Modifying Goodness-of-Fit Indicators to Incorporate Both Measurement and Model Uncertainty in Model Calibration and Validation

R. D. Harmel, P. K. Smith, K. W. Migliaccio

Abstract. *Because of numerous practical implications of uncertainty in measured data and model predictions, improved techniques are needed to analyze and understand uncertainty and incorporate it into hydrologic and water quality evaluations. In the present study, a correction factor was developed to incorporate measurement uncertainty and model uncertainty in evaluations of model goodness-of-fit (predictive ability). The correction factor, which was developed for pairwise comparisons of measured and predicted values, modifies the typical error term calculation to consider both sources of uncertainty. The correction factor was applied with common distributions and levels of uncertainty (represented by coefficients of variation ranging from 0.026 to 0.256) for each measured value and each predicted value from five example data sets. The modifications resulted in inconsequential changes in goodness-of-fit conclusions for example data sets with very good and poor model simulations, which is both logical and appropriate because very good model performance should not improve greatly and poor model performance should not become satisfactory when uncertainty is considered. In contrast, incorporating uncertainty in example data sets with initially moderate goodness-of-fit resulted in important improvements in indicator values and in model performance ratings. A model evaluation matrix was developed to present appropriate model performance conclusions, considering both model accuracy and precision, based on various levels of measurement and model uncertainty. In cases with highly uncertain calibration/validation data, definitive "good" fit conclusions are cautioned against even with "good" indicator values because of the uncertain standard of comparison; however, in these cases, poor model accuracy can be confidently concluded from "unsatisfactory" indicator values. In contrast, model accuracy can be confidently concluded from goodness-of-fit indicator values in cases with low measurement uncertainty. It is hoped that the modified goodness-of-fit indicators and the model evaluation matrix contribute to improved goodness-of-fit conclusions and to more complete assessments of model performance.*

Keywords. *Index of agreement, Model evaluation, Nash-Sutcliffe coefficient of efficiency, Watershed models.*

The scientific and societal value in determining the uncertainty in hydrologic and water quality modeling and communicating that uncertainty to scientific, regulatory, policy, and public interests has been recently emphasized (Kavetski et al., 2002; Reckhow, 2003; Muñoz-Carpena et al., 2006; Pappenberger and Beven, 2006; Beven, 2006; Shirmohammadi et al., 2006). Presenting uncertainty estimates for model predictions and measured data allows decision-makers to assess and quantify their confidence in the measured and predicted values, which facilitates informed analysis, communication, and decision-making.

Model uncertainty (excluding components related to measurement uncertainty) may be attributed to parameterization (parameter uncertainty), algorithm selection and ability to represent the natural process, and natural process variability based on temporal and spatial scales (Vicens et al., 1975; Beven, 1989; Haan, 1989). Model uncertainty assessment methods include: first-order approximation (Haan, 2002), mean value first-order reliability (Madsen et al., 1986), Monte Carlo simulation (Haan et al., 1995), generalized likelihood uncertainty estimation (Beven and Freer, 2001), dynamically dimensioned search-approximation of uncertainty (Tolson and Shoemaker, 2008), importance sampling (Gelman et al., 1995; Kuczera and Parent, 1998), Latin hypercube sampling (McKay et al., 1979), Markov chain Monte Carlo (Metropolis et al., 1953; Gelman et al., 1995), parameter solution (Duan et al., 1992; Van Griensven and Meixner, 2006), and sequential uncertainty fitting algorithm (Abbaspour et al., 2007). These methods all address parameter uncertainty using different underlying assumptions and thus produce method-specific results, which are difficult to compare quantitatively (see Shirmohammadi et al., 2006; Yang et al., 2008). The present research does not address these alternatives but focuses instead on model evaluation when

Submitted for review in November 2008 as manuscript number SW 7813; approved for publication by the Soil & Water Division of ASABE in February 2010.

Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

The authors are **R. Daren Harmel, ASABE Member Engineer,** Agricultural Engineer, USDA-ARS Grassland Soil and Water Research Laboratory, Temple, Texas; **Patricia K. Smith, ASABE Member Engineer,** Associate Professor, Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas; and **Kati W. Migliaccio, ASABE Member Engineer,** Assistant Professor and Extension Specialist, Department of Agricultural and Biological Engineering, IFAS, University of Florida, Homestead, Florida. **Corresponding author:** R. Daren Harmel, USDA-ARS, 808 E. Blackland Rd., Temple, TX 76502; phone: 254-770-6521; fax: 254-770-6561; e-mail: daren.harmel@ars.usda.gov.

the distributions of uncertainty about each measured value and each modeled value can be reasonably assumed and incorporated into calibration and/or validation.

Model Evaluation

Willmott (1981), Legates and McCabe (1999), Moriasi et al. (2007), and Jain and Sudheer (2008) provide thorough discussions regarding common indicators for evaluating model performance, and others illustrate the application of these indicators (Loague and Green, 1991; Santhi et al., 2001; Van Liew et al., 2003). The degree to which models achieve their primary goal of adequately representing real-world processes is typically judged by pairwise comparison of measured data and model output (Legates and McCabe, 1999) and with graphical comparison of measured and predicted values. As traditionally applied, most quantitative goodness-of-fit indicators use the simple difference to represent the deviation between observed and predicted data (Legates and McCabe, 1999). Harmel and Smith (2007) modified this error calculation to evaluate model goodness-of-fit considering uncertainty in measured calibration and validation data (e.g., discharge or constituent loads). However, no indicators are currently available that consider both measurement uncertainty and model uncertainty in goodness-of-fit evaluation.

