Evaluation of the Nash–Sutcliffe Efficiency Index

Richard H. McCuen[1]; Zachary Knight[2]; and A. Gillian Cutter[3]

Abstract: The Nash–Sutcliffe efficiency index (E_f) is a widely used and potentially reliable statistic for assessing the goodness of fit of hydrologic models; however, a method for estimating the statistical significance of sample values has not been documented. Also, factors that contribute to poor sample values are not well understood. This research focuses on the interpretation of sample values of E_f. Specifically, the objectives were to present an approximation of the sampling distribution of the index; provide a method for conducting hypothesis tests and computing confidence intervals for sample values; and identify the effects of factors that influence sample values of E_f including the sample size, outliers, bias in magnitude, time-offset bias of hydrograph models, and the sampling interval of hydrologic data. Actual hydrologic data and hypothetical analyses were used to show these effects. The analyses show that outliers can significantly influence sample values of E_f. Time-offset bias and bias in magnitude can have an adverse effect on E_f. The time step at which the data are recorded appears to be an insignificant factor unless the sample size is small. The Nash–Sutcliffe index can be a reliable goodness-of-fit statistic if it is properly interpreted.

DOI: 10.1061/(ASCE)1084-0699(2006)11:6(597)

CE Database subject headings: Hydrologic models; Statistics; Hydrographs; Research; Time series; Evaluation.

Introduction

Hydrologic models often require calibration prior to application. Traditionally, the correlation coefficient and standard error of estimate have been used to measure the goodness of fit of the model calibration. While the correlation coefficient is a useful goodness-of-fit index, it is theoretically applicable only to linear models that include an intercept. Even for the commonly used power model, $\hat{y}=ax^b$, the computed correlation coefficient can be a poor estimator of goodness of fit because of model bias. The correlation coefficient assumes that the model being tested is unbiased, i.e., the sum of the errors is equal to zero, and a fitted power model can be significantly biased (McCuen et al. 1990).

Recognizing the limitations of the correlation coefficient, Nash and Sutcliffe (1971) proposed an alternative goodness-of-fit index, which is often referred to as the efficiency index (E_f)

$$E_f = 1 - \frac{\sum^{n} (\hat{Y}_i - Y_i)^2}{\sum (Y_i - \bar{Y})^2} \tag{1}$$

in which $\hat{Y}_i$ and Y_i=predicted and measured values of the criterion (dependent) variable Y, respectively; $\bar{Y}$=mean of the measured values of Y; and n=sample size. If the predictions of a linear model are unbiased, then the efficiency index will lie in the interval from 0 to +1. For biased models, the efficiency index may actually be algebraically negative. For nonlinear models, which most hydrologic models are, negative efficiencies can result even when the model is unbiased.

One advantage of the Nash–Sutcliffe index is that it can be applied to a variety of model types. The ASCE Watershed Management Committee (ASCE 1993) recommends the Nash–Sutcliffe index for evaluation of continuous moisture accounting models. Erpul et al. (2003) used the index to assess nonlinear regression models of sediment transport. Merz and Bloschl (2004) used the index in the calibration and verification of catchment model parameters. Kalin et al. (2003) used the index as a goodness-of-fit indicator for a storm event model. It is also widely used with continuous moisture accounting models (Birikundavyi et al. 2002; Johnson et al. 2003; Downer and Ogden 2004). The use of the index for a wide variety of model types indicates its flexibility as a goodness-of-fit statistic.

While the Nash–Sutcliffe index is widely used as a goodness-of-fit index, values are not easily interpreted because the sampling distribution of E_f has not been presented. For this reason, users of E_f are only able to provide subjective interpretations of their sample values. Many factors influence a sample value of E_f, and high values of E_f may result even when the fit is relatively poor, such as when the variance of Y is very large. Values of E_f also depend on the sample size, such that the interpretation of "good" versus "bad" fit depends on the sample size. A value of 0.7 may or may not be indicative of a good fit. Therefore, if the Nash–Sutcliffe index is to be used with some sense of reliability, more knowledge about sample values of E_f is needed.

The objectives of this study were to present an approximate sampling distribution of E_f and to assess factors that influence computed values of E_f. Methods for computing confidence intervals on E_f and testing hypotheses with sample values are provided. These tools can assist those who use the index to more consistently assess the goodness of fit of hydrologic model predictions.

[1]Professor, Dept. of Civil and Environmental Engineering, Univ. of Maryland, College Park, MD 20742-3021.

