Fitting of Hydrologic Models: A Close Look at the Nash–Sutcliffe Index

Sharad K. Jain[1] and K. P. Sudheer[2]

Abstract: Quantitative assessments of the degree to which the modeled behavior of a system matches with the observations provide an evaluation of the model's predictive abilities. In this context, the Nash–Sutcliffe efficiency index is widely used in water resources sector to assess the performance of a hydrologic model. Through a series of results, this technical note demonstrates that this index alone is not adequate in describing the performance of a model. It is shown that relatively poor models can give a high value of the index and vice-versa. Thus, it is advisable to employ the other statistical measures before arriving at a definite conclusion about the performance of a hydrologic model.

DOI: 10.1061/(ASCE)1084-0699(2008)13:10(981)

CE Database subject headings: Hydrologic models; Statistics; Simulation models; Water resources.

Introduction

Simulation models are increasingly being used in problem solving and decision making in water resources engineering. The developers and users of these models, the decision makers using the results of the models, and people affected by the decisions based on such models would like to be sure that the model results are "correct." This concern is generally addressed through model validation, which is usually defined to mean "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Schlesinger et al. 1979). Model validation is an important step in model application.

In the field of hydrology, as knowledge of the physical processes has improved, models have become more complex (Legates and McCabe 1999). A model may have numerous parameters that are calibrated through optimization which minimizes the deviations between simulated data and the observed data (Nash and Sutcliffe 1970; Song and James 1991; Hay 1998). The initial estimates of parameters are adjusted during model calibration so that the modeled response is close to the observed response. Many times, the parameters of the hydrologic models are not measurable in the field or there may be a dearth of field measurements. In such cases, initial parameter values are assigned based on relevant measurable catchment properties (e.g., soil properties, vegetation characteristics, etc.) or by experience.

Basically, two approaches are employed to ascertain the goodness of fit of mathematical models: visual judgment and statistical measures. Although the visual inspection of the observed and simulated response of the system provides a useful insight into the model behavior and adequacy, it is subjective. Further, when the performance of different models or of different parameter sets of the same model is nearly the same, ranking the alternatives using visual judgment is difficult. Statistical measures, which are objective in nature, can be employed in automatic calibration of models and in model ranking but statistical measures for which sampling distributions are unknown can be misleading if the interpretation is subjective rather than objective.

Many principal measures that are used by the hydrologists to validate models were reviewed by Legates and McCabe (1999). Among the statistical measures that are used in hydrology, the commonly adopted criteria are the Nash–Sutcliffe (NS) efficiency index, the root mean square error, the coefficient of correlation, the coefficient of determination, and the mean absolute error. The mean absolute error is not a good statistic because it does not indicate direction of the error. In hydrologic analysis, a 5% overprediction will likely have different implications from a 5% underprediction. This is the reason bias—the deviation from observed value—is frequently used. Out of these indices, the NS efficiency index is one of the most commonly used measures. However, the use of this index suffers from a major drawback. At times, a relatively high value of the index can be obtained with a poor model and a model with a high value of the index may not be extracting all the information from the input data. As a result, the user of the model may mistakenly believe that the model is performing very well. Obviously, any decision based on poor modeling will be suboptimal.

The focus of this technical note is to critically look at the NS efficiency index for performance evaluation of models and to illustrate the limitations of this statistic. To illustrate the discussion points, four cases are considered in this study—three dealing

[1]Scientist "F," National Institute of Hydrology, Roorkee – 247667, India. E-mail: s_k_jain@yahoo.com

[2]Assistant Professor, Dept. of Civil Engineering, Indian Institute of Technology Madras, Chennai – 600036, India. E-mail: sudheer@iitm.ac.in

Note. Discussion open until March 1, 2009. Separate discussions must be submitted for individual papers. The manuscript for this technical note was submitted for review and possible publication on March 14, 2007; approved on December 7, 2007. This technical note is part of the ***Journal of Hydrologic Engineering***, Vol. 13, No. 10, October 1, 2008. ©ASCE, ISSN 1084-0699/2008/10-981–986/$25.00.

