How well does your model fit the data?

M. J. Hall

ABSTRACT

Despite almost five decades of activity on the computer modelling of input–output relationships, little general agreement has emerged on appropriate indices for the goodness-of-fit of a model to a set of observations of the pertinent variables. The coefficient of efficiency, which is closely allied in form to the coefficient of determination, has been widely adopted in many data mining and modelling exercises. Values of this coefficient close to unity are taken as evidence of good matching between observed and computed flows. However, studies using synthetic data have demonstrated that negative values of the coefficient of efficiency can occur both in the presence of bias in computed outputs, and when the computed volume of flow greatly exceeds the observed volume of flow. In contrast, the coefficient of efficiency lacks discrimination for cases close to perfect reproduction. In the latter case, a coefficient based upon the first differences of the data proves to be more helpful.

Key words | modelling, data mining, model calibration, model verification, goodness-of-fit indices

M. J. Hall
International Institute for Infrastructural,
Hydraulic and Environmental Engineering,
P.O. Box 3015,
2601 DA Delft,
The Netherlands

INTRODUCTION

Of the many input–output relationships that are encountered in hydrology, ecology and hydraulics, the modelling of the land phase of the hydrological cycle in general, and the relationship between rainfall and runoff in particular, continues to attract widespread attention. The variety of such models that have been developed is legion: lumped and distributed, physically based and conceptual, linear and non-linear, to list but a few. The scope of rainfall-runoff modelling has recently been extended by the application of models composed of Artificial Neural Networks (e.g. Minns & Hall 1996; Shamseldin 1997; Dawson & Wilby 1998), and has been subsumed into the wider activity of *data mining*, i.e. the processes of knowledge discovery and, ultimately, data reduction. Owing to the wide availability of software, data mining techniques are generally relatively easy to implement. However, the process of knowledge discovery tends to break down at the stage of interpreting results, since the analyst may not be fully versed in the necessary *domain knowledge*. The latter is particularly important in comparing model outputs to the observations selected for training and validation. The question arises as to which features of the computed and observed outputs should be emphasised in determining the efficacy of the model. This problem is unfortunately not always accorded the attention that it deserves. Dimensionless indices employed for the assessment of goodness-of-fit are often standardised using a function involving the variance of the observed data set. Indices that apply to the comparison of different models on the same set of observations therefore do not need to be as sophisticated as those for evaluating the performance of the same model on data sets of different length and variability. However, such indices tend to emphasise only a limited set of features in the data, and for a model of (say) daily streamflows, a series of measures might encompass those outlined in Table 1. This compilation, adapted and expanded from that presented by Gupta *et al.* (1998) for the calibration of a specific model, serves to illustrate the multifarious aspects of model behaviour which could, and should, be addressed in any model application.

The overall objective of applying the above criteria is the identification of a set of parameters that is capable of

Table 1 | Goodness-of-fit measures for a typical daily rainfall-runoff model.

Feature	**Definition**
Daily root mean square error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(q_i-\hat{q}_i)^2}$
Total mean square error in monthly volume	$\sum_{i=1}^{n_{\text{month}}}\left(\frac{1}{n_{\text{day}}(i)}\sum_{k=1}^{n_{\text{day}}(i)}(q_i-\hat{q}_i)\right)^2$
Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert q_i-\hat{q}_i\rvert$
Maximum absolute error	$\max\lvert q_i-\hat{q}_i\rvert;\ 1\le i\le n$
Coefficient of efficiency	$1-\frac{\frac{1}{n}\sum_{i=1}^{n}(q_i-\hat{q}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(q_i-\overline{q})^2}$
Mean daily error (average bias)	$\frac{1}{n}\sum_{i=1}^{n}(q_i-\hat{q}_i)$
Maximum difference in the highest peak discharge	$\max\ (q_i)-\max\ (\hat{q}_i);\ 1\le i\le n$
Mean error in time-to-peak for all storm events	$\frac{1}{n_e}\sum_{j=1}^{n_e}(T_j-\hat{T}_j)$
Root mean square error in time-to-peak	$\sqrt{\frac{1}{n_e}\sum_{j=1}^{n_e}(T_j-\hat{T}_j)^2}$
First serial correlation coefficient	$\frac{\sum_{i=1}^{n-1}(q_i-\overline{q})(q_{i+1}-\overline{q})}{\sum_{i=1}^{n}(q_i-\overline{q})^2}$