[2]Research Assistant, Dept. of Civil and Environmental Engineering, Univ. of Maryland, College Park, MD 20742-3021.

[3]Research Assistant, Dept. of Civil and Environmental Engineering, Univ. of Maryland, College Park, MD 20742-3021.

Note. Discussion open until April 1, 2007. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on March 15, 2005; approved on April 20, 2006. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 11, No. 6, November 1, 2006. ©ASCE, ISSN 1084-0699/2006/6-597–602/$25.00.

Rationale of the Nash–Sutcliffe Index

The Nash–Sutcliffe index uses three quantities: the measured values of the random variable Y_i, the mean of the measured values $\bar{Y}$, and the predicted or modeled values ($\hat{Y}$) of the random variable. These three variables are used to form two sums of squares. First, the denominator of Eq. (1) is the sum of squares of the measured values about the mean of the measured values. In statistical terms, this is a quantity that reflects the total variation of the observed values about the mean. It is essentially that the total variation of the random variable Y potentially can be explained by the model that will be used to predict values of the random variable.

The second term, i.e., the numerator term of Eq. (1), is the sum of the squares of the errors, where an error is the difference between a predicted value and the corresponding measured value. This summation measures the variation in the data that has not been explained by the model used to provide the predicted values. If the numerator were divided by the degrees of freedom and the square root taken, the resulting value would equal the standard error of estimate.

For a linear model, E_f has a direct relationship to the commonly used correlation coefficient, which is more accurately called the Pearson product–moment correlation coefficient. The total variation (TV) can be divided into two parts, the variation explained (EV), and the variation not explained (UV) by the model

$$\mathrm{TV} = \mathrm{EV} + \mathrm{UV} \tag{2}$$

The square of the correlation coefficient (R^2) is the ratio EV/TV. The Nash–Sutcliffe index is 1-UV/TV. Therefore, under the condition of a linear model, the E_f is related to R by

$$E_f = R^2 \tag{3}$$

Hypothesis Tests on the Efficiency Index

Computed values of the efficiency index are sample values. As with any random variable, a sample value of E_f may differ from the true, but usually unknown, value. Therefore, it has an underlying probability distribution. The distribution of the index E_f depends on both the sample size n and the underlying population value (ε_0). As ε_0 increases toward 1, the distribution becomes more skewed, with the long tail on the lower side of ε_0. As the sample size increases, the spread of the distribution decreases. Fig. 1 shows the distribution of E_f for ε_0 equal to 0.5 and 0.7 and for sample sizes of 10, 25, and 50. The characteristics of the spread of the distribution as a function of n and ε_0 are evident in Fig. 1. As n increases, the spread decreases, which indicates that the sample value of E_f is a better estimator of the population value.

While the efficiency index E_f does not have an exact distribution, the distribution can be approximated. The approximate sampling distribution can be used as the basis for hypothesis tests on sample values of E_f. To test whether or not a sample estimate of E_f is likely to have been drawn from a population based on a true value of ε_0, the following null hypothesis is tested:

$$H_0: \varepsilon = \varepsilon_0 \tag{4}$$

An alternative hypothesis H_A can be stated for a one-tailed upper, a one-tailed lower, or a two-tailed test.

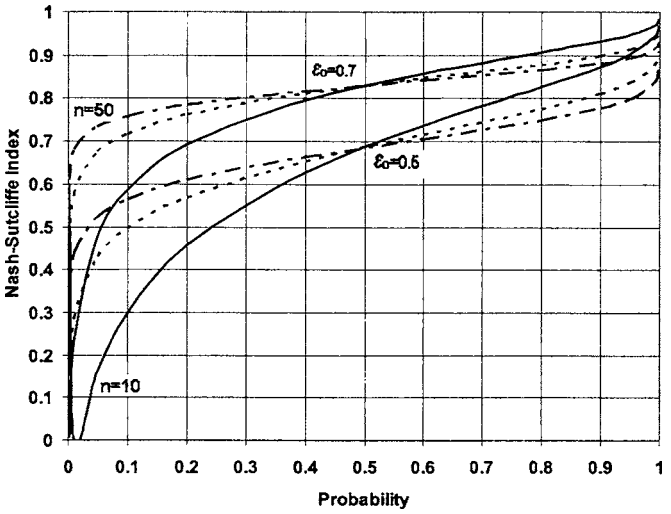

Fig. 1. Probability distributions of the Nash–Sutcliffe efficiency index for population values of ε_0 of 0.7 (top set) and 0.5 (bottom set), for $n=10$ (solid line), $n=25$ (small dash), and $n=50$ (large dash)