Table 1. Summary of Modeling Results

Feature	Establishment of rating curves for the site			Rainfall-runoff modeling
	Chebes	Satrana	Thebes	
Number of values of y (sample size)	241	497	273	36
Mean of y (m^3/s)	7,678.5	19.6	6,930	82.15
SD of y (m^3/s)	2,570	27.47	3,901	91.64
Coefficient of variation (Cov) of y	0.335	1.4	0.563	1.116
Bias	87.5	−1.448	−312.3	−1.025
Relative bias	0.011	−0.073	−0.0451	−0.0125
Nash–Sutcliffe efficiency (η)	0.9821	0.9115	0.8568	0.6443

with fitting rating curves and one dealing with rainfall-runoff modeling.

Background

Typically, parameters of a model (conceptual or empirical) of hydrologic systems are estimated by minimization of sum of squares of errors

$$\text{Min } z = \min \sum_{t=1}^{n} [q_{t.\text{obs}} - q_{t.\text{sim}}(x_t;A)]^2 \quad (1)$$

where $q_{t.\text{obs}}$ and $q_{t.\text{sim}}$=observed and simulated system responses at time t, respectively. The difference of these two is the model residual error ε_t. x_t=vector of inputs (such as rainfall, snow, and exogenous variables such as evaporation, etc.) and A=parameter vector about which the inference is sought. The use of Eq. (1) as an objective function to be minimized for parameter estimation implies certain assumptions about the residuals ε_t (Clarke 1973; Xu 2001):

1. ε_t have zero mean and constant variance σ_ε^2 [i.e. $E(\varepsilon_t)=0$, $E(\varepsilon_{t\varepsilon}^2)=\sigma_\varepsilon^2$]; and
2. ε_t are mutually uncorrelated [i.e. $E(\varepsilon_t, \varepsilon_{t-k})=0 \forall k \neq 0$].

The above-presented assumptions need to be tested for the residuals of the model. Further, a sequence of random variables is homoscedastic if the variables have the same finite variances. When using the least squares technique, one of the assumptions is that the error term has a constant variance. Many studies in hydrology do not check whether these assumptions are satisfied and commonly a visual inspection of predicted and measured hydrographs together with a "goodness of fit" statistic is presented.

The coefficient of efficiency (η), proposed by Nash and Sutcliffe (1970) or the NS index, is one of the widely employed statistics in hydrologic literature. Let y denote the output variable which in most hydrological application is flow. NS index η is defined as

$$\eta = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (2)$$

where $\hat{y}_i$ and y_i=observed and computed flow values at time t, respectively, and $\bar{y}_i$=mean of the observed and computed flow values corresponding to n patterns. The denominator of Eq. (2) is the total variance of the observed values about the mean and the numerator is the variance of the data that has not been explained by the model. A value of 1.0 of η represents a perfect match and a model with η of 0.0 is no more accurate than predicting the mean value. Under the condition of a linear model, η is equal to the square of the coefficient of determination.

Ever since the NS index was advanced, it has been used in numerous studies in hydrology. The popularity of NS index arises from the fact that it is a useful single-value index. However, one needs to exercise caution when using the Nash–Sutcliffe index η. When the variance of y is very large, high values of η may be obtained even when the fit is relatively poor. Some investigations have attempted to look at the advantages and shortcomings of this index. Garrick et al. (1978) stated two criticisms of NS index: (1) the alternative or "no model" forecast is unnecessarily primitive; and (2) no distinction is made between different kinds of error. As regards the interpretation of η, ASCE task committee (1993) noted that when the measured values of y approach the average value, the denominator of Eq. (2) approaches zero and η approaches $-\infty$ with only minor model mispredictions.

Recently, McCuen et al. (2006) pointed out that the sample values of NS index are the values of a random variable and are subject to sampling variations. Further, low values of η may be due to model bias produced by calibration. McCuen et al. (2006) emphasized that when interpreting the values of NS index, sample size, bias, effects of outliers, etc., must be kept in mind. ASCE (1993) also noted a similar observation to that of McCuen et al. (2006) that this statistic works best when the coefficient of variation of the observed data set is large.

