WATER RESOURCES RESEARCH, VOL. 35, NO. 1, PAGES 233–241, JANUARY 1999

Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation

David R. Legates
Department of Geography and Anthropology, Louisiana State University, Baton Rouge

Gregory J. McCabe Jr.
Water Resources Division, U.S. Geological Survey, Denver, Colorado

Abstract. Correlation and correlation-based measures (e.g., the coefficient of determination) have been widely used to evaluate the "goodness-of-fit" of hydrologic and hydroclimatic models. These measures are oversensitive to extreme values (outliers) and are insensitive to additive and proportional differences between model predictions and observations. Because of these limitations, correlation-based measures can indicate that a model is a good predictor, even when it is not. In this paper, useful alternative goodness-of-fit or relative error measures (including the coefficient of efficiency and the index of agreement) that overcome many of the limitations of correlation-based measures are discussed. Modifications to these statistics to aid in interpretation are presented. It is concluded that correlation and correlation-based measures should not be used to assess the goodness-of-fit of a hydrologic or hydroclimatic model and that additional evaluation measures (such as summary statistics and absolute error measures) should supplement model evaluation tools.

1. Introduction

A primary goal of modeling physical processes in the atmospheric and hydrologic sciences is the prediction of a variable in time and/or space from a given set of inputs. How well a model fits the observed data (referred to as model evaluation, or sometimes as model validation) usually is determined by pairwise comparisons of model-simulated (or model-predicted) values with observations. Quantitative assessments of the degree to which the model simulations match the observations are used to provide an evaluation of the model's predictive abilities.

Frequently, evaluations of model performance utilize a number of statistics and techniques. Usually included in these tools are "goodness-of-fit" or relative error measures (bounded statistics, usually between 0.0 and 1.0) to assess the ability of a model to simulate reality. Often these statistics are based on the familiar Pearson's product-moment correlation coefficient (r) or its square, the coefficient of determination (R^2). These two statistics describe the degree of collinearity between the observed and model-simulated variates. They are almost always discussed in basic statistics texts and, consequently, are familiar to virtually all scientists. Unfortunately, both r and R^2 suffer from limitations that make them poor measures of model performance. Although these statistics continue to be used to determine how well a model simulates the observed data, they nevertheless provide a biased view of the efficacy of a model [*Willmott*, 1981; *Willmott et al.*, 1985; *Kessler and Neas*, 1994; *Legates and Davis*, 1997].

As knowledge of physical processes has increased, models have become more complex. Often these models include numerous parameters that are calibrated through optimization procedures, where a range in model parameters is sampled until the differences between the observed and model-simulated data are minimized [*Nash and Sutcliffe*, 1970; *Song and James*, 1991; *Hay*, 1998]. Stochastic calibration procedures are usually employed, which limits graphical analyses of scatterplots, for example, so that statistical analyses must be solely used. Consequently, statistics other than r and R^2 have been developed to describe better the degree of association between the observed and model-simulated data. The objectives of this paper are to (1) examine various goodness-of-fit measures and to identify limitations associated with each, and (2) suggest viable alternative measures for the evaluation of hydrologic and hydroclimatic models.

Paper number 1998WR900018.
0043-1397/99/1998WR900018$09.00

2. Statistics for Evaluation of Hydrologic and Hydroclimatic Models

In this paper, three basic methods for model evaluation will be discussed: the coefficient of determination R^2, the coefficient of efficiency E [*Nash and Sutcliffe*, 1970], and the index of agreement d [*Willmott et al.*, 1985]. In general, this paper addresses comparisons of model-simulated data (P) with the observed data (O) for the same set of conditions (i.e., a pairwise comparison) over a given time period divided into N time increments that can be of arbitrary duration (e.g., monthly or daily time steps).

2.1. Coefficient of Determination R^2

The coefficient of determination is the square of the Pearson's product-moment correlation coefficient (i.e., $R^2 = r^2$) and describes the proportion of the total variance in the observed data that can be explained by the model. It ranges from 0.0 to 1.0, with higher values indicating better agreement, and is given by

$$R^2 = \left\{ \frac{\sum_{t=1}^{N} (O_t - \bar{O})(P_t - \bar{P})}{\left[\sum_{t=1}^{N} (O_t - \bar{O})^2\right]^{0.5} \left[\sum_{t=1}^{N} (P_t - \bar{P})^2\right]^{0.5}} \right\}^2 \quad (1)$$

where the overbar denotes the mean for the entire time period of the evaluation. Note, however, that the coefficient of determination is limited in that it standardizes for differences between the observed and predicted means and variances since it only evaluates linear relationships between the variables. It can be easily demonstrated that if $P_t = (AO_t + B)$ for any nonzero value of A and any value of B, then $R^2 = 1.0$. Thus R^2 is insensitive to additive and proportional differences between the model simulations and observations [see *Willmott*, 1984]. Large values of R^2 can be obtained even when the model-simulated values differ considerably in magnitude (i.e., values of B that differ significantly from 0.0) and variability (i.e., values of A that differ significantly from 1.0). Clearly, in such cases, a model would exhibit serious flaws that should preclude the attribution of a "perfect" designation. These limitations in the coefficient of determination and other correlation-based measures are well documented [cf. *Willmott*, 1981; *Moore*, 1991; *Kessler and Neas*, 1994; *Legates and Davis*, 1997], although such measures still have been used recently to provide, for example, an assessment of climate change detection [e.g., *Santer et al.*, 1995; *Hegerl et al.*, 1996; *Santer et al.*, 1996] and hydrological and hydroclimatological applications (see *McCuen and Snyder* [1975] and *Willmott* [1984] for some examples).