The hypothesis of Eq. (4) can be tested using a standard normal transform. For moderate-sized samples the test statistic z has an approximately standard normal distribution

$$z = (\varepsilon - m_\varepsilon)/S_\varepsilon \tag{5}$$

$$\varepsilon = 0.5 \ln_e\left(\frac{1 + E_f^{0.5}}{1 - E_f^{0.5}}\right) \tag{6}$$

$$m_\varepsilon = 0.5 \ln_e\left(\frac{1 + \varepsilon_0^{0.5}}{1 - \varepsilon_0^{0.5}}\right) \tag{7}$$

$$S_\varepsilon = (n-3)^{-0.5} \tag{8}$$

where n=sample size and the value of z computed with Eq. (5) is compared to a critical value obtained from a standard normal distribution table for a level of significance of α or $\alpha/2$.

Confidence Intervals on the Efficiency Index

The distribution of E_f depends on the sample size n and the underlying population value ε_0. Confidence intervals can be based on the approximate sampling distribution developed by Fisher (1928)

$$\left(\frac{e^x - 1}{e^x + 1}\right)^2 \tag{9}$$

in which

$$x = \ln_e\left(\frac{1 + E_f}{1 - E_f}\right) + \frac{2z}{(n-3)^{0.5}} \tag{10}$$

in which z=standard normal deviate. For a one-sided lower $\gamma\%$ (=100−$\alpha\%$) confidence interval, z of Eq. (10) will be negative for $\alpha\%$ of the standard normal distribution in the lower tail. For a one-sided upper $\gamma\%$ confidence interval, z of Eq. (10) will be positive for a $\alpha\%$ in the upper tail. For two-sided confidence intervals, use z values for $\alpha/2\%$ from each tail. The sampling distributions of Fig. 1 indicate that two-sided intervals defined by

Eqs. (9) and (10) will not be symmetric, with the shape depending on both E_f and n.

Interpretation of Sample Efficiency Indices

Despite the frequent application of the Nash–Sutcliffe index to hydrologic models, there exists a degree of disregard of the implications in using the index for assessing model accuracy in cases for which it may not be appropriate. The applications presented herein illustrate the use and misuse of the index in assessing the accuracy of empirical studies that involve commonly used hydrologic models. Two types of hydrologic models will be used to demonstrate important issues in the application of the Nash–Sutcliffe index: (1) mathematical functions such as regression models and (2) single-event hydrograph analyses. The general conclusions outlined for these two types of models will also apply to the assessment of continuous moisture accounting models.

Effect on E_f of Bias in Magnitude

Hydrologic models do not perfectly replicate measured data, with the error variation reflecting the potential prediction accuracy, or inaccuracy, of the model. The error variation in predicted values of a random variable can be due to both systematic and nonsystematic causes. Calibration is performed to reduce the error variation to a minimum, but even calibrated models may be characterized by considerable error variation.

Systematic error variation is referred to as a bias, with a positive bias indicating overprediction. Even calibration does not ensure that a model will be unbiased. For example, power models are often biased when calibrated using logarithms (McCuen et al. 1990). Model bias is estimated using the average error, where an error is the difference between the predicted and measured values. The bias ($\bar{e}$) has the same units as the criterion variable (Y) and is computed by

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i) \tag{11}$$

in which n=sample size; and $\hat{Y}_i$ and Y_i=predicted and measured values of the criterion variable, respectively. A bias is more easily interpreted when it is stated in relative terms (R_b), which is the ratio of the bias to the mean of the measured values of the criterion variable ($\bar{Y}$)

$$R_b = \frac{\bar{e}}{\bar{Y}} \tag{12}$$

R_b is dimensionless and takes the sign of $\bar{e}$. A relative bias greater than 5% in absolute value may be considered significant.