Details of Case Examples and Discussions

As stated earlier, the first three cases that are explored in this technical note pertain to the development of rating curves of the following form:

$$Q = aH^b \quad (3)$$

where Q=discharge; H=river stage; and a and b=parameters of regression. This relation was established by log transforming the data and fitting a regression line. The data pertaining to three gauge-discharge stations were used for this analysis. The summary statistics, along with the commonly employed performance evaluation measures are presented in Table 1. Note that no correction for bias (McCuen et al. 2006) was made in any of the cases as relative bias is quite small in the cases. If bias is considered, η will be slightly less than the value reported in Table 1.

In the first case, where a rating curve was established for a station named Chebes, the sample size was 241 (Table 1). In this case, the variance of the data was quite small as is evident from a small value of Cov. A very small positive relative bias was

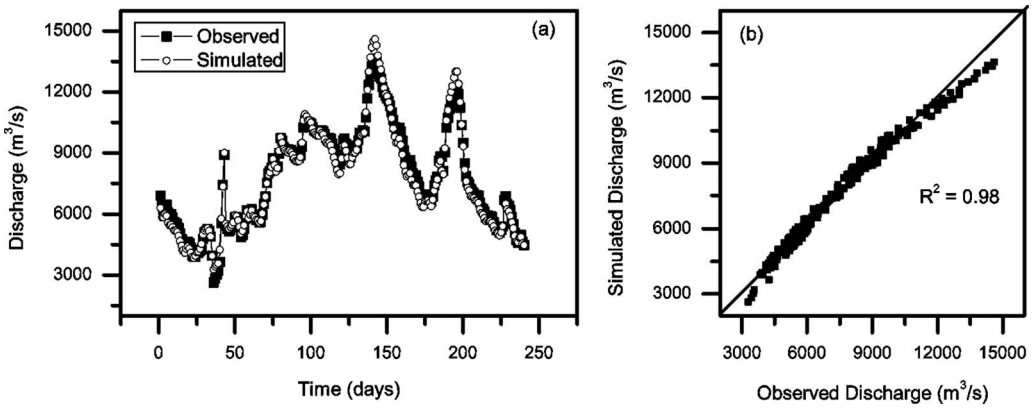

Fig. 1. Computed discharge along with its observed counterpart for Chebes (a) time series; (b) scatter diagram

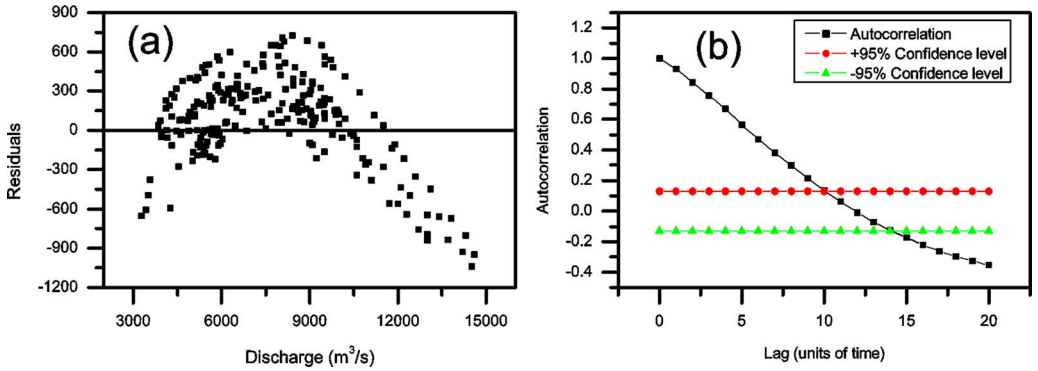

Fig. 2. Plots of residuals for Chebes to illustrate (a) homoscedasticity; (b) independence

present in the modeling results and the NS efficiency index η was 0.9821. Such a high value of the NS index suggests that the model has been able to explain the variance in the data to a very high degree.

