Journal of Hydrology (2007) 337, 326–336

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling

R. Daren Harmel [a,*], Patricia K. Smith [b]

[a] USDA-ARS, 808 E. Blackland Road, Temple, TX 76502, United States
[b] Biological and Agricultural Engineering Department, Texas A&M University, College Station, TX 77843, United States

Received 24 August 2006; received in revised form 29 January 2007; accepted 31 January 2007

KEYWORDS
Model calibration;
Model validation;
Statistics;
Nash–Sutcliffe;
Index of agreement

Summary As hydrologic and water quality (H/WQ) models are increasingly used to guide water resource policy, management, and regulation, it is no longer appropriate to disregard uncertainty in model calibration, validation, and evaluation. In the present research, the method of calculating the error term in pairwise comparisons of measured and predicted values was modified to consider measurement uncertainty with the goal of facilitating enhanced evaluation of H/WQ models. The basis of this method was the theory that H/WQ models should not be evaluated against the values of measured data, which are uncertain, but against the inherent measurement uncertainty. Specifically, the deviation calculations of several goodness-of-fit indicators were modified based on the uncertainty boundaries (Modification 1) or the probability distribution of measured data (Modification 2). The choice between these two modifications is based on absence or presence of distributional information on measurement uncertainty. Modification 1, which is appropriate in the absence of distributional information, minimizes the calculated deviations and thus produced substantial improvements in goodness-of-fit indicators for each example data set. Modification 2, which provides a more realistic uncertainty estimate but requires distributional information on uncertainty, resulted in smaller improvements. Modification 2 produced small goodness-of-fit improvement for measured data with little uncertainty but produced modest improvement when data with substantial uncertainty were compared with both poor and good model predictions. This limited improvement is important because poor model goodness-of-fit, especially due to model structure deficiencies, should not appear satisfactory simply by including measurement uncertainty.

* Corresponding author. Tel.: +1 254 770 6521, fax: +1 254 770 6561.
E-mail address: dharmel@spa.ars.usda.gov (R. Daren Harmel).

0022-1694/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2007.01.043

Introduction

Because of the uncertainty associated with hydrologic and water quality (H/WQ) modeling, uncertainty should be accounted for in model application and evaluation (Kavetski et al., 2002; Pappenberger and Beven, 2006; Beven, 2006b). The analysis and consideration of uncertainty is particularly important because decisions regarding water resource policy, management, regulation, and program evaluation are increasingly based on H/WQ modeling (Beck, 1987; Haan et al., 1995; Sharpley et al., 2002; Shirmohammadi et al., 2006). In the US, the issue of uncertainty is especially relevant in several current Federal water quality programs; two examples are Total Maximum Daily Load projects (TMDLs) and the recently initiated USDA conservation program evaluation titled the Conservation Effects Assessment Project (CEAP). In typical TMDL projects, measured H/WQ data are used to calibrate watershed models and evaluate their ability to reproduce the system being modeled. Source load allocation, which is often based on model results, is required by Federal regulation (40 CFR 130.7) to include a margin of safety to account for uncertainty. In CEAP, H/WQ models are being applied to estimate conservation benefits in selected watersheds and at the national scale. The CEAP objectives acknowledge the importance of uncertainty in H/WQ modeling and include an assessment of model prediction uncertainties at multiple scales (USDA-ARS, 2004). In Europe, the realization of substantial uncertainty related to water characterization and modeling has been cited to support adaptive management in the Water Framework Directive (e.g. Galaz, 2005; UK Administrators, 2005; Harris and Heathwaite, 2005) and to promote the use of uncertainty estimation as routine in H/WQ science (Pappenberger and Beven, 2006; Beven, 2006b).

Uncertainty in H/WQ modeling has been classified by Vicens et al. (1975) into three categories: model uncertainty, parameter uncertainty, and uncertainty inherent in natural processes. The uncertainty introduced by model structure and parameterization has received much attention in recent years (e.g. Haan, 1989; Kuczera and Parent, 1998; Beven and Freer, 2001; Bashford et al., 2002; Haan and Skaggs, 2003a,b; Beven, 2006a). Simply speaking, model uncertainty arises from incomplete understanding of the system being modeled and/or the inability to accurately reproduce H/WQ processes with mathematical and statistical techniques. In contrast, parameter uncertainty results from incomplete knowledge of parameter values, ranges, physical meaning, and temporal and spatial variability. But parameter uncertainty also reflects the incomplete model representation of H/WQ processes (model uncertainty) and inadequacies of parameter estimation techniques in light of uncertain, and often limited, measured data. For additional information on model and parameter uncertainty, which is beyond the scope of this paper, the reader is referred to Beck (1987), Haan (1989), Reckhow (1994), Haan et al. (1995), Hession and Storm (2000), Kavetski et al. (2002), and Beven (2006a).

Although the uncertainty inherent in measured data used to calibrate and validate model predictions is commonly acknowledged, measurement uncertainty is rarely included in the evaluation of model performance. One reason for this omission is the lack of data on the uncertainty inherent in measured H/WQ data. Although sources such as Pelletier (1988) and Sauer and Meyer (1992) provide excellent reviews of errors associated with streamflow measurement, information on errors associated with water quality data has only recently been published (e.g. Robertson and Roerish, 1999; Haggard et al., 2003; Harmel and King, 2005). Another reason for this omission is the previous lack of scientific data and guidance on analysis of uncertainty in measured data. Recently, however, Harmel et al. (2006) provided fundamental data and recommendations for the consideration of uncertainty in measured H/WQ data. The authors documented common sources of error in measured streamflow and water quality data on small watersheds and presented a method for determining cumulative probable error resulting from the various procedural steps required in data collection. In spite of this advancement, more research and attention on estimating measurement uncertainty for current data and retrospectively for past data are needed. In addition, uncertainty estimation to accompany all measured H/WQ data would be a beneficial outcome leading to improved scientific and stakeholder understanding and decision-making (Beven, 2006b; Pappenberger and Beven, 2006).