The effect of bias on E_f can be shown by assuming that a biased predicted value $\hat{Y}$ is the sum of an unbiased estimate $\hat{Y}_U$ and the bias $\bar{e}$. Then, it is easily shown that a computed efficiency index E_{fb} for a biased model is

$$E_{fb} = 1 - \frac{\sum (\hat{Y}_U - Y)^2 + n\bar{e}^2}{\sum (Y_i - \bar{Y})^2} \tag{13}$$

Even if the bias is negative, the second term in the numerator, $n\bar{e}^2$, will cause the efficiency to be less than that for an unbiased model. The exact effect of model bias will depend on the magnitude of the two terms in the numerator of Eq. (13). Based on the implications of Eq. (13), it is always important to report the bias and relative bias along with the efficiency index E_f of Eq. (1).

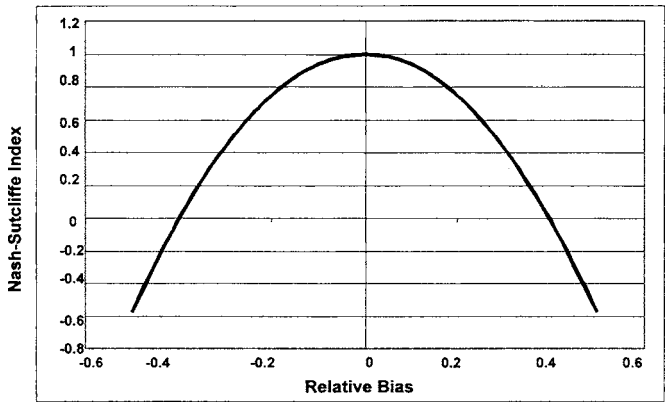

Fig. 2. Effect of bias in the magnitude of a hydrograph on the Nash–Sutcliffe index

To illustrate the potential effect of bias in magnitude on E_f, a gamma distribution was used as the measured or actual hydrograph. To avoid any effects of time-offset error, each predicted hydrograph was computed by multiplying each ordinate of the gamma hydrograph by a constant percentage, from 50 to 150%. The Nash–Sutcliffe efficiency index was computed by comparing the predicted hydrograph with the gamma distribution hydrograph. Fig. 2 shows the relationship between E_f and the relative bias. Fig. 2 indicates that the value of E_f is the same for positive and negative biases, as would be expected from Eq. (13). Fig. 2 also illustrates that E_f is very sensitive to model bias. For a relative bias of 40%, either positive or negative, E_f is zero. Fig. 2 also indicates that bias can cause E_f to become negative. The exact relationship between E_f and bias will vary with the problem. These results of bias in gamma hydrograph models show that the Nash–Sutcliffe index is greatly influenced by model bias.

Effect on E_f of Bias and Outliers in Regression Models

Power models are widely used in hydrologic engineering (e.g., Jennings et al. 1994), and the Nash–Sutcliffe index can be appropriately used as a goodness-of-fit index for such models. Models that relate discharge to drainage area are the most commonly used example of the power model in hydrology. Models that relate water quality parameters to streamflow discharge rates are also common. To illustrate the use of the efficiency index with power models, seven measurements of turbidity T [5.0, 3.1, 2.0, 3.5, 3.9, 0.7, and 20.0 NTU (nethlometric turbidity unit)] were regressed on simultaneous measurements of discharge Q (17, 37, 41, 43, 53, 63, 160) from the Choptank River near Greensboro, Md., with the following model calibrated using a logarithmic transformation:

$$\hat{T} = 0.4557Q^{0.5234} \tag{14}$$

This equation yields the following goodness-of-fit statistics for estimating turbidity: bias=−1.80 NTU; relative bias=−33%; standard error of estimate S_e=6.38 NTU; standard error ratio S_e/S_y=0.973; a correlation coefficient R=0.459; and E_f=0.211. The bias is very significant and indicates that the model will

underpredict turbidities by 33% of the mean. The three goodness-of-fit statistics (S_e/S_y, R, and E_f) indicate that the error variation is very large. Applying the test of significance of Eqs. (4) and (5) with $\in_0=0.8$ yields a z value of −1.790, which corresponds to a rejection probability of 3.7%. Thus, a null hypothesis for a population efficiency of 0.8 would be rejected for the one-tailed lower test at a 5% level of significance. This means that it is unlikely that the data reflect a relationship with an E_f of 0.8.

