received 8 May 2003; accepted in revised form 14 August 2003

Abstract Synthetic groundwater flow models with one unknown parameter, the average log transmissivity of the flow domain, and with Gaussian log-transmissivity error structure were used to study the nature and the accuracy of Generalized Likelihood Uncertainty Estimation (GLUE) intervals. The uniform prior distribution of $\log_{10}$ transmissivities was sampled uniformly and 1000 values per $\log_{10}$-transmissivity cycle were required to produce unbiased GLUE results. Because the errors in hydraulic head resulting from the log-transmissivity errors are known to be Gaussian for a linear model for heads, the Gaussian likelihood function was used as the GLUE goodness-of-fit function in most cases studied. The GLUE interval computed for the hydraulic head at different locations within the domain has the characteristics of a confidence interval for the hydraulic heads computed using the spatial average log transmissivity. The GLUE interval does not have the characteristics of a prediction interval, which is a probability interval for an uncertain observation of some variable such as the hydraulic head. The goodness-of-fit function can be corrected so that the resulting GLUE interval has the characteristics of a prediction interval. However, neither the original nor the corrected GLUE interval account for the uncertainty caused by small-scale model errors, which is always present in practical groundwater flow modelling. It is therefore concluded that one should be careful with using the GLUE interval to evaluate the validity of a model's structure. The structure may be valid even though observations fall significantly outside the GLUE interval if the observations are uncertain and the goodness-of-fit function is not corrected to account for this, or if small-scale model error is significant. If small-scale model error does not significantly bias model predictions, the predictions will be useful although uncertain. Small-scale model error did not bias the predictions in the examples studied here except near a strong sink. Changing the goodness-of-fit function from the Gaussian likelihood function to a similar but different function made the resulting GLUE intervals very inaccurate. Changing to a third function based on a fitted model rejection value produced results that were somewhat better than those obtained with the Gaussian likelihood function. However, the results were sensitive to the model rejection value, which in the ideal case should be adjusted for each predicted variable individually. Thus, the third function is not attractive for practical applications with the GLUE methodology.

Keywords Accuracy; confidence interval; GLUE; model error; prediction interval

Introduction

Suppose a groundwater flow model is used to make predictions of not yet observed variables such as hydraulic head, groundwater seepage to streams, etc. The predictions will usually be in error, because the model generally only uses the drift values (large-scale trend) of the hydrogeologic variables such as hydraulic conductivity, recharge, well discharge, boundary conditions, etc., that are truly unknown functions of space and/or time. Variability within a certain spatial scale always has to be ignored in a model, and this causes the model predictions to be in error to some extent. This type of error is here termed small-scale model error because it is caused by error in the small-scale structure of the model. A further source of prediction error is model parameter uncertainty. Values of the model parameters that allow a realistic model description of the drift of the unknown hydrogeologic variables are usually unknown and have to be estimated from uncertain and sparse observations of hydraulic head, groundwater discharge, prior information, etc. The estimated parameter values will therefore be uncertain (in error), which will add to the uncertainty (error) of the

Nordic Hydrology **Vol 35 No 1 pp 45–59** © IWA Publishing 2003

model predictions. Other sources of error will also be present in the modelling of groundwater flow in field cases but for simplicity we only focus on small-scale model error and error due to parameter uncertainty.

Several methods have been developed to quantify the uncertainty of model predictions. Here we study the ability of the Generalized Likelihood Uncertainty Estimation (GLUE) method (Binley and Beven 1991, Beven and Binley 1992, Beven and Freer 2001) to quantify prediction uncertainty of a groundwater flow model. The GLUE method is in principle easy to understand and easy to apply (although it may cost a lot of computer time). It is a Monte Carlo based method where parameter values sampled within parameter space and a subjectively chosen goodness-of-fit measure are used to estimate likelihood distributions for model predictions. The estimated distribution is used to compute the upper and lower bounds (called prediction limits by Beven and Freer (2001)) of intervals that quantify the prediction uncertainty. This type of interval is here termed a GLUE interval.

Binley and Beven (1991) were the first to demonstrate the application of the GLUE method. They considered the task of applying a semi-distributed Darcian flow model to a synthetically generated fully distributed three-dimensional system. For this example wide GLUE intervals were found, particularly for water-table levels. Binley and Beven (1991) suggested that this result was mainly caused by model structure errors. Beven and Binley (1992) demonstrated the application of the GLUE method for a similar but real problem. Since then the GLUE method has been used to quantify model prediction uncertainty in rainfall-runoff modelling (Freer *et al.* 1996, Lamb *et al.* 1998), soil erosion modeling (Brazier *et al.* 2001), modeling of tracer dispersion in a river reach (Hankin *et al.* 2001), well capture zone delineation via groundwater modeling (Feyen *et al.* 2001, Jensen 2003), and groundwater modelling of hydraulic head and stream discharge (Jensen 2003). Some of the results from these studies have general interest with respect to characterising aspects of the GLUE method.

Freer *et al.* (1996) and Feyen *et al.* (2001) demonstrated that the uncertainty bounds and likelihood distributions estimated by GLUE depend on the subjective choice of the goodness-of-fit measure. This was also foreseen by Beven and Binley (1992, p. 285).

Freer *et al.* (1996) found for some periods in their rainfall-runoff modelling study marked departures of the observed discharge from GLUE-computed uncertainty bounds and suggested this could be because the GLUE procedure cannot compensate completely for errors in either the input data or model structure. In a similar study Lamb *et al.* (1998) found that GLUE intervals were wide but still failed to enclose many observed water table depths. They thought this to be consistent with the simplified nature of the applied model that assumes, for example, soil properties to be uniform. The model is thus expected to contain model errors because heterogeneity of soil properties is ignored.