Goodness-of-fit indicators

Pairwise comparison of measured data and model predictions is typically used to judge the ability of H/WQ models to achieve their primary goal of adequately representing processes of interest (Legates and McCabe, 1999). Many of the common quantitative goodness-of-fit indicators, simply use the absolute value of the difference or the squared value of the difference to represent the deviation between paired measured and predicted data. Four common goodness-of-fit indicators for H/WQ model evaluation, Nash–Sutcliffe coefficient of efficiency, index of agreement, root mean square error, and mean absolute error, are no exception. These selected indicators are briefly described in the following section but are thoroughly discussed in Willmott (1981), Legates and McCabe (1999), and Moriasi et al. (in review).

Nash–Sutcliffe coefficient of efficiency

The Nash–Sutcliffe coefficient of efficiency, E, is a dimensionless indicator widely used to evaluate H/WQ models (Nash and Sutcliffe, 1970). E is better suited to evaluate model goodness-of-fit than the coefficient of determination, R^2, because R^2 is insensitive to additive and proportional differences between model simulations and observations. However, like R^2, E is overly sensitive to extreme values because it squares the values of paired differences, as shown in Eq. (1) (Legates and McCabe, 1999). These deficiencies are diminished in a modified version that uses the absolute value of the deviations to the 1st power (Legates and McCabe, 1999).

$$E = 1.0 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (1)$$

where: O_i = measured (observed) data, P_i = modeled (predicted) data, $\bar{O}$ = mean of measured data.

Index of agreement

The index of agreement, d, which was developed by Willmott (1981), is another widely used dimensionless indicator of H/WQ model goodness-of-fit Eq. (2). The index of agreement was not designed to be a measure of correlation but of the degree to which a model's predictions are error free. According to Legates and McCabe (1999), d is also better suited for model evaluation than R^2, but it too is overly sensitive to extreme values. In a manner similar to that of E, this sensitivity is alleviated in a modified version that uses the absolute value of the deviations instead of the squared deviations.

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

Root mean square error and mean absolute error

The root mean square error, RMSE, and mean absolute error, MAE, 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). They are calculated as shown (Eqs. (3) and (4)).

$$\mathrm{RMSE} = \sqrt{N^{-1}\sum_{i=1}^{N}(O_i - P_i)^2} \quad (3)$$

$$\mathrm{MAE} = N^{-1}\sum_{i=1}^{N}|O_i - P_i| \quad (4)$$

Objective

With the increased relevance of uncertainty to H/WQ modeling and water resource decision-making, goodness-of-fit indicators require modification to appropriately compare model predictions to measured H/WQ data. Therefore, the objective of this paper is to facilitate improved evaluation of H/WQ models by developing modifications to the deviation term in goodness-of-fit indicators based on the uncertainty of measured data. The modifications were designed to apply directly to the Nash–Sutcliffe coefficient of efficiency, the index of agreement, root mean square error, and mean absolute error, but they may be appropriate for other pairwise comparisons of measured and predicted data. The selected indicators are widely applied and accepted but, as traditionally applied, fail to explicitly represent uncertainty in measured data. Although, these modifications were developed for H/WQ modeling, they are valid for any comparison of measured versus predicted data where measurement uncertainty can be estimated or assumed.

Methods

All of the selected indicators (E, d, RMSE, and MAE) contain the same error term, e_i, which is the difference between each pair of measured and predicted values Eq. (5). It is this deviation that the present modifications address. Traditionally, deviations are determined simply as the difference between observed and predicted data, but this does not account for measurement uncertainty in model calibration, evaluation, and validation data sets.

$$e_i = O_i - P_i \quad (5)$$

where: e_i = deviation between paired observed and predicted values.

In the presence of measurement uncertainty, it is more appropriate to evaluate paired measured and predicted data against the uncertainty boundaries or the probability distribution of measured data than against individual data values. Thus, two modifications were developed to more appropriately calculate the deviation between each pair of measured and predicted values based on the measurement uncertainty. Modification 1, which applies when the uncertainty boundary but not the distribution of uncertainty around each measured data point is known, minimizes each calculated deviation. Modification 2, which applies when the probability distribution is known or assumed for each measured value, produces a more practical estimate of the deviation. It is important to note that the probability distributions in Modification 2 are the distribution of possible measured data values for each individual data point, O_i, *not* for the entire population of measured data. Summary descriptions and design requirements for these modifications are presented subsequently followed by detailed descriptions.

Summary description and design requirements

The deviation calculations for Modifications 1 and 2 are summarized in Table 1. The modified deviations are substituted for the traditional deviation term, $e_i = O_i - P$, in the E, d, RMSE, and MAE goodness-of-fit indicators to consider measurement uncertainty. The modifications were designed to meet the following requirements.

- $e_i = eu1_i = eu2_i$ if the uncertainty in measured data are either ignored or assumed to be 0,
- $eu1_i \leqslant eu2_i \leqslant e_i$,
- the modified deviations, $eu1_i$ and $eu2_i$, decrease as the measured data uncertainty increases for a given pair of measured and predicted values,
- the modified deviations, $eu1_i$ and $eu2_i$, decrease as the absolute value of the difference between measured and predicted data decreases for a given uncertainty range or probability distribution for each measured value.

Modification 1 — probable error range, no distributional assumption

Modification 1 is applicable when the probable error range is known or assumed for each measured data point and no distributional assumptions are made. To use Modification 1, probable error range, PER, for each measured data point is first determined. The probable error range can be estimated by the root mean square method in Eq. (6) (Topping, 1972) or can be estimated by professional judgment or from literature values. The root mean square method is widely accepted and has been used for uncertainty estimates related to discharge measurements (Sauer and Meyer, 1992) and water quality constituents (Cuadros-Rodriquez et al., 2002). This method, which was designed to combine all potential errors and produce realistic estimates of cumulative uncertainty, assumes error sources are independent and bi-