The time series plot of the observed and computed discharge is shown in Fig. 1(a) and the same are presented in a scatter diagram in Fig. 1(b). It appears from Fig. 1 that the match between the observed and computed discharge is very good but the scatter plot shows that the simulated discharge is consistently higher than the observed. Further, a close examination of the residuals and their autocorrelation structure presented in Fig. 2 reveals that the developed model violates the fundamental assumptions (stated earlier) that are involved in parameter estimation using the least squares error method. The residuals have significant correlation at higher lags, which indicates that the model was not able to completely extract the information contained in the data. Also, it is to be noted that the residuals are highly biased toward the magnitude of flows. These are positive in low and medium flow ranges and are negative in the high flow ranges. Thus a model that appears to be close to a perfect fit, based on NS index, fares rather poorly when the behavior of residuals is examined. Hence, a conclusion on model performance based on the NS efficiency index alone may be misleading.

In the second case, a rating curve was established for the Sa-

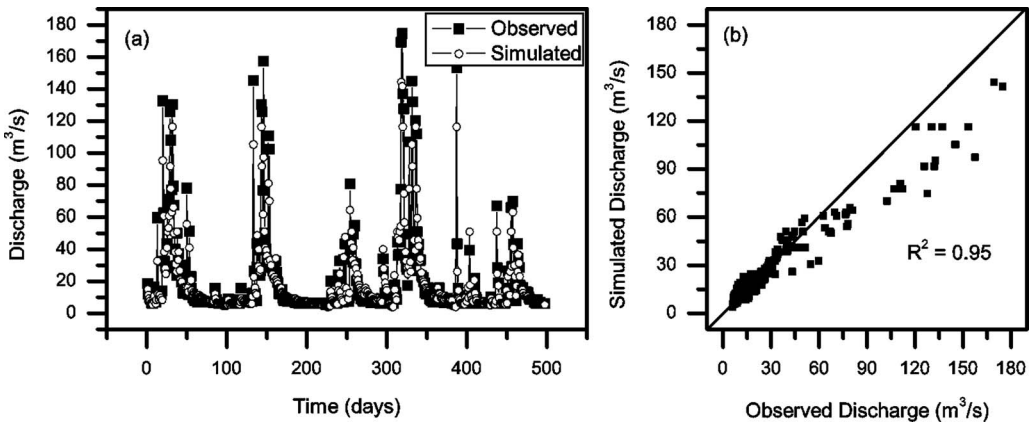

Fig. 3. Computed discharge along with its observed counterpart for Satrana (a) time series; (b) scatter diagram

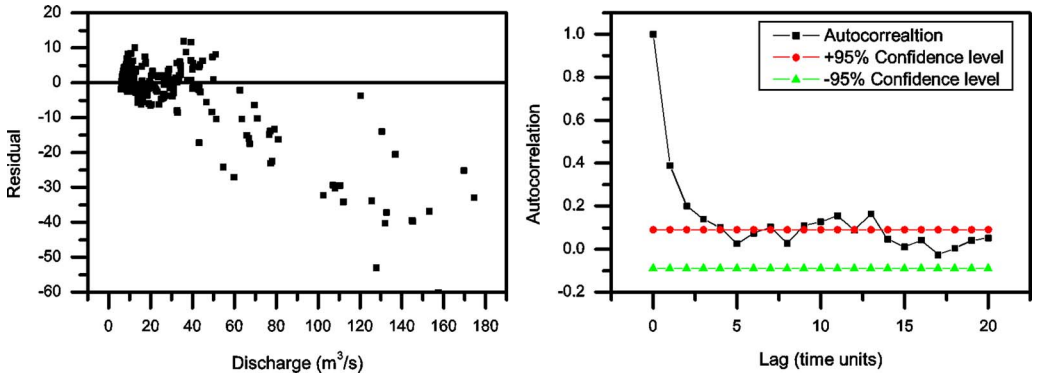

Fig. 4. Plots of residuals for Satrana to illustrate (a) homoscedasticity; (b) independence