In a modelling study for a Danish catchment Jensen (2003) used the GLUE procedure together with the groundwater model of Christensen *et al.* (1998) to compute GLUE intervals for hydraulic heads that were not used for the interval estimation (i.e. these heads were not used to compute the goodness-of-fit measure). He found that, no matter the choice of the goodness-of-fit measure, the GLUE intervals were too narrow to contain the expected number of observations and suggested that this may be due to conceptual model errors. However, this viewpoint seems to be in contradiction with the fact that Christensen and Cooley (1999) for the same model found regression-based prediction intervals to contain the expected number of head observations and also found no apparent bias in head predictions.

As mentioned above it has been suggested in studies that model structural errors can cause GLUE intervals to be sometimes wide (Binley and Beven 1991, Lamb *et al.* 1998), sometimes narrow (Jensen 2003), and sometimes offset (Freer *et al.* 1996, Lamb *et al.* 1998) when compared to observations. It could be tempting to use a comparison of GLUE intervals

and observations to test for model structural errors and if, for example, the intervals fail to contain many of the observations then to use this as an indication that the model structure is invalid (Beven and Binley 1992, p. 285). In this context one could ask when is a model structure invalid? For many purposes a structural error that significantly biases predictions would invalidate a model, whereas small-scale error that does not significantly bias the predictions in many cases would be acceptable even though it adds to the uncertainty of the predictions. Therefore, if small-scale model error causes observations to fall significantly outside their GLUE intervals, and if GLUE intervals do not account for the uncertainty caused by small-scale model error, then it would be (in many cases) a mistake to conclude from such a comparison that the model structure is invalidating for the purpose of making useful predictions with the model.

This paper studies some problems with the GLUE method that can significantly affect the accuracy of GLUE intervals, or at least affect their interpretation and use. The main objectives are to study the influence of small-scale model error on GLUE interval accuracy and the evaluation of GLUE intervals compared against observations that are in themselves uncertain. Previously recognised problems that are also briefly studied here include the problem of how to sample parameter space to avoid inaccuracies and the problem that the accuracy may depend upon the chosen goodness-of-fit measure. A synthetic groundwater flow model is used for the study so that computed values can be compared with the actual (true) values that are never known in practice. The sequence of sections following gives an overview of the GLUE methodology, describes the synthetic model, describes the method used for testing interval accuracy, presents the results and draws the main conclusions from the study.

The GLUE method as applied here

The procedure used for the GLUE method is the following.

1) Define a function (measure) that quantifies the goodness of fit between observations and corresponding simulated values. This function is termed a likelihood function by Beven and Binley (1992, p. 283), who do not require it to be derived from the probability density function for the differences between observed and simulated values, as would follow from the commonly used statistical definition of the term. This disparity may often be critical, as is shown later. If it is known that the errors $\boldsymbol{e}$, defined as differences between observed values and corresponding simulated values, are normally distributed it would be obvious to use the Gaussian likelihood function

$$L(\boldsymbol{\theta}|\boldsymbol{Y}) = \left((2\pi)^n|\boldsymbol{V}|\right)^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{e}^T\boldsymbol{V}^{-1}\boldsymbol{e}\right) \quad (1)$$

where $\boldsymbol{\theta}$ is the parameter vector, $\boldsymbol{Y}$ is the vector of observations by which the model simulations are compared, $\boldsymbol{e} \sim N(\boldsymbol{0},\boldsymbol{V})$ is the vector of normally distributed errors with a mean vector of zero's and covariance $\boldsymbol{V}$ and n is the dimension of $\boldsymbol{Y}$ and $\boldsymbol{e}$. Usually $\boldsymbol{V}$ is unknown but in the following synthetic case studies it is known. Two other functions than Eq. (1) are also used as goodness-of-fit functions in the following case studies.

2) Define the prior distribution of parameters. In the following case studies there is only one parameter, the $\log_{10}$ transmissivity. The prior distribution is assumed to be uniform and cover the range from -5 to 5 (it is made this wide to apply to all the following examples and realisations). The distribution is sampled by using uniform increments. The increment size is changed from example to example. Applications of GLUE have used this sampling technique (Beven and Freer 2001, p. 17).

3) Calculate the estimated distributions for the predicted variables. For each predicted variable this is done by ranking the predicted values g_i by value, then for each g_i, cumulating the goodness-of-fit function values producing predicted values less than or equal to g_i. Before accumulation the goodness-of-fit function values are rescaled to give a cumulative sum of 1.0 by division by the sum of values over all simulations. In the following case studies the predicted variables are hydraulic head values at different locations within the flow domain.

4) Compute uncertainty bounds for each predicted variable (here the hydraulic head) using its estimated distribution. In the following the 2.5 and 97.5 percentiles of the estimated cumulative goodness-of-fit distribution are used as bounds for the 95% GLUE interval for the predicted variable.

Synthetic models

The models used in the following are somewhat similar to the model used by Guadagnini and Neuman (1999). The dimensions of the two-dimensional flow domain are 18×8, divided into 90×40 uniform structural elements (Figure 1). The transmissivity is constant within each structural element. There is a pumping well in the centre of the domain where groundwater is abstracted at a rate of 1. There is no flow across the top and bottom boundaries, a constant head of 0 along the right boundary and a constant flux across the left boundary (simulated as recharge equal to 3.1076 over the leftmost column of cells).