In an excellent overview, Engel et al. (2007) defined the content necessary to develop model application protocols (or modeling quality assurance plans), which are needed to enhance the scientific validity of models and to increase the defensibility of model applications in light of regulatory, programmatic, and research implications. In discussing calibration and validation procedures, Engel et al. (2007) emphasized the need to assess model goodness-of-fit and to assess the uncertainty in model results and measured data; however, no methods for assessing goodness-of-fit considering both measurement uncertainty and model uncertainty were provided.

Therefore, the primary objective of this research was to develop a correction factor to incorporate both measurement uncertainty and model uncertainty into goodness-of-fit evaluation in calibration and validation of hydrologic and water quality models. The correction factor was developed to modify the error term between pairs of individual measured and predicted values, not for the comparison of the populations of measured and predicted values. Correction factor development is described, and application results for five example data sets are presented. In addition, a model evaluation matrix was developed and is presented to assist modelers in drawing appropriate goodness-of-fit conclusions. In contrast to the typically employed model evaluation mindset, this matrix considers the inherent measurement uncertainty and model uncertainty in model performance evaluation.

Correction Factor Development

Background

Several goodness-of-fit indicators, which are commonly utilized in pairwise comparison of measured and predicted values, were selected for this study. The selected indicators include the Nash-Sutcliffe coefficient of efficiency (E_{NS}), the index of agreement (d), root mean square error (RMSE), and mean absolute error (MAE). Basic information on each indicator is presented subsequently; more detailed information can be found in Nash and Sutcliffe (1970), Willmott (1981), Legates and McCabe (1999), and Moriasi et al. (2007). The Nash-Sutcliffe coefficient of efficiency was developed as a dimensionless indicator to better evaluate hydrologic and water quality model goodness-of-fit than the coefficient of determination (R^2), which is insensitive to additive and proportional differences between measured and simulated values (Nash and Sutcliffe, 1970). However, both R^2 and E_{NS} are overly sensitive to extreme values because each squares the values of paired differences (Legates and McCabe, 1999). McCuen et al. (2006) and Jain and Sudheer (2008) provide in-depth discussion of the appropriate interpretation and application of E_{NS}. Another widely used dimensionless indicator of hydrologic and water quality model goodness-of-fit is the index of agreement, which was designed by Willmott (1981) to be a measure of the degree to which a model's predictions are error free, not to be a measure of correlation. According to Legates and McCabe (1999), d is better suited for model evaluation than R^2, but it too is overly sensitive to extreme values. The root mean square error and mean absolute error are well-accepted absolute error goodness-of-fit indicators that describe differences in observed and predicted values in the appropriate units (Legates and McCabe, 1999).

As shown in table 1, these indicators all contain an identical error term (deviation calculation), which is the difference between each pair of observed and predicted values (eq. 1). As such, this calculation does not consider the uncertainty in measured data or model output:

$$e_i = O_i - P_i \tag{1}$$

where

e_i = deviation between paired observed and predicted data

O_i = observed (measured) value

P_i = predicted (modeled) value.

Table 1. Traditional calculations for selected goodness-of-fit indicators used in pairwise comparison of measured and predicted values.

Indicator	Equation	Eq. No.
E_{NS}[a]	$E_{NS} = 1 - \dfrac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2}$	(2)
d[b]	$d = 1 - \dfrac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left(\lvert P_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert\right)^2}$	(3)
RMSE[c]	$\text{RMSE} = \sqrt{N^{-1}\sum_{i=1}^{N}(O_i - P_i)^2}$	(4)
MAE[c]	$\text{MAE} = N^{-1}\sum_{i=1}^{N}\lvert O_i - P_i\rvert$	(5)

[a] Nash and Sutcliffe (1970).

[b] Willmott (1981).

[c] Legates and McCabe (1999).

A correction factor developed by Harmel and Smith (2007) modified the error term in equation 1 by incorporating the distribution of measurement uncertainty (eq. 6). This modification enhances goodness-of-fit evaluation in model calibration and validation in the presence of measurement uncertainty, but it does not consider the effect of model (prediction) uncertainty:

$$e(meas)_i = \frac{CF(meas)_i}{0.5} \cdot \left(O_i - P_i\right) \tag{6}$$

where

$e(meas)_i$ = modified deviation considering only measurement uncertainty

$CF(meas)_i$ = non-dimensional correction factor for each measured (O_i) and predicted (P_i) pair based on the probability distribution of each measured value

0.5 = One-sided probability for O_i at mean value assuming a symmetric distribution.

Thus, in the presence of both measurement uncertainty and model uncertainty, each of which can be considerable (Beck, 1987; Harmel et al., 2006), it is more appropriate to evaluate model predictions considering both sources of uncertainty. A correction factor was, therefore, developed to incorporate measurement uncertainty and prediction uncertainty and enhance goodness-of-fit evaluation by producing realistic estimates of the deviations between measured values and model predictions.