Table 1 Modified deviation calculations based on uncertainty in measured data

Deviation method	Affect of measurement uncertainty on deviation	Calculation	Eq.
Traditional	None	$e_i = O_i - P_i$	(5)
Modification 1	Deviation modified based on the probable error range of measured data	$eu1_i = 0$ if $\mathrm{UO}_i(l) \leqslant P_i \leqslant \mathrm{UO}_i(u)$ $eu1_i = \mathrm{UO}_i(l) - P_i$ if $P_i < \mathrm{UO}_i(l)$ $eu1_i = \mathrm{UO}_i(u) - P_i$ if $P_i > \mathrm{UO}_i(u)$	(8)
Modification 2	Deviation modified based on the probability distribution of measured data	$eu2_i = \frac{CF_i}{0.5} \times (O_i - P_i)$	(9)

directional (therefore non-additive). The procedure for applying this method to H/WQ measurements appears in Harmel et al. (2006), along with uncertainty estimates for measured flow, nutrient, and sediment data.

$$\mathrm{PER} = \sqrt{\sum_{i=1}^{n} (E_1^2 + E_2^2 + E_3^2 + \ldots + E_n^2)} \tag{6}$$

where: PER = probable error range (±%), n = number of potential error sources, E_1, E_2, $E_3 \ldots E_n$ = uncertainty associated with each potential error source (±%).

The procedural steps necessary in H/WQ data collected are categorized by Harmel et al. (2006) as streamflow (discharge) measurement, sample collection, sample preservation/storage, and laboratory analysis. It is these procedural categories and their corresponding sources of error that the present modifications consider in goodness-of-fit evaluation. Other issues related to model uncertainty, such as commensurability of errors and model structure, parameterization, and input data errors, are not included. It is the combined effect of these factors, which are reflected in model output data, that the present modifications were designed to evaluate but with the enhancement that allows consideration of measurement uncertainty.

A single PER can be applied to all measured data, or a unique value can be calculated for each measured point, depending on the variation of uncertainty throughout the range of measured data. The uncertainty boundaries for each measured value are determined based on estimated PER.

$$\mathrm{UO}_i(u) = O_i + \frac{\mathrm{PER}_i \times O_i}{100} \quad \mathrm{UO}_i(l) = O_i - \frac{\mathrm{PER}_i \times O_i}{100} \tag{7}$$

where: $\mathrm{UO}_i(u)$ = upper uncertainty boundary for each measured data point, $\mathrm{UO}_i(l)$ = lower uncertainty boundary for each measured data point, PER_i = probable error range for each measured data point, O_i.

To use Modification 1 to calculate the modified deviation, $eu1_i$, it is necessary to determine whether each model predicted value is within the uncertainty boundaries of the corresponding measured value. For predicted values that lie within the uncertainty boundaries, the deviation is set equal to 0. For predicted values that lie outside the boundaries, the deviation is determined as the difference between the predicted data point and the nearest uncertainty boundary. Thus, Modification 1 minimizes the error estimate for each measured and predicted data pair. The calculation of $eu1_i$ is shown numerically in Eq. (8) and graphically in Fig. 1.

$$\begin{aligned} eu1_i &= 0 \quad \text{if } \mathrm{UO}_i(l) \leqslant P_i \leqslant \mathrm{UO}_i(u) \\ eu1_i &= \mathrm{UO}_i(l) - P_i \quad \text{if } P_i < \mathrm{UO}_i(l) \\ eu1_i &= \mathrm{UO}_i(u) - P_i \quad \text{if } P_i > \mathrm{UO}_i(u) \end{aligned} \tag{8}$$

where: $eu1_i$ = modified deviation (Modification 1) between paired measured and predicted data.

Modification 2 — probability distribution

In contrast to Modification 1, which is necessary in the absence of distributional information, Modification 2 can produce more practical error estimates when the probability distribution about each measured value is known or assumed. With Modification 2, deviations between paired measured and predicted data are modified based on the properties of the probability distribution of each measured data value. Either the probability density function (pdf), in the case of the normal distribution, or the continuous distribution function (cdf), in the case of the triangular distribution, which account for uncertainty about each measured value, are used to calculate a correction factor (CF) for each paired deviation Eq. (9). A single distribution can be applied to all measured data or a unique distribution can be applied for each measured value, depending on the variation of distributional properties throughout the range of measured data.

$$eu2_i = \frac{\mathrm{CF}_i}{0.5} \times (O_i - P_i) \tag{9}$$

where: CF_i = correction factor based on the probability distribution of each measured value, $eu2_i$ = modified deviation (Modification 2) between paired measured and predicted data.

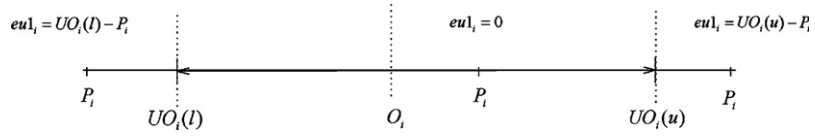

Figure 1 Graphical representation of Modification 1 to calculate the deviation between paired measured and predicted H/WQ data based on the probable error range of measured data but no distributional assumptions.

The assumptions made in designing Modification 2 were that the probability distribution for each measured value is symmetric and that each measured value represents the mean and median value of that distribution. Under these assumptions, the CF is used to adjust each deviation based on measurement uncertainty. As shown in Eq. (9), CF is divided by 0.5, which is the maximum probability for one-sided pdfs, to represent the proportion (0–1.0) of deviation that is accounted for by the probability distribution of each O_i. The calculations to determine CF for two common symmetric probability distributions (normal and triangular) are presented subsequently.

Modification 2 for measured data with a normal probability distribution

When the probability distribution about a data point is known or assumed to be represented by the normal distribution, probabilities calculated from the standard normal distribution, $z \sim N(0,1)$, are used to determine the CF. To calculate the CF for normally distributed data, the mean and the variance are required. With Modification 2, the measured value, O_i, represents the mean, μ, and median of the distribution. The variance, σ^2, can either be input directly from project specific data or estimated by the following procedure.

