Advances in Geosciences, 5, 89–97, 2005
SRef-ID: 1680-7359/adgeo/2005-5-89
European Geosciences Union

Advances in Geosciences

Comparison of different efficiency criteria for hydrological model assessment

P. Krause[1], **D. P. Boyle**[2], **and F. Bäse**[1]

[1]Department for Geoinformatics, Hydrology and Modelling, Friedrich-Schiller-University, Jena, Germany
[2]Department of Hydrologic Sciences, Desert Research Institute, Reno, Nevada, USA

Received: 7 January 2005 – Revised: 1 August 2005 – Accepted: 1 September 2005 – Published: 16 December 2005

Abstract. The evaluation of hydrologic model behaviour and performance is commonly made and reported through comparisons of simulated and observed variables. Frequently, comparisons are made between simulated and measured streamflow at the catchment outlet. In distributed hydrological modelling approaches, additional comparisons of simulated and observed measurements for multi-response validation may be integrated into the evaluation procedure to assess overall modelling performance. In both approaches, single and multi-response, efficiency criteria are commonly used by hydrologists to provide an objective assessment of the "closeness" of the simulated behaviour to the observed measurements. While there are a few efficiency criteria such as the Nash-Sutcliffe efficiency, coefficient of determination, and index of agreement that are frequently used in hydrologic modeling studies and reported in the literature, there are a large number of other efficiency criteria to choose from. The selection and use of specific efficiency criteria and the interpretation of the results can be a challenge for even the most experienced hydrologist since each criterion may place different emphasis on different types of simulated and observed behaviours. In this paper, the utility of several efficiency criteria is investigated in three examples using a simple observed streamflow hydrograph.

1 Introduction

There are a number of reasons why hydrologists need to evaluate model performance: (1) to provide a quantitative estimate of the model's ability to reproduce historic and future watershed behaviour; (2) to provide a means for evaluating improvements to the modeling approach through adjustment of model parameter values, model structural modifications, the inclusion of additional observational information, and representation of important spatial and temporal characteristics of the watershed; (3) to compare current modeling efforts with previous study results.

The process of assessing the performance of a hydrologic model requires the hydrologist to make subjective and/or objective estimates of the "closeness" of the simulated behaviour of the model to observations (typically of streamflow) made within the watershed. The most fundamental approach to assessing model performance in terms of behaviours is through visual inspection of the simulated and observed hydrographs. In this approach, a hydrologist may formulate subjective assessments of the model behaviour that are generally related to the systematic (e.g., over- or under-prediction) and dynamic (e.g., timing, rising limb, falling limb, and base flow) behaviour of the model. Objective assessment, however, generally requires the use of a mathematical estimate of the error between the simulated and observed hydrologic variable(s) – i.e. objective or efficiency criteria.

Efficiency criteria are defined as mathematical measures of how well a model simulation fits the available observations (Beven, 2001). In general, many efficiency criteria contain a summation of the error term (difference between the simulated and the observed variable at each time step) normalized by a measure of the variability in the observations. To avoid the canceling of errors of opposite sign, the summation of the absolute or squared errors is often used for many efficiency criteria. As a result, an emphasis is placed on larger errors while smaller errors tend to be neglected. Since errors associated with high streamflow values tend to be larger than those associated with errors for lower values, calibration (both manual and automatic) attempts aimed at minimizing these types of criteria often lead to fitting the higher portions of the hydrograph (e.g., peak flows) at the expense of the lower portions (e.g., baseflow). Further, different efficiency criterion may place emphasis on different systematic and/or dynamic behavioural errors making it difficult for a hydrologist to clearly assess model performance.

There have been several studies (e.g. Bastidas et al., 1999; Boyle et al., 2000, 2001; Yapo et al., 1998) aimed at utilizing efficiency measures to more closely estimate the subjective

Correspondence to: P. Krause
(p.krause@uni-jena.de)

process of visually inspecting the hydrograph. In these studies, the observed streamflow time series at the watershed outlet was partitioned, based on the idea that the real watershed system may exhibit modal behaviour – streamflow rapidly rises when there is precipitation (rising limb), quickly decreases after the precipitation ends (falling limb), and slowly decreases long after precipitation ends (baseflow). While these studies demonstrate the advantages of using multiple efficiency measures over a single measure, they do not provide much guidance to the selection of the actual efficiency measure for use with each modal behaviour.

In the next sections of this paper, different efficiency criteria are described and compared through a series of three simple examples involving an observed streamflow hydrograph.

2 Efficiency criteria

In this section, the efficiency criteria used in this study are presented and evaluated. These are the five criteria: coefficient of determination, Nash-Sutcliffe efficiency, Nash-Sutcliffe efficiency with logarithmic values, index of agreement, together with four modified forms that may prove to provide more information on the systematic and dynamic errors present in the model simulation.

2.1 Coefficient of determination r^2

The coefficient of determination r^2 is defined as the squared value of the coefficient of correlation according to Bravais-Pearson. It is calculated as:

$$r^2 = \left(\frac{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)\left(P_i - \bar{P}\right)}{\sqrt{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)^2} \sqrt{\sum_{i=1}^{n} \left(P_i - \bar{P}\right)^2}} \right)^2 \quad (1)$$

with O observed and P predicted values.

R^2 can also be expressed as the squared ratio between the covariance and the multiplied standard deviations of the observed and predicted values. Therefore it estimates the combined dispersion against the single dispersion of the observed and predicted series. The range of r^2 lies between 0 and 1 which describes how much of the observed dispersion is explained by the prediction. A value of zero means no correlation at all whereas a value of 1 means that the dispersion of the prediction is equal to that of the observation. The fact that only the dispersion is quantified is one of the major drawbacks of r^2 if it is considered alone. A model which systematically over- or underpredicts all the time will still result in good r^2 values close to 1.0 even if all predictions were wrong.