Incorporating Both Measurement and Model Uncertainty

The correction factor and resulting error term modification were based on the idea that the deviation between measured and predicted values should be adjusted to incorporate their respective assumed uncertainty distributions. The theoretical basis of the correction factor is that of Haan et al. (1995), which stated that the degree of overlap between corresponding probability density functions (pdfs) for measured and predicted values is indicative of model predictive ability. The closer the measured and predicted values are to one another and/or the greater the variance in their pdfs, the more their respective uncertainty distributions overlap. The degree of overlap is represented by their intersection or joint probability. Assuming that O_i and P_i are independent, the degree of overlap can be determined with basic probability methods, as shown in equation 7 (Haan, 2002) and illustrated in figure 1. Since this method assumes independence between individual observed and predicted values, the method is not applicable for models in which future predictions of the variable of interest are based on historic realizations of that same variable:

$$DO_i = prob\ (O_i\ \mathrm{I}\ \ P_i) = prob(O_i) \cdot\ prob(P_i)$$

$$DO_i = \int_{P_i\ \min}^{P_i\ \max} p_O(o_i)do \cdot \int_{O_i\ \min}^{O_i\ \max} p_P(p_i)dp$$

$$DO_i = [prob(o_i < P_{i\max}) - prob(o_i < P_{i\min})]$$

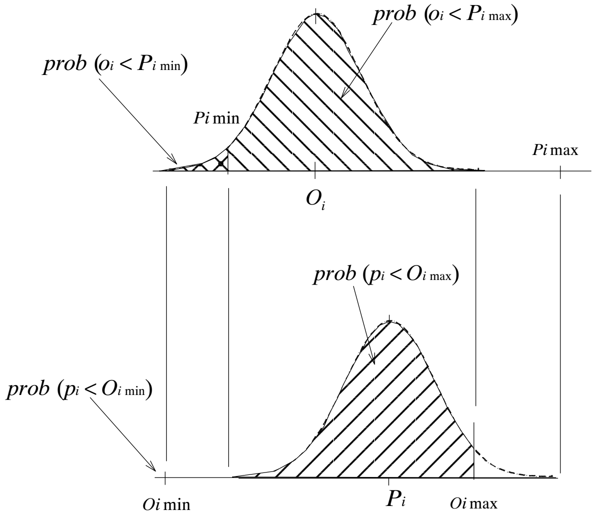
$$\cdot\ [prob(p_i < O_{i\max}) - prob(p_i < O_{i\min})] \tag{7}$$

where

DO_i = degree of overlap for distributions for each measured (O_i) and predicted (P_i) pair

$p_O(o_i)$ = probability density function for the observed value O_i

$p_P(p_i)$ = probability density function for the predicted value P_i.



$prob\ (o_i < P_{i\,\max})$
$prob\ (o_i < P_{i\,\min})$
$P_i\min$
$P_i\max$
O_i
$prob\ (p_i < O_{i\,\max})$
$prob\ (p_i < O_{i\,\min})$
$O_i\min$
P_i
$O_i\max$

Figure 1. Graphical representation of the degree of overlap and resulting correction factor determination for $O_i < P_i$ assuming normal distributions for measurement uncertainty and model uncertainty.

The degree of overlap for each pair of measured and predicted values is then used to determine the correction factor, which ranges from 0 to 1.0 (eq. 8). The smaller the degree of overlap, the larger the correction factor becomes, and vice versa:

$$CF(meas + pred)_i = 1 - DO_i \tag{8}$$

where $CF(meas+pred)_i$ is the correction factor that incorporates measurement and prediction uncertainty for each measured (O_i) and predicted (P_i) pair.

Thus, the resulting error term modification (eq. 9) considers both measurement uncertainty and model uncertainty. This modified error term can then be substituted for the simple O_i - P_i error term to modify the traditional calculations for goodness-of-fit indicators (eqs. 2, 3, 4, and 5):

$$e(meas + pred)_i = CF(meas + pred)_i \cdot\ \left(O_i - P_i\right) \tag{9}$$

where $e(meas+pred)_i$ is the modified deviation for each measured (O_i) and predicted (P_i) pair.

This correction factor was designed for pairwise comparisons of individual measured and predicted values. The current application assumed a single value for each measured data point and assumed deterministic (single value) model output for the corresponding model prediction. However, repeated measurements and/or stochastic modeling that produce multiple measured and/or predicted values or a distribution of predicted values are also appropriate for estimating mean and standard deviations with which to determine the uncertainty boundaries for correction factor calculation. In this case, pairwise comparisons are made between the measures of central tendency for each measured and/or predicted value.

It is important to note that in either case, the probability distributions utilized are those for individual values of O_i or P_i, not for the entire population of measured or predicted values. It is also important to note that the error calculation in equation 6 includes only measurement uncertainty and that the error calculation in equation 9 includes both measurement uncertainty and model uncertainty. Thus, equation 6 should be used only when measurement uncertainty alone is considered, and equation 9 should be used only when both measurement uncertainty and model uncertainty are considered. The values of $e(meas)_i$ and $e(meas{+}pred)_i$ are not comparable and were not designed to be so.

Application to Example Data Sets

Correction Factor Determination for Selected Uncertainty Distributions

The procedures to calculate the correction factor for several common distributions (normal, uniform, lognormal) are summarized subsequently. Little if any information is available regarding the distributional properties of measurement uncertainty for individual values, although the uncertainty associated with various procedures is fairly well documented (e.g., Pelletier, 1988; Kotlash and Chessman, 1998; Harmel et al., 2009). Model uncertainty information is also limited, but distributional properties for the population of predicted values (e.g., Haan and Skaggs, 2003a, 2003b; Shirmohammadi et al., 2006; Migliaccio and Chaubey, 2008) and upper and lower boundaries for each predicted value (e.g., Muleta and Nicklow, 2005; Shen et al., 2008) have been presented. Thus, in the absence of comprehensive information describing appropriate distributions for measurement and model uncertainty for individual values, the normal, uniform, and lognormal distributions were assumed to be appropriate for the current application; however, any appropriate distribution may be used in subsequent applications of the correction factor method (eqs. 7, 8, and 9).