Since the bias was very large, the intercept was adjusted to produce the following model, which yields unbiased estimates of T:

$$\hat{T} = 0.6804Q^{0.5234} \tag{15}$$

Eq. (15) produces the following goodness-of-fit statistics: bias=0; $S_e/S_y=0.840$, $R=0.642$, and $E_f=0.416$. Therefore, both the bias and overall accuracy have improved. Applying the same hypothesis test used for Eq. (14) to the model of Eq. (15) yields a value for z of −1.250, which corresponds to a rejection probability of 10.6%. Therefore, the null hypothesis that E_f is from a population with ε_0 equal to 0.8 cannot be rejected at the 5% level. A comparison of these two analyses indicates that E_f was significantly influenced by model bias, not just the precision component of accuracy. The accuracy improved when the model bias was eliminated. While the biased model failed the hypothesis test, the unbiased model led to the acceptance of the null hypothesis. This illustrates that sample estimates of E_f are sensitive to model bias.

The Choptank River turbidity data also include one extreme event, i.e., the pair 20 and 160. The turbidity value was tested using both the Dixon–Thompson and Chauvenets outlier tests (McCuen 2003). Both tests indicated that it is an outlier. Therefore, it was censored. Using the remaining six pairs, the following power model was developed:

$$\hat{T} = 106.4Q^{-1.013} \tag{16}$$

which had a bias of −0.178 NTU, a relative bias of −5.9%, a standard error of 1.37 NTU, a standard error ratio of 0.909, a correlation coefficient of −0.582, and $E_f=0.339$. The E_f indicates a poor fit, although the accuracy is better than that from Eq. (14). The detection and censoring of outliers is important, as the efficiency index indicates that model accuracy can be influenced by outliers. Also, the E_f of 0.339 fails to indicate that the relationship is negative, which is shown by the negative sign on the correlation coefficient and the exponent of Eq. (16).

This example illustrates that both model bias and outliers can affect sample values of E_f. The elimination of both the bias and the outlier increased the sample values of E_f. Of course, every case will be different, and the effect of unbiasing and outliers will vary case by case. However, when poor values of E_f occur, the data should be checked for both bias and outliers.

Efficiency Assessment of Unit Hydrograph Analyses

Unit hydrograph modeling is important because of the widespread use of unit hydrographs for hydrologic design. The Natural Resource Conservation Service and Snyder's unit hydrographs are widely used as part of the computer software TR-20 and Hydrologic Engineering Center models. Unit hydrograph modeling differs from regression-type models in that time is a factor and adjacent ordinates can be correlated. Unit hydrographs are developed from actual storm event data, and the shape and time scale can be distorted because of a lack of synchronization between the measured rainfall and runoff records. Temporal disharmony between the time distributions of rainfall and runoff can reduce the fitting efficiency.

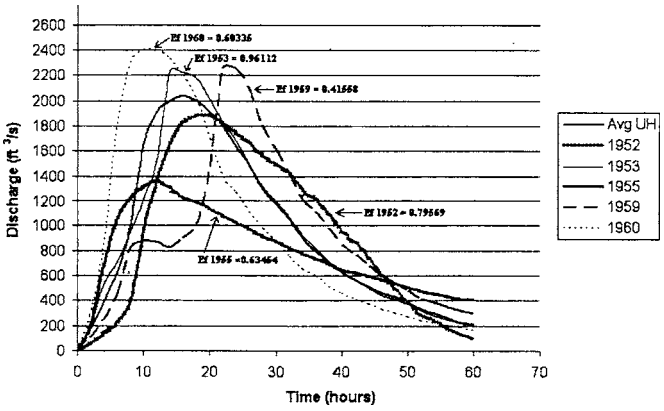

Fig. 3. Assessment of the Nash–Sutcliffe efficiency index using a storm-event unit hydrograph for the White Oak Watershed

An analysis of actual data was conducted using unit hydrographs from the White Oak Bayou, Texas. Unit hydrographs (UHs) were developed (Hare 1970) for five storm events. The five UHs were then used to develop an average unit hydrograph for the watershed (McCuen 2005). Then, the storm event unit hydrographs were individually compared with the average UH and values of E_f computed. The six UHs are shown in Fig. 3. The five values of E_f vary from 0.42 to 0.97. Reasons for the individual values differ from event to event. The unit hydrograph for the 1955 event produced a relatively low value of E_f (0.63) because of the bias in magnitude, with the peak of the actual UH being about 65% of the peak of the average UH. The E_f for the 1960 UH ($E_f=0.60$) was about the same as that for the 1955 event but the two UHs were quite different. The 1960 UH was more peaked than the average UH but also suffered because it was offset in time, with the peak occurring 6 h earlier than the peak of the average UH. This represents a time-offset bias of about 40% of the time to peak. The 1953 UH had the largest E_f (0.96) as it was not offset in time and differed in the magnitude of the peak by only 10%. The 1959 unit hydrograph showed a very significant time offset in the peak, as well as a flat area on the rising limb, which produced a very poor index of 0.42. In these examples, both magnitude bias and time-offset bias contributed to poor accuracy. The Nash–Sutcliffe index was able to detect the effects of these factors, but a value of E_f cannot identify which factor is the principal problem. Therefore, other analyses are necessary such as graphical assessments and computations of the magnitude and time-offset biases.