If r^2 is used for model validation it therefore is advisable to take into account additional information which can cope with that problem. Such information is provided by the gradient b and the intercept a of the regression on which r^2 is based. For a good agreement the intercept a should be close to zero which means that an observed runoff of zero would also result in a prediction near zero and the gradient b should be close to one. In example 1 the intercept is zero but the gradient is only 0.7 which reflects the underprediction of 30% at all time steps.

For a proper model assessment the gradient b should always be discussed together with r^2. To do this in a more operational way the two parameters can be combined to provide a weighted version (wr^2) of r^2. Such a weighting can be performed by:

$$wr^2 = \begin{cases} |b| \cdot r^2 \text{ for } b \leq 1 \\ |b|^{-1} \cdot r^2 \text{ for } b > 1 \end{cases} \quad (2)$$

By weighting r^2 under- or overpredictions are quantified together with the dynamics which results in a more comprehensive reflection of model results.

2.2 Nash-Sutcliffe efficiency E

The efficiency E proposed by Nash and Sutcliffe (1970) is defined as one minus the sum of the absolute squared differences between the predicted and observed values normalized by the variance of the observed values during the period under investigation. It is calculated as:

$$E = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)^2} \quad (3)$$

The normalization of the variance of the observation series results in relatively higher values of E in catchments with higher dynamics and lower values of E in catchments with lower dynamics. To obtain comparable values of E in a catchment with lower dynamics the prediction has to be better than in a basin with high dynamics. The range of E lies between 1.0 (perfect fit) and $-\infty$. An efficiency of lower than zero indicates that the mean value of the observed time series would have been a better predictor than the model.

The largest disadvantage of the Nash-Sutcliffe efficiency is the fact that the differences between the observed and predicted values are calculated as squared values. As a result larger values in a time series are strongly overestimated whereas lower values are neglected (Legates and McCabe, 1999). For the quantification of runoff predictions this leads to an overestimation of the model performance during peak flows and an underestimation during low flow conditions. Similar to r^2, the Nash-Sutcliffe is not very sensitive to systematic model over- or underprediction especially during low flow periods.

2.3 Index of agreement d

The index of agreement d was proposed by Willmot (1981) to overcome the insensitivity of E and r^2 to differences in the observed and predicted means and variances (Legates and McCabe, 1999). The index of agreement represents the ratio

of the mean square error and the potential error (Willmot, 1984) and is defined as:

$$d = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(|P_i - \bar{O}| + |O_i - \bar{O}| \right)^2} \quad (4)$$

The potential error in the denominator represents the largest value that the squared difference of each pair can attain. With the mean square error in the numerator d is also very sensitive to peak flows and insensitive for low flow conditions as it is E. The range of d is similar to that of r^2 and lies between 0 (no correlation) and 1 (perfect fit).

Practical applications of d show that it has some disadvantages: (1) relatively high values (more than 0.65) of d may be obtained even for poor model fits, leaving only a narrow range for model calibration; and (2) despite Willmot's intention, d is not sensitive to systematic model over- or underprediction.

2.4 Nash-Sutcliffe efficiency with logarithmic values ln E

To reduce the problem of the squared differences and the resulting sensitivity to extreme values the Nash-Sutcliffe efficiency E is often calculated with logarithmic values of O and P. Through the logarithmic transformation of the runoff values the peaks are flattened and the low flows are kept more or less at the same level. As a result the influence of the low flow values is increased in comparison to the flood peaks resulting in an increase in sensitivity of ln E to systematic model over- or underprediction.

2.5 Modified forms of E and d

The logarithmic form of E is widely used to overcome the oversensitivity to extreme values, induced by the mean square error in the Nash-Sutcliffe efficiency and the index of agreement, and to increase the sensitivity for lower values. In addition to this modification, a more general form of the two equations can be used for the same purpose:

$$E_j = 1 - \frac{\sum_{i=1}^{n} |O_i - P_i|^j}{\sum_{i=1}^{n} |O_i - \bar{O}|^j} \text{ with } j \in \mathrm{N} \quad (5)$$

$$d_j = 1 - \frac{\sum_{i=1}^{n} |O_i - P_i|^j}{\sum_{i=1}^{n} \left(|P_i - \bar{O}| + |O_i - \bar{O}| \right)^j} \text{ with } j \in \mathrm{N} \quad (6)$$

In particular, for j=1, the overestimation of the flood peaks is reduced significantly resulting in a better overall evaluation. Based on this result, it can be expected that the modified forms are more sensitive to significant over- or underprediction than the squared forms. In addition, the modified forms with j=1 always produce lower values than the forms with squared parameters. This behaviour can be viewed in two ways: (1) The lower values leave a broader range for model calibration and optimisation, but (2) the lower values might be interpreted as a worse model result when compared to the squared forms.

A further increase in the value of j results in an increase in the sensitivity to high flows and could be used when only the high flows are of interest, e.g. for flood prediction.

2.6 Relative efficiency criteria E_{rel} and d_{rel}

All criteria described above quantify the difference between observation and prediction by the absolute values. As a result, an over- or underprediction of higher values has, in general, a greater influence than those of lower values. To counteract this efficiency measures based on relative deviations can be derived from E and d as:

$$E_{rel} = 1 - \frac{\sum_{i=1}^{n} \left(\frac{O_i - P_i}{O_i} \right)^2}{\sum_{i=1}^{n} \left(\frac{O_i - \bar{O}}{\bar{O}} \right)^2} \quad (7)$$

$$d_{rel} = 1 - \frac{\sum_{i=1}^{n} \left(\frac{O_i - P_i}{O_i} \right)^2}{\sum_{i=1}^{n} \left(\frac{|P_i - \bar{O}| + |O_i - \bar{O}|}{\bar{O}} \right)^2} \quad (8)$$

Through this modification, the differences between the observed and predicted values are quantified as relative deviations which reduce the influence of the absolute differences during high flows significantly. On the other hand the influence of the absolute lower differences during low flow periods are enhanced because they are significant if looked at relatively. As a result, it can be expected that the relative forms are more sensitive on systematic over- or underprediction, in particular during low flow conditions.