In addition to this obvious limitation of correlation-based measures, *Legates and Davis* [1997] illustrate that correlation-based measures are more sensitive to outliers than to observations near the mean [see also *Moore*, 1991]. Statistical texts frequently illustrate that the correlation can be greatly influenced by the relationship between the two variables for one extreme outlier. This oversensitivity to outliers leads to a bias toward the extreme events if correlation-based measures are employed in model evaluation. A model that can follow the observed data during extreme events will have an artificially higher value of R^2, which may obscure the true relationship between the model-simulated and observed data over most of the remainder of the domain. *Legates and Davis* [1997] illustrate further limitations in correlation-based statistics when derived data (e.g., differences from a standardized mean) are used.

McCuen and Snyder [1975] recognized these limitations in correlation-based measures and developed an adjusting factor equal to

$$\left[\sum_{t=1}^{N} (O_t - \bar{O})^2 \sum_{t=1}^{N} (P_t - \bar{P})^{-2}\right]^{-0.5}.$$

The correlation between the observed and predicted time series is multiplied by this adjusting factor to account for differences in the observed and predicted standard deviations. This adjustment, however, does not account for differences in the mean of the two time series and assumes [see *McCuen and Snyder*, 1975] that the observed variance is less than the model-predicted variance. If, in fact, the model-predicted variance were greater, then the application of the McCuen and Snyder adjusting factor would result in an increase in the correlation, possibly causing it to exceed 1.0 in extreme cases. Consequently, we do not advocate the use of such adjusting factors.

It should be noted that nonparametric or rank correlation methods also exist (e.g., Spearman's rho or Kendall's tau). As nonparametric statistics, they are less sensitive to outliers in the data and generally provide a more robust characterization of the correlation between observed and predicted values. Unfortunately, rank correlation measures are associated with a loss of information as interval/ratio data are converted to ordinal (ranked) form [see *Burt and Barber*, 1996], and, like their parametric counterparts, they are not sensitive to additive and proportional differences between the observed and model-simulated values.

2.2. Coefficient of Efficiency E

The coefficient of efficiency E has been widely used to evaluate the performance of hydrologic models [e.g., *Leavesley et al.*, 1983; *Wilcox et al.*, 1990]. *Nash and Sutcliffe* [1970] defined the coefficient of efficiency which ranges from minus infinity to 1.0, with higher values indicating better agreement, as

$$E = 1.0 - \frac{\sum_{t=1}^{N} (O_t - P_t)^2}{\sum_{t=1}^{N} (O_t - \bar{O})^2}. \quad (2)$$

Physically, E is the ratio of the mean square error,

$$\text{MSE} = N^{-1} \sum_{t=1}^{N} (O_t - P_t)^2,$$

to the variance in the observed data, subtracted from unity. For example, if the square of the differences between the model simulations and the observations is as large as the variability in the observed data, then $E = 0.0$, and if it exceeds it, then $E < 0.0$ (i.e., the observed mean is a better predictor than P_i). Thus a value of zero for the coefficient of efficiency indicates that the observed mean $\bar{O}$ is as good a predictor as the model, while negative values indicate that the observed mean is a better predictor than the model [*Wilcox et al.*, 1990].

The coefficient of efficiency represents an improvement over the coefficient of determination for model evaluation purposes in that it is sensitive to differences in the observed and model-simulated means and variances; that is, if $P_i = (AO_i + B)$, then E decreases as A and B vary from 1.0 and 0.0, respectively. Because of the squared differences, however, E is overly sensitive to extreme values, as is R^2.

2.3. Index of Agreement d

Willmott [1981] sought to overcome the insensitivity of correlation-based measures to differences in the observed and model-simulated means and variances by developing the index of agreement d, given by

$$d = 1.0 - \frac{\sum_{t=1}^{N} (O_t - P_t)^2}{\sum_{t=1}^{N} (|P_t - \bar{O}| + |O_t - \bar{O}|)^2} = 1.0 - N \frac{\text{MSE}}{\text{PE}}. \quad (3)$$

The index of agreement varies from 0.0 to 1.0, with higher values indicating better agreement between the model and observations, similar to the interpretation of the coefficient of determination R^2. *Willmott* [1984] argued that the index of agreement represented the ratio between the mean square error and the "potential error" (PE), multiplied by N and then subtracted from unity. Potential error was defined as

$$\mathrm{PE} = \sum_{i=1}^{N} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2$$

(i.e., the sum of the squared absolute values of the distances from P_i to $\bar{O}$ to O_i) and represents the largest value that $(O_i - P_i)^2$ can attain for each observation/model-simulation pair. As with the coefficient of efficiency, the index of agreement represents a decided improvement over the coefficient of determination but also is sensitive to extreme values, owing to the squared differences.

3. Discussion

Even a cursory examination of (2) and (3) shows similarities between the coefficient of efficiency E and the index of agreement d. Specifically, d additionally includes the difference between the model-simulated values (P_i) and the observed mean in the denominator of the second term. This leads to the difference in the range of the two statistics, $-\infty$–1.0 for the coefficient of efficiency and 0.0–1.0 for the index of agreement, which provides unique advantages to both. Ranging from 0.0 to 1.0, the index of agreement is similar in interpretation to R^2, while the meaningful value of 0.0 for the coefficient of efficiency provides a convenient reference point to compare the model with the predictive abilities of the observed mean.