The true $\log_{10}$-transmissivity field, $\boldsymbol{\beta}$, is assumed to be a stationary random field in the same way as assumed by Guadagnini and Neuman (1999). That is, the vector $\boldsymbol{\beta}$ has the dimension equal to the number of structural elements and is generated so that

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\bar{\theta}\boldsymbol{1}, \boldsymbol{\sigma}_\beta^2 \boldsymbol{V}_\beta\right) \tag{2}$$

where $\bar{\theta}$ is a scalar equal to the expected value of the $\log_{10}$ transmissivity, which is constant for the entire field, $\boldsymbol{1}$ is a vector with all elements equal to one, σ_β is a scalar and $\sigma_\beta^2 \boldsymbol{V}_\beta$ is the covariance of $\boldsymbol{\beta}$. In the examples $\bar{\theta} = 0$, whereas two different values of σ_β were used ($\sigma_\beta = 0.0$ and $\sigma_\beta = 0.75$). The covariance of $\boldsymbol{\beta}$ is assumed to be exponential with a correlation scale of 1 and $\boldsymbol{\beta}$ was generated using a simulator based on the matrix decomposition technique (identical to the simulator described by Kitanidis (1997, p. 237)).

The n-vector of observations, $\boldsymbol{Y}$, used to compute the goodness-of-fit values consists of 36 observations of hydraulic head observed at homogeneously distributed grid points (Figure 1). It is assumed that

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \tag{3}$$

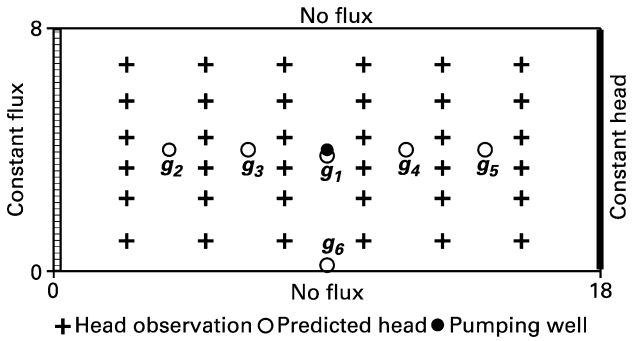

Figure 1 Domain, boundary conditions, observations and predictions for flow models

where $\boldsymbol{f}(\boldsymbol{\beta})$ is the vector of hydraulic heads corresponding to $\boldsymbol{Y}$ computed using the true transmissivity field and $\boldsymbol{\varepsilon}$ is a vector of observation errors for which

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}) \tag{4}$$

where $\boldsymbol{0}$ is a vector of zeros, σ_ε^2 is a scalar equal to the variance of the measurement error and $\boldsymbol{I}$ is the identity matrix. In the examples the observations, $\boldsymbol{Y}$, are computed by adding a random error, $\boldsymbol{\varepsilon}$, to hydraulic head values, $\boldsymbol{f}(\boldsymbol{\beta})$, computed by the numerical model using the true log-transmissivity field, $\boldsymbol{\beta}$. Two values of standard error, σ_ε, were used: $\sigma_\varepsilon = 0.0$ and $\sigma_\varepsilon = 1.0$. That is, in some examples the observation errors are assumed to be zero, but in most examples they are assumed to be significant (of the order of 10% of the head difference between the left and right boundaries).

In the model used to compute GLUE intervals for predictions it is assumed that the log-transmissivity field can be approximated by one constant parameter value, θ (one zonal constant), because $\boldsymbol{\beta}$ is a stationary random variable. This value is sampled from its prior parameter distribution as mentioned above. Because the actual variable log-transmissivity field, $\boldsymbol{\beta}$, is replaced in the model by $\theta \boldsymbol{I}$, small-scale model errors are introduced in the model simulations. Notice that in this paper the term 'small-scale model error' stands for the error in model prediction that arises because only the large-scale features of the hydrogeology are represented in the model used for prediction.

Cooley (2003) has shown that for linear models of $\boldsymbol{f}(\theta \boldsymbol{I})$ and $\boldsymbol{f}(\boldsymbol{\beta})$,

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{f}(\theta_* \boldsymbol{I}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}) \tag{5}$$

where $\boldsymbol{f}(\theta_* \boldsymbol{I})$ is the vector of hydraulic heads computed by a model where each element of $\boldsymbol{\beta}$ is replaced by a constant value θ_* equal to the spatial average value of $\boldsymbol{\beta}$. The mean vector of $\boldsymbol{e}$ is $\boldsymbol{0}$ and the covariance matrix, $\boldsymbol{V}$, depends on both σ_ε^2 and $\sigma_\varepsilon^2 \boldsymbol{V}_\beta$ that are known for the following synthetic examples. $\boldsymbol{V}$ can either be computed accurately for nonlinear models (Cooley 2003), which would not be straightforward, or approximated using linearized models (Cooley 2003), or computed from Monte Carlo simulations. The latter method is used in this study where $\boldsymbol{V}$ is computed from 1000 independent realizations of $\boldsymbol{Y} - \boldsymbol{f}(\theta_* \boldsymbol{I})$ in which $\boldsymbol{Y}$ and $\boldsymbol{f}(\theta_* \boldsymbol{I})$ are produced from nonlinear models (here numerical groundwater flow models).

To emphasise the distinction between the symbols $\bar{\theta}$, θ_* and θ their definition is repeated: $\bar{\theta}$ is the expected value of the $\log_{10}$ transmissivity (i.e. the expectation of the underlying stochastic process), θ_* is the spatial average of $\boldsymbol{\beta}$ (i.e. the spatial average $\log_{10}$ transmissivity of the individual realisation, $\boldsymbol{\beta}$, of the stochastic process) and θ is a $\log_{10}$-transmissivity value sampled from the uniform prior distribution.

As an approximation (the quality of which is investigated here) Eq. (1) is used as a likelihood function for the GLUE methodology. Some of the increased dispersion due to model nonlinearity is accounted for by using $\boldsymbol{V}$ computed as given above. The consequence of choosing another function similar to Eq. (1) but with $\boldsymbol{V}$ substituted by $\boldsymbol{I}$ is also investigated. A third function taken from Feyen et al. (2001, Eq. (7a)) is used in the last example.