3 Methods

An observed streamflow hydrograph from the Wilde Gera catchment in Germany was selected for this study (Fig. 1). The observed values were daily records measured at the outlet of the 13 km^2 large basin in the period of November 1990 to April 1991. A description of the basin and its hydrological dynamics can be found in Krause and Flügel (2005). The hydrograph shows a flood peak at the end of November resulting from a rainfall event and two peaks in January resulting from a mixture of snowmelt and rainfall. The application of a hydrologic model to simulate the observed streamflow hydrograph from other observed hydrologic variables (e.g., precipitation and temperature) was not performed in this study. Rather, three different approaches were used to create synthetic model simulations based on simple modifications to the observed streamflow hydrograph. Each of the three approaches was selected to emphasize specific types of errors

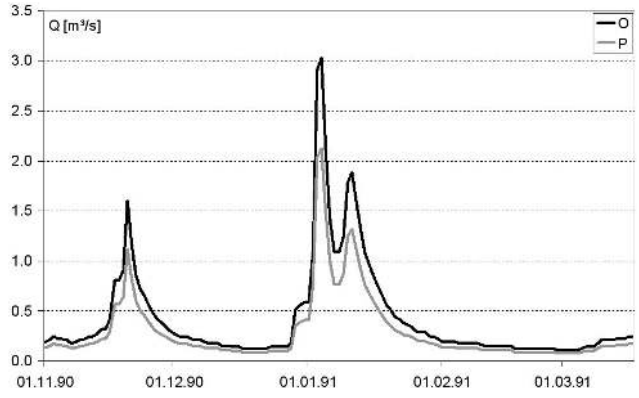

Fig. 1. The systematically underpredicted runoff for the assessment of different efficiency measurements of example 1.

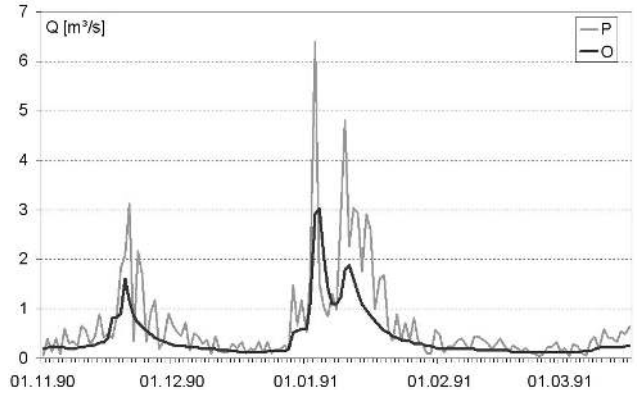

Fig. 2. Example plot of example 2, showing the observed runoff as black and the random prediction as gray line.

frequently encountered in real-world hydrologic model applications and to facilitate the testing and analysis of the selected efficiency criteria. The details of each approach are described in each of the three examples in the remainder of this section.

3.1 Example 1

In the first example, a single synthetic model simulation with a systematic under prediction (poor water balance) but good system response dynamics for the entire observation period was generated by multiplying each ordinate of the observed hydrograph by a factor of 0.7. From Fig. 1 it can be seen that the dynamics of the observed hydrograph are predicted very well while the observed value is never matched by the model simulation (i.e., the model simulation is incorrect at every time step).

3.2 Example 2

In the second example, 10 000 separate synthetic model simulations were generated by multiplying each ordinate of the observed hydrograph with a random value of range 0.1 to 3.0 (each ordinate in a given model simulation can have a maximum underprediction of factor 10 up to maximal overprediction of factor 3.0). An example of one of the model simulations is shown in Fig. 2 and is representative of the generally poor water balance and system response dynamic compared with the observed hydrograph.

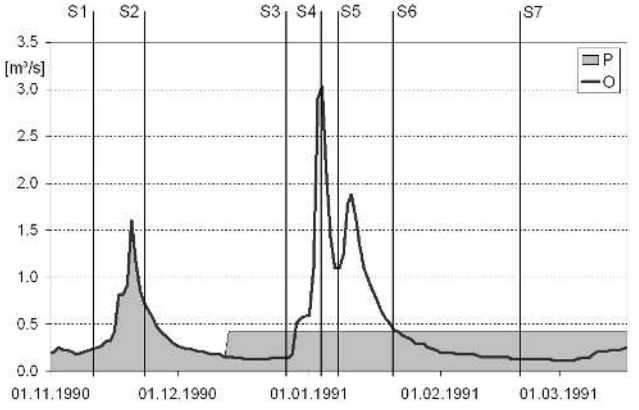

Fig. 3. Observed (black line) and predicted (gray area) hydrograph and the times steps S1 to S7 for example 3.

3.3 Example 3

In the third example, 136 separate synthetic model simulations were generated to simulate a range of possible model predictions with varying degrees of good water balance and poor to good system response dynamics. This was accomplished as follows: for model simulation number 1 each ordinate of the first model simulation was simply the arithmetic mean of the entire observed hydrograph; for the second model simulation, the first ordinate was the same as the first ordinate of the observed hydrograph and the remaining ordinates of the simulation were the arithmetic mean of the entire observed; and in the remaining model simulations (3 to 136), the observed hydrograph values were progressively substituted for the arithmetic mean of the entire observed hydrograph until the last model simulation (number 136) was the actual observed hydrograph. Figure 3 shows the observed and predicted hydrographs for model simulation number 45 (the first 45 time steps are the same as the observed and the remaining values are the arithmetic mean of the entire observed hydrograph).