Ironically, both statistics have been criticized for a similar interpretational difficulty. *Garrick et al.* [1978, p. 376] noted that for the coefficient of efficiency, "... even poor models produce relatively high values (80 or 90%), and the best models do not produce values which, on first examination, are impressively higher." *Willmott et al.* [1985] also observed a similar problem with the index of agreement. In particular, these difficulties lead to an assumption that the model is, in fact, better than the statistic indicates because interpretation of any statistic that ranges between 0.0 and 1.0 usually follows intuitively from R^2 (i.e., a value of 0.5 is usually interpreted as a mediocre model, since for R^2 the variance in the observed data explained by the model is only 50%). Although the meaning is clearly different for E and d, the usual perception of the values from these statistics is the same. Such perceptions can be misleading, since a value of 0.5, for example, has substantially different meanings for R^2, E, or d. Reasons cited by *Garrick et al.* [1978] and *Willmott et al.* [1985] for the relatively high values of the respective statistics actually are quite different, and we believe the issues raised by both authors should be addressed to provide a proper measure of model performance.

The sensitivity to outliers is associated with both E and d (as well as with R^2) and leads to relatively high values of both statistics. This arises due to the squaring of the difference terms [*Willmott*, 1981]. *Willmott et al.* [1985] noted that a more generic index of agreement could be developed as

$$d_j = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|^j}{\sum_{i=1}^{N} (|P_i - \overline{O'}| + |O_i - \overline{O'}|)^j} \tag{4}$$

where j represents an arbitrary power (i.e., a positive integer). Note that the original index of agreement d developed by *Willmott* [1981] becomes d_2 using this notation.

Of particular interest in this discussion is d_1, known also as the modified index of agreement. The advantage of d_1 is that errors and differences are given their appropriate weighting, not inflated by their squared values. Squaring in statistics is useful because squares are easier to manipulate mathematically than are absolute values, but use of squares forces an arbitrarily greater influence on the statistic by way of the larger values. Experience using both d_2 and d_1 shows that, in general, $d_2 \geq d_1$ for the range of most values, although this relationship does not hold for extremely low values of both statistics. Similarly, the coefficient of efficiency can be adjusted to reduce the effect of squared terms by rewriting a more generic form of the coefficient of efficiency (following that of the index of agreement in equation (4)) as

$$E_j = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|^j}{\sum_{i=1}^{N} |O_i - \bar{O}|^j} \tag{5}$$

where the statistic E_1 (termed here the modified coefficient of efficiency) has the desired properties (not inflated by squared values) and is commensurate with d_1.

Garrick et al. [1978, p. 376] further argued that the assumption of comparing the model to the observed mean was "unnecessarily primitive." Better methods exist to define the baseline against which a model should be compared. For example, persistence or averages that vary by season or another time period (i.e., a climatology) may provide a more appropriate baseline for most hydrological or hydroclimatological studies than simply the average of the entire time series. Thus both E_1 and d_1 can be rewritten in a "baseline adjusted" form as

$$E_1' = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|}{\sum_{i=1}^{N} |O_i - \overline{O'}|} \tag{6}$$

$$d_1' = 1.0 - \frac{\sum_{i=1}^{N} |O_i - P_i|}{\sum_{i=1}^{N} (|P_i - \overline{O'}| + |O_i - \overline{O'}|)} \tag{7}$$

where $\overline{O'}$ is the baseline value of the time series against which the model is to be compared. It usually is a function of time and, in some applications, may be a function of other variables as well. Consequently, d_1' is useful in that its interpretation is more conventional, as it more closely follows the interpretation

of R^2 for the range of most values encountered. However, the meaningful evaluation of 0.0 for E'_1 in cases where the model is compared against a more appropriate baseline causes us to recommend E'_1 slightly over d'_1 for most applications. Determining that the baseline (e.g., climatology or persistence) is a better predictor than the model can prove useful in many applications where the model estimates still explain a statistically significant proportion of the observed variance. In estimating air temperature in a spatial context, for example, *Willmott and Robeson* [1995] found that climatology was a better estimate of air temperature at unsampled locations than was traditional spatial interpolation, even though traditional interpolation provided estimates that captured a significant proportion of the seasonal variation in air temperature.

One of the useful properties of correlation, and consequently R^2, is that its statistical distribution has been well-defined. Thus it is easy to evaluate statistical significance (a null hypothesis of no correlation) or to assess whether two correlations (obtained from different models, for example) are statistically different from one another. However, no such simple equations exist to determine statistical significance of either the coefficient of efficiency or the index of agreement. Nevertheless, it is possible to determine statistically significant or statistically different values of both these statistics using bootstrap methods. If it is assumed that the observations represent the population from which they were sampled (a reasonable assumption), then it is possible to repeatedly sample from the observed/model-simulated pairs to provide a distribution of the statistic from which confidence intervals can be developed. For a complete discussion of the use of the bootstrap to determine significance estimation, the reader is urged to consult *Efron* [1981a, b], *Efron and Gong* [1983], and *Willmott et al.* [1985].

Although dimensionless measures (e.g., E'_1 and d'_1) that provide a relative assessment of model performance have been the focus of this paper thus far, these measures should not be used exclusively, as *Willmott* [1981] correctly argued. In addition, it is appropriate to quantify the error in terms of the units of the variable. These measures, or absolute error measures (nonnegative statistics that have no upper bound), include the square root of the mean square error, or RMSE (RMSE = $\sqrt{\text{MSE}}$), and the mean absolute error, or MAE, given by

$$\text{MAE} = N^{-1} \sum_{i=1}^{N} |O_i - P_i| \tag{8}$$

which describe the difference between the model simulations and observations in the units of the variable. As with d_2 and d_1, experience using MAE and RMSE shows that, in general, RMSE $\geq$ MAE for the range of most values. The degree to which RMSE exceeds MAE is an indicator of the extent to which outliers (or variance in the differences between the modeled and observed values) exist in the data.