To estimate the variance, the PER must be determined for each measured data point. As stated in the description of Modification 1, the PER can be estimated with the literature values, professional judgment, or Eq. (6). Then, the uncertainty boundaries, $UO_i(l)$ and $UO_i(u)$, are calculated with Eq. (7). This calculation is not strictly valid because the normal distribution is infinite in both directions; however, it is appropriate because the mean ±3.9 standard deviations contains >99.99% of the normal probability distribution (Haan, 2002). Thus for Modification 2, the uncertainty boundaries are assumed to represent the measured value (mean) ± 3.9 standard deviations (Eq. (10), Fig. 2). The variance is then calculated with Eq. (11).

$$UO_i(l) = O_i - 3.9\sigma \text{ and } UO_i(u) = O_i + 3.9\sigma \tag{10}$$

where: σ = standard deviation about measured data value O_i.

$$\sigma^2 = \left(\frac{O_i - UO_i(l)}{3.9}\right)^2 \text{ or } \sigma^2 = \left(\frac{UO_i(u) - O_i}{3.9}\right)^2 \tag{11}$$

where: σ^2 = variance about measured data value O_i.

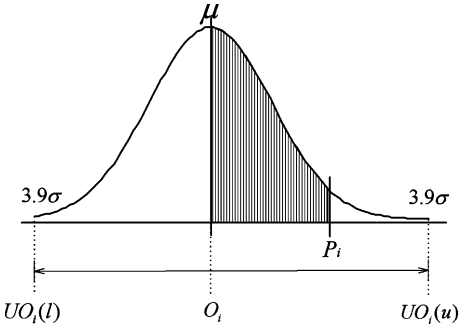

Figure 2 Graphical representation of Modification 2 to calculate the deviation between measured and predicted H/WQ data for measured values with a normal probability distribution.

With the two distributional parameters, the normal distribution can be transformed to the standard normal distribution. The Z value in the standard normal distribution is calculated with Eq. (12). Then, CF is calculated as the area under the standard normal distribution, which represents the probability that the transformed P_i value is between 0 and z_i [CF = prob ($Z_i < z$)]. The CF value for the normal distribution can range from 0 (for $O_i = P_i$) to 0.5 (for $\sigma = 0$ or for $|Z_i| \geqslant 3.9$, which occurs if P_i is farther then 3.9σ from O_i).

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$
$$Z_i = \frac{P_i - O_i}{\left(\frac{O_i - UO_i(l)}{3.9}\right)} \text{ or } Z_i = \frac{P_i - O_i}{\left(\frac{UO_i(u) - O_i}{3.9}\right)} \tag{12}$$

where: Z_i = linearly transformed value of P_i in the standard normal distribution.

Modification 2 for measured data with a symmetric triangular probability distribution

When the probability distribution about a data point is known or assumed to be represented by a symmetric triangular distribution, resulting probabilities are used to determine the CF. To calculate the CF in this case, the mean and upper and lower limits must be known or assumed. To apply Modification 2 to symmetrical triangular distributions, the mean and median are represented by O_i. The upper and lower limits represented by the uncertainty boundaries, $UO_i(l)$ and $UO_i(u)$, are calculated from the PER with Eq. (7) (Fig. 3).

With the relevant distributional parameters, CF is calculated with Eq. (13). For the triangular distribution, CF can range from 0 (for $O_i = P_i$) to 0.5 (for $\sigma = 0$ or for P_i outside the uncertainty boundaries).

$$CF = 0.5 - \frac{[P_i - UO_i(l)]^2}{[UO_i(u) - UO_i(l)] \times [O_i - UO_i(l)]}$$
$$\text{if } UO_i(l) \leqslant P_i \leqslant O_i$$
$$CF = 0.5 - \frac{[UO_i(u) - P_i]^2}{[UO_i(u) - UO_i(l)] \times [UO_i(u) - O_i]} \tag{13}$$
$$\text{if } O_i \leqslant P_i \leqslant UO_i(u)$$
$$CF = 0.5 \quad \text{if } P_i > UO_i(u), P_i < UO_i(l)$$

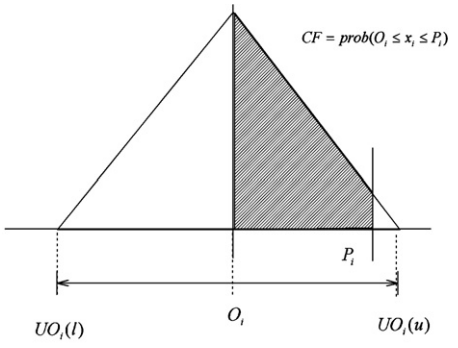

Figure 3 Graphical representation of Modification 2 to calculate the deviation between measured and predicted H/WQ data with a triangular probability distribution.

Results

Application examples

Several measured hydrologic and water quality data sets and associated model application output sets were selected to illustrate the application of these modifications. These data sets were selected to include various temporal and spatial scales, data types, and observed-predicted fits. The examples are not meant to compare the performance of specific models but to illustrate the inclusion of measurement uncertainty in the evaluation of goodness-of-fit. The uncertainty estimates for measured data are based on information in Harmel et al. (2006). The examples include selected graphical and quantitative (both dimensionless and absolute error) goodness-of-fit indicators as recommended for proper model evaluation (Willmott, 1981; Legates and McCabe, 1999; Moriasi et al., in review).

Monthly runoff (Riesel watershed Y6)

In this example, measured monthly runoff data for a small watershed (Y6) located near Riesel, Texas, were compared to corresponding data simulated by the EPIC model (Williams and Sharpley, 1989). This data set was selected to represent a good fit between measured and modeled values. As shown in Fig. 4, the model performed fairly well throughout the range of measured data, except for very low values of measured runoff. Thus, the influence of uncertainty is not readily apparent.

All of the mathematical indicators, as traditionally applied, also suggested that EPIC was able to reproduce observed runoff quite well (Table 2). The E and d indicators in their squared and absolute deviation forms were close to 1.0 indicating good agreement between measured and modeled values. The RMSE and MAE values were also quite low in comparison to the magnitude of measured values ($\bar{O} = 23.33$ mm, $\bar{P} = 22.63$ mm, $n = 48$). When the indicators were modified to account for measurement uncertainty, their goodness-of-fit improvement was not drastic and suggested modest improvement in already strong agreements. For Modification 1 (PER = ±10%, 42%), which produces the best possible agreement between measured and predicted values, E and d improved 1–15%, and RMSE and MAE improved 34–68%. For Modification 2 with normal distributed measured data, E and d improved 0–2%, and RMSE and MAE improved 0–8%.

