1-2004

Statistical Procedures for Evaluating Daily and Monthly Hydrologic Model Predictions

Marilyn E. Coffey
AMEC Earth and Environmental

Stephen R. Workman
University of Kentucky, steve.workman@uky.edu

Joseph L. Taraba
University of Kentucky, joseph.taraba@uky.edu

Alex W. Fogle
Kentucky Geological Survey

STATISTICAL PROCEDURES FOR EVALUATING DAILY AND MONTHLY HYDROLOGIC MODEL PREDICTIONS

M. E. Coffey, S. R. Workman, J. L. Taraba, A. W. Fogle

ABSTRACT. *The overall study objective was to evaluate the applicability of different qualitative and quantitative methods for comparing daily and monthly SWAT computer model hydrologic streamflow predictions to observed data, and to recommend statistical methods for use in future model evaluations. Statistical methods were tested using daily streamflows and monthly equivalent runoff depths. The statistical techniques included linear regression, Nash-Sutcliffe efficiency, nonparametric tests, t-test, objective functions, autocorrelation, and cross-correlation. None of the methods specifically applied to the non-normal distribution and dependence between data points for the daily predicted and observed data. Of the tested methods, median objective functions, sign test, autocorrelation, and cross-correlation were most applicable for the daily data. The robust coefficient of determination (CD*) and robust modeling efficiency (EF*) objective functions were the preferred methods for daily model results due to the ease of comparing these values with a fixed ideal reference value of one. Predicted and observed monthly totals were more normally distributed, and there was less dependence between individual monthly totals than was observed for the corresponding predicted and observed daily values. More statistical methods were available for comparing SWAT model-predicted and observed monthly totals. The 1995 monthly SWAT model predictions and observed data had a regression R_r^2 of 0.70, a Nash-Sutcliffe efficiency of 0.41, and the t-test failed to reject the equal data means hypothesis. The Nash-Sutcliffe coefficient and the R_r^2 coefficient were the preferred methods for monthly results due to the ability to compare these coefficients to a set ideal value of one.*

Keywords. *Dependent data, Hydrologic time series, Hypothesis testing, Model validity, Non-normality.*

The increased use of physically based models has exacerbated the evidence of two problems associated with computer modeling: lack of methods for adequate calibration of model parameters, and limited means for assessing model performance. Comparing model results to observed data is critical for model performance evaluation (Haan et al., 1995). The development of increasingly complex models has resulted in more model parameters being defined, increased need for calibration, and increased uncertainty of model results. With technological advances, physical data are more readily available for parameterizing models and for comparison to model output than ever before. However, improved measurement methods and increased amounts of data alone will not entirely eliminate the need to improve methods for calibrating model parameters and evaluating model performance since physically based models do not include all of the relationships and components of the actual system.

Due to model parameter and output uncertainty, Haan et al. (1995) used Monte Carlo simulations to determine the output probability density function (pdf) and to establish model prediction confidence intervals. The observed data pdf and mean were compared to the model output pdf. If the observed data pdf and mean were within the output confidence interval, then the model was deemed statistically satisfactory. Increasing the number of uncertain model parameters can make the confidence interval so wide that statistically sound model results are unacceptable for the desired applications. The model output pdf should include uncertain model parameters, but for models with many uncertain parameters, calculation time and efforts may exceed benefits in determining how well the model simulates the actual situation. Haan et al. (1995) recommended establishing other quantitative criteria for deeming model results acceptable.

Gupta et al. (1998) proposed a multi-objective parameter calibration approach for use in conjunction with the statistical methods traditionally used to evaluate model performance. Multi-objective calibration and evaluation involves determining which parameters are most important for the particular case and attempting to define a Pareto solution space that encompasses the best ranges of parameter values. This solution space is defined and narrowed by selecting the appropriate objective functions to test the hypothesis, and by using a population estimate of values to determine any patterns in the Pareto solution space. Duan et al. (1992) proposed the shuffled complex evolution algorithm to conduct global optimizations of hydrologic models.

Article was submitted for review in March 2002; approved for publication by the Soil & Water Division of ASAE in October 2003.

The information reported in this article is part of the Kentucky Agricultural Experiment Station (Paper No. 02-05-31) and is published with approval of the Director.

The authors are **Marilyn E. Coffey, ASAE Member Engineer,** Water Resources Specialist, AMEC Earth and Environmental, Knoxville, Tennessee; **Stephen R. Workman, ASAE Member Engineer,** Associate Professor, and **Joseph L. Taraba, ASAE Member Engineer,** Extension Professor, Department of Biosystems and Agricultural Engineering, University of Kentucky, Lexington, Kentucky; and **Alex W. Fogle,** Hydrologist, Kentucky Geological Survey, Lexington, Kentucky. **Corresponding author:** Steve Workman, 105 C. E. Barnhart Building, University of Kentucky, Lexington, KY 40546-0276; phone: 859-257-3000; fax: 859-257-5671; e-mail: sworkman@bae.uky.edu.

Whether defining parameter confidence intervals or Pareto solution space, statistical evaluation approaches require the model user to determine which statistical techniques provide the most accurate descriptions of the fit between modeled values and observed data. There are no standard statistical criteria available for evaluating model results (ASCE, 1993; Gupta et al., 1998).

Qualitative graphs of predicted and observed data provide preliminary model performance assessment (ASCE, 1993). Time series graphs are useful for determining whether a model systematically over- or under-predicts at certain time periods as well as for viewing observed data-model results synchronization. However, evaluating a model's ability to re-create complex system interactions requires more objective testing methods.

Many of the quantitative tests developed to compare model results with observed data assume that both data sets are from normally distributed populations (Shapiro and Wilk, 1965). Although some statistical tests are robust enough to apply to certain types of non-normal data, knowing how well data meet the normality assumption is necessary for using test results (Shapiro and Wilk, 1965). Therefore, normality testing is critical for determining appropriate statistical techniques.

