Environmental Modelling & Software 57 (2014) 40–51

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Position paper

Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations☆

R.D. Harmel [a,*], P.K. Smith [b], K.W. Migliaccio [c], I. Chaubey [d], K.R. Douglas-Mankin [e], B. Benham [f], S. Shukla [g], R. Muñoz-Carpena [h], B.J. Robson [i]

[a] USDA-ARS Grassland, Soil and Water Research Laboratory, Temple, TX, USA
[b] Dept. of Biological and Agricultural Engineering, Texas A&M Univ., College Station, TX, USA
[c] Dept. of Agricultural and Biological Engineering, IFAS, Univ. of Florida, Homestead, FL, USA
[d] Dept. of Earth, Atmospheric, and Planetary Sciences, Dept. of Agricultural and Biological Engineering, Purdue Univ., West Lafayette, IN, USA
[e] USFWS, Boynton Beach, FL, USA
[f] Dept. of Biological Systems Engineering, Virginia Tech Univ., Blacksburg, VA, USA
[g] Dept. of Agricultural and Biological Engineering, Univ. of Florida, Immokalee, FL, USA
[h] Dept. of Agricultural and Biological Engineering, Univ. of Florida, Gainesville, FL, USA
[i] CSIRO Land and Water, Black Mountain, ACT, Australia

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form
14 February 2014
Accepted 17 February 2014
Available online 17 March 2014

Keywords:
Water quality
Calibration
Validation
Goodness-of-fit
Stakeholders

ABSTRACT

Previous publications have outlined recommended practices for hydrologic and water quality (H/WQ) modeling, but limited guidance has been published on how to consider the project's purpose or model's intended use, especially for the final stage of modeling applications – namely evaluation, interpretation, and communication of model results. Such guidance is needed to more effectively evaluate and interpret model performance and more accurately communicate that performance to decision-makers and other modeling stakeholders. Thus, we formulated a methodology for evaluation, interpretation, and communication of H/WQ model results. The recommended methodology focuses on interpretation and communication of results, not on model development or initial calibration and validation, and as such it applies to the modeling process following initial calibration. The methodology recommends the following steps: 1) evaluate initial model performance; 2) evaluate outliers and extremes in observed values and bias in predicted values; 3) estimate uncertainty in observed data and predicted values; 4) re-evaluate model performance considering accuracy, precision, and hypothesis testing; 5) interpret model results considering intended use; and 6) communicate model performance. A flowchart and tables were developed to guide model interpretation, refinement, and proper application considering intended model uses (i.e., ***Exploratory***, ***Planning***, and ***Regulatory/Legal***). The methodology was designed to enhance application of H/WQ models through conscientious evaluation, interpretation, and communication of model performance to decision-makers and other stakeholders; it is not meant to be a definitive standard or a required protocol, but together with recent recommendations and published best practices serve as guidelines for enhanced model application emphasizing the importance of the model's intended use.

Published by Elsevier Ltd.

1. Introduction

Hydrologic/water quality (H/WQ) models are commonly applied for regulatory/legal, planning, and exploratory purposes, each of which necessitates a different level of confidence in model output. The standard by which models are judged may be quite high in legal or regulatory applications for example, but a lesser standard is likely acceptable in exploratory analyses or preliminary applications. Loague and Green (1991) stated that evaluation protocols are

☆ Position papers aim to synthesise some key aspect of the knowledge platform for environmental modelling and software issues. The review process is twofold – a normal external review process followed by extensive review by EMS Board members. See the Editorial in Volume 21 (2006).

* Corresponding author. USDA-ARS, 808 E. Blackland Rd., Temple, TX 76502, USA. Tel.: +1 254 770 6521; fax: +1 254 770 6561.

E-mail address: daren.harmel@ars.usda.gov (R.D. Harmel).

http://dx.doi.org/10.1016/j.envsoft.2014.02.013
1364-8152/Published by Elsevier Ltd.

needed for models used in environmental management, and ASCE (1993) concluded that modeling publications generally lack adequate discussion of model performance. Thus in recent years, Refsgaard et al. (2005), Jakeman et al. (2006), Engel et al. (2007), Moriasi et al. (2007), Biondi et al. (2012), Bennett et al. (2013), Black et al. (2014), and others have recommended protocols for some aspects of H/WQ modeling based on the need for formal model development, application, and communication guidelines rather than *ad hoc* approaches. Such formalized methodologies were rightly deemed necessary because modeling applications often have human health, regulatory, ecological, socio-economic, and policy implications, which necessitate improved communication of modeling techniques, limitations, and performance.