Also shown in Fig. 3, are different time periods S1 to S7 (vertical lines) that were visually selected to partition different dominate behaviours (e.g., rising limb, falling limb, base flow) in the observed hydrograph. For each model simulation, the different efficiency measures described in Sect. 2 were calculated to examine the behaviour on the specific parts of the hydrograph. With the layout of this example, the behaviour of the different measures on different parts of the hydrograph (peaks, low flows, rising and falling limbs) were examined and quantified.

4 Results

The next sections show the values and results of the different efficient criteria obtained for the three examples.

4.1 Results of example 1

In example 1, the value of the coefficient of determination r^2, is 1.0 while the value of the weighted coefficient, wr^2, is 0.7 reflecting the poor simulation better than r^2 alone. The value of the Nash-Sutcliffe efficiency E is 0.85, indicating that this criterion is not very sensitive to the quantification of systematic underprediction errors. The calculation of the index of agreement resulted in a value of 0.95 also indicating that d is not sensitive to systematic over- or underprediction. The value of the logarithmic efficiency, ln E, in example 1 was 0.81, a little lower than E, r^2, and d but still very high considering that all runoff values were predicted incorrectly. The calculation of the modified form of E and d, with j=1, resulted in values of E_1=0.62 and d_1=0.80. The lower values give an indication that the modified forms seem to be more sensitive to the significant underprediction than the squared forms. The results from the relative forms of E and d (E_{rel}=0.94 and d_{rel}=0.94) demonstrate that this modification is also not sensitive to the systematic underprediction in example 1.

4.2 Results of example 2

The 10 000 model realisations of example 2 resulted in values for r^2 between 0.23 and 0.93. A closer inspection revealed that the gradient b in the best realisation of r^2 was 2.3 with an intercept a of -1.7 – both significantly different from 1.0 and 0, respectively. The results for the weighted coefficient were between 0.13 and 0.67, reflecting the generally poor model results much more accurately. The highest value of wr^2 resulted from an original r^2 of 0.68 and a gradient b of 1.03 and an intercept a of 1.3. The range of values for E in example 2 was calculated between -2.75 and 0.44, reflecting the poor model behaviour very well. A comparison of E with r^2 (Fig. 4, upper plot) shows the interesting fact that the two criteria were only weakly correlated. The best r^2 value (0.93) was found in the realisation with a bad value for E (-1.66). This again demonstrates the limited value of r^2 alone for model performance quantifications. The correlation between E and the weighted coefficient of determination wr^2 was dramatically different (Fig. 4, lower plot) where a much closer and positive correlation was identified. In this case, the realisation with the best value of wr^2 (0.67) was accompanied with the best value for E (0.44).

The range of values for d in example 2 were between 0.65 and 0.89. The narrow range of only 0.24 of d for all of the 10 000 realisations highlights the problems associated with using d – relatively high values which make the criterion insensitive for smaller model enhancements. The comparison of d and E for the 10 000 random samples of example 2 provides an interesting picture (Fig. 5), showing the Nash-Sutcliffe efficiency on the y-axis and the index of agreement on the x-axis. The nearly linear lower border of the point cloud in the plot which marks the realisations with the worst values both for d and E, indicates that the two criteria seem to evaluate the same behaviour but with a considerable amount of scatter above over the whole range. Therefore, the best values of d and E were found in very different realisations.

The values of ln E for the 10 000 random realisations of example 2 were generally lower than those for E, with a range between -0.70 and 0.28. The realisation with the best

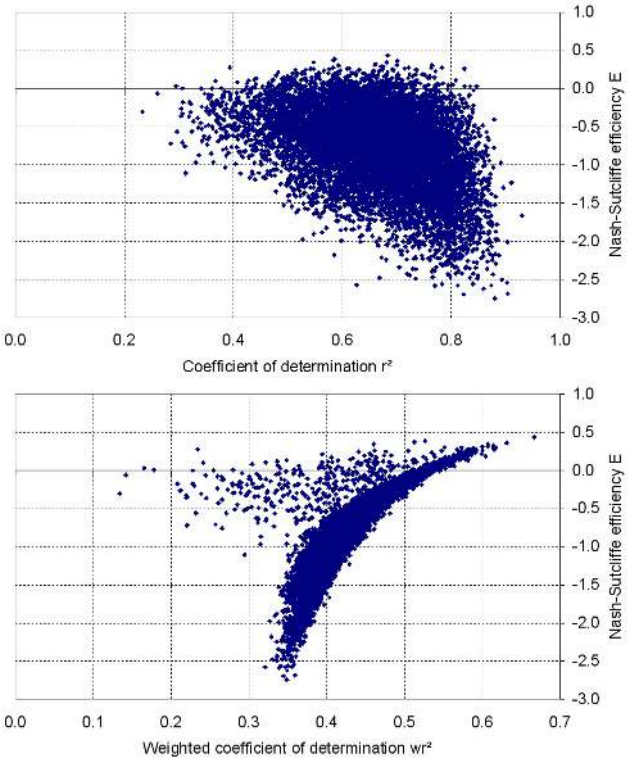

Fig. 4. Nash-Sutcliffe efficiency (y-axis) vs. coefficient of determination (x-axis, upper plot) and the weighted coefficient of determination (x-axis, lower plot) for the 10 000 random samples.

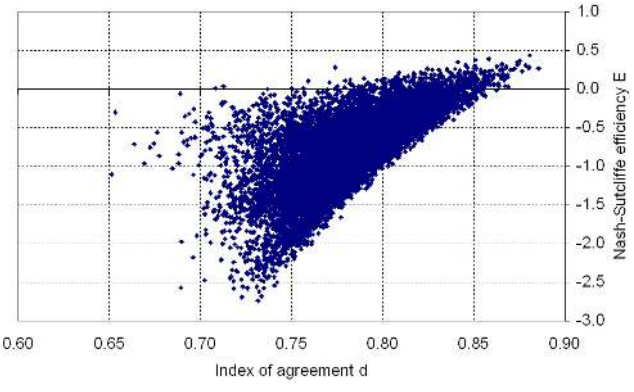

Fig. 5. Nash-Sutcliffe efficiency (y-axis) vs. index of agreement (x-axis) for the 10 000 random samples.