Preliminary quantitative analysis involves central tendency and variation calculations (Zacharias et al., 1996). Mean and standard deviation are preferred central tendency and variation measures for approximately normal data. For data of non-normal or unknown distribution, median and median absolute deviations are recommended central tendency and variation measures. Median estimators are less sensitive to data contamination effects (and thus non-normality) than mean estimators (Rousseeuw and Leroy, 1987).

A plethora of quantitative tests have been utilized for comparing model results to observed data, but the underlying assumptions must not be ignored. Spruill et al. (2000) used average absolute deviation (α) for calibration testing of a watershed model. Linear least-squares regression of a plot of predicted versus observed values is another evaluation technique (Arnold and Allen, 1996; Bingner, 1996; Arnold et al., 1998). The regression correlation coefficient (R_r^2) relatively compares the model regression to the ideal case. An ASCE Task Committee (ASCE, 1993) recommended using Nash-Sutcliffe model efficiency (R^2) and average runoff volume deviation (D_V) for gauging hydrologic model performance. Legates and McCabe (1999) found these correlation methods to be sensitive to extreme values and insensitive to additive differences, which are cases commonly found in hydrologic data. Mean and median objective functions were also implemented for evaluating model fit (Loague et al., 1988; Legates and McCabe, 1999; Zacharias et al., 1996). Nonparametric methods allow statistical analysis in the presence of non-normality, outlying data points, skewed distributions, and truncated data (Hirsch et al., 1991; Bilisoly et al., 1997). Comparison of time series with dependent data points may require autocorrelation and cross-correlation (Haefner, 1997).

While all of the hydrologic model statistical analysis articles reviewed included appraisal of model results, limited work was available on assessing which statistical techniques were best suited for evaluating hydrologic model effectiveness. This research involved testing the applicability of various statistical methods in comparing daily and monthly results obtained by Spruill et al. (2000) using the Soil and Water Assessment Tool (SWAT) model developed by Arnold et al. (1999) to observed site data. The statistical techniques tested included average absolute deviation, least-squares regression, Nash-Sutcliffe efficiency coefficient, average deviation, product moment correlation coefficient, goodness-of-fit objective functions, hypothesis testing, and correlation analysis. The overall study objective was to determine which statistical methods were most appropriate for gauging hydrologic model performance, particularly SWAT model performance, and to provide recommendations for evaluating model results in future studies.

Description of Statistical Tests

Various statistical methods were used for the quantitative portion of the model performance evaluation. The quantitative tests included methods for evaluating the properties of individual data sets (e.g., checking for normal distribution and autocorrelations) as well as tests to gauge model performance by comparing model results to the observed data. The calculations within individual data sets were used to evaluate how well the data satisfied the assumptions of the statistical measures used to evaluate model performance. Many of the study statistics were dimensionless. If a statistic has units associated with the calculated value, then the units are presented with the statistical results.

Normality Statistical Testing

Three measures were used to assess normality of observed and predicted data sets. Kurtosis describes distribution peakedness and is 3 for a normal distribution, with kurtosis greater than 3 indicating a less peaked (more heavily tailed) distribution than the normal distribution, and vice versa for values less than 3 (Haan, 2002). The skewness coefficient shows data skew direction, with symmetric distributions having a skewness coefficient of zero (Haan, 2002). The Shapiro-Wilk *W*-test uses the sample variance and size to form the *W* statistic (Shapiro and Wilk, 1965). The Shapiro-Wilk test equations are presented in equations 1-4. The null hypothesis of normally distributed data is evaluated using P-values based on the *W* statistic.

$$n = \begin{cases} 2k & \text{if sample size is even} \\ 2k+1 & \text{if sample size is odd} \end{cases} \quad (1)$$

$$S^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (2)$$

$$b = \sum_{i=1}^{k} a_{n-i+1} (y_{n-i+1} - y_i) \quad (3)$$

$$W = \frac{b^2}{S^2} \quad (4)$$

where

S^2 = sample variance
y_i = sample observation (in ascending order)
$\bar{y}$ = sample mean
k = test statistic summation index
a = normalized coefficient

b = linear sample order statistic
W = test statistic (small values indicate non-normality)
n = sample size.

MODEL PERFORMANCE STATISTICS

The average absolute deviation or mean absolute error (α) uses absolute deviations between model values and observed data to prevent opposite-signed error cancellation (eq. 5). Average absolute deviations can be effective for model calibration to assess result differences associated with changing a model parameter (Spruill et al., 2000). Model parameters can be optimized by minimizing α values.

$$\alpha = \frac{\sum_{i=1}^{n} |Y_i - X_i|}{n} \tag{5}$$

where
X_i = predicted value
Y_i = observed value.

The regression correlation coefficient (R_r^2) gauges how closely the observed-predicted regression line approaches an ideal fit. This coefficient usually ranges from zero to one, with an R_r^2 of one indicating a perfect fit (eqs. 6-8). For the best fit, regression slope and intercept are one and zero, respectively. Three least-squares regression assumptions involving error terms and data points are constant variance, independence, and approximate normal distribution (Freund and Wilson, 1997).

$$R_r^2 = 1 - \frac{SSE}{SST} \tag{6}$$

$$SSE = \sum \left(X_i - \overline{Y_{i,r}}\right)^2 \tag{7}$$

$$SST = \left(\sum X_i^2\right) - \frac{\left(\sum X_i\right)^2}{n} \tag{8}$$

where
R_r^2 = regression correlation coefficient
SSE = sum of squares errors
SST = total sum of squares
$\overline{Y_{i,r}}$ = value predicted by regression equation.

The Nash-Sutcliffe coefficient (R^2) shown in equation 9 was developed as a sum of squares relative model efficiency measure (Nash and Sutcliffe, 1970):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \tag{9}$$

where $\overline{Y}$ is the average of observed values.

The R^2 coefficient is less than or equal to one, with R^2 of one showing ideal model fit, and R^2 of zero indicating that the model results are no better than the observed data mean (ASCE, 1993). The R^2 value can also be negative because the coefficient is calculated using actual differences and not absolute values for the differences.