Notation:
q_i and $\hat{q}_i$ represent the observed and computed flows for day i, $1\le i\le n$.
$\bar{q}$ is the mean of the observed flows.
$n_{\text{day}}\ (i)$ is the number of days in a month.
n_{month} is the number of months in the time series.
n_e is the number of storm events in the observed and computed series.
T_j and $\hat{T}_j$ are the observed and estimated times-to-peak of the jth storm event.

reproducing as closely as possible the recorded streamflow outputs, given the rainfall and possibly other inputs, such as evaporation. Ideally, the modeller would wish to express the goodness-of-fit of the model to the data in terms of a single index or objective function that could be optimised objectively in fitting the model. However, as amply demonstrated by Diskin & Simon (1977), there is no such index that is of universal application. Indeed, the objective function should be selected according to the purpose for which the model is to be applied; a flood

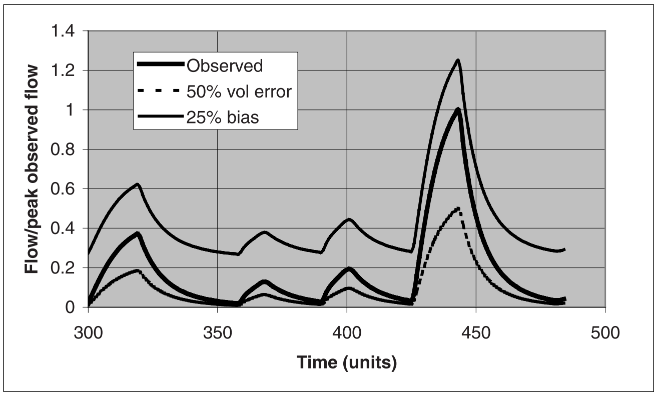

Figure 1 | Segments of the synthetic flow series used in the numerical experiments: (a) the 'observed' series; (b) the 'computed' series with a 50% underestimation of runoff volume; and (c) the 'computed' series with a constant bias of 25% of the observed peak flow.

model should emphasise the peak flows, but a resources model should be orientated towards the low flow sequences. A multi-criteria calibration procedure based upon a global optimisation algorithm has recently been suggested by Gupta *et al*. (1998) in which several different objective functions may be satisfied simultaneously (see also Yapo *et al*. 1998). For example, the hydrograph might be divided into periods with or without rainfall. The rain-free periods might then be further divided into periods dominated by either throughflow or baseflow processes, and model performance assessed separately for each (see, for example, Wagener *et al*. 2000). Nevertheless, the separate objective functions tend to be based upon the root mean square error or related forms of criteria.

The use of a single, all-embracing criterion, such as the coefficient of efficiency (see Table 1), is very attractive because the process of calibration or training is greatly simplified. In addition, fellow modellers tend to have (or *believe* they have) a general appreciation of the relative performance of the model based upon such single measures. For example, Shamseldin (1997) has written that:

> 'A value of [the coefficient of efficiency] of 90% indicates a very satisfactory model performance while a value in the range 80–90% indicates a fairly good model. Values of [the coefficient of efficiency] in the range 60–80% would indicate unsatisfactory model fit.'