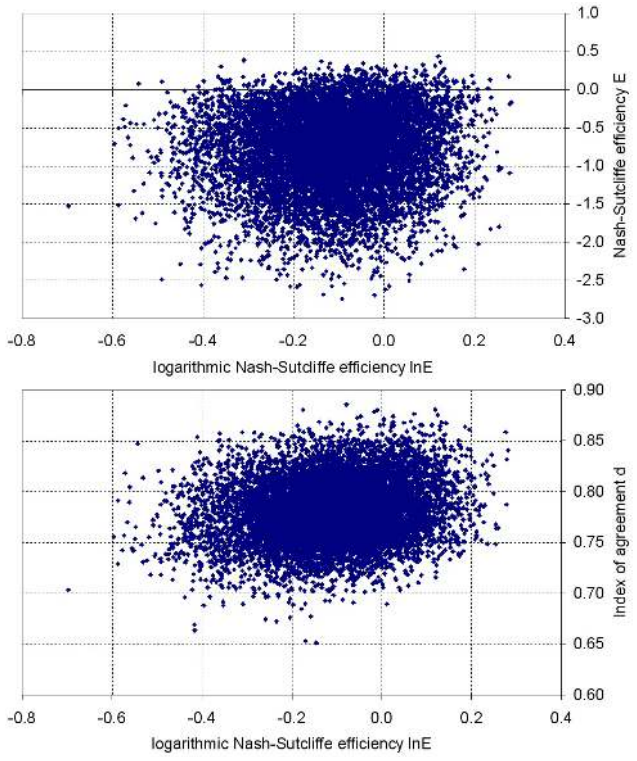

Fig. 6. Comparison of $\ln E$ (x-axis) with E (upper plot) and d (lower plot) on the y-axis for the 10 000 random realisations.

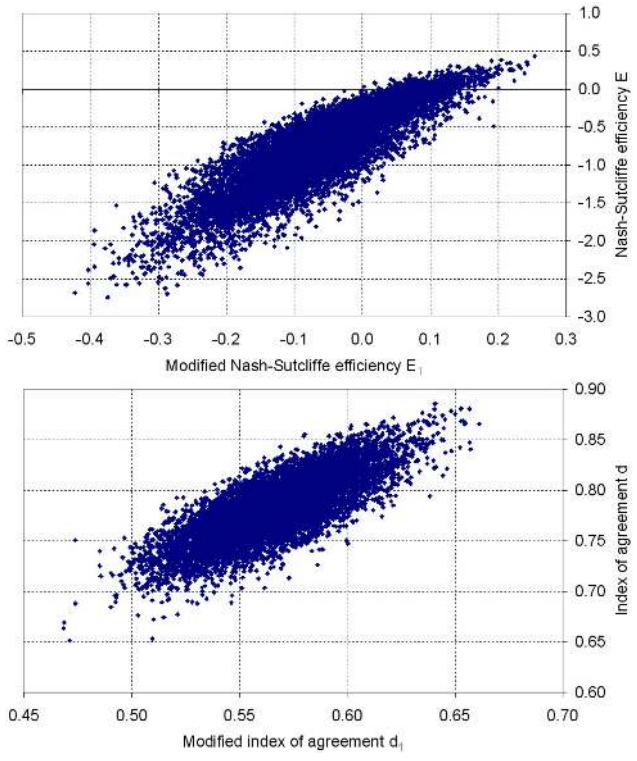

Fig. 7. Comparison of E_1, d_1 (x-axis) with E, d (y-axis) for the 10 000 random samples.

value was not correlated with good values of E (-0.17) or d (0.84). The corresponding value of r^2 was relatively high (0.77) but the gradient b in this realisation was also high (1.5) which leads to a low wr^2 value of 0.51. The fact that the realisations with good values of $\ln E$ were not accompanied by good values for the other criteria is an indicator that $\ln E$ is sensitive to errors of other parts of the predicted and observed time series. The low correlation can also be seen by the graphical comparison of $\ln E$ with E and d for the samples of example 2 in Fig. 6.

The larger scatter and the round outline in Fig. 6 are indicators that the two criteria plotted against each other show different sensitivities (E and d on extreme values and $\ln E$ on the lower ones) on different parts of the hydrograph. By the combined use of two criteria model realisations can be found which produce relatively good results not only for the peak flows but also during low flow conditions. Such a realisation shows values of $\ln E$=0.27, E=0.17, d=0.86, r^2=0.87 and wr^2=0.56.

The range of values for E_1 in example 2 was between -0.42 and 0.25 and between 0.47 and 0.66 for d_1, both measures showing narrower ranges than the squared forms. Figure 7 shows the comparison of the criteria with j=1 on the x-axis and j=2 on the y-axis for the 10 000 random samples.

The comparison shows that the values of E_1 (Fig. 7, upper plot) have a smaller range (0.7) than the range of values (3.2) of E. The highest values of E_1 (0.25) is significantly lower than that of E (0.44). The same is true for the high values of d_1 which are lower than those of d but not for the lower values of d_1 which are smaller than the d values. The reduced high values imply that for comparable high values of the modified form of E and d a better representation of the observed data by the prediction is needed than it is necessary for the squared values. The shape of the point clouds shows that both modified forms have a somewhat linear relationship (with considerable scatter) compared to the squared forms. The realisation with the best value for d_1 (0.66) has relative good values for the other measures (r^2=0.66, wr^2=0.61, E=0.32, d=0.87, E_1=0.24) but a low value for $\ln E$ (0.09). This is even clearer for the realisation with the best value for E_1 (also the same realisation with the best combination of E and wr^2) with the corresponding values for the remaining measures: r^2=0.68, wr^2=0.67, E=0.44, d=0.86, d_1=0.65, $\ln E$=0.12. These results may indicate, that d_1 and E_1 are more integrative efficiency measures that quantify the average behaviour better without being influenced as much by extreme values as the other criteria.