While these previous publications outline recommended modeling practices, additional guidance is needed to assist modelers in considering the project's purpose or model's intended use in the final stage of model applications – evaluation, interpretation, and communication of model results. Wagener et al. (2001) did note the importance of first defining the modeling purpose and considering the appropriate complexity needed in model development with respect to performance and associated uncertainty. Guidelines developed by Refsgaard et al. (2005) and Jakeman et al. (2006) both begin with defining the modeling project purpose and end with tailoring reporting to various types of users. Bennett et al. (2013) established a comprehensive model evaluation method, which briefly discusses communication with the user community throughout modeling applications. Additionally, Augusiak et al. (2014) reviewed the terminology used to describe model validation and evaluation and suggested that steps in model evaluation should be separately recognized: data evaluation, conceptual model evaluation, implementation verification, model output verification, model analysis (including sensitivity analysis), and model output corroboration (*post hoc* testing of predictions). However, the authors stop short of model evaluation and communication based on intended use and outcomes. Additional guidance would enhance evaluation and interpretation of model performance and assist modelers in more effectively communicating performance to stakeholders (e.g., agricultural producers, environmental organizations, elected officials, public interest groups) as well as researchers, regulators, and local/state/federal agencies.

The methodology focuses on interpretation and communication of model results, which is the final stage of modeling applications, not on model development and validation or initial calibration (e.g., objective function identification, sensitivity analysis, parameter estimation techniques). This does not mean that these topics are not important but that they are relatively well addressed in published modeling literature and are not the focus of this manuscript. In other words, the methodology applies to the modeling process following initial calibration and addresses final model refinement, evaluation, interpretation, and communication of model performance for a specific model application. As such, selected model post-development and refinement steps (e.g., outlier evaluation, prediction bias, and uncertainty evaluation), which are not commonly utilized, are emphasized. In addition, the recommendations in this manuscript focus on deterministic model applications. Probabilistic applications, which are increasing in importance as more regulations are being written with regard to risk rather than discrete outcomes, would require different evaluation, interpretation, and communication methods.

The recommendations were designed to expand guidance from foundational work and "good modeling practice" manuscripts such as Refsgaard et al. (2005), Jakeman et al. (2006), Engel et al. (2007), Moriasi et al. (2007), Bennett et al. (2013), Black et al. (2014), and others. Specifically, they were developed as recommendations that contribute to improved modeling practice through conscientious application of H/WQ models and enhanced evaluation, interpretation, and communication of modeling results considering intended use.

2. Recommended methodology

Procedures within the following steps were developed from existing literature and best professional judgment to guide post-development model refinement, evaluation, interpretation, and communication: 1) evaluate initial model performance; 2) evaluate outliers and extremes in observed values and bias in predicted values; 3) estimate uncertainty in observed data and predicted values; 4) re-evaluate model performance considering accuracy, precision, and hypothesis testing; 5) interpret model results considering intended use; and 6) communicate model performance. A flowchart outlining the methodology is presented in Fig. 1, and detailed information for each of the steps is presented subsequently. The recommended methodology is not meant to be prescriptive but to present recently-developed "good modeling practices" and emphasize existing ones, many of which are commonly neglected.

Upon initiation of all modeling projects, it is critical to clearly define the model's intended use (Fig. 1). To present this topic and discuss general cases of modeling applications, three categories of intended model use were established: ***Exploratory***, ***Planning***, and ***Regulatory/Legal***. The ***Exploratory*** category includes modeling projects in which initial or approximate comparisons or beta model development is desired, and therefore, reduced confidence in predictions is acceptable. This category also includes models applied to explore the implications of alternative conceptual models (hypotheses regarding system function), to integrate scientific process studies (concretize a conceptual model), or to share knowledge with stakeholders in a participatory process. The ***Planning*** category includes modeling for planning purposes (e.g., urban development and watershed planning), conservation implementation (e.g., management practice placement), and policy formulation (e.g., incentives for land conversion to biofuel production) in which confidence in the model's ability to capture scenario differences is important, but very high accuracy and precision in model predictions is less critical. The ***Regulatory/Legal*** category includes modeling projects with regulatory (e.g., violation of regulatory standards), legal (e.g., lawsuits or criminal cases) and/or human health (i.e., chronic or acute) implications in which very high confidence in model predictions is essential from both accuracy and precision standpoints. Although these intended use categories may not be mutually exclusive nor cover the entire spectrum of modeling applications, they represent general categories that warrant differing expectations related to model performance.

2.1. Step 1. Evaluate initial model performance

A crucial step in evaluating and interpreting model results is creating graphs to display model results and calculating goodness-of-fit indicator values to quantify model performance in terms of prediction accuracy (Fig. 1, Step 1a, 1b, 1c). The value of utilizing multiple evaluation methods, including graphical techniques and quantitative indicators to assess overall model performance, is widely accepted (e.g., Willmott, 1981; Loague and Green, 1991; ASCE, 1993; Legates and McCabe, 1999; Moriasi et al., 2007; Jain and Sudheer, 2008; Harmel et al., 2010; Bennett et al., 2013). Although utilization of multiple techniques requires additional effort, it produces a more comprehensive evaluation of model performance.