In contrast, Beran (1999) has criticised the null hypothesis implicit in the structure of the coefficient of efficiency (outlined below), and has concluded that this criterion provides '. . . an exaggerated impression of the presumed skill in prediction'. According to the same author, a modeller should not be satisfied with a coefficient of efficiency lower than the mid-to-high 90s in percentage terms. These somewhat conflicting statements raise the question as to the overall sensitivity of this criterion to differences in the observed and modelled time series. This paper summarises the results from a series of simulation experiments designed to explore this problem.

THE COEFFICIENT OF EFFICIENCY

One of the most widely used forms of fitting criterion has indeed been the coefficient of efficiency introduced by Nash & Sutcliffe (1970). Those authors drew an analogy with the coefficient of determination familiar from the analysis of variance. This coefficient may be developed as follows. Given a sequence of observed flows, q_i, $i = 1, 2, \ldots, n$, with a mean $\bar{q}$, and a sequence of computed flows, $\hat{q}_i$, $i = 1, 2, \ldots, n$, with the same mean, the sums of the squares of the deviations of the observations from their overall mean may be partitioned approximately into two parts: the sums of the squares of the differences between the observed and the computed values, and the sums of the squares of the deviations of the computed values from their overall mean, i.e.

$$\sum(q_i - \bar{q})^2 \approx \sum(q_i - \hat{q}_i)^2 + \sum(\hat{q}_i - \bar{q})^2 \tag{1}$$

where all summations are taken over the n terms of the sequence. As proposed by Nash & Sutcliffe (1970), the term on the left-hand side of Equation (1) may be regarded as a *no-model* or a *no-skill* variance, i.e. the sum of the squares of the difference between the computed and observed values when the model was simply taken as the average of the recorded flows. (This is the null hypothesis that has been criticised by Beran (1999), as noted above.) The second term on the right-hand side of Equation (1) is the sum of the squares attributable to an actual model, so that the fraction of the total sum of the squares of the observations (or the *no-model* case) explained by that model is given by the ratio:

$$E=\frac{\sum(\hat{q}_i-\overline{q})^2}{\sum(q_i-\overline{q})^2}=1-\frac{\sum(q_i-\hat{q}_i)^2}{\sum(q_i-\overline{q})^2}\ . \qquad (2)$$

In the case of perfect agreement, obviously $E=1$. However, inexperienced users, no doubt with the analogy of the coefficient of determination in mind, often assume that E has a lower limit of zero. However, if the mean square error exceeds the variance of the observed flows, Equation (2) may assume negative numbers. The lower limit of zero only applies if the $\hat{q}_i$ are derived from a simple linear regression of the q_i on an independent variable, in which case Equation (1) may easily be shown to become an identity. The occurrence of negative E values, usually at an early stage in a modelling exercise, is not always interpreted correctly. The question as to the most common circumstances in which such negative values might be encountered gave rise to a more detailed study using a series of numerical experiments.

SYNTHETIC DATA GENERATION

The numerical experiments performed in exploring the behaviour of the coefficient of efficiency were based on a generated time series of streamflows. These data were derived from a sequence of synthetic storm events of varying duration, total depth and profile, occurring at irregular intervals, which were routed through a simple conceptual hydrological model. The storm events were produced using Monte Carlo methods based upon the following assumptions:

1. storm durations were normally distributed, with a mean of 20 time units and a standard deviation of 6 units;
2. storm depths were lognormally distributed, with a mean of 25 mm and a standard deviation of 2 mm (implying a distribution of depths with a coefficient of variation of 0.785 and a skewness coefficient of 2.84);
3. the time variations of depths within each event were defined by one of six storm profiles, each of which was described by a simple polynomial function, broadly based upon those of the *UK Flood Studies Report* (Natural Environment Research Council, 1975), and including early-peaked and late-peaked as well as symmetrical events with a constant intensity profile as an extreme case; and
4. the inter-event times were taken as double the previous storm duration minus one time unit.