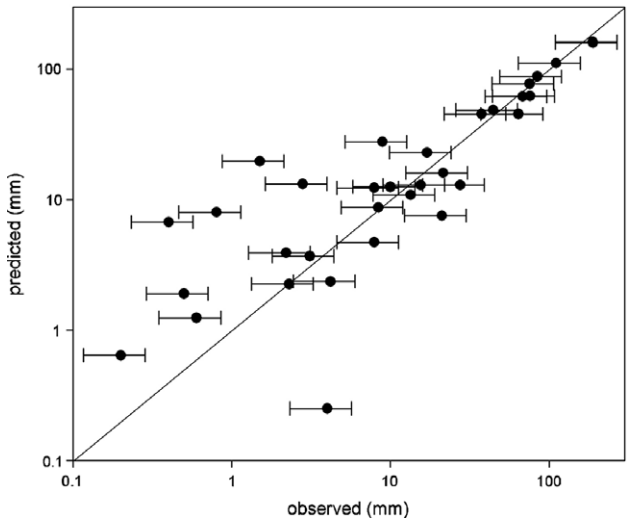

Figure 4 Scatterplot of measured and predicted monthly runoff (mm) for Riesel watershed Y6, modified with Modification 1 to include the uncertainty range for each measured value plotted as an error bar (PER = ±42%).

Monthly dissolved *P* loss (Riesel watershed Y6)

In this example, measured monthly dissolved P losses for Riesel watershed Y6 were compared to corresponding data simulated by the EPIC model (Williams and Sharpley, 1989). As shown in Fig. 5, EPIC did a reasonable job of reproducing measured P loss for most months but drastically overestimated P loss in other months. This overestimation occurred even when the uncertainty boundaries were expanded to represent a PER = ±104%, which is a reasonable range for dissolved P measured under less than ideal monitoring and laboratory conditions (Harmel et al., 2006).

As expected from the visual observation, the mathematical indicators confirmed a relatively poor reproduction of measured dissolved P losses (Table 3). The unmodified E values were either negative or only slightly positive indicating poor agreement between measured and modeled values. The d values were substantially lower than in the previous example. The RMSE and MAE values were quite similar to the measured and predicted mean value ($\bar{O} = 0.03$ kg/ha, $\bar{P} = 0.06$ kg/ha, $n = 48$), which is another indication of poor performance. When the indicators were modified to account for measurement uncertainty, their improvement was quite variable. Modification 1 improved indicators 7–236% for PER = ±23% and 22–690% for PER = ±104%. In contrast, indicators modified with Modification 2 (triangularly distributed measured data), produced no improvement for PER = ±23% and 1–82% for PER = ±104%. The modest improvement in goodness-of-fit indicators for Modification 2 occurred because many of the deviations between paired measured and predicted data were quite large even when modified. The modified indicators did however, as per their design, suggest improved model performance when the uncertainty increased from ±23% to ±104%.

Daily streamflow (Reynolds Creek watershed)

In this example, measured daily streamflow data for the 239 km^2 Reynolds Creek watershed in Idaho were compared to corresponding data simulated by the SWAT model (Arnold et al., 1998). These data represent a typical hydrologic data set with reasonable agreement between measured and modeled values ($\bar{O} = 0.68$ cms, $\bar{P} = 0.71$ cms, $n = 1827$). Without consideration of uncertainty, the flow exceedance curves in Fig. 6 appear somewhat different. However, predicted values are well within the uncertainty boundaries of the measured flow (PER = ±42%).

As traditionally applied, the E, d, RMSE, and MAE indicators suggested that SWAT was able to reproduce observed runoff reasonably well (Table 4). For a PER of ±42%, Modification 1 indicated that the model did quite well in predicting flow within the uncertainty boundaries of measured

Table 2 Results of traditional and modified deviations in selected goodness-of-fit indicators for comparison of measured and EPIC predicted monthly runoff (mm) for Riesel watershed Y6, assuming PER = ±10% and ±42% and a normal distribution for Modification 2

O_i PER=	0%	10%	10%	42%	42%
O_i dist.=	n.a.	n.a.	Normal	n.a.	Normal
Indicator	Trad.	Mod 1 value (% inc)	Mod 2 value (% inc)	Mod 1 value (% inc)	Mod 2 value (% inc)
E	0.96	0.98 (2%)	0.96 (0%)	0.99 (3%)	0.97 (1%)
E^1	0.82	0.89 (8%)	0.82 (0%)	0.94 (15%)	0.84 (2%)
d	0.99	1.00 (1%)	0.99 (0%)	1.00 (1%)	0.99 (0%)
d^1	0.91	0.94 (4%)	0.91 (0%)	0.97 (7%)	0.92 (1%)
RMSE	8.77	5.76 (34%)	8.77 (0%)	4.02 (54%)	8.06 (8%)
MAE	5.24	3.37 (36%)	5.21 (1%)	1.68 (68%)	4.81 (8%)

The magnitude of improved fit is represented by the % increase in the indicator values.

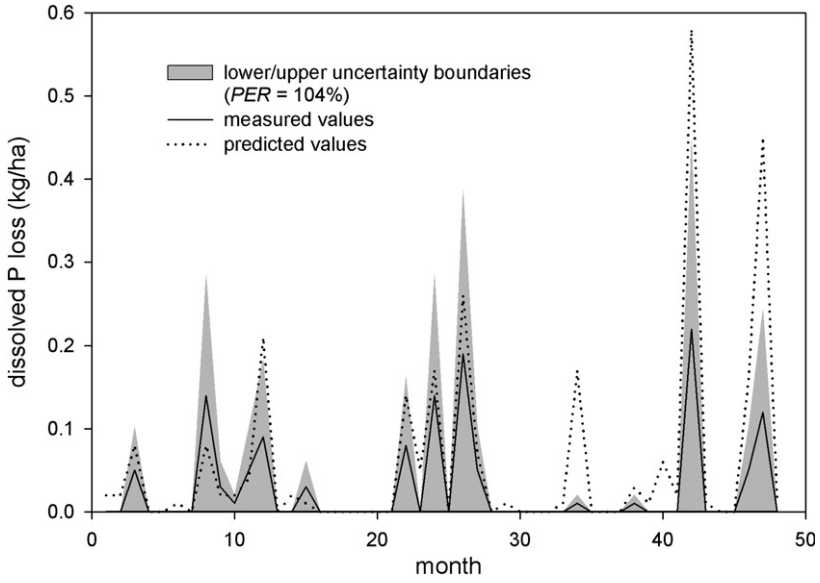

Figure 5 Monthly dissolved *P* loss (kg/ha) predicted by EPIC compared to measured losses from Riesel watershed Y6.