For each individual realization of the $\log_{10}$-transmissivity field, $\boldsymbol{\beta}$, the GLUE method described above is used to compute the bounds of the 95% GLUE interval for hydraulic head predictions at the six locations given in Figure 1. Each pair of limits is compared with $g_i(\theta_* \boldsymbol{I})$, $g_i(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ (for $i = 1,6$) where g_i is a computed head value and ε_{gi} is a random error with variance σ_g^2. The variable $g_i(\theta_* \boldsymbol{I})$ is the model-computed value of the hydraulic head at location i, whereas $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ corresponds to an uncertain observation of the true head value, $g_i(\boldsymbol{\beta})$, at location i. If the random error $\varepsilon_{gi} = 0$ (e.g. if $\sigma_g = 0$) the observation, $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$, equals the true head value, $g_i(\boldsymbol{\beta})$.

The flow computations that produced $\boldsymbol{f}(\boldsymbol{\beta})$, $\boldsymbol{f}(\theta\boldsymbol{1})$, $\boldsymbol{f}(\theta_*\boldsymbol{1})$, $g_i(\theta\boldsymbol{1})$, $g_i(\theta_*\boldsymbol{1})$ and $g_i(\boldsymbol{\beta})$ were made by a finite-difference model having 99×49 numerical elements. Most numerical elements coincide with the structural elements, but numerical grid refinement was used around the pumping well to produce accurate results in this area. The vector of observations $\boldsymbol{Y}$ was computed from Eq. (3) by adding random error to $\boldsymbol{f}(\boldsymbol{\beta})$.

Method for testing interval accuracy

For each of the following examples, the accuracy of the GLUE intervals is tested using the following procedure.

(1) Generate a realisation of log transmissivity, $\boldsymbol{\beta}$, and errors $\boldsymbol{\varepsilon}$, and $\boldsymbol{\varepsilon}_{gi}$. Use $\boldsymbol{\beta}$ in the numerical model to compute $\boldsymbol{f}(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta})$, for $i = 1,6$, and add error to compute the observations $\boldsymbol{Y}$ and the uncertain head observations used to test interval accuracy, $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$.

(2) Compute the spatial average $\log_{10}$ transmissivity θ_* as the average of all values in $\boldsymbol{\beta}$. Use this value as a constant $\log_{10}$-transmissivity value in the numerical model to compute modelled hydraulic heads $g_i(\theta_*\boldsymbol{1})$ at the six locations.

(3) Use $\boldsymbol{Y}$ to compute 95% GLUE intervals for the hydraulic head at the six locations. The GLUE intervals are computed by the method described above. The goodness-of-fit function used to compute the interval bounds is either Eq. (1) or one of two others mentioned later. In the goodness-of-fit function $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{f}(\theta\boldsymbol{1})$ where the head values $\boldsymbol{f}(\theta\boldsymbol{1})$ are computed by the numerical model using a constant $\log_{10}$-transmissivity value θ sampled from the prior parameter distribution.

(4) For $i = 1,6$, register (by counters) whether $g_i(\theta_*\boldsymbol{1})$ falls inside, below or above the GLUE uncertainty interval. Similar checking is also done (using other counters) for the true head values $g_i(\boldsymbol{\beta})$ and for the uncertain head observations $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$, respectively.

(5) If 1)–4) have been done less than 1000 times, go back to 1).

Based on the 1000 results, the number of realisations for which each predicted head, $g_i(\theta_*\boldsymbol{1})$, falls inside, below and above the corresponding 95% GLUE interval is compared with the expected number. If the number of times $g_i(\theta_*\boldsymbol{1})$ falls inside is equal (or close) to 950 then the interval is accurate, because the experimental coverage probability is equal (or close) to the nominal probability of 95%. If the number of times $g_i(\theta_*\boldsymbol{1})$ falls below the interval is significantly different from the number of times $g_i(\theta_*\boldsymbol{1})$ falls above the interval then the interval is said to be skew. Similar checking is done for the intervals using $g_i(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$, respectively.

In the first of the following examples it is assumed that there is no small-scale model error (i.e. $\boldsymbol{\beta} = \theta_*\boldsymbol{1} = \boldsymbol{0}$). This case demonstrates that the GLUE interval produced by the procedure given above is an accurate interval for $g_i(\boldsymbol{\beta})$ but an inaccurate interval for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$. The interval for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ can be corrected to account for the observation error in the following way. We can write $g_i(\boldsymbol{\beta}) + \varepsilon_{gi} = g_i(\theta\boldsymbol{1}) + e_{gi} + \varepsilon_{gi}$ where e_{gi} is the simulation error when $\theta\boldsymbol{1}$ is used for the log-transmissivity field instead of $\boldsymbol{\beta}$. Because the errors e_{gi} and ε_{gi} are independent, the estimated probability for $Z_i = e_{gi} + \varepsilon_{gi}$ is

$$p_i(Z_i = z) = \sum_{j=1}^{N} p_{ei}(x_j) p_\varepsilon(z - x_j) \tag{6}$$

and the estimated cumulative distribution is

$$P_i(Z_i < z) = \int_{-\infty}^{z} p_i(x)\mathrm{d}x \tag{7}$$

where N is the number of sampled values of θ, $p_{ei}(x_j)$ is the probability that $e_{gi} = x_j = g_i(\boldsymbol{\beta}) - g_i(\theta_j \boldsymbol{1})$ (which is obtained using the jth sampled value of θ, θ_j, in step 3 of the procedure described above) and $p_\varepsilon(z - x_j)$ is the probability that $\varepsilon_{gi} = z - x_j$ (which is assumed to be Gaussian). The distribution Eq. (7) is used to compute what in the following is called the corrected 95% GLUE interval for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$. The integration in Eq. (7) was done numerically using the trapezoidal rule. The accuracy of corrected intervals is checked as described above for the uncorrected GLUE intervals.