Durations averaged 19.2 time units with a standard deviation of 6.95 units, and mean storm depth was 31.6 mm with a standard deviation of 1.9 mm. These data were then routed through a single non-linear reservoir using the RORB model (Mein *et al*. 1974) with a storage constant of 20 and an exponent of 0.8, the latter value being typical for a wide range of catchments (Laurenson & Mein 1988). For convenience, the generated flows were standardised using the largest peak ordinate. A sample sequence of storm hydrographs, being roughly a quarter of the total time series but including the event with the largest flow ordinate, is shown in Figure 1(a).

NUMERICAL EXPERIMENTS

For the purposes of the numerical experiments, the generated time series of flows was assumed to be the sequence of observed flows upon which a hydrological model was to be calibrated. The sequence of model outputs was assumed to be similar in basic form, but subject to the following different types of error:

1. **volume error**: all observed ordinates were multiplied by a constant factor k, $0.5\leq k\leq 1.5$, to give the computed flows (see Figure 1(b));
2. **bias**: a constant displacement, b, $0\leq b\leq 0.25$ standardised flow units, was applied to all observed ordinates in order to form the computed model output (see Figure 1(c)); and
3. **timing error**: the computed flows were displaced by a constant number of time units, t, $-6\leq t\leq 6$, relative to the observed flows.

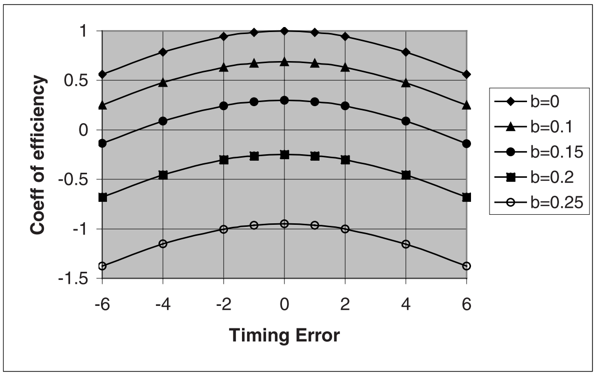

Figure 2 | The effect of bias and timing error on the coefficient of efficiency; the amount of bias ranges from 0–25% of the observed peak flow.

Results are presented below for the separate cases of timing errors combined with either bias or volume error, although the former can also be considered a special case of the latter. A displacement $b = 0.15$ units almost doubles the computed runoff volume, and at $b = 0.25$, the computed volume is 2.64 times the observed volume.

RESULTS

Figure 2 summarises the values of the coefficient of efficiency, as defined in Equation (2), for cases of combined bias and timing error. For any given timing error, the coefficient is seen to decrease non-linearly with the increment in b. More significantly, at a bias of 0.15, time displacements above ±4 time units result in negative coefficients. With $b = 0.2$, all E values are negative. A further point to note is the relatively small changes in coefficients that arise from small timing errors of ±2 time units at all levels of b.

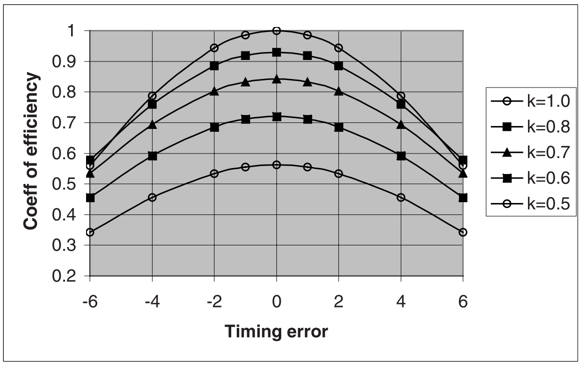

Figure 3 | The effect of volumetric and timing errors on the coefficient of efficiency; volumetric errors range from 0–50% underestimation of the observed runoff volume.