Table 3 Results of traditional and modified deviations in selected goodness-of-fit indicators for comparison of measured and EPIC predicted monthly dissolved *P* loss (kg/ha) for Riesel watershed Y6, assuming PER = ±23% and ±104% and a triangular distribution for Modification 2

O_i PER=	0%	23%	23%	104%	104%
O_i dist.=	n.a.	n.a.	Triangular	n.a.	Triangular
Indicator	Trad.	Mod 1 value (% inc)	Mod 2 value (% inc)	Mod 1 value (% inc)	Mod 2 value (% inc)
E	−1.40	−0.90 (38%)	−1.40 (0%)	0.28 (>100%)	−1.34 (4%)
E^1	0.07	0.23 (>100%)	0.07 (0%)	0.54 (>100%)	0.12 (82%)
d	0.76	0.81 (7%)	0.76 (0%)	0.93 (22%)	0.77 (1%)
d^1	0.64	0.70 (10%)	0.64 (0%)	0.82 (28%)	0.66 (3%)
RMSE	0.08	0.07 (12%)	0.08 (0%)	0.05 (45%)	0.08 (1%)
MAE	0.04	0.03 (17%)	0.04 (0%)	0.02 (50%)	0.03 (6%)

The magnitude of improved fit is represented by the % increase in the indicator values.

flow; however, Modification 2 produced only slight improvement (<2%) when measurement uncertainty was included. In the second example for this data set, the uncertainty in measured data were not uniform but were adjusted based on measurement uncertainty differences throughout the range of measured values. This example was presented to illustrate the flexibility of the modifications based on the available information on measurement uncertainty (as recommended by Beven, 2006a). In this example, PER was set at ± 100—200% for very low stages to represent the difficulty

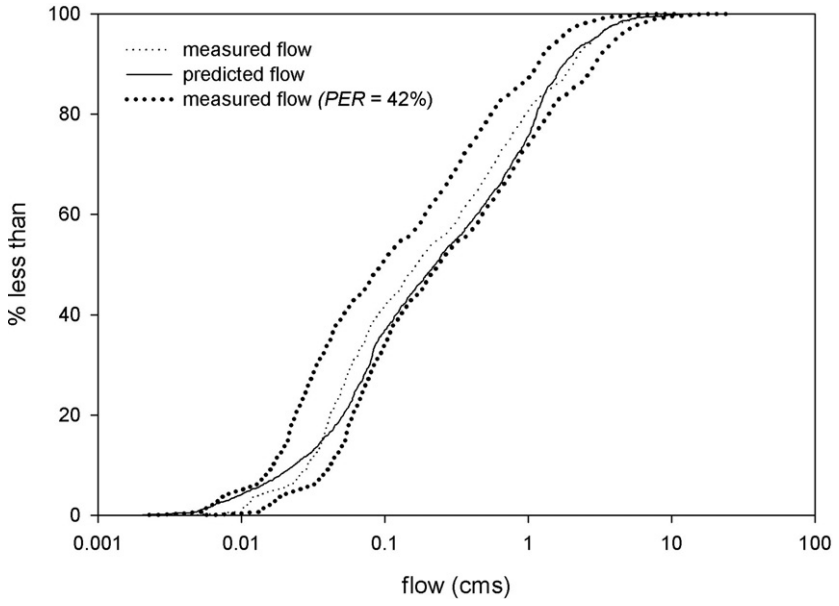

Figure 6 Comparison of measured and predicted percent exceedence curves for daily streamflow (m^3/s) for Reynolds Creek, assuming measured data with a PER = ±42%.

Table 4 Results of traditional and modified deviations in selected goodness-of-fit indicators for comparison of measured and SWAT predicted daily streamflow (m^3/s) for the Reynolds Creek watershed, assuming PER = ±42% and a variable PER and a normal distribution for Modification 2

O_i PER=	0%	42%	42%	Vary %[a]	Vary %
O_i dist.=	n.a.	n.a.	Normal	n.a.	Normal
Indicator	Trad.	Mod 1 value (% inc)	Mod 2 value (% inc)	Mod 1 value (% inc)	Mod 2 value (% inc)
E	0.73	0.91 (26%)	0.73 (0%)	0.92 (27%)	0.75 (3%)
E^1	0.53	0.81 (51%)	0.54 (1%)	0.83 (56%)	0.55 (3%)
d	0.93	0.98 (6%)	0.93 (0%)	0.98 (6%)	0.93 (1%)
d^1	0.76	0.90 (18%)	0.77 (0%)	0.92 (20%)	0.77 (1%)
RMSE	0.66	0.37 (44%)	0.65 (0%)	0.35 (46%)	0.63 (4%)
MAE	0.35	0.15 (58%)	0.35 (1%)	0.13 (64%)	0.34 (4%)

The magnitude of improved fit is represented by the % increase in the indicator values.
[a] The PER values for measured data in this example were assumed to vary from ±42% to 200% based on flow rate.

in accurate stage measurement. Similarly, PER was assumed to be ± 100% for high flows when floodplain flow, rapid morphological changes, and stage estimation procedures increase measurement error. A similar variable adjustment in measurement error would be appropriate for laboratory analysis, as determination of very low constituent concentrations can produce inflated relative error (Kotlash and Chessman, 1998). As expected, increasing the uncertainty for selected data further improved perceived model performance (Table 4). The indicators improved 6—64% for Modification 1 and 1—4% for Modification 2.