An average deviation (D_V) of zero indicates ideal model fit (eq. 10). Absolute differences avoid canceling opposite-signed errors (Martinec and Rango, 1989). The D_V calculation is weighted by the actual observed values.

$$D_V(\%) = \frac{\sum_{i=1}^{n} |Y_i - X_i|}{\sum_{i=1}^{n} Y_i} (100) \tag{10}$$

where D_V is the average percent deviation.

The product moment correlation coefficient (r) shown in equation 11 is another statistic for determining the relationship between two data sets (Addiscott et al., 1995). The correlation coefficient measures the linear relationship between observed data and model values (Haan, 2002).

$$r = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\left(\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right)^{1/2} \left(\sum_{i=1}^{n} (X_i - \overline{X})^2\right)^{1/2}} \tag{11}$$

where $\overline{X}$ is the average of predicted values.

For linearly correlated sets, r is one. The r coefficient only tests for linear correlation, so an r value of zero does not mean that no correlation exists.

Model goodness-of-fit objective functions were presented by Loague et al. (1988). The mean-based functions include maximum error (ME), normalized root mean square error (RMSE), coefficient of determination (CD), and modeling efficiency (EF):

$$ME = max\left(|Y_i - X_i|\right)_{i=1}^{n} \tag{12}$$

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i)^2\right]^{0.5} \left(\frac{100}{\overline{Y}}\right) \tag{13}$$

$$CD = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{Y})^2} \tag{14}$$

$$EF = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 - \sum_{i=1}^{n} (Y_i - X_i)^2}{\sum_{i=1}^{n} (X_i - \overline{Y})^2} \tag{15}$$

Zacharias et al. (1996) presented three robust objective function modifications for use where median is the better central tendency estimator (i.e., non-normal data) (eqs. 16-19). The statistics include normalized median absolute error (MdAE), robust coefficient of determination (CD*), and robust modeling efficiency (EF*), corresponding to normalized RMSE, CD, and EF, respectively. For perfect model fit, ME = RMSE = MdAE = 0, and CD = CD* = EF = EF* = 1.

$$MdAE = median\left\{|Y_i - X_i| : i = 1, 2, ..., n\right\} \times \left(\frac{100}{\overline{Y^*}}\right) \tag{16}$$

$$CD^* = \frac{median\left\{\left|Y_i - \overline{Y^*}\right| : i = 1,2,...,n\right\}}{median\left\{\left|X_i - \overline{Y^*}\right| : i = 1,2,...,n\right\}} \quad (17)$$

$$EF^* = \frac{median\left\{\left|Y_i - \overline{Y^*}\right| : i = 1,2,...n\right\} - median\left\{\left|Y_i - X_i\right| : i = 1,2,...n\right\}}{median\left\{\left|Y_i - \overline{Y^*}\right| : i = 1,2,...n\right\}} \quad (18)$$

$$\overline{Y^*} = median\{Y_i : i = 1,2,...,n\} \quad (19)$$

Hypothesis tests are more rigorous methods to compare data sets (Haefner, 1997). The null hypothesis for comparing approximately normal data is that the two data set means are equal, so the sample *t* statistic shown in equation 20 can be used (Freund and Wilson, 1997):

$$t = \frac{\left(\overline{d} - \mu\right)}{s/\sqrt{n}} \quad (20)$$

where

$\overline{d}$ = sample mean of differences
μ = population mean (null hypothesis)
s = sample estimate of standard deviation.

The sign test and Wilcoxon sign rank test are two nonparametric paired data methods useful with data of unknown distribution. Nonparametric statistics use median for central tendency and do not rely on distribution-specific assumptions, but they have other assumptions to consider (Hollander and Wolfe, 1973). Hirsch et al. (1991) demonstrated that nonparametric testing had small efficiency and testing power advantages over parametric tests when data were slightly non-normal, but the advantage increased as data moved farther from normality. Each test statistic is based on differences between the values in each observed-predicted data pair. The null hypothesis is that the data sets share the same median (the median of the differences is zero). The sign test assumes that errors are random, independent variables from a continuous population. The data pair differences (Z_i) are used to form the *B* statistic for hypothesis testing:

$$\varphi_i = \begin{cases} 1 \text{ if } Z_i > 0 \\ 0 \text{ if } Z_i < 0 \\ \text{discard } Z_i = 0 \end{cases} \quad (21)$$

$$B = \sum_{i=1}^{n} \varphi_i \quad (22)$$

where

φ_i = indicator variable
n = number nonzero Z_i values.

The Wilcoxon sign rank test shares the sign test assumptions along with an additional normally distributed errors assumption. The more powerful sign rank test has more limited applicability than the sign test. The Wilcoxon test involves ranking the absolute differences (Z_i) and computing the T^+ statistic or the sum of the positive ranks:

$$T^+ = \sum_{i=1}^{n} \varphi_i R_i \quad (23)$$

where R_i is the rank of absolute differences (Z_i).

Data correlations in computer-simulated and observed daily time series (e.g., streamflow) can violate the data point and error term independence assumptions of many statistical tests (Haefner, 1997). Autocorrelation and cross-correlation techniques for testing time series model fit require other statistical assumptions. These correlation methods assume that the time series are stationary with no deterministic components (Haan, 2002). Removal of trends from deterministic time series produces residual autocorrelation between unrelated variables (Haan, 2002; Diggle, 1990). Autocorrelation calculations shown in equations 24-26 define correlation between two data points within a single time series for specified lag times (Haan, 2002). For purely random processes, all of a given time series' lag time autocorrelation estimators are zero, indicating no linear dependence between the values within a data set.

$$\overline{x_n} = \frac{1}{n}\sum_{t=1}^{n} X_t \quad (24)$$

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}\left(X_t - \overline{x_n}\right)\left(X_{t+h} - \overline{x_n}\right) \quad (25)$$

$$\hat{r}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \quad (26)$$

where

$\overline{x_n}$ = mean value of time series
X = observation within time series
h = lag time (days)
$\hat{\gamma}(h)$ = autocovariance estimator
$\hat{r}(h)$ = autocovariance estimator.