Results

The following subsections present results that illustrate how the accuracy of GLUE intervals can be affected by the sampling scheme, by the choice of goodness-of-fit function and by small-scale model error. An evaluation of GLUE intervals for use with uncertain observations is also made.

Effect of sampling scheme

In the first example there is no small-scale variability in the log transmissivity (i.e. $\boldsymbol{\beta} = \boldsymbol{0}$), which means that $g_i(\theta_* \boldsymbol{1}) = g_i(\boldsymbol{\beta})$ and $\boldsymbol{V} = \sigma_\varepsilon^2 \boldsymbol{I}$. For $\boldsymbol{Y}$ the observation errors, $\boldsymbol{\varepsilon}$, have the variance $\sigma_\varepsilon^2 = 1.0$ and the predicted head values have observation errors, ε_{gi}, with the variance $\sigma_g^2 = 1.0$. The goodness-of-fit function is Eq. (1), the Gaussian likelihood function.

Table 1 shows that the scheme used to sample the prior distribution can be extremely important. Uniform sampling of 1000 values per $\log_{10}$-transmissivity cycle produces results for $g_i(\theta_* \boldsymbol{1})$ with close to 95% of the realizations of each $g_i(\theta_* \boldsymbol{1})$ value being inside its GLUE interval, close to 2.5% being below the interval and close to 2.5% being above the interval (Table 1). The GLUE interval is thus close to having the characteristics of a 95% confidence interval for $g_i(\theta_* \boldsymbol{1})$. Uniform sampling of 100 values per $\log_{10}$-transmissivity cycle gives similar results except that the number of realisations falling below the interval is a little different than the number of realisations falling above the interval (Table 1). This indicates that the computed intervals are skewed by the reduced sampling frequency. Finally, uniform sampling of only 10 values per $\log_{10}$ cycle is clearly insufficient and produces erroneous results, with far too few realisations (only 258) of each prediction being inside its uncertainty interval (Table 1). The results also indicate that the GLUE intervals are highly skewed in this case. Analysis shows that the skewness arises because the estimated hydraulic head distributions shift for increasing increments of parameter sampling (in a plot of head versus cumulated probability, the distribution estimated using the larger sampling increment plots to the left of that estimated using the smaller increment; the former distribution is also the more edgy).

GLUE intervals used with uncertain observations

Table 1 shows that with the highest sampling frequency the relative number of times that $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ is inside the 95% GLUE interval is smaller than 36% for the six hydraulic heads studied. The GLUE interval does not quantify the full uncertainty of the random variable $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ because it is not dependent on the observation error, ε_{gi}. (In the statistics literature the uncertainty interval for a random variable is called a prediction interval. The GLUE interval is not an accurate prediction interval.) If the observation error variance, σ_g^2, had been smaller than 1.0, the GLUE interval would probably have been a better approximation for the 95% uncertainty interval of $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$. This leads to the conclusion that one should be careful with using a GLUE interval to gauge the uncertainty of uncertain observations of a variable.

Table 1 also shows how many times that $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ is inside, below and above the corrected 95% GLUE interval obtained using Eq. (7) instead of using step 3 in the standard

Table 1 Number of realisations where the predicted variables fell below, inside and outside the 95% GLUE intervals for $\boldsymbol{\beta} = \boldsymbol{0}$, $\sigma_\varepsilon^2 = 1.0$ and $\sigma_g^2 = 1.0$ The results were obtained using uniform sampling with a frequency of 1000, 100 and 10 per $\log_{10}$-transmissivity cycle, respectively

Sampling frequency	i	(x, y)	$g_i(\theta 1) = g_i(\beta)$			$g_i(\beta) + \varepsilon_{gi}$					
						Uncorrected interval			Corrected interval		
			Below	Inside	Above	Below	Inside	Above	Below	Inside	Above
1000 per log cycle	1	(9.0, 3.8)	22	958	20	398	200	402	26	948	26
	2	(3.8, 4.0)	22	958	20	322	358	320	18	959	23
	3	(6.4, 4.0)	22	958	20	351	284	365	20	955	25
	4	(11.6, 4.0)	22	958	20	439	162	399	24	954	22
	5	(14.2, 4.0)	22	958	20	432	94	474	32	941	27
	6	(9.0, 0.2)	22	958	20	390	227	383	18	971	11
100 per log cycle	1	(9.0, 3.8)	18	952	30	390	195	415	25	949	26
	2	(3.8, 4.0)	18	952	30	312	358	330	17	960	23
	3	(6.4, 4.0)	18	952	30	340	282	378	20	955	25
	4	(11.6, 4.0)	18	952	30	430	164	406	24	953	23
	5	(14.2, 4.0)	18	952	30	430	95	475	32	941	27
	6	(9.0, 0.2)	18	952	30	385	227	388	18	970	12
10 per log cycle	1	(9.0, 3.8)	7	258	735	313	207	480	24	952	24
	2	(3.8, 4.0)	7	258	735	200	358	442	15	972	13
	3	(6.4, 4.0)	7	258	735	249	283	468	17	967	16
	4	(11.6, 4.0)	7	258	735	365	179	456	24	955	21
	5	(14.2, 4.0)	7	258	735	399	96	505	31	943	26
	6	(9.0, 0.2)	7	258	735	292	255	453	17	975	8

GLUE procedure described above. The results indicate that five out of six corrected GLUE intervals in this case are accurate prediction intervals for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$, whereas the GLUE interval for the sixth observation appears to be a little conservative.

Effect of small-scale model error

In the second modelling example the $\log_{10}$ transmissivity is a correlated random field with exponential covariance, a correlation scale of 1.0, and $\sigma_\beta^2 = 0.75$. The observation error variance for $\boldsymbol{Y}$ is either (a) $\sigma_\varepsilon^2 = 1.0$ or (b) $\sigma_\varepsilon^2 = 0.0$ and the observation error variance for the predicted head values is $\sigma_g^2 = 1.0$. The Gaussian likelihood function (1) is used as goodness-of-fit function and $\boldsymbol{V}$ is computed by the Monte Carlo procedure as described above.