Other measures, such as the slope and intercept of the predicted-versus-observed regression line and the systematic and unsystematic components of RMSE, also are quite useful for diagnostic purposes. The reader is urged to consult *Willmott* [1984] or *Willmott et al.* [1985] for a more complete discussion of these additional measures; discussion of them here is beyond the scope of this paper. Nevertheless, a complete assessment of model performance should include at least one "goodness-of-fit" or relative error measure (e.g., E'_1 or d'_1) and at least one absolute error measure (e.g., RMSE or MAE) with additional supporting information (e.g., a comparison between the observed and simulated mean and standard deviations).

4. Two Illustrative Examples

To illustrate the usefulness and the interpretation of E'_1 and d'_1, two examples are given: (1) Model simulations of monthly potential evapotranspiration are compared with observations from 1981 through 1983 for Baton Rouge, Louisiana, and (2) a simulation of runoff is compared with observations from October 1972 through September 1989 for the East River Basin in southwestern Colorado. These two disparate examples are used to show how these statistics may be used for model evaluation of hydrologic and hydroclimatic data. Please note that the results from these model evaluations are presented here solely for illustrating the utility of the relative error statistics. Overall efficacy of the potential evapotranspiration or runoff-estimating models used here should not be gleaned from the simple examples.

4.1. Potential Evapotranspiration

Monthly potential evapotranspiration was estimated using methods developed by *Thornthwaite* [1948], *Jensen and Haise* [1963], and *van Bavel* [1966] (hereinafter referred to as the Thornthwaite, Jensen-Haise, and van Bavel methods) using data from 1981 through 1983 for Baton Rouge, Louisiana (see *McCabe and Muller* [1987] for a complete discussion of the data and methods). Inputs for these two methods are quite disparate, as the Thornthwaite method uses only mean monthly air temperature (as well as station latitude) and the Jensen-Haise and van Bavel methods use daily data: mean daily air temperature and solar radiation for the Jensen-Haise method and mean daily air temperature, wind speed, solar radiation, air pressure, and the vapor pressure deficit for the van Bavel method. Monthly time series for the observations (measured pan evaporation multiplied by a pan coefficient of 0.76 [*Kohler et al.*, 1959; *McCabe and Muller*, 1987]) and the three models are shown in Figure 1, while scatterplots for these models are shown in Figure 2.

A comparison of the 3-year mean potential evapotranspiration with the model-simulated values (Table 1) indicates that the estimate using the Thornthwaite method is just over 1 mm lower than the observed, while both the Jensen-Haise and van Bavel estimates are larger by more than 6 and 5 mm, respectively. However, a comparison of the observed and modeled standard deviations shows that only the van Bavel method has a standard deviation that is near the observed. The standard deviation estimated from the Thornthwaite method is nearly twice that of the observed, while the Jensen-Haise method is almost 20 mm higher. Overestimation of potential evapotranspiration in the summer and underestimation in the winter are the reasons for these differences (see Figure 1). Note also that the absolute error measures (MAE and RMSE) indicate that the mean error for the Jensen-Haise method is nearly twice that of the van Bavel method, while the mean error for the Thornthwaite method is nearly 3 times larger than the error in the van Bavel method. Inclusion of more meteorological data on a daily time step is largely responsible for the better estimates from the van Bavel method. The fact that Thornthwaite's method is best suited for long-term estimates and the indirect adjustment of the potential evapotranspiration for precipitation gage measurement biases [see *Legates and Mather*, 1992] are obvious limitations in the Thornthwaite