The upper boundary of the range of values for E_{rel} (0.42) in example 2 was very similar to that of E (0.44), however, the lower boundary was lower (E_{rel} -0.19; E -2.75). The comparison of E_{rel} with E is shown in Fig. 8, upper plot.

The larger scatter and the round outline in Fig. 8 are similar to the plots of $\ln E$ (Fig. 6) and are an indicator that the relative criteria are more sensitive to errors during low flow and less sensitive to peak flow errors. Such behaviour can also be seen in the comparison of E_{rel} with $\ln E$ which shows

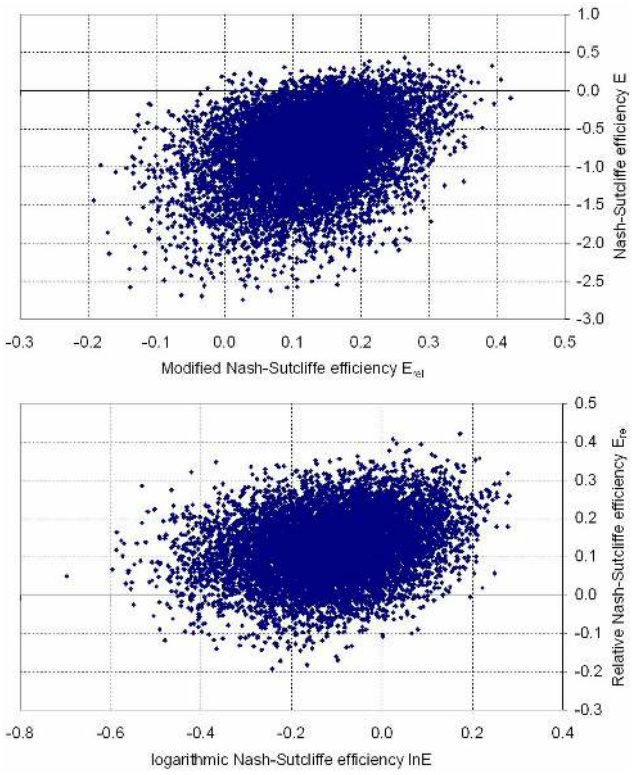

Fig. 8. Comparison of E_{rel} (x-axis) with E (y-axis) (upper plot) and $\ln E$ (x-axis) with E_{rel} (y-axis) (lower plot) for the 10 000 random samples.

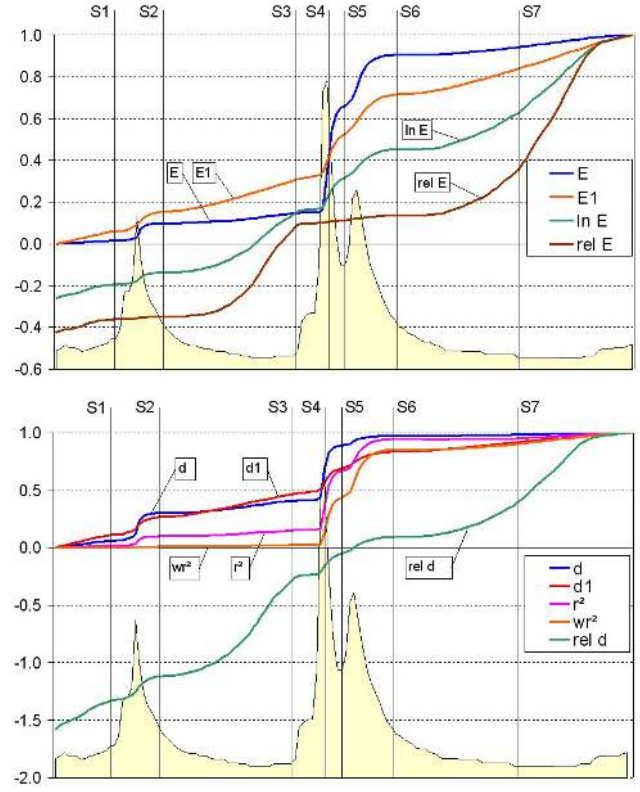

Fig. 9. Evolution of different efficiency measures discussed in Sect. 2 during example 3.

a weak linear relationship (Fig. 8, lower plot) with a considerable amount of scatter. This indicates that the two criteria are evaluating all in all the same model behaviour but seems to be sensitive on different parts of the hydrograph.

4.3 Results of example 3

The values of the efficiency measures described in section 2 for each of the 7 time steps (S1–S7) are shown in Fig. 9 and Table 1. In addition, the table also contains the absolute (absVE in m^3/s) and relative (relVE in %) volume errors calculated during the example.

The upper plot shows the Nash-Sutcliffe efficiency and its modified forms, the lower plot the index of agreement, its modified forms and the coefficient of determination and its weighted form. Table 1 shows the time steps (columns) 0 to 7 and the values (rows) of the different efficiency criteria at these points together with the absolute and relative volume errors.

Step 0 reflects the well known behaviour of E and r^2 that the arithmetic mean as predictor results in a value of zero. This is also true for the index of agreement and the modified versions E_1 and d_1. The relative forms E_{rel}, d_{rel} as well as $\ln E$ have negative values and do not show this well defined lower boundary. The volume errors are, of course, also zero at time step 0 of example 1.