Unlike autocorrelation, which checks for correlations within a single time series, cross-correlation gauges correlation between two given time series (Fuller, 1996). The cross-correlation method relates cross-covariance to the autocovariance at lag time zero (eqs. 27-28). Cross-correlation shows agreement between observed and simulated time series and can be calculated for specific lag times.

$$\hat{\gamma}_{ij}(h) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-h}\left(X_{it} - \overline{x_{in}}\right)\left(X_{j,t+h} - \overline{x_{jn}}\right) & h = 0,1,...n-1 \\ \frac{1}{n}\sum_{t=-h}^{n}\left(X_{it} - \overline{x_{in}}\right)\left(X_{j,t+h} - \overline{x_{jn}}\right) & h = -1,-2,...-(n-1) \end{cases} \quad (27)$$

$$\hat{r}_{ij}(h) = \frac{\hat{\gamma}_{ij}(h)}{\sqrt{\hat{\gamma}_{ii}(0)\,\hat{\gamma}_{jj}(0)}} \quad (28)$$

where

$\hat{\gamma}_{ij}(h)$ = cross-covariance estimator
X_{ii} = observation from observed data
x_{in} = observed data mean
$X_{j,t+h}$ = observation from model results at lag time h
$\overline{x_{jn}}$ = model data mean
$\hat{r}_{ij}(h)$ = cross-correlation estimator

$\hat{\gamma}_{jj}(0)$ = model results autocorrelation estimate at lag time zero

$\hat{\gamma}_{ii}(0)$ = observed data autocorrelation estimate at lag time zero.

Materials and Methods

The SWAT watershed model (Arnold et al., 1999) was used to simulate daily streamflow for 1995 and 1996 for a basin at the University of Kentucky Animal Research Center (ARC) site (Spruill, 1998). Streamflow data from 1995 and 1996 collected at the ARC located near Versailles in central Kentucky were used to evaluate the SWAT model results obtained by Spruill (1998) using statistical procedures. The ARC site covers 5.5 km^2 and is located in the Inner Bluegrass geologic region of Kentucky, which is characterized by prevalent karst geologic features such as sinkholes and springs. The ARC is used for growing tobacco, hay, small grains, and row crops and will become the University's primary animal research location. The soil series are predominantly Maury (Typic Paleudalf) and McAfee (Mollic Hapludalf) that have moderate available water content and permeability. Water from the ARC eventually enters the Kentucky River.

Weirs have been placed at various locations on the ARC site to measure streamflow. The main weir (inlet to a box culvert) for streamflow measurement is at the edge of the property. The outlet weir data referenced in this article are for the first six and a half months of 1995 and nearly all of 1996. For 1995, the data collected from mid-July through the rest of the year were removed from the analysis due to equipment malfunctions. The 1996 flow data start at the beginning of January and end at the beginning of December, when equipment problems also occurred. The flow data (collected at a 5 min interval) were summarized to obtain daily values, and these daily values were compared to the model predictions.

Streamflow data collected in 1996 were used to calibrate the SWAT model in the Spruill study, and 1995 flow data were used to test SWAT's performance in predicting ARC streamflow. Both 1995 and 1996 flow data records included gaps in the data where streamflow was not recorded. The Spruill study noted that the model tended to over-predict peak flow values for summer months.

This study not only compared predicted and observed values from the Spruill study but also involved comparison between daily and monthly data sets. The daily observed and SWAT-predicted values were streamflow amounts (cms). For monthly observed and model-predicted values, monthly equivalent runoff depths (m) were used for the study. The equivalent runoff depths were calculated by summing the average daily volumes to get a monthly volume and then dividing the monthly volume by the watershed drainage area.

Results and Discussion

Qualitative Analysis

Daily Average Flows

A graph of observed and SWAT-simulated results over time showed that SWAT daily average flows were often not synchronized with observed averages (figs. 1 and 2). Overall, SWAT tended to under-predict flows and showed quicker recession than the observed data (Spruill et al., 2000). Some flow over-prediction occurred in the 1995 winter months and within certain periods of high precipitation. The Spruill study results were affected by late afternoon or evening storms. The storms were modeled on the day of the storm; however, the stream hydrograph usually peaked shortly after midnight and thus occurred on the day following the storm event. Smithers and Engel (1996) also reported that the SWAT model over-predicted flows and simulated little or no recession between event peaks for one of the two watersheds modeled in a separate study.

Monthly Equivalent Runoff Depths

The monthly totals graph was scrutinized for overall agreement between observed data and SWAT results (fig. 3). Monthly totals for March, May, and June of 1995 and for March, April, May, and September of 1996 showed model under-prediction of water exiting the site, corresponding with the daily graph recession and peak flow difficulties (figs. 1 and 2). Over half of the monthly totals were under-predicted by the SWAT model. For the ARC location,

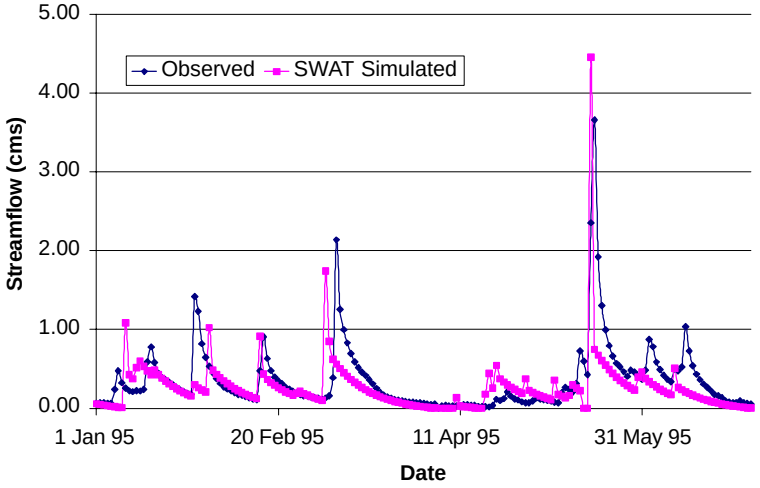

Figure 1. Average daily streamflow values for January through June 1995. The observed data were recorded at the watershed outlet of the University of Kentucky Animal Research Center.