Table 2(a) shows that when $\sigma_\varepsilon^2 = 1.0$ the 95% GLUE interval contains $g_i(\theta_*\boldsymbol{I})$ for 94% of the realisations. The results indicate that GLUE intervals in this case are accurate confidence intervals for hydraulic heads computed using the spatial average log transmissivity, θ_*. The intervals are skew.

Table 2(a) also shows that the GLUE uncertainty interval in this case is not an accurate interval for the true hydraulic head, $g_i(\boldsymbol{\beta})$, or for an observation of the hydraulic head, $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$, at the six observed locations. At five positions the relative number of realisations lying inside the GLUE interval is significantly smaller than 95%, and at one position the relative number of realisations lying inside is larger than 95%. Larger numbers of $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ are inside the corrected GLUE intervals but, because of the small-scale model error, the corrected intervals are found to be inaccurate prediction intervals.

Table 2(b) shows similar results for the case when the observations, $\boldsymbol{Y}$, used to compute the GLUE intervals were made without uncertainty ($\sigma_\varepsilon^2 = 0.0$). The accuracy of the GLUE is slightly worse in this case compared to case (a) where the observation uncertainty is significant ($\sigma_\varepsilon^2 = 1.0$).

The conclusion for this example is that because $\boldsymbol{V}$ is known then (1) can be used to compute close to accurate confidence intervals for $g_i(\theta_*\boldsymbol{I})$. Neither the uncorrected nor the corrected GLUE intervals account for small-scale model error. They are therefore inaccurate intervals for $g_i(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$.

It could be tempting to conclude from the results in Table 2 that, because the GLUE limits are narrow compared to observations of hydraulic head, the model structure must be invalid. However, Christensen and Cooley (2003) show that for this case the hydraulic heads at positions 2–5 are nearly unbiased predictions of the true hydraulic heads if the heads are computed using a least-squares estimate of θ_*, whereas the predicted head at position 1 next to the pumping well is somewhat biased (10% error). It would therefore be fair to say that the model structure is valid except that small-scale variability in transmissivity is not (and never can be in practice) accounted for in the model. The GLUE interval does not account for the uncertainty caused by small-scale model error, which can never be avoided. One should therefore be careful with basing a rejection of a model structure on observations falling outside GLUE intervals. The model structure, which is used to define θ_*, should not be rejected from such results alone.

Effect of chosen goodness-of-fit function

The third and fourth modelling examples illustrate some consequences of using a goodness-of-fit function that is different from Eq. (1), which is considered the best one for the present example. The assumed conditions are the same as in example 2, i.e. the $\log_{10}$-transmissivity has exponential covariance with a correlation scale of 1.0 and $\sigma_\beta^2 = 0.75$, and the observation error variances are $\sigma_\varepsilon^2 = 1.0$ and $\sigma_g^2 = 1.0$, respectively.

Table 3 shows results obtained using a goodness-of-fit function that equals (1) with the identity matrix $\boldsymbol{I}$ substituted for $\boldsymbol{V}$. Note that $\boldsymbol{I}$ does not quantify the true covariance of $\boldsymbol{e}$.

Table 2 Number of realisations where the predicted variables fell below, inside and outside the 95% GLUE intervals for $\sigma_\beta^2 = 0.75$, $\sigma_g^2 = 1.0$ and (a) $\sigma_\varepsilon^2 = 1.0$ or (b) $\sigma_\varepsilon^2 = 0.0$. The goodness-of-fit function used is (1). The results were obtained using uniform sampling with a frequency of 1000 per $\log_{10}$-transmissivity cycle

i	$g_i(\theta \cdot 1)$			$g_i(\beta)$			$g_i(\beta) + \varepsilon_{gi}$					
							Uncorrected interval			Corrected interval		
	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above
(a)												
1	55	940	5	118	769	113	136	715	149	111	815	74
2	55	940	5	8	990	2	11	981	8	12	986	2
3	55	940	5	41	924	35	53	895	52	50	926	24
4	55	940	5	118	726	156	155	658	187	99	814	87
5	55	940	5	179	645	176	269	502	229	110	814	76
6	55	940	5	80	819	101	110	742	148	89	837	74
(b)												
1	55	930	15	130	747	123	147	699	154	120	801	79
2	55	930	15	9	975	16	19	955	26	19	970	11
3	55	930	15	56	895	49	66	871	63	61	906	33
4	55	930	15	141	686	173	167	632	201	111	798	91
5	55	930	15	195	619	186	284	474	242	114	806	80
6	55	930	15	87	793	120	115	727	158	93	825	82

Table 3 Number of realisations where the predicted variables fell below, inside and outside the 95% GLUE intervals for $\sigma_\beta^2 = 0.75$, $\sigma_\varepsilon^2 = 1.0$ and $\sigma_g^2 = 1.0$. The goodness-of-fit function is a modification of (1) where $\boldsymbol{I}$ is substituted for $\boldsymbol{V}$. The results were obtained using uniform sampling with a frequency of 1000 per $\log_{10}$-transmissivity cycle

i	$g_i(\theta \cdot 1)$			$g_i(\beta)$			$g_i(\beta) + \varepsilon_{gi}$					
							Uncorrected interval			Corrected interval		
	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above
1	357	221	422	439	143	418	456	118	426	159	728	113
2	357	221	422	269	385	346	314	265	421	79	819	102
3	357	221	422	275	333	392	348	197	455	92	836	72
4	357	221	422	441	135	424	447	73	480	131	756	113
5	357	221	422	528	76	396	486	76	438	131	782	87
6	357	221	422	352	181	467	394	123	483	132	753	115