The first time step (S1) ends at day 16 at the beginning of the first rising limb of the hydrograph. Up to this point, only low flow predictions were affected. From Table 1, it can be seen that the criteria react differently over S1: E and r^2 increase only slightly by 0.02, whereas $\ln E$, E_1, E_{rel}, and d increase moderately by 0.06 and d_{rel} by 0.26.

At time step 2 (S2 – 25 days after start) after the first peak has passed, the immediate reaction of d, r^2 and E to the improved prediction for this period is obvious. E_1 and $\ln E$ also show a reaction but with a smoother gradient, whereas E_{rel} remains more or less at the same value. The highest increase of 0.21 can be observed for d_{rel} during this time period.

The adaptation was continued during the succeeding low flow period until the rise of the second peak at time step S3. Criteria E, d and r^2 showed only minor increases of 0.05, 0.11, and 0.08 in the prediction improvement. The modified forms E_1 (0.15) and d_1 (0.20) did exhibit a stronger reaction. The largest increases were in the relative forms E_{rel} (0.44) and d_{rel} (0.85).

The next time step (S4) included only the next 7 days and marks the highest runoff peak of the hydrograph. The line plots of Fig. 9 show a sharp increase of E (0.19), d (0.27) and r^2 (0.19) during time step 4. Moderate increases could be investigated for $\ln E$ (0.06), E_1 (0.09), d_1 (0.09), d_{rel} (0.10) and wr^2 (0.10). Only the criteria E_{rel} exhibited almost no reaction (0.02).

Table 1. Efficiency values and volume errors for the 7 time steps of example 3.

Step	0	1	2	3	4	5	6	7
E	0.00	0.02	0.10	0.15	0.34	0.66	0.91	0.95
$\ln E$	−0.26	−0.20	−0.14	0.16	0.22	0.32	0.45	0.68
E_1	0.00	0.06	0.16	0.31	0.40	0.53	0.72	0.86
E_{rel}	−0.42	−0.36	−0.35	0.09	0.10	0.12	0.14	0.45
d	0.00	0.06	0.30	0.41	0.68	0.89	0.97	0.99
d_1	0.00	0.11	0.27	0.47	0.57	0.69	0.84	0.92
d_{rel}	−1.58	−1.32	−1.11	−0.26	−0.16	−0.05	0.09	0.46
r^2	0.00	0.02	0.10	0.16	0.34	0.67	0.94	0.96
wr^2	0.00	0.00	0.01	0.02	0.12	0.44	0.85	0.91
absVE	0.00	−2.79	1.57	−5.48	−1.98	3.98	12.87	6.60
relVE	0.00	−4.86	2.73	−9.53	−3.45	6.92	22.38	11.48

In time step S5, 4 days after the rising limb of the peak has passed in S4, there were major increases of E (0.32), d (0.21), r^2 (0.32) and wr^2 (0.32) with moderate increases of $\ln E$ (0.10), E_1 (0.13), d_1 (0.12) and d_{rel} (0.11). E_{rel} (0.01) was even less affected than in the antecedent time step.

The next break, time step S6, was made after the third peak has passed and the falling limb reaches the mean runoff value of 0.42 after 14 days. The improvements of the efficiency criteria for this period did show comparable increases to that of the antecedent runoff peak but on a lower level. E (0.24), r^2 (0.28), wr^2 (0.41) again showed a strong reaction and so did E_1 (0.19). Moderate increases were observed for $\ln E$ (0.14), d_1 (0.15) and d_{rel} (0.14). E_{rel} (0.02) and d (0.09) only showed minor increases but for different reasons. E_{rel} was clearly not sensitive to the improvement during this step, whereas the value of d was already very high so that only a minor increase was possible.

Example 3 was continued for 30 days during the following low flow period until time step S7. The different efficiency criteria showed an inverse behaviour compared to the antecedent time steps. Major increases were observed for $\ln E$ (0.23), E_{rel} (0.31) and d_{rel}(0.37), moderate increases for E_1 (0.14) and d_1 (0.09) whereas E (0.04), d (0.01), r^2 (0.02) and wr^2 (0.06) show only minor improvements.

The results of example 3 confirmed the findings from Sect. 2 concerning the different ranges of sensitivity for each efficiency criteria on different parts of the hydrograph. The frequently used Nash-Sutcliffe efficiency E, the coefficient of determination r^2, as well as the index of agreement d, all based on squared deviations of prediction from observation, all exhibited high sensitivity on peak flows and only minor reactions during improvements of the low flow conditions. In the 36 days between time step 1 to 2 and 3 to 6, which is 1/4 of the whole period E was increased by 0.84, d by 0.80, r^2 by 0.86 and wr^2 by 0.84 which implies that only about 20% of the remaining efficiency is induced by correct low flow values.

An opposite behaviour was observed with the relative criteria E_{rel} and d_{rel}. Here the increase during peak flow conditions was 0.06 for E_{rel} and 0.56 of d_{rel}. If the negative starting values are taken into account 22% of the d_{rel} increase was achieved during peak flow and only 4% for E_{rel}. Similar values were calculated for $\ln E$ with 0.35 increase during peak flow which is 27% of the whole efficiency range.

The modified forms E_1 and d_1 exhibited a stronger reaction than the relative forms and $\ln E$ but more moderate than E and d. The increase of E_1 during the peak flow conditions amounted to 0.51 and 0.53 for d_1, meaning that half of the efficiency range was achieved during the peak flows and the other half during the low flow conditions.

5 Discussion and conclusions

Nine different efficiency measures for the evaluation of model performance were investigated with three different examples. In the first example efficiency values were calculated for a systematically underpredicted runoff hydrograph. The systematic error was not reflected by all of the measures – values between 1.0 (r^2) and 0.81 ($\ln E$) were calculated. Only the weighted form wr^2 and the modified form E_1 produced lower values of 0.7 and 0.62 and therefore proved to be more sensitive to the model error in this example. Since most of the criteria investigated are primarily focused on the reproduction of the dynamics compared to the volume of the hydrograph, it is advisable to quantify volume errors with additional measures like absolute and relative volume measures or the mean squared error for a thorough model evaluation.