Effect of Time Interval Size in Hydrograph Analyses

Single-event hydrograph models are widely used in hydrologic design. The unit hydrograph is the basis for many of these models. The time interval of the hydrologic data that is used to calibrate a unit hydrograph varies with the analysis. However, the time interval is inversely related to the sample size. As the time interval is decreased, the sample size increases. Statistical theory suggests that accuracy improves with increasing sample size, which is supported by the sampling distribution of E_f [Eq. (8)]. Therefore, the potential effect of the time interval on values of E_f needed to be assessed.

Table 1. Variation of Nash–Sutcliffe Efficiency Index with the Gamma Distribution Shape Parameter (C) and the Time Offset as a Percentage of the Time to the Peak of the Hydrograph

C	4	8	12	16	20	24	28	32	36	40
3.00	0.995	0.982	0.960	0.929	0.892	0.848	0.800	0.747	0.692	0.635
3.50	0.995	0.981	0.958	0.926	0.886	0.839	0.785	0.727	0.665	0.600
4.00	0.995	0.980	0.955	0.921	0.878	0.827	0.769	0.705	0.637	0.565
4.50	0.995	0.979	0.952	0.916	0.870	0.815	0.752	0.683	0.608	0.530
4.70	0.994	0.978	0.951	0.913	0.866	0.810	0.745	0.674	0.597	0.516
5.00	0.994	0.977	0.949	0.910	0.860	0.802	0.735	0.660	0.580	0.495

While rainfall hyetographs and runoff hydrographs may be recorded on one time scale, analyses of the data may be carried out on a different time scale. Thus, if the magnitude of the time interval is changed, the efficiency index could change as well. A gamma distribution with a shape parameter of 4.75 exactly matches the Soil Conservation Service dimensionless unit hydrograph and was used as the true hydrograph. Keeping the scale parameter constant, the shape parameter was varied, which yielded a "predicted" hydrograph that could be compared with the true hydrograph. Twenty-four ordinates were defined on the rising limb and the time base was set at 120 time increments. The predicted hydrograph was based on the same scale parameter but the shape parameter was varied. The efficiency index was computed for time increments of 1, 2, 3, 4, 6, 8, 10, 12, and 15, which yields sample sizes (i.e., number of ordinates) of 120, 60, 40, 30, 24, 20, 15, 12, 10, and 8, respectively. For a predicted hydrograph based on a shape parameter of 4, the efficiency was high and did not vary very much as long as 12 or more ordinates were used. For the very small sample sizes of 8 and 10, the E_f decreased. However, the hypothesis test based on Eqs. (4)–(8) showed greater variation. When the predicted hydrograph was based on a shape parameter of 2.4, the E_f values varied as the sample size was changed but all of the values were poor. Additionally, the z statistic of Eq. (5) showed little variation, even as the sample size decreased. These analyses indicate that the Nash–Sutcliffe index is not very sensitive to the time interval as long as the sample size is moderate. However, the statistical significance of a sample E_f can be influenced by the sampling interval in hydrograph analyses.

The White Oak Bayou data base of Fig. 3 was also used to examine the effect of the time interval. Data were recorded on a 1 h interval, which gives 60 values. By interpolating between the measured points, the time interval used to define values of the unit hydrographs was cut in half, which doubled the number of discharge measurements. For the five unit hydrographs, the values of E_f changed very little, on average by only 0.5%. These changes in the computed E_f indices is insignificant because it is much less than the sampling variation. This shows that unless the change in interval is going to be significant, the sampling interval has a minor effect on values of the Nash–Sutcliffe index. In conclusion, the time interval should be kept as small as possible for the most accurate index in cases where a bad fit is suspected; otherwise, it is assumed that the values of the index are not sensitive to the time interval. However, if a hypothesis test is to be made with the data, a test of significance can be sensitive to the sampling interval when the sample size is small.