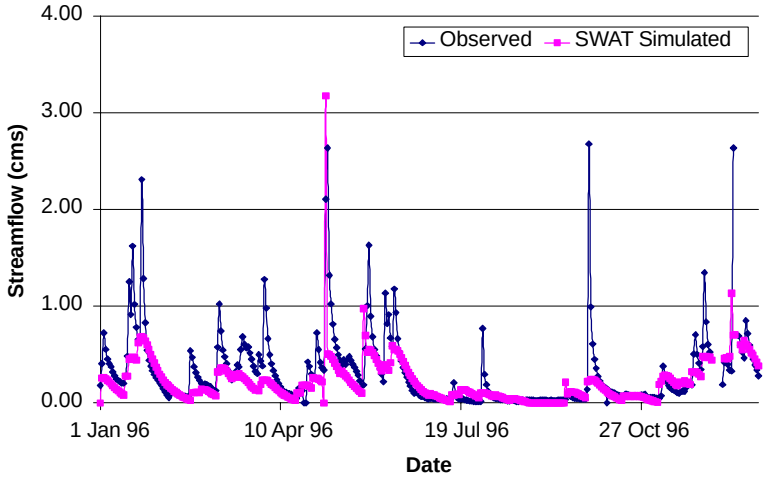

Figure 2. Average daily streamflows and SWAT results for 1996. The observed data were recorded at the watershed outlet of the University of Kentucky Animal Research Center.

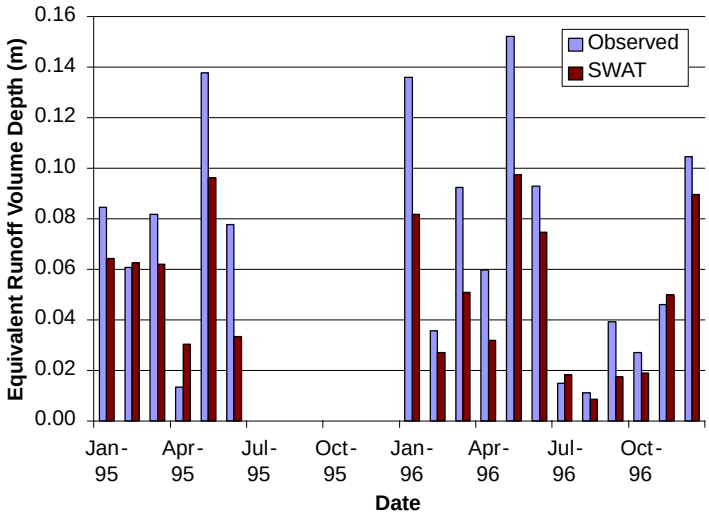

Figure 3. Monthly equivalent runoff depths for 1995 and 1996. The equivalent depths were obtained by summing average daily volumes to get a monthly volume, and then dividing the monthly volume by the 12 km^2 optimal drainage area used by Spruill (1998).

model under-prediction may have been due to the regional karst geology, which could not be explicitly modeled in SWAT. Kosky and Engel (1997) also found that the SWAT model generally under-predicted runoff volume for 15 rainfall events where measured runoff volumes were compared to SWAT-predicted runoff volumes.

Quantitative Analysis

Daily Average Flows

Normally distributed and independent data and error assumptions were scrutinized to determine the most applicable statistical tests. Kurtosis and skewness coefficient calculations were performed for daily and monthly values including the observed data and SWAT predictions (table 1). The daily observed and predicted kurtosis calculations were all much greater than 3, indicating that the distributions were less peaked than the normal distribution. All of the daily skewness coefficients were much greater than zero, meaning that the observed and predicted data sets were skewed when compared to the normal distribution. The Shapiro-Wilk test gauged whether each set of values was normally distributed by using the null hypothesis that the data within each set were normally distributed for confidence level $\alpha = 0.5$. None of the daily data were approximately normally distributed (table 1). Each data set showed autocorrelations, but as lag time distance from zero increased, the autocorrelations quickly approached zero, indicating low system error persistence (eqs. 24-26 and table 2).

Table 1. Normality testing results.

			Shapiro-Wilk Test	
Data	Kurtosis	Skewness Coefficient	P-value	Conclusion
1995 daily				
Observed	21.5	3.9	0.00	Reject h_0
SWAT	77.5	7.6	0.00	Reject h_0
1996 daily				
Observed	12.8	3.0	0.00	Reject h_0
SWAT	63.5	5.7	0.00	Reject h_0
1995 monthly				
Observed	1.8	-0.05	0.65	Do not reject h_0
SWAT	0.12	0.40	0.35	Do not reject h_0
1996 monthly				
Observed	-0.90	0.57	0.32	Do not reject h_0
SWAT	-1.5	0.43	0.16	Do not reject h_0

Daily time series analysis combined autocorrelation and cross-correlation criteria for model fit evaluation (eqs. 24-28). Ideally, both autocorrelation patterns would be identical, and the cross-correlation at lag time zero would be one with a symmetric cross-correlation pattern. Box and Jenkins (1976) recommend at least 50 time series observations for autocorrelation and cross-correlation calculations, which was met by all the daily observed data. The observed and SWAT autocorrelation patterns for each year were not similar, indicating a lack of symmetry between the two data sets (table 2). For both 1995 and 1996, the lag time zero cross-correlations showed correlation between observed data and model results, but the lag time zero values were well below one, and the patterns showed limited symmetry about zero (table 2). While the 1995 cross-correlation at lag time zero was larger than the corresponding 1996 value, the 1996 observed data set was larger than the 1995 observed data set. Therefore, the 1995 and 1996 cross-correlation coefficients should not be directly compared.