The normal, lognormal, and uniform distributions are all two-parameter distributions either described by or derived from the mean and standard deviation. The means for O_i and P_i were set at each measured and predicted value, respectively. The standard deviations were calculated from the coefficients of variation (Cv), as shown in equation 10 (Haan, 2002):

$$\mathrm{Cv} = \frac{s_x}{\bar{x}} \tag{10}$$

where

s_x = sample standard deviation

$\bar{x}$ = sample mean.

The uncertainty boundaries ($O_{i\,\max}$, $O_{i\,\min}$, $P_{i\,\max}$, and $P_{i\,\min}$) or limits in equation 7 were estimated for the normal and lognormal distributions assuming that they occur at the 0.0001 and 0.9999 probabilities. For the uniform distribution, the uncertainty boundaries (the uniform distribution parameters α and β) were estimated by equation 11:

$$\alpha = \bar{x} - \sqrt{3}s_x$$

$$\beta = \bar{x} + \sqrt{3}s_x \tag{11}$$

The coefficient of variation is one of the most common ways to express uncertainty in measured and modeled data (Haan et al., 1995; Hession et al., 1996; Tyagi and Haan, 2001; Haan and Skaggs, 2003a, 2003b), although other expressions of uncertainty such as the probable error range (±%) used in Harmel et al. (2006, 2009) and Harmel and Smith (2007) are also appropriate. Expressing uncertainty with Cv values also allows negative values in the distributions to be avoided, as recommended by Tyagi and Haan (2001). Four Cv values (0.026, 0.085, 0.192, and 0.256) were selected to evaluate the goodness-of-fit impact of various levels of uncertainty (from low, Cv = 0.026; to moderate, Cv = 0.085 and 0.192; to high, Cv = 0.256) about each measured value and each predicted value. These Cv values represent the uncertainty about each measured value and each predicted value, and as such they are expected to be smaller than for the populations of measured and predicted values. Only results for the maximum and minimum uncertainty estimates (Cv = 0.026 and 0.256) are presented in table 3 because these results encompass those for Cv = 0.085 and Cv = 0.192.

Description of Example Data Sets

The correction factor method was applied to determine goodness-of-fit indicator values for five example data sets selected to represent a range of time scales, data types, simulation models, and goodness-of-fit results (table 2). The

Table 2. Summary information for example data sets.

Study Site (Location)	Data Type (units)	Model Used (Reference)	$\bar{O}$ [a]	$\bar{P}$ [b]	n[c]
Riesel Field Y6 (Texas)	Monthly runoff (mm)	EPIC (Williams and Sharpley, 1989)	23.3	22.6	48
Riesel Field Y6 (Texas)	Monthly dissolved P load (kg/ha)	EPIC (Williams and Sharpley, 1989)	0.03	0.06	48
Reynolds Creek watershed (Idaho)	Daily streamflow (m^3/s)	SWAT (Arnold et al., 1998)	0.68	0.71	1827
South Fork watershed (Iowa)	Daily streamflow (ft^3/s)	SWAT (Arnold et al., 1998)	0.56	0.55	552
Medina River watershed (Texas)	Daily streamflow (ft^3/s)	HSPF (Bicknell et al., 1997)	4.62	3.98	730

[a] $\bar{O}$ = mean of observed (measured) values.

[b] $\bar{P}$ = mean of predicted values.

[c] Number of paired measured and predicted values.

Table 3. Goodness-of-fit indicator values as traditionally calculated and as modified with the correction factor to account for measurement uncertainty and model uncertainty. Results are shown for all three selected distributions (normal, uniform, and log normal) but only for the maximum and minimum uncertainties (Cv = 0.026 and 0.256).

Study Site (Data Type)	Indicator	Traditional Calculation	Normal Cv = 0.026	Uniform Cv = 0.026	Log Normal Cv = 0.026	Normal Cv = 0.256	Uniform Cv = 0.256	Log Normal Cv = 0.256
Field Y6 (monthly runoff)	*d*	0.99	0.99	0.99	0.99	1.00	0.99	1.00
	E_{NS}	0.96	0.96	0.96	0.96	0.99	0.98	0.99
	E_{NS} (rating)[a]	Very good	Very good	Very good	Very good	Very good	Very good	Very good
	RMSE (mm)	8.78	8.72	8.76	8.72	4.81	5.95	4.74
	MAE (mm)	5.24	5.05	5.18	5.05	2.16	3.45	2.03
Field Y6 (monthly dissolved P load)	*d*	0.76	0.76	0.76	0.76	0.81	0.77	0.82
	E_{NS}	-1.40	-1.40	-1.40	-1.40	-0.89	-1.34	-0.83
	E_{NS} (rating)	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.
	RMSE (kg/ha)	0.08	0.08	0.08	0.08	0.07	0.08	0.07
	MAE (kg/ha)	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Reynolds Creek watershed (daily streamflow)	*d*	0.92	0.92	0.92	0.92	0.96	0.94	0.97
	E_{NS}	0.73	0.73	0.73	0.73	0.87	0.78	0.88
	E_{NS} (rating)	Good	Good	Good	Good	Very good	Very good	Very good
	RMSE (m^3/s)	0.66	0.66	0.66	0.66	0.46	0.58	0.43
	MAE (m^3/s)	0.35	0.35	0.35	0.35	0.20	0.30	0.17
South Fork watershed (daily streamflow)	*d*	0.72	0.72	0.72	0.72	0.77	0.73	0.78
	E_{NS}	-0.01	0.00	-0.01	0.00	0.17	0.04	0.20
	E_{NS} (rating)	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Unsatis.
	RMSE (ft^3/s)	0.77	0.77	0.77	0.77	0.70	0.75	0.68
	MAE (ft^3/s)	0.43	0.43	0.43	0.43	0.33	0.40	0.31
Medina River watershed (daily streamflow)	*d*	0.88	0.88	0.88	0.88	0.93	0.90	0.93
	E_{NS}	0.49	0.49	0.49	0.49	0.72	0.58	0.72
	E_{NS} (rating)	Unsatis.	Unsatis.	Unsatis.	Unsatis.	Good	Satisfactory	Good
	RMSE (ft^3/s)	8.65	8.65	8.65	8.65	6.49	7.85	6.42
	MAE (ft^3/s)	1.77	1.75	1.77	1.75	0.88	1.44	0.75