In the second experiment 10 000 random predictions were created by modifying the values of an observed hydrograph to compare the behaviour of different efficiency measures against each other. It was found that E and r^2 are not very correlated and the realisation with the best value for r^2 exhibited the worst value of E. To improve the sensitivity of r^2, a weighted form wr^2 of r^2 was proposed which takes the deviation of the gradient from 1.0 into account. With wr^2, a good and positive correlation with E was found, stressing the improved applicability of wr^2 over r^2 for model evaluation.

The comparison of the index of agreement d with E revealed that only the very good values for both measures were found in the same model realisations. In the range of lower values an increasing amount of scatter did occur. From the comparisons and the fact that E, r^2, wr^2 and d are based

on squared differences, it is fair to say that these efficiency measures are primarily focused on the peaks and high flows of the hydrograph at the expense of improvements to the low flow predictions.

For a better quantification of the error in fitting low flows, the logarithmic Nash-Sutcliffe efficiency ($\ln E$) was tested. The comparison of $\ln E$ with E and d showed nearly no correlation which is an evidence that $\ln E$ is sensitive to other parts of the model results. With the findings of example 3, it was shown that $\ln E$ reacts less on peak flows and stronger on low flows than E.

To increase the sensitivity of efficiency measures to low flow conditions even more, relative forms of E and d were proposed. The results from the three different examples showed that neither E_{rel} nor d_{rel} were able to reflect the systematic underprediction of example 1. The comparison in example 2 demonstrated that the correlation of E_{rel} and E was similar to the that of $\ln E$ and E. This could be underpinned by the comparison of E_{rel} with $\ln E$ which showed a linear trend but also a considerable amount of scatter. In example 3, the scatter was explained by the fact that E_{rel} did show nearly no reaction on model enhancement during peak flow and therefore was mostly sensitive for better model realisation during low flow conditions.

A more overall sensitivity measure for the quality of the model results during the entire period was found in the two modified forms E_1 and d_1. Both parameters showed linear correlations with E and d, but also with $\ln E$. These findings could be underpinned by the evolution of E_1 and d_1 during example 3 where they showed average values between the extremes of E and d on the one side and $\ln E$, E_{rel} and d_{rel} on the other side.

Overall, it can be stated that none of the efficiency criteria described and tested performed ideally. Each of the criteria has specific pros and cons which have to be taken into account during model calibration and evaluation. The most frequently used Nash-Sutcliffe efficiency and the coefficient of determination are very sensitive to peak flows, at the expense of better performance during low flow conditions. This is also true for the index of agreement because all three measures are based on squared differences between observation and prediction. Additionally it was shown that r^2 alone should not be used for model quantification, because it can produce high values for very bad model results, because it is based on correlation only. To counteract this a weighted form wr^2 was proposed which integrates the gradient b in the evaluation.

The Nash-Sutcliffe efficiency calculated with logarithmic values showed that it is more sensitive to low flows but it still reacts to peak flows. The reaction to peak flows could be suppressed by the derivation of the relative form E_{rel}. E_{rel} proved to be sensitive on low flows only and not reactive on peak flows at all. Based on this behaviour E_{rel} could be used for calibration of model parameters which are responsible for low flow conditions. The use of E or r^2 for such a task often results in the statement that the parameter under consideration is not sensitive.

As more global measures the modified forms of E_1 and d_1 were identified. They stand always in the middle between the squared forms on the one side and the relative forms on the other side. One drawback of these two criteria is that it is more difficult to achieve high values, which makes them less attractive on the first view.

For scientific sound model calibration and validation a combination of different efficiency criteria complemented by the assessment of the absolute or relative volume error is recommended. The selection of the best efficiency measures should reflect the intended use of the model and should concern model quantities which are deemed relevant for the study at hand (Janssen and Heuberger 1995). The goal should be to provide good values for a set of measures, even if they are lower than single best realisations, to include the whole dynamics of the model results.

Edited by: P. Krause, K. Bongartz, and W.-A. Flügel
Reviewed by: anonymous referees

References

Bastidas, L. A., Gupta, H. V., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Sensitivity analysis of a land surface scheme using multicriteria methods, J. Geophys. Res., 104, 19 481–19 490, 1999.

Beven, J. K.: Rainfall-Runoff Modelling – The Primer, John Wiley & Sons Ltd., Chichester, 319, 2001.

Blackie, J. R. and Eeles, C. W. O.: Lumped catchment models, in: Hydrological Forecasting, edited by: Anderson, M. G. and Burt, T. P., John Wiley, Chichester, 311–346, 1995.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., AGU, 36(12), 3663–3674, 2000.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z., and Smith, M.: Towards improved streamflow forecasts: The value of semi-distributed modeling, Water Resour. Res., AGU, 37(11), 2749–2759, 2001

Janssen, P. H. M. and Heuberger, P. S. C: Calibration of process-oriented models, Ecological Modelling, 83, 55–66, 1995.

Krause, P. and Flügel, W.-A.: Integrated research on the hydrological process dynamics from the Wilde Gera catchment in Germany; Headwater Control VI: Hydrology, Ecology and Water Resources in Headwaters, IAHS Conference, Bergen 2005.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 1, 233–241, 1999.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Willmot, C. J.: On the validation of models, Physical Geography, 2, 184–194, 1981.

Willmot, C. J.: On the evaluation of model performance in physical geography, in: Spatial Statistics and Models, edited by: Gaile, G. L. and Willmot, C. J., D. Reidel, Dordrecht, 443–460, 1984.

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