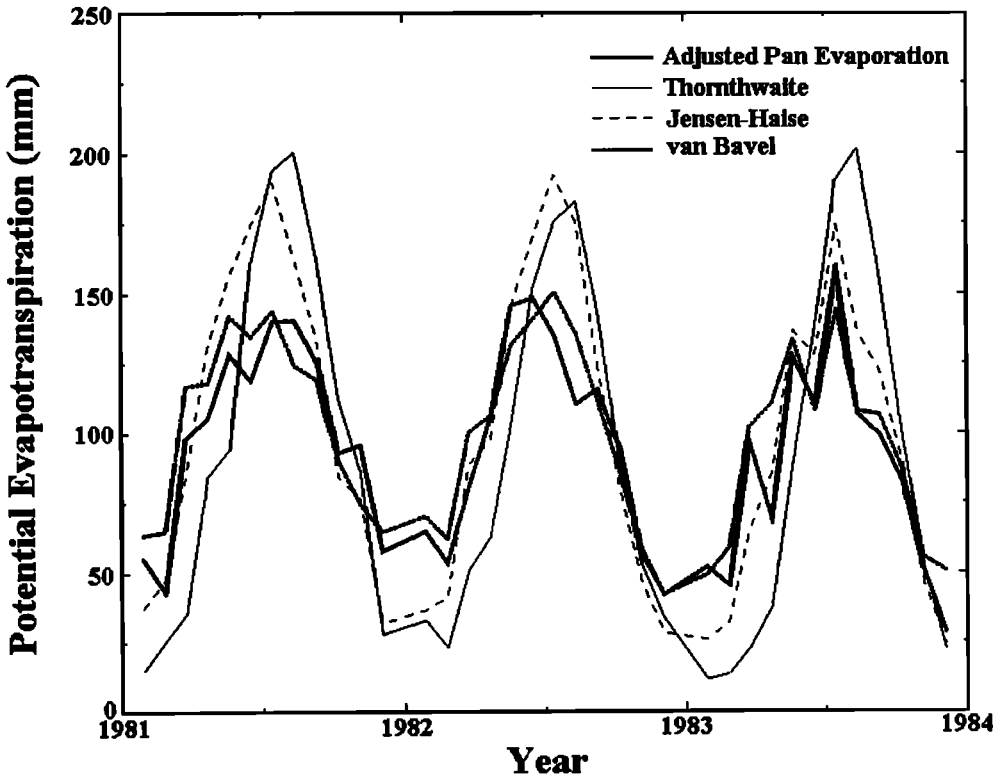

Figure 1. Time series of monthly potential evapotranspiration estimated by the *Thornthwaite* [1948], *Jensen and Haise* [1963], and *van Bavel* [1966] methods compared with observations (measured pan evaporation multiplied by a pan coefficient of 0.76). Data were taken from January 1981 through December 1983 for southern Louisiana (data from *McCabe and Muller* [1987]).

method, although *Willmott* [1981] demonstrated that much of the error in the Thornthwaite method is systematic (most likely due to the indirect adjustment for precipitation gage biases) and can be easily adjusted to produce more accurate estimates.

An examination of Pearson's correlation coefficient (r) or the coefficient of determination (R^2) yields a misleading picture. The van Bavel and Jensen-Haise methods have identical values of r and R^2, while the values for the Thornthwaite method, albeit somewhat lower, imply a reasonably good association. Remember that the correlation coefficient standardizes the variance; thus the differences in the modeled and observed standard deviations do not influence these two statistics. As a result, use of the correlation coefficient or the coefficient of determination yields a rather biased picture of the three methods, as it implies that all three methods are relatively good with no notable differences between the Jensen-Haise and the van Bavel methods.

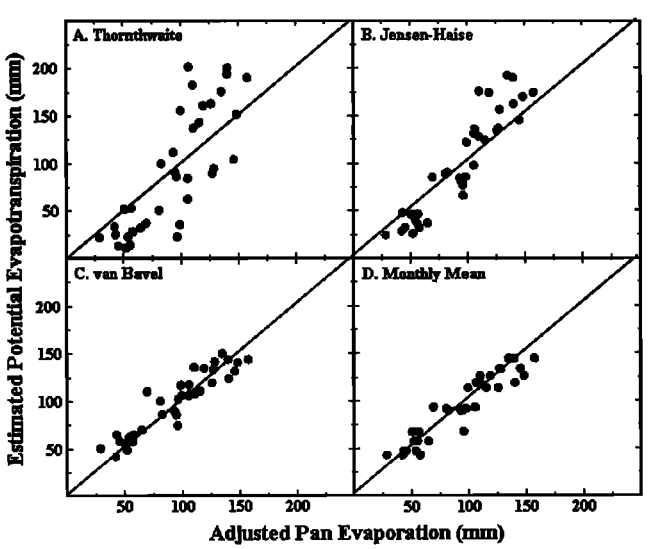

Figure 2. Scatterplots of the Thornthwaite, Jensen-Haise, van Bavel, and climatological (monthly mean) estimates versus the observed (coefficient adjusted pan evaporation) data presented in Figure 1.

Use of the index of agreement, the coefficient of efficiency, or their modified counterparts gives a different, and more accurate, representation of the three models. From these statistics, the van Bavel method has a higher value than (i.e., is better than) the Jensen-Haise method, which, in turn, has a higher value than the Thornthwaite method (note that bootstrap methods could be used to determine if these values are statistically different). This conclusion clearly is warranted from an examination of the time series (Figure 1) and from the absolute error measures.

The issue raised earlier was one of interpretation: Both the index of agreement and the coefficient of efficiency tend to make mediocre models look good by their relatively high values. An index of agreement of 0.82 would lead one to believe that the Thornthwaite method produces reasonably good estimates. A MAE of more than 30 mm per month (more than a third of the mean monthly potential evapotranspiration), however, indicates otherwise. When the squared terms are replaced by absolute values (i.e., d_1 is used rather than d_2), the

Table 1. Means and Standard Deviations of Observed and Simulated Monthly Potential Evapotranspiration for Baton Rouge, Louisiana, for the Period From January 1981 Through December 1983, and Statistics Comparing the Observed and Simulated Time Series

Statistic	Observed	Thornthwaite	Jensen-Haise	van Bavel
Mean, mm	94	93	100	99
Standard deviation, mm	36	64	54	33
Mean absolute error, mm		34	20	11
Root mean square error, mm		40	25	14
Pearson's correlation		0.81	0.93	0.93
Coefficient of determination		0.66	0.86	0.86
Index of agreement		0.82	0.92	0.96
Modified index of agreement		0.60	0.74	0.82
Baseline-adjusted index of agreement		0.24	0.32	0.46
Coefficient of efficiency		−0.30	0.48	0.85
Modified coefficient of efficiency		−0.14	0.32	0.64
Baseline-adjusted coefficient of efficiency		−2.32	−0.98	−0.05

interpretation of the efficacy of the three models (0.60, 0.74, and 0.82, respectively) is much more appropriate. Both the coefficient of efficiency and the modified coefficient of efficiency exhibit considerable differences between the three models, and interpretation does not appear to be a problem. Since these statistics are lower unbounded, they cannot be interpreted like the coefficient of determination or the indices of agreement.