[a] Qualitative model performance ratings for E_{NS} (unsatisfactory = $E_{NS} \leq 0.50$, satisfactory = $0.50 < E_{NS} \leq 0.65$, good = $0.65 < E_{NS} \leq 0.75$, and very good = $E_{NS} > 0.75$) were determined from Moriasi et al. (2007).

Erosion Productivity Impact Calculator (EPIC) example represented very good results (Riesel Y6 runoff) and poor results (Riesel Y6 dissolved P load). One of the Soil and Water Assessment Tool (SWAT) examples represented poor model performance due to structural deficiencies (South Fork streamflow), which were corrected by Green et al. (2006). The other SWAT example (Reynolds Creek streamflow) and the Hydrologic Simulation Program Fortran (HSPF) example (Medina River streamflow) represented typical hydrologic conditions with moderate predictions.

Results of Application to Example Data Sets

Model Performance Evaluation

The results for application of the correction factor with four levels of uncertainty (as represented by four Cv values from 0.026 to 0.256) and three distributions about each measured and predicted value for five example data sets appear subsequently. The discussion focuses on goodness-of-fit determination based on indicator values as traditionally calculated and as calculated with a correction factor that incorporates both measurement uncertainty and model uncertainty (table 3). However, graphical analysis is also presented for one example data set (figs. 2 and 3) because of the value of coupling quantitative and visual evaluations of model performance.

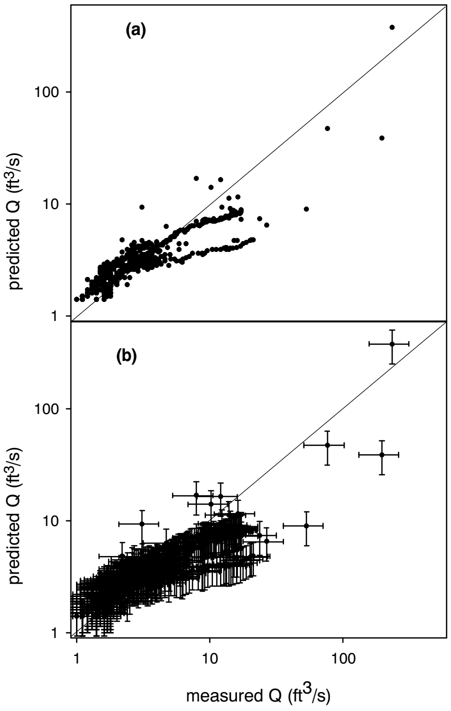


Figure 2. Measured versus simulated daily streamflow for the Medina River: (a) as a simple scatter plot, and (b) as a scatter plot with uncertainty boundaries assuming a uniform distribution and a Cv of 0.192 for each measured and simulated value.

In the example with very good fit between measured and predicted data (EPIC - field Y6 monthly runoff), the correction factor produced inconsequential changes in the E_{NS}, *d*, RMSE, and MAE indicator values compared to the traditional calculations. As shown in table 3, little to no improvement in indicator values relative to their traditional calculations occurred with low uncertainty (Cv = 0.026), but noticeable improvement in already good indicator values occurred as Cv values increased. All E_{NS} values, as

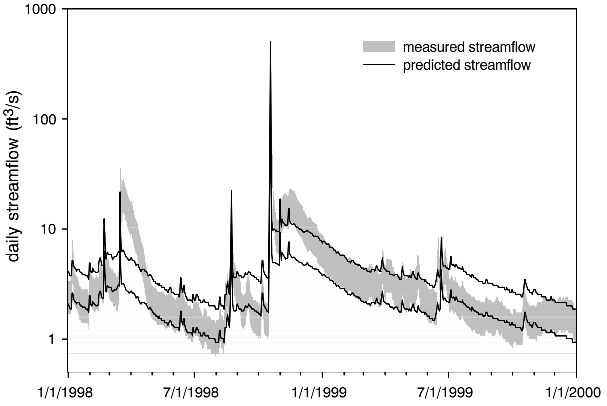


Figure 3. Measured and simulated daily streamflow hydrograph for the Medina River example assuming a uniform distribution and a Cv of 0.192 for each measured and simulated value. The uncertainty boundaries are presented as the shaded area for measured streamflow and as the upper and lower boundary lines for simulated streamflow.

traditionally calculated and as modified in the current application, produced very good model performance ratings based on Moriasi et al. (2007). Since this example was chosen for its very good model performance, it was logical and appropriate that incorporating measurement uncertainty and model uncertainty did not change any goodness-of-fit conclusions.