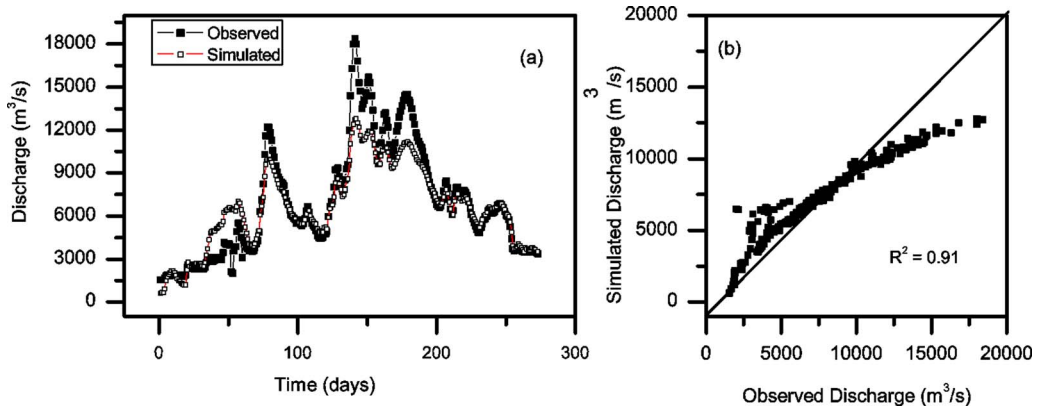

Fig. 5. Computed discharge along with its observed counterpart for Thebes (a) time series; (b) scatter diagram

trana site. Here, the sample size was 497 and the data had high variance as reflected by Cov of 1.4. The results [Fig. 3(a)] show that the match between the observed and the computed discharges was not very good in this case. In the lower range, discharge is modeled well but in the middle and higher ranges, discharges are undersimulated. The scatter plot between the observed and computed discharges [Fig. 3(b)] also shows a small negative bias in the middle and higher ranges of discharges. For this case, the NS index η is quite high at 0.9115. McCuen et al. (2006) have also noted that when the variance is high, even a poor fit may give high value of η. This case belongs to a category where the fit is not very good but even then, quite high value of index η is obtained. If the plots showing homoscedasticity and dependence structure of the residuals (Fig. 4) alone are compared with Fig. 3, it reveals that the fitted model is better compared to that developed for the Chebes site. It is also noted that the residuals are spread equally on the positive and negative side of the expected value (zero) only in the lower ranges of flow, indicating that the model predictions are good only in low ranges of flow and the higher flows are underpredicted. The autocorrelation structure of the residual is found to decay rapidly. Overall, although the NS index in this case is much smaller than the first case, the behavior of the residuals shows this model to be better.

The third case is also related with the development of rating curve for another site named Thebes. At this site, the data displayed moderate variance with Cov of 0.563. As can be seen from

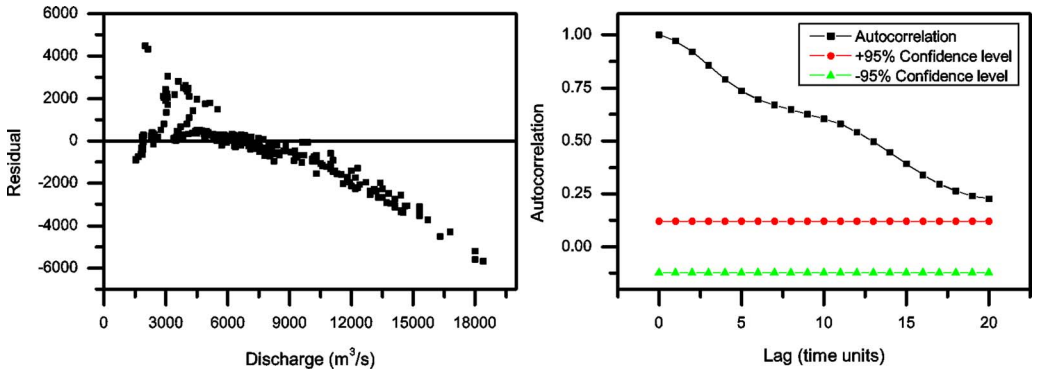

Fig. 6. Plots of residuals for Thebes to illustrate (a) homoscedasticity; (b) independence