The average absolute deviations showed average model differences but provided little information regarding how well model values corresponded with observed data (eq. 5 and table 3). For ideal model fit, α would be zero. The α statistic is useful in making rapid comparisons between successive model simulations for model calibration (e.g., Spruill, 1998). The average absolute deviation provides a statistic in the units of the variable, which is useful in making quick assessments of model capabilities (Legates and McCabe, 1999).

Table 2. Daily time series correlation results.

Year	Lag Time (days)	Autocorrelation		Cross-correlation
		Observed	SWAT	
1995	-7	0.12	0.06	0.13
	-6	0.15	0.09	0.16
	-5	0.22	0.14	0.19
	-4	0.31	0.16	0.25
	-3	0.40	0.19	0.36
	-2	0.51	0.19	0.40
	-1	0.76	0.25	0.65
	0	1.00	1.00	0.52
	1	0.76	0.25	0.25
	2	0.51	0.19	0.25
	3	0.40	0.19	0.25
	4	0.31	0.16	0.17
	5	0.22	0.14	0.10
	6	0.15	0.09	0.08
	7	0.12	0.06	0.05
1996	-7	0.24	0.31	0.23
	-6	0.29	0.34	0.26
	-5	0.36	0.37	0.30
	-4	0.36	0.41	0.38
	-3	0.41	0.44	0.47
	-2	0.51	0.47	0.66
	-1	0.71	0.48	0.64
	0	1.00	1.00	0.48
	1	0.71	0.48	0.46
	2	0.51	0.47	0.45
	3	0.41	0.44	0.43
	4	0.36	0.41	0.42
	5	0.36	0.37	0.34
	6	0.29	0.34	0.31
	7	0.24	0.31	0.29

Table 3. Partial summary of parametric statistic results.

Data	R^2	D_V (%)	*r*	α (cms)
Daily				
1995	0.09	61.73	0.51	0.22
1996	0.15	54.25	0.48	0.17
Monthly				
1995	0.41	31.80	0.84	0.02
1996	0.61	32.02	0.94	0.03

Both Nash-Sutcliffe coefficients comparing SWAT results to observed data showed model efficiencies below 20%, which was much less than the ideal fit value of one (eq. 9 and table 3). The R^2 efficiency statistic uses the observed values mean, a potential shortcoming for non-normal data. A value of R^2 greater than zero indicates that the model is a better predictor of the data than simply using the mean (Legates and McCabe, 1999). The small time lag in observed flows versus simulated flows resulting from late afternoon storms caused the lower model efficiencies. Other studies involving the SWAT model also used the Nash-Sutcliffe coefficient for comparison between predicted and observed values (Arnold et al., 1993; Srinivasan and Arnold, 1994; King et al., 1999; Peterson and Hamlett, 1997; Kosky and Engel, 1997). The range of values for R^2 results was from -1.89 to 0.86, indicating that the data mean was sometimes a better predictor of the observed data than the SWAT model (i.e., when the Nash-Sutcliffe coefficient was negative). The R^2 efficiencies for the 1995 and 1996 data were within this range of values.

The D_V statistic showed how well model and generated data represented observed daily runoff volumes with an ideal D_V of zero (eq. 10 and table 3). The SWAT deviations were over 50%. Periods of low runoff show higher D_V results than periods of high runoff due to the D_V statistic (Martinec and Rango, 1989), with this effect contributing to model and generated data D_V values. Peterson and Hamlett (1997) obtained an overall daily D_V of 40% with the D_V value lowered to 4% for daily and monthly results where snowfall events were assumed negligible. The daily D_V values from the Spruill results were also much higher than the corresponding monthly values (table 3).

The *r* correlation showed that the model results and observed data were positively related (eq. 11 and table 3); however, it was more sensitive to the timing discrepancy between the predicted and observed values than previous techniques. The *r* statistic used the mean for measuring central tendency. Ideal linear correlation produces a product moment correlation of one (Addiscott et al., 1995). Both SWAT *r* values were approximately 0.5. Two other SWAT model studies utilized the *r* correlation coefficient (Smithers and Engel, 1996; Kosky and Engel, 1997). The product moment correlation values reported in these two studies were between 0.35 and 0.84. The *r* values using the Spruill study SWAT results also fell within this range.

Regression line slope and intercept would ideally be one and zero, respectively, with an R_r^2 coefficient of one (eqs. 6-8). The R_r^2 value for 1996 was closer to one than the 1995 value (table 4). Other SWAT model studies using the R_r^2 statistic produced regression correlation coefficients between -0.22 and 0.95 for watersheds of various sizes in different geographic regions (e.g., Arnold et al., 1993; Srinivasan and Arnold, 1994; Arnold and Allen, 1996;

Bingner, 1996; Smithers and Engel, 1996; Kosky and Engel, 1997). Even though the 1996 R_r^2 value of 0.40 was less than the ideal coefficient value of one, the result was within the range of values produced from other studies involving the SWAT model. Regression techniques can withstand some constant error variance assumption infractions, but are sensitive to error independence violations (Haefner, 1997).

The mean objective function and robust (median) objective functions results generally showed that SWAT predictions matched the observed data (eqs. 12-19 and table 5). For ideal fit, MdAE = ME = RMSE = 0, and CD = CD* = EF = EF* = 1. However, ME, RMSE, CD, and EF are most accurate for approximately normally distributed data since they are mean-based functions. Median objective functions are better suited for non-normal data since median is the better central tendency estimator for non-normal data (Rosenberger and Gasko, 1983). The SWAT model MdAE values were consistently lower than mean-based RMSE counterparts, and the SWAT CD* values were greater than SWAT CD values for 1995 but not for 1996. The SWAT EF statistic showed lower efficiencies than SWAT EF*. No single trend was observed when comparing SWAT EF and EF* values. For this study, the goal was to obtain model results that produced CD, CD*, EF, and EF* values within ±0.5 of the ideal value of one. The 1995 CD, 1996 CD*, and 1996 EF* results were within this desired range. Overall, median-based objective functions were considered the better model fit estimators for the non-normal data. However, even the median objective functions did not achieve the desired performance level.