Note that the negative values for the coefficients of efficiency for the Thornthwaite model (particularly that of E and E_1) indicate, at least for these data, that the mean monthly potential evapotranspiration estimates using the Thornthwaite method are less in agreement with the observations than if a constant value (i.e., the observed annual mean value of 94 mm) were used for each month of the year. This arises because the absolute error statistics (MAE and RMSE) are more than one half of the observed monthly standard deviation. A model with no variability (i.e., a nonvarying, climatological estimate) will produce a RMSE of 36 mm (the observed standard deviation), which is lower than the RMSE produced by the Thornthwaite method. Thus the first two coefficients of efficiency (E and E_1) are negative. A value of 0.81 for Pearson's product-moment correlation coefficient obscures the real inadequacies of the Thornthwaite estimates.

As *Garrick et al.* [1978] indicate, the obvious ability of the models to reproduce the strong seasonality in potential evapotranspiration may not be of interest. That is, we may wish to ask, Are these models able to reproduce the interannual variability in potential evapotranspiration without regard to the intra-annual variability? In such a case, the question is whether the models' predictive abilities are better than simply using the long-term mean monthly potential evapotranspiration (here defined by the 3-year average monthly potential evapotranspiration). If this is our interest, then the long-term mean monthly potential evapotranspiration, $\bar{O}$, in the modified index of agreement and the modified coefficient of efficiency should be replaced with a monthly varying (subscript i) baseline average, $\bar{O}' = O_i$.

Examining the baseline-adjusted values (Table 1) for the three models indicates that using climatology (i.e., mean monthly values) may be a better predictor of potential evapotranspiration for this data set than using any of the three methods, including even the van Bavel method. Although this conclusion cannot be extrapolated outside of this data set (which we wish to stress), it nevertheless sheds light on how these goodness-of-fit statistics can be used to evaluate the efficacy of the three models in this example.

4.2. Runoff

As a second example, runoff from October 1972 through September 1989 for the East River Basin (750 km^2) in southwestern Colorado was modeled using the precipitation runoff modeling system (PRMS) [*Leavesley et al.*, 1983] and compared with observations (see *McCabe and Hay* [1995] for a complete discussion of the data and methods). Monthly observed and simulated runoff time series are shown in Figure 3, and the scatterplots between the simulated and observed series are shown in Figure 4.

PRMS is a modular-design, distributed-parameter, physical-process watershed model. Distributed-parameter capabilities are provided by partitioning a watershed into relatively physically homogeneous units. Each unit is assumed to be homogeneous with respect to its hydrologic response and is called a hydrologic response unit (HRU). Both a water balance and an energy balance are computed daily for each HRU. The sum of the responses of all HRUs, weighted on a unit-area basis, produces the daily watershed response [*Leavesley et al.*, 1992]. Meteorological inputs include daily precipitation, maximum and minimum air temperature, and solar radiation. Daily meteorological data were obtained for each HRU or were extrapolated to each HRU using data from meteorological stations and a set of user-defined adjustment coefficients developed from regional climate data [*Leavesley et al.*, 1983, 1992]. PRMS parameters were estimated for each HRU using relations between parameter values and measurable basin and climatic characteristics that were defined in previous studies [*Leavesley et al.*, 1992]. Daily runoff estimates were summed to produce monthly values for comparison with monthly runoff measured at Almont, Colorado.

The difference between the simulated and observed mean monthly runoff for the 17 years is just over 1 mm, while the difference between the simulated and observed standard deviations is virtually negligible (Table 2). While the small values of these differences are encouraging, the absolute error statistics indicate that the differences between the monthly simulated and observed runoff are large relative to the observed mean and standard deviation, although these differences do

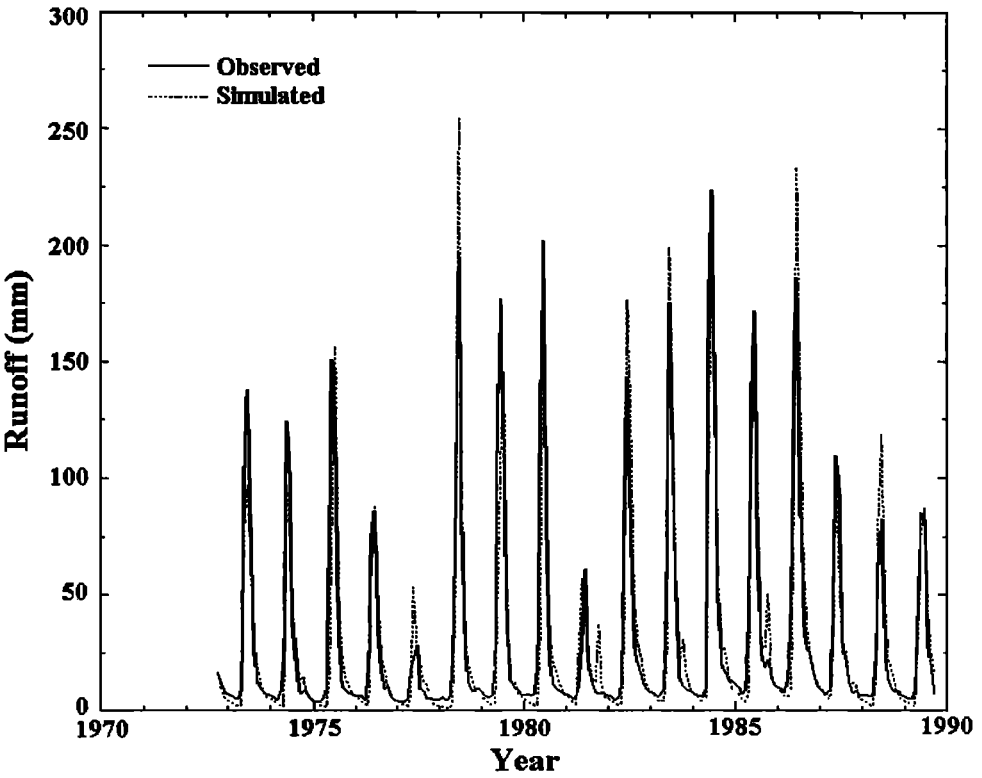

Figure 3. Observed and modeled monthly runoff (precipitation runoff modeling system) for the East River Basin in southwestern Colorado from October 1972 through September 1989 (data from *McCabe and Hay* [1995]).

not appear to exhibit a systematic bias, as is evidenced by the strong agreement in the mean and standard deviation.