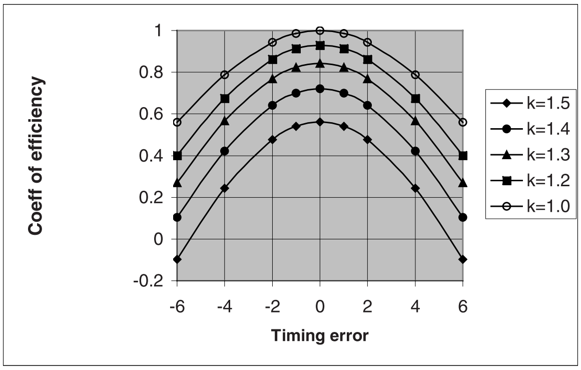

Figure 4 | The effect of volumetric and timing errors on the coefficient of efficiency; volumetric errors range from 0–50% overestimation of the observed runoff volume.

Figure 3 shows the combined effects of timing errors and underestimated runoff volumes, i.e. $0 \le k \le 0.5$. Even when the volumetric error is one-half the observed volume, all coefficients are positive for $-6 \le t \le 6$ time units. However, when the computed runoff volume exceeds the observed, the changes are more marked, with negative coefficients appearing at time displacements of ±6 units for a 50% increase in volume (see Figure 4). A comparison between Figures 3 and 4 shows that, for any given timing error, volume overestimation has more effect than underestimation, with E values falling more rapidly as t increases in either direction.

An initial reaction to the occurrence of negative coefficients of efficiency might be to resort to the use of the formal statistical coefficients of correlation and determination, if only because of their ease of use in widely available spreadsheet software. The correlation coefficient is defined as the ratio between the covariance of the dependent and independent variables (in this case, the modelled and observed flows) divided by the square root of the product of the variances of these variables. However, if the modelled output contains a bias, b, both its

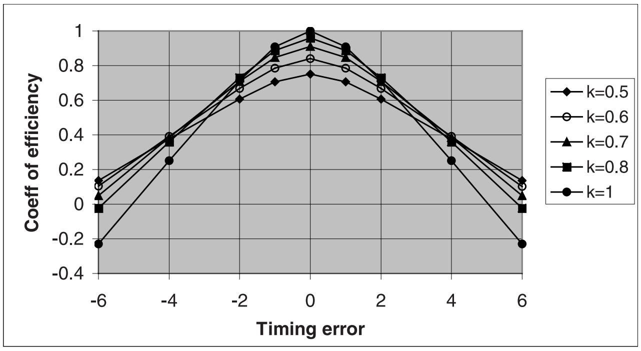

Figure 5 | The effect of volumetric and timing errors on the coefficient of efficiency based upon the first differences of the computed and observed flows; volumetric errors range from 0–50% underestimation of the observed runoff volume.

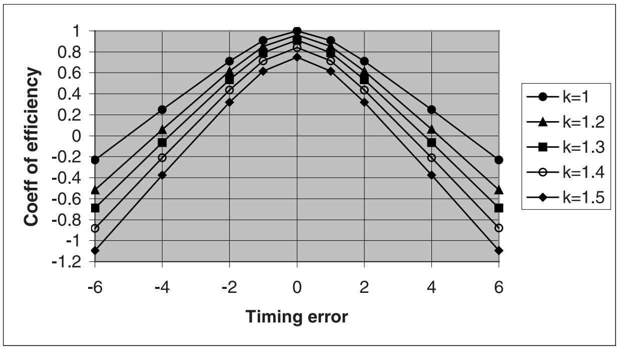

Figure 6 | The effect of volumetric and timing errors on the coefficient of efficiency based upon the first differences of the computed and observed flows; volumetric errors range from 0–50% overestimation of the observed runoff volume.

variance and its covariance with the observed series are unchanged. Similarly, if the model outputs are a constant multiplier, k, times the observed values, then both its variance and its covariance with the observed series are multiplied by k, and the correlation coefficient again remains unchanged. The coefficients of correlation and determination are therefore incapable of reflecting either bias or volumetric error in model results, although they are sensitive to the magnitude, but not the direction, of a timing error. For errors in timing up to ± 4 units but no bias or volumetric error, the coefficients of determination and efficiency are virtually identical.