Daily streamflow (South Fork watershed)

In this example, measured daily streamflow data for the 780 km^2 South Fork watershed in Iowa were compared to corresponding data simulated by the SWAT model (Arnold et al., 1998). These data represent a situation where predicted values do not match well with measured data due to model structure error ($\bar{O} = 0.56$ cfs, $\bar{P} = 0.66$ cfs, $n = 552$). The poor performance of SWAT resulted from inadequate representation of tile drainage flow and prairie pothole hydrology (SWAT was subsequently revised to better represent these processes, see Green et al., 2006). As shown in Fig. 7, the model performed poorly for much of the measurement period.

All of the mathematical indicators, as traditionally applied, also illustrated the poor performance of the model (Table 5). The unmodified E values were either negative or only slightly positive, the d values were lower than in the other examples, and RMSE and MAE were similar to the mean measured value. When the indicators were modified to account for measurement uncertainty, their improvement was quite variable. For Modification 1 the indicators improved substantially because of the large uncertainty in measured data (PER = ±42%, 84%), but the goodness-of-fit generally remained poor inspite of this improvement. In contrast, the indicators improved only slightly for Modification 2 assuming normally distributed

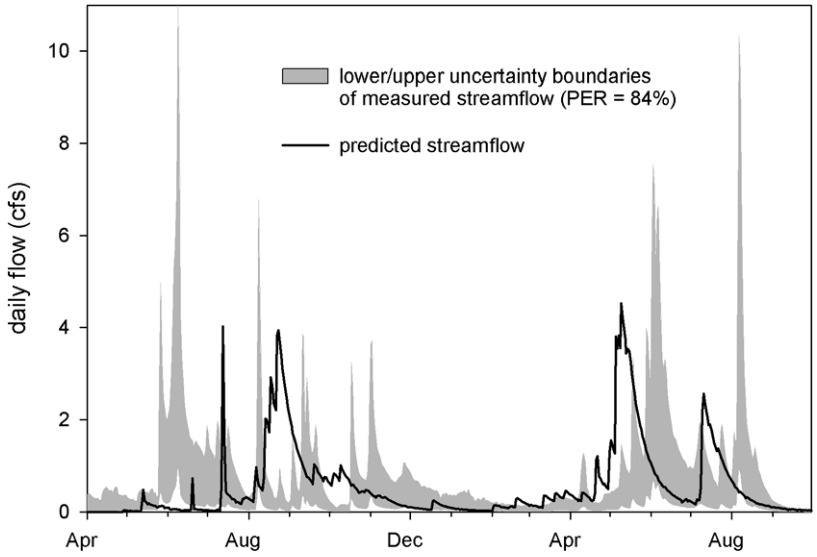

Figure 7 Comparison of measured and predicted daily streamflow (ft^3/s) for the South Fork watershed; the uncertainty boundary for predicted values (PER = ±84%) are presented.

Table 5 Results of traditional and modified deviations in selected goodness-of-fit indicators for comparison of measured and SWAT predicted daily streamflow (ft^3/s) for the South Fork watershed, assuming PER = ±42% and 84% and a normal distribution for Modification 2

O_i PER=	0%	42%	42%	84%	84%
O_i dist.=	n.a.	n.a.	Normal	n.a.	Normal
Indicator	Trad.	Mod 1 value (% inc)	Mod 2 value (% inc)	Mod 1 value (% inc)	Mod 2 value (% inc)
E	−0.01	0.56 (>100%)	−0.00 (31%)	0.79 (>100%)	0.00 (>100%)
E^1	0.10	0.51 (>100%)	0.10 (6%)	0.75 (>100%)	0.12 (24%)
d	0.72	0.88 (22%)	0.72 (0%)	0.94 (30%)	0.73 (1%)
d^1	0.58	0.77 (33%)	0.58 (0%)	0.88 (52%)	0.59 (2%)
RMSE	0.77	0.51 (34%)	0.77 (0%)	0.36 (54%)	0.77 (1%)
MAE	0.43	0.24 (45%)	0.43 (1%)	0.12 (72%)	0.42 (3%)

The magnitude of improved fit is represented by the % increase in the indicator values.

measured data. With Modification 2, the values of d, RMSE, and MAE improved <1% for PER = ±42% and only 1−3% for PER = ±84%. For both Modification 1 and 2, the relative improvement of E was exaggerated because E was <0.01 as traditionally calculated. Increasing the measurement uncertainty resulted in only modest improvements in goodness-of-fit indicators because the model predictions were so poor (Table 5, Fig. 7). This result is important because consideration of measurement uncertainty should not prevent poor goodness-of-fit conclusions in the presence of model inadequacies.

Conclusions

The present modifications were designed to consider measurement uncertainty in the evaluation of H/WQ models. Specifically, the error term calculations of several accepted and commonly used goodness-of-fit indicators (E, d, RMSE, and MAE) were modified based on the uncertainty boundaries (Modification 1) or the probability distribution (Modification 2) of measured data. As traditionally applied, goodness-of-fit indicators simply use the difference between paired measured and predicted data. It is, however, more appropriate to calculate deviations based on the uncertainty boundaries or the probability distribution of measured data. Thus, the modifications were based on the theory that H/WQ models should be evaluated against the measurement uncertainty instead of the values of measured data, which are inherently uncertain.

Modification 1 was developed to provide enhanced goodness-of-fit information when distributional information on measured data uncertainty is not available and not reasonably assumed. Because of its design, Modification 1 minimizes the calculated deviations and thus produces the minimum estimate of error. Thus, when applied to the example data sets, the selected goodness-of-fit indicators improved considerably.

Modification 2 was designed to provide a more realistic calculation of paired deviations when distributional information regarding measurement uncertainty is known or rea-

sonably assumed. Thus, Modification 2 is more appropriate if information is available regarding uncertainty distribution. When applied to four example data sets, the improvements were much smaller than observed for Modification 1 for the selected goodness-of-fit indicators. In general, goodness-of-fit increased only slightly for measured data with little uncertainty, but Modification 2 resulted in modest improvement when data with substantial uncertainty were compared with both poor and good model predictions. The modest improvement for poor model performance is an important result, as poor predictions — especially in the presence of large model structure errors — should not appear satisfactory simply because of measurement uncertainty.