In examples with poor simulation of measured data (EPIC - field Y6 monthly dissolved P load; SWAT - South Fork daily streamflow), the correction factor produced even smaller indicator value changes than were produced in the very good fit example (table 3). Incorporating measurement and model uncertainty in E_{NS} calculations did not change the unsatisfactory model performance rating (Moriasi et al., 2007) for any distribution at any uncertainty level. The minimal indicator value changes and the unchanging performance ratings are very important because incor–porating measurement uncertainty and model uncertainty should not prevent poor goodness-of-fit conclusions. Similarly, uncertainty should not be used to justify the utility of a model application with obviously poor representation of measured data.

In contrast to the poor simulation examples, the correction factor produced important changes in goodness-of-fit indicator values when applied to example data sets with moderate model performance (SWAT–Reynolds Creek daily streamflow; HSPF–Medina River daily streamflow). Noticeable improvement occurred for all indicator values, especially as uncertainty increased to Cv = 0.256. These indicator value changes were important because their magnitudes tended to be greater than for the other examples, especially the poor simulation examples, and because the changes crossed model performance rating thresholds (e.g., satisfactory to good) based on Moriasi et al. (2007). The modified E_{NS} values improved model performance ratings from good to very good for SWAT prediction of daily streamflow on Reynolds Creek and from unsatisfactory to satisfactory or good performance for HSPF prediction of daily streamflow on the Medina River. The other examples either presented such very good or very poor model results that incorporating measurement and model uncertainty had little effect on goodness-of-fit conclusions. In contrast, incorporating uncertainty in these two examples with fair (moderate) model performance produced important and appropriate changes in goodness-of-fit conclusions.

The graphical illustration of model performance for the Medina River data set (figs. 2 and 3) was included to emphasize the value of additional insight not provided by the quantitative goodness-of-fit indicators. In this example data set, reasonable goodness-of-fit was observed throughout the range of measured values, although higher flows were not simulated as accurately as lower flows (fig. 2a). When uncertainty was included, the uncertainty boundaries overlapped the 1:1 line for most values between 1 and 5 ft^3/s, for fewer values between 5 and 20 ft^3/s, but for very few values larger than 20 ft^3/s (fig. 2b). Thus, model performance became poorer as stream flow increased, which is an important insight not provided by the quantitative indicators. Figure 3 illustrates good model fit in most flow conditions, except for very high flow and for flow recession following high flow.

Value of Multiple Model Evaluation Methods

Another important consideration in model evaluation is the value of utilizing multiple evaluation methods, including graphical techniques and quantitative goodness-of-fit indicators, to assess overall model performance (Willmott, 1981; Legates and McCabe, 1999; Moriasi et al., 2007; Jain and Sudheer, 2008), the benefits of which are clearly shown in the Medina River example (figs. 2 and 3). Ideally, model assessment performance should include graphical techniques, at least one relative indicator (e.g., E_{NS} or d), and at least one absolute indicator (e.g., RMSE or MAE). Although applying multiple model evaluation techniques requires additional calculations and can produce mixed results, it produces more complete model assessments. Such assessments would no doubt benefit from development of model performance ratings for all of the typically applied indicators, similar to Moriasi et al. (2007) for E_{NS}, to reduce the subjectivity in rating performance as good, satisfactory, unsatisfactory, etc. It is important to note that the present methodology incorporates measurement uncertainty and model uncertainty, and thus provides valuable supplemental information to be used in conjunction with, not instead of, traditionally applied statistical and graphical model evaluation methods.

Effects of Distributional Choices

As shown in table 3, the selected uncertainty distribution affected the magnitude of improvement in goodness-of-fit indicator values. For equal uncertainty estimates (e.g., Cv = 0.256), the magnitude of goodness-of-fit improvement was consistently greater for the normal and log normal distributions than for the uniform distribution. This occurred because the uniform distribution bounds are narrower than the normal distribution (theoretically unbounded in both tails of the distribution) and the lognormal distribution (theoretically unbounded in the right tail). However, the distributional assumption is not as important in uncertainty analysis as good estimates of the means and standard deviations (Haan et al., 1998). The magnitude of the uncertainty in the form of increasing Cv values (and subsequently increasing standard deviations) had a much larger impact on goodness-of-fit improvement.

Table 4. Model evaluation matrix for appropriate model performance conclusions in model calibration/validation considering both measurement uncertainty and prediction uncertainty.

Case	Uncertainty in Measured Data	Uncertainty in Model Predictions	Overall Model Performance Conclusions Based on Model Accuracy (goodness-of-fit) and Model Precision	
			"Good" Indicator Values	"Unsatisfactory" Indicator Values
1	High	Low	High model precision, but high measurement uncertainty prevents definitive model accuracy conclusion in spite of good fit indication.	Unsatisfactory model performance due to poor accuracy in spite of high model precision.
2	High	High	Low model precision, but high measurement uncertainty prevents definitive model accuracy conclusion in spite of good fit indication.	Unsatisfactory model performance due to low precision and poor accuracy.
3	Low	High	Low model precision, but good model accuracy.	Unsatisfactory model performance due to low precision and poor accuracy.
4	Low	Low	Good model performance in terms of high precision and good accuracy.	Unsatisfactory model performance due to poor accuracy in spite of high model precision.

Discussion

Model Evaluation Matrix

Because of the ever present, yet rarely considered, effect of uncertainty on judgments of model performance, a model evaluation matrix was developed to assist modelers in making appropriate model performance conclusions (table 4). The matrix presents appropriate conclusions for overall model performance that include both model accuracy (goodness-of-fit) and model precision, as both are important considerations in evaluating and reporting model performance. The model evaluation matrix presents four general cases based on qualitative indications of the uncertainty in measured calibration/validation data and in model predictions.