This has the consequence that the GLUE interval in all six cases is very inaccurate as a probability interval for any of the functions $g_i(\theta_* \boldsymbol{I})$, $g_i(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$. For $g_i(\theta_* \boldsymbol{I})$ only 221 out of 1000 realisations are inside the GLUE interval, for $g_i(\boldsymbol{\beta})$ between 76 and 385 realisations are inside and for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ between 73 and 265 realisations are inside. For all variables the GLUE interval is much too narrow. For the corrected GLUE intervals between 728 and 836 realisations of $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ are inside. The inaccuracy of the uncorrected and corrected GLUE intervals are partly because of small-scale model error, which is not accounted for, and partly because $\boldsymbol{I}$ is used instead of $\boldsymbol{V}$ in the goodness-of-fit function (1). The latter reason is the most significant. This is noticed by comparing the results in Table 2 (obtained using $\boldsymbol{V}$) with the results in Table 3 (obtained using $\boldsymbol{I}$).

Table 4 shows results obtained using a goodness-of-fit function taken from Feyen *et al.* (2001, Eq. (7a)):

$$L(\theta|\mathbf{Y}) = \begin{cases} 0, & P > r_{\lim} \\ \frac{1}{P}, & P \leq r_{\lim} \end{cases} \tag{8}$$

where

$$\frac{1}{P} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} (Y_i - f_i(\theta \mathbf{1}))^2}.$$

Note that P equals the sum of squared errors divided by the number of observations. The function is equal to $1/P$ in cases where P does not exceed a model rejection value, $r_{\lim}$. In cases where P exceeds the rejection value the function is set equal to 0. The model rejection value can be tuned to produce GLUE intervals of a width that the modeller finds realistic. The results in Table 4 were computed using $r_{\lim} = 25.0$ which was found to produce a relative number of realisations of $g_i(\theta_* \boldsymbol{I})$ lying inside the GLUE interval close to 95%, the expected number for a 95% probability interval. Table 4 shows that for the six hydraulic heads 948 out of 1000 realisations of $g_i(\theta_* \boldsymbol{I})$ are inside the GLUE interval, while 42 realisations are below and 10 realisations are above the interval, respectively, while for $g_i(\boldsymbol{\beta})$ between 752 and 986 realisations are inside and for $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ between 580 and 983 realisations are inside. Between 858 and 990 realisations of $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$ are inside the corrected GLUE intervals. Thus the GLUE intervals are inaccurate intervals for $g_i(\boldsymbol{\beta})$ and $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$. A comparison of the results in Table 4 with those in Table 2 indicates that Eq. (8) produces intervals that are somewhat more accurate than those produced from Eq. (1), which was considered the best goodness-of-fit function because it is based on the true error structure. However, fitting of the model rejection value, $r_{\lim}$, was necessary to obtain the results in Table 4. Experiments showed that the model rejection value has to be adjusted for each predicted hydraulic head, $g_i(\boldsymbol{\beta})$ (or $g_i(\boldsymbol{\beta}) + \varepsilon_{gi}$), individually to produce a relative number of realisations being inside the GLUE interval close to 95%. In other words, there is no unique tuning that works for all hydraulic head variables.

Summary and conclusions

Synthetic groundwater flow models have been used to study some problems with the Generalized Likelihood Uncertainty Estimation (GLUE) method that can significantly affect the accuracy of GLUE intervals, or at least affect their interpretation. Synthetic models were preferred because in such cases the true error structure is either known or can be accurately estimated using the Monte Carlo method. Monte Carlo estimation was used in this study.

Table 4 Number of realisations where the predicted variables fell below, inside and outside the 95% GLUE intervals for $\sigma_\beta^2 = 0.75$, $\sigma_\varepsilon^2 = 1.0$ and $\sigma_g^2 = 1.0$. The goodness-of-fit function used is (8) with $r_{\text{lim}} = 25.0$. The results were obtained using uniform sampling with a frequency of 1000 per $\log_{10}$-transmissivity cycle

I	$g_i(\theta,1)$			$g_i(\beta)$			$g_i(\beta) + \varepsilon_{gi}$					
							Uncorrected interval			Corrected interval		
	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above	Below	Inside	Above
1	42	948	10	51	909	40	66	856	78	55	911	34
2	42	948	10	4	986	10	5	983	12	2	990	8
3	42	948	10	7	981	12	12	970	18	8	984	8
4	42	948	10	44	867	89	82	777	18	53	883	64
5	42	948	10	106	752	142	224	580	196	80	858	62
6	42	948	10	18	951	31	35	902	63	25	948	27

The groundwater flow models studied only have one unknown parameter, the spatial average log transmissivity of the flow domain. Observed hydraulic head values were computed at points homogeneously distributed within the domain.

Uniform sampling of 1000 values per $\log_{10}$-transmissivity cycle were required to produce unbiased GLUE results. Using uniform sampling of 100 values per $\log_{10}$ cycle gave slightly skewed results, while using uniform sampling of only 10 values per $\log_{10}$ cycle produced erroneous results. If these results can be extended to multi-parameter groundwater models in general, which may not be the case, then it will be computationally expensive to use uniform sampling to obtain accurate GLUE results in practical groundwater modelling.

The Gaussian likelihood function was used as the GLUE goodness-of-fit function in most cases studied here because it is approximately correct for the distribution of the true errors. Using this, it was demonstrated that the GLUE interval computed for the hydraulic head at different locations within the domain has the characteristics of a confidence interval for the hydraulic head computed using the average value of log transmissivity. The GLUE interval does not have the characteristics of a prediction interval, which is a probability interval for an uncertain observation of, for example, hydraulic head. This leads to the conclusion that one should be careful with using a GLUE interval to gauge the uncertainty in observations of a variable. As a consequence, one should be careful with using the interval as recommended by Beven and Binley (1992, p. 285): "if the uncertainty limits are drawn too narrowly, then a comparison with observations will suggest that the model structure is invalid". The structure may be valid even though observations fall significantly outside the GLUE interval if the observations are uncertain; they are always uncertain to some extent.