Although these example data sets were analyzed assuming normal or triangular distributions for each individual measured data point, others such as the uniform distribution are probably equally valid. Research regarding typical distributions for individual measurements of streamflow and water quality indicators would be a valuable contribution and enhance the application of the present modifications.

As a result of increased knowledge on measured data uncertainty (e.g. Pelletier, 1988; Sauer and Meyer, 1992; Kotlash and Chessman, 1998; Jarvie et al., 2002; Slade, 2004; Harmel et al., 2006), modelers now have the capability to consider the ''quality'' of their calibration, validation, and evaluation data sets to more realistically judge model performance. As H/WQ models are becoming guiding factors in water resource policy, management, and regulatory decision-making, it is no longer appropriate to discuss but not consider uncertainty in model evaluation (Pappenberger and Beven, 2006; Beven, 2006b). The modified deviation calculations for the selected indicators should facilitate this advancement.

Acknowledgements

Keith Beven (Lancaster University, UK) and Jim Bonta (USDA-ARS, Coshocton, OH) provided ideas, discussion, and review that substantially enhanced this work. Their contribution is sincerely appreciated. Thanks also to the anonymous reviewers for their very insightful comments. Mike Van Liew (Montana DEQ, Helena, MT), Cole Green (USDA-ARS, Temple, TX), and Susan Wang (Blackland REC, Temple, TX) graciously provided the example data sets and model results.

References

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part I: model development. J. Am. Water Resour. Assoc. 34 (1), 73—89.

Bashford, K., Beven, K.J., Young, P.C., 2002. Model structures, observational data and robust, scale dependent parameterisations: explorations using a virtual hydrological reality. Hydrol. Process. 16 (2), 293—312.

Beck, M.B., 1987. Water quality modeling: a review of the analysis of uncertainty. Water Resour. Res. 23 (8), 1393—1442.

Beven, K., 2006a. A manifesto for the equifinality thesis. J. Hydrol. 320, 18—36.

Beven, K., 2006b. On undermining the science? Hydrol. Process. 20, 3141—3146.

Beven, K.J., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11—29.

Cuadros-Rodriquez, L., Hernandez Torres, M.E., Almansa Lopez, E., Egea Gonzalez, F.J., Arrebola Liebanas, F.J., Martinez Vidal, J.L., 2002. Assessment of uncertainty in pesticide multiresidue analytical methods: main sources and estimation. Anal. Chim. Acta 454 (2), 297—314.

Galaz, V., 2005. Does the EC water framework directive build resilience? Harnessing Socio-Economic Complexity in European Water Management. Policy Paper, Swedish Water House.

Green, C.H., Tomer, M.D., Di Luzio, M., Arnold, J.G., 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, second ed. Iowa State Press, Ames, IA.

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

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

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

Haggard, B.E., Soerens, T.S., Green, W.R., Richards, R.P., 2003. Using regression methods to estimate stream phosphorus loads at the Illinois River, Arkansas. Appl. Eng. Agric. 19 (2), 187—194.

Harmel, R.D., King, K.W., 2005. Uncertainty in measured sediment and nutrient flux in runoff from small agricultural watersheds. Trans. ASAE 48 (5), 1713—1721.

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

Harris, G., Heathwaite, A.L., 2005. Inadmissible evidence: knowledge and prediction in land and riverscapes. J. Hydrol. 304, 3—19.

Hession, W.C., Storm, D.E., 2000. Watershed-level uncertainties: implications for phosphorus management and eutrophication. J. Environ. Qual. 20, 1172—1179.

Jarvie, H.P., Withers, P.J.A., Neal, C., 2002. Review of robust measurement of phosphorus in river water: sampling, storage, fractionation, and sensitivity. Hydrol. Earth Syst. Sci. 6 (1), 113—132.

Kavetski, D., Franks, S.W., Kuczera, G., 2002. Confronting input uncertainty in environmental modelling. In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), Calibration of Watershed Models, AGU Water Science and Applications Series, vol. 6, pp. 49—68.

Kotlash, A.R., Chessman, B.C., 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., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. J. Hydrol. 211, 69—85.

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

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T., in review. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE.

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

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: Or 7 reasons not to use uncertainty analysis. Water Resour. Res. 42 W05302, doi:10.1029/2005WR004820.

Pelletier, P.M., 1988. Uncertainties in the single determination of river discharge: a literature review. Can. J. Civil Eng. 15 (5), 834–850.

Reckhow, K.H., 1994. Water quality simulation modeling and uncertainty analysis for risk assessment and decision making. Ecol. Modell. 72, 1–20.

Robertson, D.M., Roerish, E.D., 1999. Influence of various water quality sampling strategies on load estimates for small streams. Water Resour. Res. 35, 3747–3759.

Sauer, V.B., Meyer, R.W., 1992. Determination of error in individual discharge measurements. USGS Open File Report 92-144. Washington, DC: USGS.

Sharpley, A.N., Kleinman, P.J.A., McDowell, R.W., Gitau, M., Bryant, R.B., 2002. Modeling phosphorus transport in agricultural watersheds: processes and possibilities. J. Soil Water Conserv. 57 (6), 425–439.

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

Slade, R.M., 2004. General Methods, Information, and Sources for Collecting and Analyzing Water-Resources Data. CD-ROM. Copyright 2004 Raymond M. Slade, Jr.

Topping, J., 1972. Errors of Observation and their Treatment, fourth ed. Chapman and Hall, London, UK.

UK Administrators, 2005. Water Framework Directive (WFD): Note from the UK administrators on the next steps of characterization. Welsh Assembly Government, Scottish Executive, Irish Department of the Environment, UK Department for Environmental and Rural Affairs.

USDA-ARS, 2004. Conservation Effects Assessment Project – The ARS Watershed Assessment Study. National Program 201 – Water Quality and Management.

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

Williams, J.R., Sharpley, A.N. (Eds.), 1989. EPIC – Erosion/Productivity Impact Calculator: 1. Model Documentation, USDA Technical Bulletin No. 1768.

Willmott, C.J., 1981. On the validation of models. Phys. Geograph. 2 (2), 184–194.