Effect on E_f of Time Offset Bias

Time-dependent models based on measured data may be subject to time-offset errors if the rainfall and runoff are not synchronized on the time scale. This could occur if the rain gauge was not located within the watershed and the rainfall hyetograph was offset from the runoff hydrograph by an amount equal to the travel time of the rainfall between the watershed and the rain gauge. In Fig. 3, time-offset errors are evident in both the 1959 and 1960 events. More-detailed hydrologic models, such as continuous moisture accounting models, such as the hydrological simulation program FORTRAN model, include storage components and parameters that control the release of water from the storages. The storage components of these models might also contribute to time-offset errors in continuous moisture accounting. If the model components and parameters are not properly calibrated, predicted discharges or pollution loads may be offset in time from the measured values. A time offset or an inaccurate modeling of the recession of flows from the storages can significantly affect the goodness of fit.

Time-offset errors will increase the numerator of the Nash–Sutcliffe index [Eq. (1)]. Specifically, the error variation term $\Sigma(\hat{Y}-Y)^2$ will be inflated by time-offset errors. In a sense, the time-offset errors are biases on the time scale, in contrast to a bias in the magnitude, as discussed previously. Even if computed hydrographs take the general shape of the measured hydrographs, but are offset in time, the error variation term of Eq. (1) can be large, which will produce low values of E_f.

To assess the effect of time-offset error, a gamma distribution hydrograph with a shape parameter of 4.7 was translated on the time axis to reflect a model that has not been properly calibrated to fit in the time domain but reproduces the magnitudes of the measured hydrograph. To provide dimensionless indicators, the time offset was scaled as a fraction of the time to peak. The effects of a time-offset bias on the Nash–Sutcliffe index are evident in Table 1.

Table 1 shows that as the offset interval increases the index and, therefore, the goodness of fit decreases. The smaller the time interval, the less significant an offset is on the E_f, as is shown by the higher values for small percentage changes. For a large interval, i.e., large changes, the decrease in the index can be dramatic, decreasing to about 0.81 for a 24% time offset for the hydrograph with a shape parameter of 4.7. These results show the importance of choosing an appropriate time interval and to know that a time offset can significantly affect the goodness of fit of hydrograph models.

The sensitivity of the efficiency index to time-offset bias can also be illustrated using two of the White Oak unit hydrographs of Fig. 3. The storm events of 1959 and 1960 produced unit hydrographs that were sensitive to rainfall characteristics. The rainfall for the 1960 event was of short duration and high intensity, which produced a relatively peaked UH. The 1959 event began with a period of low-intensity rainfall followed by a short period of intense rainfall, which produced a unit hydrograph with a relatively flat rising limb followed by a peak that exceeded the peak of the

average unit hydrograph. The computed Nash–Sutcliffe efficiencies were 0.60 and 0.42 for the 1960 and 1959 unit hydrographs, respectively. These values of E_f may suggest relatively low accuracy. Fig. 3 also suggests poor agreement with the watershed average unit hydrograph.

The two unit hydrographs were individually lagged both forward and backward in time and compared to the average UH using the efficiency index. A compilation of E_f versus the lag, which will be referred to as an efficiogram, is similar to a cross correlogram used in time series modeling. Using a 2 h lag for each efficiogram, the following E_f values resulted for lags from −10 to +6 h for the 1959 event: 0.81, 0.82, 0.75, 0.65, 0.52, 0.42, 0.05, −0.31, −0.48. Thus, if the unit hydrograph is lagged by 8 h, E_f would increase from 0.42 to 0.82. This indicates that a time-offset bias is very significant in the 1959 unit hydrograph. For the 1960 UH, the efficiogram for a 2 h lag is: −1.40, −1.17, −0.89, −0.44, 0.12, 0.60, 0.82, 0.91, 0.80. This shows that a time offset of 4 h, i.e., two time lags, would increase the efficiency from 0.60 to 0.91. In both cases, the time-offset bias in the actual unit hydrographs caused a significant loss of accuracy. In the assessment of the accuracy of a time-dependent model, the efficiogram should be computed and examined for a time-offset bias. If such a bias is evident, it may be useful to either revise the model or adjust the data to account for the underlying cause of the time-offset bias. Time-offset biases should be investigated before the data analysis, but they are often difficult to detect, such as where they are caused by highly variable storm cells in an area with a low density of rain gauges. However, the efficiogram can be a useful aide in detecting the errors.