The sign and Wilcoxon sign rank tests checked the equal medians null hypothesis for observed and model values at the 0.05 confidence level (eqs. 21-23 and table 6). The nonparametric tests accounted for non-normal data but not for data dependence. There were no null hypothesis rejections for the sign test and for the Wilcoxon sign rank test.

Monthly Equivalent Runoff Depths

Monthly runoff depth statistics were compared with daily results for comparing statistic performance. Normality testing (kurtosis, skewness coefficient, and Shapiro-Wilk test) showed that all monthly data sets could be assumed approximately normally distributed, increasing the number of statistical options available for comparing predicted and observed values (eqs. 1-4 and table 1). Monthly summaries also reduce data point dependence at the price of fewer data points. Lacking Box and Jenkins' (1976) 50-point minimum for correlation, no monthly autocorrelations and cross-correlations were calculated.

Table 4. Least-squares regression results.

Data	Slope	Intercept (cms)	R_r^2
Daily			
1995	0.45	0.11	0.26
1996	0.38	0.10	0.40
Monthly			
1995	0.50	0.02	0.70
1996	0.63	0.00	0.88

Table 5. Mean and median objective function values.

Data	ME	MdAE	RMSE	CD	CD*	EF	EF*
Daily							
1995	2.91	42.69	119.5	1.23	2.08	0.09	0.37
1996	2.84	34.85	114.9	2.22	0.57	0.15	1.32
Monthly							
1995	0.15	25.05	37.2	1.68	0.68	0.41	-0.68
1996	0.26	31.39	41.9	1.53	1.24	0.61	0.57

Table 6. Nonparametric test results.

	Sign Test		Wilcoxon Sign Rank Test	
Data	P-value	Conclusion	P-value	Conclusion
Daily				
1995	0.41	Do not reject h_0	0.84	Do not reject h_0
1996	0.13	Do not reject h_0	0.83	Do not reject h_0
Monthly				
1995	0.22	Do not reject h_0	0.09	Do not reject h_0
1996	0.39	Do not reject h_0	0.62	Do not reject h_0

Overall improved model and generated data fit for monthly totals versus daily flows was partly due to better compliance with normality and data independence assumptions (table 3). The improved model fit results may also have been due to lumping the daily values into monthly totals (i.e., using fewer data points). For monthly SWAT data sets, all Nash-Sutcliffe coefficients (R^2) were closer to one, all had lower D_V statistics, and all *r* correlations were closer to one than the corresponding daily statistics (eqs. 9-11 and table 4). The SWAT monthly α values were much lower than SWAT daily α values (eq. 5 and table 3). All monthly regression and R_r^2 values showed better fit than the corresponding daily regression results (eqs. 6-8 and table 4).

Monthly mean and median objective function results (eqs. 12-19 and table 5) did not exhibit a general trend. Dimensional daily and monthly error terms (ME, MdAE, RMSE) could not be readily compared, but monthly MdAE values were lower than RMSE values. No single pattern emerged comparing monthly CD versus CD* values, but these statistics were generally closer to one than the daily CD and CD* results. For most cases, monthly EF and EF* values were closer to ideal fit than daily EF and EF* values, and monthly CD* and EF* were lower than corresponding CD and EF amounts. The 1995 CD*, 1996 CD*, 1996 EF, and 1996 EF* values were within the target range of ±0.5 away from the perfect fit value of one. The monthly sign test and Wilcoxon sign rank test at the 0.05 confidence level did not show any differences between daily and monthly data (eqs. 21-23 and table 6). There were no null hypothesis rejections for either year's monthly totals.

Combining daily values for monthly totals decreased the data dependence between the individual data points noted for the daily data. Decreased dependence did not prove that the error independence assumption was valid, but the error independence assumption was more valid for monthly values than for daily figures. For approximately normal monthly totals, *t*-tests checked the equal data set means null hypothesis at the 0.05 confidence level (eq. 20). The 1995 *t*-test results showed no null hypothesis rejections. For 1996, the model P-value was slightly less than the required 0.05, but the *t*-test may lack sensitivity to smaller fit errors.

SUMMARY AND CONCLUSIONS

The increased numbers of parameters and outputs from physically based models require additional attention when

assessing model results. Statistical evaluation should carefully consider daily and monthly data characteristics as well as other properties of the model output to customize statistical analysis. Qualitative daily and monthly analysis should be performed to look for general fit problems. More than one statistical test should be implemented to evaluate model performance. Multiple statistical test methods help confirm model fit (or lack thereof). Data normality should be considered in statistical analysis of model results where the statistics used invoke underlying assumptions of normally distributed data.

The daily SWAT results showed good model fit, but timing, peak flow, and recession curve estimations needed improvements. Daily autocorrelation and cross-correlation patterns showed that model predictions were correlated with the observed data, but the correlation could be strengthened. Monthly totals were closer to meeting required statistical assumptions than daily values.

The major pitfalls for daily results analysis were both non-normal and dependent data sets. Many of the statistical techniques evaluated were based on the assumptions of normality and/or independence between the data values. Nonparametric methods and median objective functions were applicable to non-normal data but required independent error terms. Only autocorrelation and cross-correlation statistics explicitly addressed dependent data. More statistical techniques were available for the monthly analysis than for the daily analysis since monthly totals could be assumed approximately normal and data point dependence was reduced. The cost for using monthly totals rather than daily values was having fewer data points.

For evaluating daily model results, median objective functions, sign test, autocorrelation, and cross-correlation were the most appropriate techniques evaluated from the standpoint of evaluating non-normal data sets containing dependence between the data points. None of the evaluated statistics was designed specifically for daily, non-normal, dependent data sets. The CD* and EF* median objective functions were especially suitable for gauging model fit for ease of comparison to a set reference point for ideal model fit. Several other SWAT studies also used the Nash-Sutcliffe coefficient and the product moment correlation coefficient for evaluating model performance versus observed data. These statistics would also be of value in comparing and contrasting a SWAT study with other SWAT modeling endeavors.