As shown in table 4, accounting for highly uncertain calibration/validation data can preclude definitive goodness-of-fit conclusions. In cases 1 and 2, although "good" indicator values indicate good fit between model predictions and measured data, definitive model accuracy conclusions are inappropriate because of the highly uncertain standard of comparison. Thus, definitive "good" conclusions are cautioned against in these cases, even though it is entirely possible that the model may actually represent true field conditions as well as or better than the measured data. In contrast, when "unsatisfactory" indicator values result even with the consideration of high measurement uncertainty, then poor model accuracy can be confidently concluded. When calibration/validation data have low uncertainty (cases 3 and 4), then model goodness-of-fit (accuracy) can be confidently concluded because the standard of comparison is relatively certain. As shown in table 4, "good" indicator values definitively indicate good model accuracy, and "unsatisfactory" indicator values definitively indicate poor model accuracy.

In either of the cases with high model prediction uncertainty (cases 2 and 3), low model precision is an important model deficiency regardless of the goodness-of-fit conclusion. This deficiency is especially important when coupled with poor fit ("unsatisfactory" indicator values), which together indicate low precision and poor accuracy and overall poor model performance. In the cases with low prediction uncertainty (cases 1 and 4), high model precision is apparent, and evaluation of model accuracy becomes more important.

Conclusions

The methodology presented herein produced a correction factor to modify the traditional error term in commonly applied model goodness-of-fit indicators by incorporating both measurement (data) uncertainty and model (prediction) uncertainty. When applied to example data sets with very good and poor model simulations, the correction factor produced inconsequential changes in goodness-of-fit results, which is exactly what should occur in these circumstances. In contrast, the correction factor produced important improvements in goodness-of-fit indicator values and in model performance ratings for data sets with moderate agreement. This again is exactly what should occur when uncertainty is considered in model simulations that are reasonable but not great.

Accounting for measurement uncertainty and model uncertainty with this methodology can improve model calibration by reducing the likelihood of "overcalibration." Overcalibration, which can occur when attempting to reach "target" indicator value thresholds (such as Moriasi et al., 2007) to conclude satisfactory model calibration, can make the model less representative of the real-world system and more dependant on the calibration data set. Application of the modified goodness-of-fit indicators and the model evaluation matrix, along with traditionally applied statistical and graphical methods, should also facilitate enhanced goodness-of-fit conclusions and more complete assessments of model performance, given that the presence of measurement uncertainty and prediction uncertainty cannot be debated.

Acknowledgements

The authors appreciate the questions and input from Devendra Amatya, Ken Reckhow, Daniel Moriasi, and others at the 2007 ASABE Watershed Management and TMDL Conference. They encouraged us to pursue this research by stressing the need for and presenting ideas about modified model evaluation techniques to consider both measurement and model uncertainty. Similarly, Martin Futter and Nikki Baggaley of the Macaulay Institute, Aberdeen, Scotland, are recognized for their input into methods to describe the relative magnitudes of measurement and model uncertainty.

References

Abbaspour, K. C., J. Yang, I. Maximov, R. Siber, K. Bogner, J. Mieleitner, J. Zobrist, and R. Srinivasan. 2007. Modelling of hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. *J. Hydrol*. 333(2-4): 413-430.

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.

Beck, M. B. 1987. Water quality modeling: A review of the analysis of uncertainty. *Water Resources Res.* 23(8): 1393-1442.

Beven, K. 1989. Changing ideas in hydrology: The case of the physically-based models. *J. Hydrol.* 105(1-2): 157-172.

Beven, K. 2006. On undermining the science? *Hydrol. Proc.* 20(14): 3141-3146.

Beven, K., and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* 249(1-4): 11-29.

Bicknell, B. R., J. C. Imhoff, J. L. Kittle Jr., A. S. Donigian Jr., and R. C. Johanson. 1997. *Hydrological Simulation Program FORTRAN: User's Manual for Version 11.* EPA/600/R-97/080. Research Triangle Park, N.C.: U.S. Environmental Protection Agency, National Exposure Research Laboratory.

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

Engel, B., D. Storm, M. White, J. Arnold, and M. Arabi. 2007. A hydrologic/water quality model application protocol. *J. American Water Resources Assoc.* 43(5): 1223-1236.

Gelman, S., J. B. Carlin, H. S. Stren, and D. B. Rubin. 1995. *Bayesian Data Analysis.* New York, N.Y.: Chapman and Hall.

Green, C. H., M. D. Tomer, M. Di Luzio, and J. G. Arnold. 2006. Hydrologic evaluation of the Soil and Water Assessment Tool for a large tile-drained watershed in Iowa. *Trans. ASABE* 49(2): 413-422.

Haan, C. T. 1989. Parametric uncertainty in hydrologic modeling. *Trans. ASAE* 32(1): 137-146.

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

Haan, C. T., B. Allred, D. E. Storm, G. J. Sabbagh, and S. Prahhu. 1995. Statistical procedure for evaluating hydrologic/water quality models. *Trans. ASAE* 38(3): 725-733.

Haan, C. T., D. E. Storm, T. Al-Issa, S. Prabhu, G. J. Sabbagh, and D. R. Edwards. 1998. Effect of parameter distributions on uncertainty analysis of hydrologic models. *Trans. ASABE* 41(1): 65-70.

Haan, P. K., and R. W. Skaggs. 2003a. Effect of parameter uncertainty on DRAINMOD predictions: I. Hydrology and yield. *Trans. ASAE* 46(4): 1061-1067.

Haan, P. K., and R. W. Skaggs. 2003b. Effect of parameter uncertainty on DRAINMOD predictions: II. Nitrogen loss. *Trans. ASAE* 46(4): 1069-1075.