J. Hydrol. Eng., 2008, 13(10): 981-986

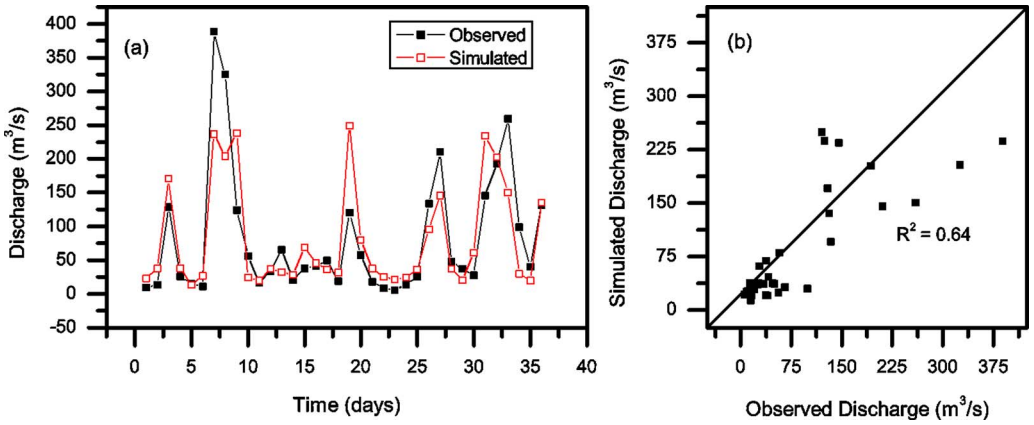

Fig. 7. Computed discharge along with its observed counterpart for Tawi (a) time series; (b) scatter diagram

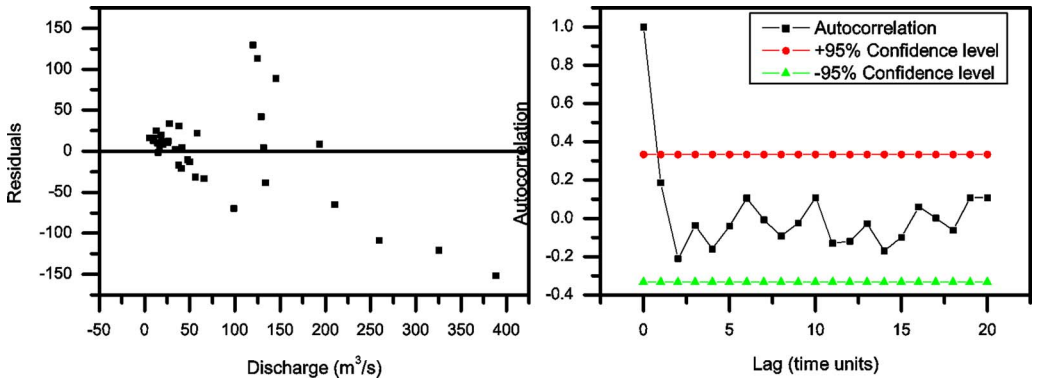

Fig. 8. Plots of residuals from rainfall runoff model for Tawi to illustrate (a) homoscedasticity; (b) independence

Fig. 5(a), here the match between the observed and computed hydrographs is somewhat inferior. In particular, the peaks are not well computed. The model has a high negative bias (the relative bias is about 4.5%), which is also evident from the scatter plot [Fig. 5(b)]. The scatter plot also confirms that in the higher ranges, discharge remains very much undersimulated. But the value of NS index η at 0.8568 was reasonably high. Thus, in this case, the sample size was quite big, data had moderate variance, the model was biased but the NS efficiency index was >0.85.

If the hydrologist assesses the performance of modeling for the Thebes site solely based on NS index, it can be termed as quite good. However, the plots of residuals [Fig. 6(a)] indicate that the model is poor in terms of its ability to capture the structure in the data. The residuals are highly correlated [Fig. 6(b)] at larger lags and the correlogram shows very slow decay.