It was demonstrated that the GLUE interval can be corrected to include uncertain observation error in a predicted observation.

Small-scale model error is always present in practical groundwater modelling because the model usually only represents the drift values of the hydrogeologic variables, whereas variability below a certain spatial (or temporal) scale has to be ignored. It has been demonstrated that neither the GLUE interval nor the corrected GLUE interval account for the uncertainty caused by small-scale model error. Also for this reason one must be careful with basing a rejection of a model structure on observations falling outside GLUE intervals. The model drift structure should not be rejected from such a comparison alone. In many cases predictions would only be useless if they were significantly biased. Small-scale model error did not bias the predictions in the examples studied here except at a strong sink. The conclusion that one must be careful with basing a rejection of a model structure on observations falling outside GLUE intervals is supported by the statement made by Beven and Freer (2001, p. 24): "prediction limits, however, are quantiles of the model predictions, not direct estimates of the probability of simulating a particular observation (which is not easily estimated given model structural error)".

It is well known that GLUE results depend on the goodness-of-fit function used. This was also demonstrated by examples in this manuscript. Changing from the Gaussian likelihood function to a similar but different function made the resulting GLUE intervals very inaccurate. Changing to a third function based on a fitted model rejection value produced results that were better than those obtained with the Gaussian likelihood function. However, the results were sensitive to the rejection value, which in the ideal case should be adjusted individually for each predicted variable. Thus, the third function is not attractive for practical application with the GLUE methodology.

Using the same test procedure for the same groundwater flow problem with small-scale model error as studied here, Christensen and Cooley (2003) tested the accuracy of non-linear regression-based prediction intervals. Their testing showed that for all six head predictions the experimental coverage probability of the 95% prediction interval varied between 93.2%

and 98.8%, which is fairly close to the nominal probability of 95%, and which is far better than the results obtained for the 95% GLUE intervals (Tables 2–4). The reason for the better results is that the regression-based prediction interval has contributions from: (1) covariance of the estimated parameters and (2) a residual model error variance. The GLUE procedure considers only the variability of the parameters. Thus, if contribution (1) is much smaller than (2) GLUE has a problem because it is not possible to explain the overall error by parameter uncertainty only. On the other hand, if (1) is the largest contributor, GLUE would probably work well. With respect to groundwater modeling it is the author's experience that contribution (2) is often very significant for hydraulic heads (Christensen 1997, Christensen *et al.* 1998, Christensen and Cooley 1999).

Acknowledgements

This research was supported by the Danish Natural Science Research Council, grant 21-00-0512. R. L. Cooley and the anonymous reviewers are thanked for their comments that improved the manuscript.

References

Beven, K. and Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Processes*, **6**, 279–298.

Beven, K. and Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology. *J. Hydrol.*, **249**, 11–29.

Binley, A.M. and Beven, K.J. (1991). Physically-based modeling of catchment hydrology: a likelihood approach to reducing predictive uncertainty. In: *Computer Modeling in Environmental Sciences*, D.G. Farmer and M.J. Rycroft (eds.). Clarendon Press, Oxford, pp. 75–88.

Brazier, R.E., Beven, K.J., Anthony, S.G. and Rowan, J.S. (2001). Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions. *Earth Surf. Processes Landforms*, **26**, 1333–1352.

Christensen, S. (1997). On the strategy of estimating regional-scale transmissivity fields. *Ground Water*, **35**(1), 131–139.

Christensen, S. and Cooley, R.L. (1999). Evaluation of prediction intervals for expressing uncertainties in groundwater flow model predictions. *Wat. Res. Res.*, **35**(9), 2627–2639.

Christensen, S. and Cooley, R.L. (2003). Experiences gained in testing a theory for modelling groundwater flow in heterogeneous media. In: *Calibration and Reliability in Groundwater Modelling: A Few Steps Closer to Reality. Proc. ModelCARE'2002 (Prague, June 2002)*, K. Kovar and Z. Hrkal (eds.), IAHS Publ. no. 277, pp. 22–27.

Christensen, S., Rasmussen, K.R. and Møller, K. (1998). Prediction of regional ground-water flow to streams. *Ground Water*, **36**(2), 351–360.

Cooley, R.L. (2003). A theory for modeling ground-water flow in heterogeneous media. *U.S. Geological Survey Professional Paper* in press.

Feyen, L., Beven, K.J., De Smedt, F. and Freer, J. (2001). Stochastic capture zone delineation within the generalized likelihood uncertainty estimation methodology: conditioning on head observations. *Wat. Res. Res.*, **37**(3), 625–638.

Freer, J., Beven, K. and Ambroise, B. (1996). Baysian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Wat. Res. Res.*, **32**(7), 2161–2173.

Guadagnini, A. and Neuman, S.P. (1999). Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains – 2. Computational examples. *Wat. Res. Res.*, **35**(10), 3019–3039.

Hankin, B.G., Hardy, R., Kettle, H. and Beven, K.J. (2001). Using CFD in a GLUE framework to model the flow and dispersion characteristics of a natural fluvial dead zone. *Earth Surf. Processes Landforms*, **26**, 667–687.

Jensen, J.B. (2003). *Parameter and Uncertainty Estimation in Groundwater Modelling.*, PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper No. 23.

Kitanidis, P. (1997). *Introduction to Geostatistics*, Cambridge University Press, New York.

Lamb, R., Beven, K. and Myrabø, S. (1998). Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. *Adv. Wat. Res.*, **22**(4), 305–317.