A notable feature of Figures 2–4 is the comparative insensitivity of the coefficient of efficiency to timing errors for $-2 \leq t \leq 2$ time units for all cases of bias and volumetric error. This performance could be improved if the first differences:

$$\Delta q_i = q_i - q_{i-1};\ \Delta \hat{q}_i = \hat{q}_i - \hat{q}_{i-1};\ i = 2, 3, \ldots, n$$

of the observed and modelled outputs were used instead of the observed and computed ordinates in Equation (2). The results are shown in Figures 5 and 6 for the cases of underestimated and overestimated runoff volumes, respectively. (Obviously, the use of first differences makes the E value insensitive to the amount of bias.) Once again, the effect of overestimation is more pronounced than that of underestimation for any given timing error. A 50% increase in the computed runoff volume at timing errors of $t = \pm 6$ units drives the coefficient of efficiency below -1. However, for small timing errors at all levels of volumetric error, the E value decreases more rapidly with the first differences than with the actual observations.

CONCLUDING REMARKS

The results presented in Figures 2–6 tend to support the conclusion of Beran (1999) that a coefficient of efficiency of 0.95 or more is required to ensure a good model performance. The figures show that the coefficient of efficiency is liable to fall below zero when a strong bias is introduced into the computed output by a hydrological model. A similar effect is possible when the model produces relatively large timing errors along with major overestimation of runoff volumes. The coefficient of efficiency appears less sensitive to the underestimation of runoff volumes, and is relatively insensitive to small timing errors. Nevertheless, the E value does reflect such discrepancies, which cannot be detected by the use of the formal statistical coefficients of correlation and determination. For the case in which the performance of different models is being assessed on the basis of the same set of observed data, identical conclusions should be reached by using either the coefficient of efficiency or the mean square error. Reference to Equation (2) shows that the two criteria differ only in the standardisation by the (constant) observed variance in the *E* value. If timing errors are particularly important to the modelling, then the use of the first differences of the observed and computed ordinates appears more effective than the actual ordinates. However, the preferred solution would be to develop a series of criteria, such as those presented in Table 1, that focuses upon the more important aspects of model behaviour rather than to rely on a single index.

REFERENCES

Beran, M. 1999 Hydrograph prediction—how much skill? *Hydrol. Earth Syst. Sci.* **3**(2), 305–307.

Dawson, C. W. & Wilby, R. 1998 An artificial neural network approach to rainfall-runoff modelling. *Hydrol. Sci. J.* **43**, 47–66.

Diskin, M. H. & Simon, E. 1977 A procedure for the selection of objective functions for hydrologic simulation models. *J. Hydrol.* **34**, 129–149.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1998 Towards improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Wat. Resour. Res.* **34**, 751–763.

Laurenson, E. M. & Mein, R. H. 1988 *RORB—Version 4 runoff routing program—user manual*. Monash University, Clayton, Australia.

Mein, R. H., Laurenson, E. M. & McMahon, T. A. 1974 Simple nonlinear model for flood estimation. *Proc. Am. Soc. Civ. Engrs., J. Hydraul. Div.* **100** (HY11), 1507–1518.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall-runoff models. *Hydrol. Sci. J.* **41**, 399–417.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models. *J. Hydrol.* **10**, 282–290.

Natural Environment Research Council. 1975 *Flood Studies Report, vol II, Meteorological Studies*, The Council, London.

Shamseldin, A. Y. 1997 Application of a neural network technique to rainfall-runoff modelling. *J. Hydrol.* **199**, 272–294.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V. & Soorooshian, S. 2000 A framework for development and application of hydrological models. *Proc. 7th National Hydrology Symposium, Newcastle-upon-Tyne* British Hydrological Society, London, 3.75–3.81.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. 1998 Multi-objective global optimisation for hydrologic models. *J. Hydrol.* **204**, 83–97.