Examination of the goodness-of-fit or relative error measures indicates that the model is fairly adequate in reproducing the observed runoff. The high value of the index of agreement ($d = 0.95$) tends to give the impression that the model is much better than it really is. Thus the modified index of agreement (d_1) is probably a better index than its original counterpart from a purely interpretational standpoint. In addition, the three coefficients of efficiency indicate that for this example, the model provides no better predictive ability than using the long-term mean monthly runoff (i.e., although the value of E'_1 is greater than zero, it is not statistically significant). It should not be construed from this simple example, however, that the PRMS model does not contribute any more predictive ability than the use of monthly means for other potential applications.

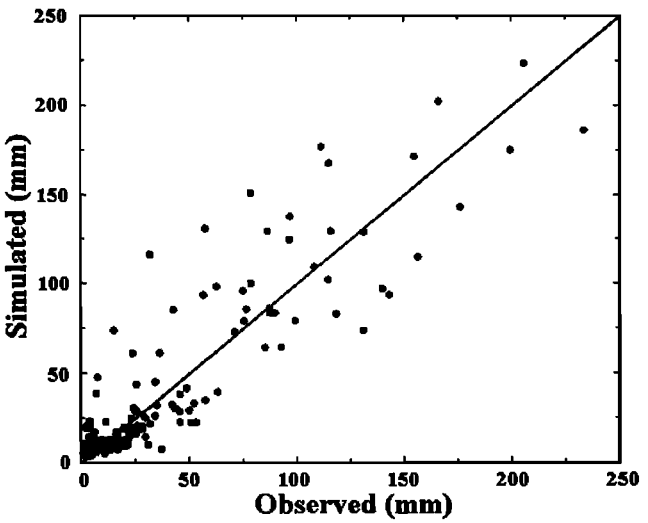

Figure 4. Scatterplot of the observed and modeled monthly runoff (precipitation runoff modeling system) data presented in Figure 3.

5. Recommendations and Final Thoughts

From our discussion and evaluation it is clear that correlation-based measures are inappropriate and should not be used

Table 2. Means and Standard Deviations of Observed and Simulated Monthly Runoff in the East River Basin in Southwestern Colorado for the Period October 1972 Through September 1989, and Statistics Comparing the Observed and Simulated Time Series

Statistic	Observed	Simulated
Mean, mm	33	32
Standard deviation, mm	46	46
Mean absolute error, mm		11
Root mean square error, mm		19
Pearson's correlation		0.91
Coefficient of determination		0.83
Index of agreement		0.95
Modified index of agreement		0.83
Baseline-adjusted index of agreement		0.60
Coefficient of efficiency		0.82
Modified coefficient of efficiency		0.66
Baseline-adjusted coefficient of efficiency		0.06

to evaluate the goodness-of-fit of model simulations. This conclusion arises from the standardization inherent in correlation-based measures and the fact that high correlations can be achieved by mediocre or poor models. Note in our evaluation of potential evapotranspiration that the correlation was identical ($r = 0.93$) for the van Bavel and Jensen-Haise methods despite the fact that other statistics and visual inspection of the time series clearly indicate the superiority of the van Bavel estimates. These concerns are of particular importance owing to the widespread use of correlation-based measures to assess the goodness-of-fit of a model, often without additional evaluation information. It is recommended therefore that either E'_1 or d'_1 be used in lieu of correlation-based measures to provide a relative assessment of model performance. These two statistics use absolute values rather than squared differences (as in their originally specified counterparts) in their computation, which makes E'_1 and d'_1 more conservative measures.

One of the biggest problems associated with all relative error or goodness-of-fit measures is one of interpretation. As implied earlier, the widespread use of correlation-based measures leads to the implicit interpretation of all relative error measures as if they were correlation-based. Interpretation of correlation-based measures is relatively straightforward; that is, a coefficient of determination (R^2) of 0.70 indicates that the model explains 70% of the variability in the observed data. With the indices of agreement, any value (excepting 0.0 and 1.0) is difficult to interpret because its physical meaning is obscure. The meaningful zero present in the coefficient of efficiency, however, yields an appropriate reference point for the interpretation of all other values. Specifically, a value of 0.70 for E, for example, indicates that the mean square error (i.e., the squared differences between the observed and model-simulated values) is 30% of the variance in the observed data (see equation (2)). Both E_1 and E'_1 can be similarly interpreted. On the basis of the efficacy of the coefficients of efficiency (over correlation-based measures) and their ability for physical interpretation owing to their meaningful comparison with 0.0 (i.e., comparison with a "base" model such as the long-term mean or climatology), E_1 and E'_1 are suggested as the most appropriate relative error or goodness-of-fit measures available. Nevertheless, d, and d'_1 have advantages due to their bounds between 0.0 and 1.0, just like correlation-based measures.