Harmel, R. D., and P. K. Smith. 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *J. Hydrol.* 337(3-4): 326-336.

Harmel, R. D., R. J. Cooper, R. M. Slade, R. L. Haney, and J. G. Arnold. 2006. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Trans. ASABE* 49(3): 689-701.

Harmel, R. D., D. R. Smith, K. W. King, and R. M. Slade. 2009. Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling applications. *Environ. Modelling Software* 24(7): 832-842.

Hession, W. C., D. E. Storm, C. T. Haan, S. L. Burks, and M. D. Matlock. 1996. A watershed-level ecological risk assessment methodology. *Water Resources Bull.* 32(5): 1039-1054.

Jain, S. K., and K. P. Sudheer. 2008. Fitting of hydrologic models: A close look at the Nash-Sutcliffe index. *J. Hydrol. Eng.* 13(10): 981-986.

Kavetski, D., S. W. Franks, and G. Kuczera. 2002. Confronting input uncertainty in environmental modelling. In *Calibration of Watershed Models*, 49-68. AGU Water Science and Applications Series, vol. 6. S. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte, eds. Washington, D.C.: American Geophysical Union.

Kotlash, A. R., and B. C. Chessman. 1998. Effects of water sample preservation and storage on nitrogen and phosphorus determinations: Implications for the use of automated sampling equipment. *Water Res.* 32(12): 3731-3737.

Kuczera, G., and E. Parent. 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *J. Hydrol.* 211(1-4): 69-85.

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 Res.* 35(1): 233-241.

Loague, K., and R. E. Green. 1991. Statistical and graphical methods for evaluating solute transport models: Overview and application. *J. Contam. Hydrol.* 7(1-2): 51-73.

Madsen, H. O., S. Krenk, and N. C. Lind. 1986. *Methods of Structural Safety.* Englewood Cliffs, N.J.; Prentice-Hall.

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

McKay, A. D., W. J. Conover, and R. J. Beckman. 1979. A comparison of three methods for selecting values on input variables in the analysis of output from a computer code. *Technometrics* 21(6): 239-245.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chemical Physics* 21(6): 1087-1092.

Migliaccio, K. W., and I. Chaubey. 2008. Spatial distributions and stochastic parameter influences on SWAT flow and sediment predictions. *J. Hydrol. Eng.* 13(4): 258-269.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50(3): 885-900.

Muleta, M. K., and J. W. Nicklow. 2005. Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. *J. Hydrol.* 306: 127-145, doi:10.1016/j.jhydrol .2004.09.005.

Muñoz-Carpena, R., G. Vellidis, A. Shirmohammadi, and W. W. Wallender. 2006. Evaluation of modeling tools for TMDL development and implementation. *Trans. ASABE* 49(4): 961-965.

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

Pappenberger, F., and K. J. Beven. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Res.* 42(5): W05302.

Pelletier, P. M. 1988. Uncertainties in the single determination of river discharge: A literature review. *Canadian J. Civil Eng.* 15(5): 834-850.

Reckhow, K. J. 2003. On the need for uncertainty assessment in TMDL modeling and implementation. *J. Water Resources Planning Mgmt.* 129(4): 245-246.

Santhi, C., J. G. Arnold, J. R. Williams, W. A. Dugas, R. Srinivasan, and L. M. Hauck. 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. *J. American Water Resources Assoc.* 37(5): 1169-1188.

Shen, Z., Q. Hong, H. Yu, and R. Liu. 2008. Parameter uncertainty analysis of the non-point source pollution in the Daning River watershed of the Three Gorges Reservoir Region, China. *Sci. Total Environ.* 405(1-3): 195-205.

Shirmohammadi, A., I. Chaubey, R. D. Harmel, D. D. Bosch, R. Muñoz-Carpena, C. Dharmasri, A. Sexton, M. Arabi, M. L. Wolfe, J. Frankenberger, C. Graff, and T. M. Sohrabi. 2006. Uncertainty in TMDL models. *Trans. ASABE* 49(4): 1033-1049.

Tolson, B. A., and C. A. Shoemaker. 2008. Efficient prediction uncertainty approximation in the calibration of environmental simulation models. *Water Resources Res.* 44: W0441 doi:10.1029/2007WR005869.

Tyagi, A., and C. T. Haan. 2001. Uncertainty analysis using corrected first-order approximation method. *Water Resources Res.* 37(6): 1847-1858.

Van Griensven, A., and T. Meixner. 2006. Methods to quantify and identify the sources of uncertainty for river basin water quality models. *Water Sci. Tech.* 53(1): 51-59.

Van Liew, M. W., J. G. Arnold, and J. D. Garbrecht. 2003. Hydrologic simulation on agricultural watersheds: Choosing between two models. *Trans. ASAE* 46(6): 1539-1551.

Vicens, G. J., I. Rodriguez-Iturbe, and J. C. Shaake. 1975. A Bayesian framework for the use of regional information in hydrology. *Water Resources Res.* 11(3): 405-414.

Williams, J. R., and A. N. Sharpley, eds. 1989. EPIC - Erosion/Productivity Impact Calculator: 1. Model Documentation. USDA Technical Bulletin No. 1768. Washington, D.C.: USDA-ARS.

Willmott, C. J. 1981. On the validation of models. *Physical Geographer* 2(2): 184-194.

Yang, J., P. Reichert, K. C. Abbaspour, J. Xia, and H. Yang. 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *J. Hydrol.* 358: 1-23, doi: 10.1016/j.jhydrol.2008.05.012.