Monthly model fit was best estimated using regression coefficients, R_r^2 coefficient, Nash-Sutcliffe coefficient, and *t*-test. Of these techniques, the Nash-Sutcliffe coefficient and the R^2 coefficient were the two methods most often used for evaluation in other SWAT studies. These two statistics also were easily comparable to a fixed reference value of one for perfect model fit.

REFERENCES

Addiscott, T., J. Smith, and N. Bradbury. 1995. Critical evaluation of models and their parameters. *J. Environ. Quality* 24(5): 803-807.

Arnold, J. G., and P. M. Allen. 1996. Estimating hydrologic budgets for three Illinois watersheds. *J. Hydrology* 176: 57-77.

Arnold, J. G., P. M. Allen, and G. Bernhardt. 1993. A comprehensive surface-groundwater flow model. *J. Hydrology* 142: 47-69.

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and J. R. Williams. 1998. Large-area hydrologic modeling and assessment: Part I. Model development. *J. American Water Resources Assoc.* 34(1): 73-89.

Arnold, J. G., L. R. Williams, R. Srinivasan, and K. W. King. 1999. SWAT Soil and Water Assessment Tool Model Theory. Temple, Texas: USDA-ARS Grassland, Soil, and Water Research Laboratory.

ASCE. 1993. Criteria for evaluation of watershed models. *J. Irrig. and Drainage Eng.* 119(3): 429-442.

Bilisoly, R. L., S. E. Nokes, and S. R. Workman. 1997. Statistical treatment of soil chemical concentration data. *J. Environ. Quality* 26(3): 877-883.

Bingner, R. L. 1996. Runoff simulated from Goodwin Creek watershed using SWAT. *Trans. ASAE* 39(1): 85-90.

Box, G. E. P., and G. M. Jenkins. 1976. *Time Series Analysis Forecasting and Control*. Oakland, Cal.: Holden-Day.

Diggle, P. J. 1990. *Time Series: A Biostatistical Introduction.* Oxford, U.K.: Clarendon Press.

Duan, Q, V. K. Gupta, and S. Sorooshian. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* 28(4): 1015-1031.

Freund, R. J., and W. J. Wilson. 1997. *Statistical Methods Revised Edition.* 2nd ed. San Diego, Cal.: Academic Press.

Fuller, W. A. 1996. *Introduction to Statistical Time Series.* 2nd ed. New York, N.Y.: John Wiley and Sons.

Gupta, H. V., S. Sorooshian, and P. O. Yapo. 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research* 34(4): 751-763.

Haan, C. T. 2002, *Statistical Methods in Hydrology.* 2nd ed. Ames, Iowa: Iowa State University Press.

Haan, C. T., B. Allred, D. E. Storm, G. J. Sabbagh, and S. Prabhu. 1995. Statistical procedure for evaluating hydrologic/water quality models. *Trans. ASAE* 38(3): 25-733.

Haefner, J. W. 1997. *Modeling Biological Systems: Principles and Applications.* New York, N.Y.: International Thomson Publishing.

Hirsch, R. M., R. B. Alexander, and R. A. Smith. 1991. Selection of methods for the detection and estimation of trends in water quality. *Water Resources Research* 27(5): 803-813.

Hollander, M., and D. A. Wolfe. 1973. *Nonparametric Statistical Methods.* New York, N.Y.: John Wiley and Sons.

King, K. W., J. G. Arnold, and R. L. Bingner. 1999. Comparison of Green-Ampt and curve number methods on Goodwin Creek watershed using SWAT. *Trans. ASAE* 42(4): 919-925.

Kosky, K. M., and B. A. Engel. 1997. Evaluation of three distributed parameter hydrologic/water quality models. Presented at the 1997 ASAE Annual International Meeting. ASAE Paper No. 972010. St. Joseph, Mich.: ASAE.

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

Loague, K. M., R. E. Green, and L. A. Mulkey. 1988. Evaluation of mathematical models of solute migration and transformation: An overview and an example. In *Proc. of the International Conf. and Workshop on the Validation of Flow and Transport Models for the Unsaturated Zone*, 231-247. Albuquerque, N.M.: University of New Mexico Press.

Martinec, J., and A. Rango. 1989. Merits of statistical criteria for the performance of hydrological models. *Water Resources Bulletin* 25(2): 421-432.

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

Peterson, J. R., and J. M. Hamlett. 1997. Hydrologic calibration of the SWAT model in a watershed containing fragipan soils and wetlands. Presented at the 1997 ASAE Annual International Meeting. ASAE Paper No. 972193. St. Joseph, Mich.: ASAE.

Rosenberger, J. L., and M. Gasko. 1983. Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis*. New York, N.Y.: John Wiley and Sons.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection.* New York, N.Y.: John Wiley and Sons.

Shapiro, S. S., and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52: 591-611.

Smithers, J. C., and B. A. Engel. 1996. An initial assessment of SWAT as a hydrological modeling tool for the mid-west USA. Presented at the 1996 ASAE International Meeting. ASAE Paper No. 962065. St. Joseph, Mich.: ASAE.

Spruill, C. A. 1998. Hydrological assessment and calibration of the SWAT model for small watersheds in central Kentucky. MS thesis. Lexington, Ky.: University of Kentucky, Department of Biosystems and Agricultural Engineering.

Spruill, C. A., S. R. Workman, and J. L. Taraba. 2000. Simulation of daily and monthly stream discharge from small watersheds using the SWAT model. *Trans. ASAE* 43(6): 1431-1439.

Srinivasan, R. and J. G. Arnold. 1994. Integration of a basin-scale water quality model with GIS. *Water Resources Bulletin* 30(3): 453-462.

Zacharias, S., C. D. Heatwole, and C. W. Coakley. 1996. Robust quantitative techniques for validating pesticide transport models. *Trans. ASAE* 39(1): 47-54.