In addition to E_1 or E'_1, it is strongly recommended that the observed and modeled means and standard deviations, as well as MAE or RMSE (and probably both), be reported. Scatterplots and residual and outlier analyses also are essential to an appropriate model assessment. Use of absolute error measures (such as MAE or RMSE) provide an evaluation of the error in the units of the variable, which often can provide more information about model efficiency than can be gleaned from the use of relative error or goodness-of-fit measures. Because of the slight bias (i.e., inflated values) in RMSE when large outliers are present, MAE is slightly preferred over RMSE. Statistically significant differences among models (or relative to 0.0 in the case of E'_1) should be assessed through bootstrap techniques discussed by *Willmott et al.* [1985]. Additional statistics and graphical tools suggested by *Willmott* [1984] and *Willmott et al.* [1985] provide useful diagnostics into systematic problems associated with a given model. Simply reporting a single goodness-of-fit measure is inappropriate; goodness-of-fit measures are but a single tool in evaluating model performance.

Acknowledgments. The authors thank Cort Willmott for providing the insight and foresight that has led to our increased and continuing interest in model evaluation issues. Helpful comments from H. V. Gupta and S. M. Robeson, USGS reviewers, and one anonymous reviewer are greatly appreciated.

References

Burt, J. E., and G. M. Barber, *Elementary Statistics for Geographers*, 2nd ed., Guilford, New York, 1996.

Efron, B., Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods, *Biometrika*, *68*, 589–599, 1981a.

Efron, B., Nonparametric standard errors and confidence intervals, *Can. J. Stat.*, *9*, 139–172, 1981b.

Efron, B., and G. Gong, A leisurely look at the bootstrap, the jackknife and cross-validation, *Am. Stat.*, *37*, 36–48, 1983.

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

Hay, L. E., Stochastic calibration of an orographic precipitation model, *Hydrol. Process.*, *12*, 613–634, 1998.

Hegerl, G. C., H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, Detecting greenhouse gas-induced climate change with an optimal fingerprint method, *J. Clim.*, *9*, 2281–2306, 1996.

Jensen, M. E., and H. R. Haise, Estimating evapotranspiration from solar radiation, *J. Irrig. Drain. Div., Am. Soc. Civ. Eng.*, *89*, 15–41, 1963.

Kessler, E., and B. Neas, On correlation, with applications to the radar and raingage measurement of rainfall, *Atmos. Res.*, *34*, 217–229, 1994.

Kohler, M. A., T. J. Nordenson, and D. R. Baker, Evaporation maps for the United States, Plate 3, *Tech. Pap. 37*, Weather Bur., U.S. Dep. of Commer., Washington, D. C., 1959.

Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon, Precipitation-runoff modeling system user's manual, *U.S. Geol. Sur. Water Resour. Invest. Rep. 83-4238*, 207 pp., 1983.

Leavesley, G. H., M. D. Branson, and L. E. Hay, Using coupled atmospheric and hydrologic models to investigate the effects of climate change in mountainous regions, in *Proceedings of the Symposium on Managing Water Resources During Global Change*, pp. 691–700, Am. Water Resour. Assoc., Bethesda, Md., 1992.

Legates, D. R., and R. E. Davis, The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches, *Geophys. Res. Lett.*, *24*, 2319–2322, 1997.

Legates, D. R., and J. R. Mather, An evaluation of the average annual water balance, *Geogr. Rev.*, *82*, 253–267, 1992.

McCabe, G. J., and L. E. Hay, Hydrological effects of hypothetical climate change in the East River Basin, Colorado, *Hydrol. Sci. J.*, *40*, 303–318, 1995.

McCabe, G. J., and R. A. Muller, An index of evaporation in southern Louisiana, *Phys. Geogr.*, *8*, 99–112, 1987.

McCuen, R. H., and W. M. Snyder, A proposed index for computing hydrographs, *Water Resour. Res.*, *11*, 1021–1024, 1975.

Moore, D. S., *Statistics: Concepts and Controversies*, 3rd ed., 439 pp., W. H. Freeman, New York, 1991.

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

Santer, B. D., K. E. Taylor, T. M. L. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, Towards the detection and attribution of an anthropogenic effect on climate, *Clim. Dyn.*, *12*, 77–100, 1995.

Santer, B. D., et al., A search for human influences on the thermal structure in the atmosphere, *Nature*, *382*, 39–46, 1996.

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

Thornthwaite, C. W., An approach toward a rational classification of climate, *Geogr. Rev.*, *38*, 55–94, 1948.

Van Bavel, C. H. M., Potential evaporation: The combination concept and its experimental verification, *Water Resour. Res.*, *2*, 455–467, 1966.

Wilcox, B. P., W. J. Rawls, D. L. Brakensiek, and J. R. Wight, Pre-

dicting runoff from rangeland catchments: A comparison of two models, *Water Resour. Res.*, *26*, 2401–2410, 1990.

Willmott, C. J., On the validation of models, *Phys. Geogr.*, *2*, 184–194, 1981.

Willmott, C. J., On the evaluation of model performance in physical geography, in *Spatial Statistics and Models*, edited by G. L. Gaile, and C. J. Willmott, pp. 443–460, D. Reidel, Norwell, Mass., 1984.

Willmott, C. J., and S. M. Robeson, Climatologically aided interpolation (CAI) of terrestrial air temperature, *Int. J. Climatol.*, *15*, 221–229, 1995.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, Statistics for the evaluation and comparison of models, *J. Geophys. Res.*, *90*, 8995–9005, 1985.

D. R. Legates, Department of Geography and Anthropology, Louisiana State University, 227 Howe/Russell Geoscience Complex, Baton Rouge, LA 70803-4105. (legates@bayamo.srcc.lsu.edu)

G. J. McCabe Jr., United States Geological Survey Water Resources Division, Box 25046, MS 412, Denver Federal Center, Denver, CO 80225.

(Received November 17, 1997; revised August 31, 1998; accepted September 9, 1998.)