The fourth example case pertains to a monthly rainfall-runoff modeling using a conceptual model. This model employs a series of storage elements and linear reservoirs to simulate the behavior

Table 2. Summary of Model Fit and Residual Behavior for Case Studies

Case	NS index η	Model fit and residual behavior
Rating curve for Chebes	0.9821	• Good match between observed and computed hydrographs. • Residuals have significant correlation at higher lags. • Residuals are biased toward flow magnitude—positive in low ranges and negative in high ranges.
Rating curve for Satrana	0.9115	• Model has slight negative bias in the middle and higher range of discharge. • Residuals are evenly spread around zero in the lower ranges. • Correlogram of residuals decays rapidly.
Rating curve for Thebes	0.8568	• Peak discharge is undersimulated. • Residuals have high negative bias. • Correlogram of residuals decays slowly.
Rainfall runoff modeling	0.6443	• Match between observed and computed hydrograph is not very good. • Residuals are widely scattered around 45° line. • Correlogram of residualsdecays rapidly.

J. Hydrol. Eng., 2008, 13(10): 981-986

of a catchment. Jain (1993) has described this model. It may be stated that the quality of input data used in this case was not good and this was the major reason behind the poor performance of the model. Here, the sample size was small and the data had high variance. Although the low flows and the hydrograph recession was simulated fairly well [Fig. 7(a)], some peaks were oversimulated and some undersimulated. As can be seen from the scatter plot [Fig. 7(b)], the data points are widely scattered around the 45° line. The results given in Table 1 also show that the model displayed a negative bias, but despite all this, the efficiency index η was 0.6443. The plot of residuals in Fig. 8(a) shows that at high discharge, the residuals have large negative values. It is worth mentioning that the residuals in this case are independent [Fig. 8(b)] and the correlogram decays sharply.

To summarize, four cases of modeling have been presented wherein the NS index ranged from 0.9821 to 0.6443. The results are summarized in Table 2. It was shown that even when the model was unable to extract all the information contained in the data and the residuals were correlated, the NS index was close to the maximum possible value. In fact, the behavior of the residuals was independent of the NS index. This shows that the NS index alone is not an adequate indicator of the adequacy of a mathematical model and additional criteria such as the scatter plot and residual behavior should be employed.

Concluding Remarks

The Nash–Sutcliffe efficiency index (η) is one of the most commonly employed indices to evaluate the performance of a hydrologic model. As shown here, one can achieve a high value of this index even with a not so good model. Therefore, it is not advisable to conclude the performance of a model solely on the basis of NS index. Other statistical tools such as a scatter plot which may reveal important information about the ability of the model to reproduce the dependent variable in different ranges need to be employed to arrive at a definite conclusion about the model performance. Even a poor model may yield η values in the vicinity of 0.60 and these cases definitely warrant a careful look at the model results before drawing any conclusion about its suitability or otherwise.

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.

Clarke, R. T. (1973). "A review of some mathematical models used in hydrology, with observations on their calibrations and their use." *J. Hydrol.*, 19(1), 1–20.

Garrick, M., Cunnane, C., and Nash, J. E. (1978). "A criterion of efficiency for rainfall-runoff models." *J. Hydrol.*, 36(3/4), 375–381.

Hay, L. E. (1998). "Stochastic calibration of an orographic precipitation model." *Hydrolog. Process.*, 12, 613–634.

Jain, S. K. (1993). "Calibration of conceptual models for rainfall runoff simulation." *Hydrol. Sci. J.*, 38(5), 431–441.

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

McCuen, R. H., Knight, Z., and Cutter, A. G. (2006). "Evaluation of the Nash-Sutcliffe efficiency index." *J. Hydrol. Eng.*, 11(6), 597–602.

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

Schlesinger, S., et al. (1979). "Terminology for model credibility." *Simulation*, 32(3), 103–104.

Song, Z., and James, L. D. (1991). "Calibration of a parametric-stochastic model." *Hydrol. Sci. Technol.*, 6(1), 93–98.

Xu, C.-Y. (2001). "Statistical analysis of parameters and residuals of a conceptual water balance model—Methodology and case study." *Water Resour. Manage.*, 15(2), 75–92.