Conclusions

Sample values of the Nash–Sutcliffe index are values of a random variable and subject to sampling variations, as is any random variable. A method that approximates the sampling distribution of E_f was presented and shown how it can be used both to test hypotheses about the underlying population value of E_f and to compute confidence intervals for sampling values. These statistical tools will enable users of the Nash–Sutcliffe index to systemically assess values of E_f, thus avoiding subjective assessments of goodness of fit.

Hydrologic models are intended to reflect the physical processes that the model is designed to represent. Less than ideal values of goodness of fit are not necessarily indicative of a poor model. Rather, they may be the result of the misuse of the goodness-of-fit index. Therefore, having reliable goodness-of-fit criteria is an important element of the modeling process. A goodness-of-fit statistic is supposed to reflect some aspect of the prediction accuracy of the calibrated model. Selection of a statistic is, therefore, important to ensure that it will reflect the characteristic that it is intended to reflect.

The structure of the Nash–Sutcliffe efficiency index E_f is very similar to the Pearson product—moment correlation coefficient. Low values of E_f may be the result of model bias produced by the calibration, with bias resulting either from differences in magnitude or time offset for time-dependent models. Thus, the model bias should always be recorded when the efficiency index is applied.

It was not the intent of these analyses to suggest that the Nash–Sutcliffe index is a poor goodness-of-fit statistic. Actually, it can be a useful index. However, it is a single-valued index that can be sensitive to a number of factors, including sample size, outliers, magnitude bias, and time-offset bias. Failure to recognize the limitations of E_f may lead to rejection of a good model solely because E_f was misapplied, such as to a biased model. To avoid misusing the efficiency index of Eq. (1), the computation of values should be accompanied by other analyses, such as the computation of biases and efficiograms. Also, the effect of outliers should be assessed.

References

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management, Irrigation, and Drainage Division (ASCE). (1993). "Criteria for evaluation of watershed models." *J. Irrig. Drain. Eng.*, 119(3), 429–442.

Birikundavyi, S., Labib, R., Trung, H. T., and Rousselle, J. (2002). "Performance of neural networks in daily streamflow forecasting." *J. Hydrol. Eng.*, 7(5), 392–398.

Downer, C. W., and Ogden, F. L. (2004). "GSSHA: Model to simulate diverse stream flow producing processes." *J. Hydrol. Eng.*, 9(3), 161–174.

Erpul, G., Norton, L. D., and Gabriels, D. (2003). "Sediment transport from interrill areas under wind-driven rain." *J. Hydrol.*, 276, 184–197.

Fisher, R. A. (1928). "The general sampling distribution of the multiple correlation coefficient." *Proc. R. Soc. London*, 121, 654–673.

Hare, G. S. (1970). "Effects of urban development on storm runoff rates." *Proc., Seminar on Urban Hydrology, Paper No. 2, HEC, Corps of Engineers*. Davis, Calif.

Jennings, M. E., Thomas, W. O., Jr., and Riggs, H. C. (1994). "Nationwide summary of U.S. Geological Survey regional regression equations for estimating magnitude and frequency of floods for ungaged sites, 1993." *USGS WRI 94-4002*, U.S. Geological Survey, Reston, Va.

Johnson, M. S., Coon, W., Mehta, V., Steenhuis, T., Brooke, E., and Boll, J. (2003). "Applications of two hydrologic models with different runoff mechanisms to a hillslope dominated watershed in the northeastern U.S.: A comparison of HSPF and SMR." *J. Hydrol.*, 284, 57–76.

Kalin, L., Govindaraju, R. S., and Hantush, M. M. (2003). "Effect of geomorphological resolution on modeling of runoff hydrograph and sedimentograph over small watersheds." *J. Hydrol.*, 276, 89–111.

McCuen, R. H. (2003). *Modeling hydrologic change*, CRC, Boca Raton, Fla.

McCuen, R. H. (2005). *Hydrologic analysis and design*, Pearson/Prentice-Hall, Upper Saddle River, N.J.

McCuen, R. H., Leahy, R. B., and Johnson, P. A. (1990). "Problems with logarithmic transformations in regression." *J. Hydraul. Eng.*, 116(3), 414–428.

Merz, R., and Bloschl, G. (2004). "Regionalization of catchment model parameters." *J. Hydrol.*, 287, 95–123.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models. Part 1: A discussion of principles." *J. Hydrol.*, 10(3), 282–290.