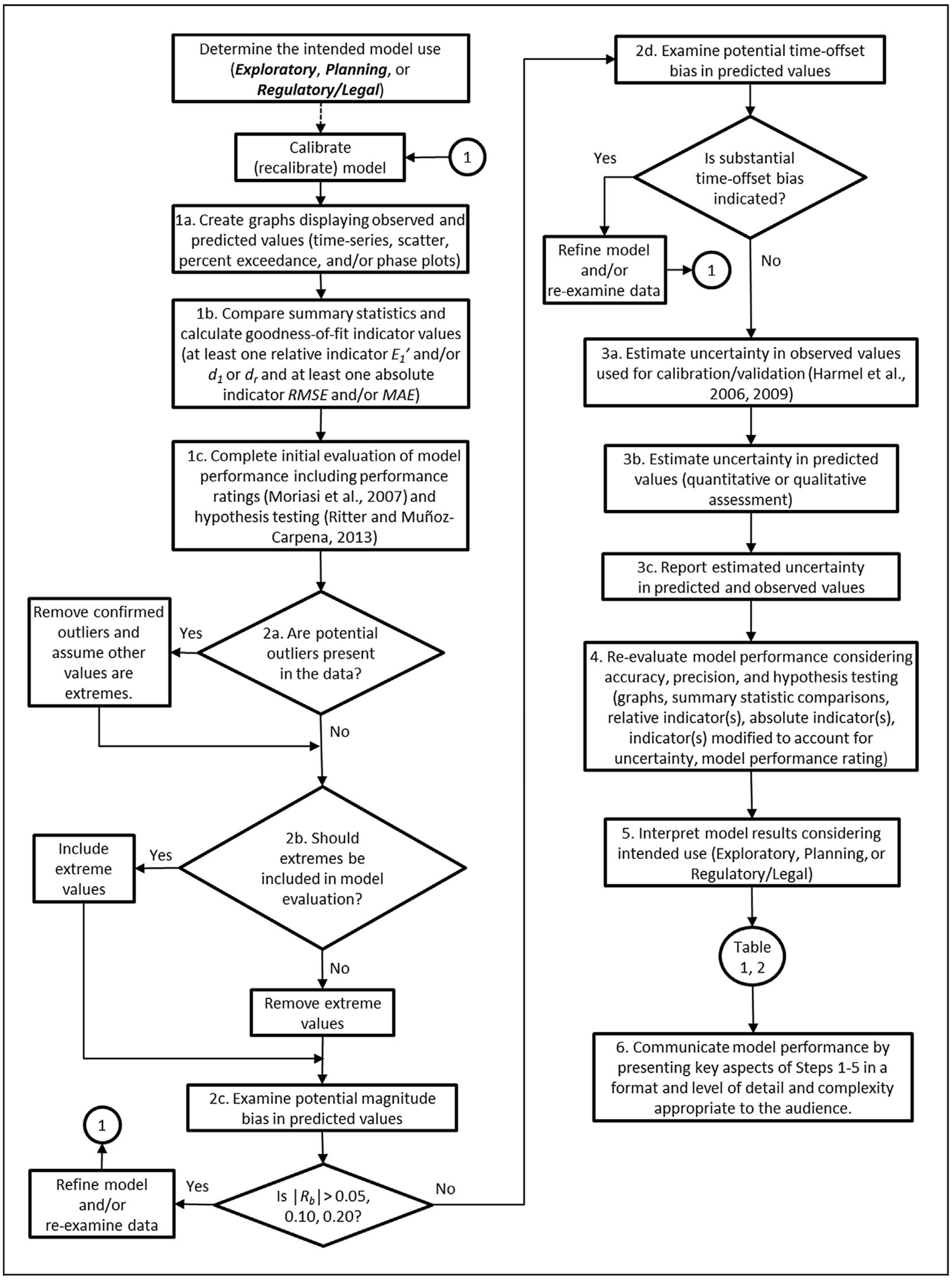

Determine the intended model use (*Exploratory*, *Planning*, or *Regulatory/Legal*)
Calibrate (recalibrate) model
1
1a. Create graphs displaying observed and predicted values (time-series, scatter, percent exceedance, and/or phase plots)
1b. Compare summary statistics and calculate goodness-of-fit indicator values (at least one relative indicator E_1' and/or d_1 or d_r and at least one absolute indicator *RMSE* and/or *MAE*)
1c. Complete initial evaluation of model performance including performance ratings (Moriasi et al., 2007) and hypothesis testing (Ritter and Muñoz-Carpena, 2013)
2a. Are potential outliers present in the data?
Yes
Remove confirmed outliers and assume other values are extremes.
No
2b. Should extremes be included in model evaluation?
Yes
Include extreme values
No
Remove extreme values
2c. Examine potential magnitude bias in predicted values
Is $|R_b| > 0.05$, 0.10, 0.20?
Yes
Refine model and/or re-examine data
1
No
2d. Examine potential time-offset bias in predicted values
Is substantial time-offset bias indicated?
Yes
Refine model and/or re-examine data
1
No
3a. Estimate uncertainty in observed values used for calibration/validation (Harmel et al., 2006, 2009)
3b. Estimate uncertainty in predicted values (quantitative or qualitative assessment)
3c. Report estimated uncertainty in predicted and observed values
4. Re-evaluate model performance considering accuracy, precision, and hypothesis testing (graphs, summary statistic comparisons, relative indicator(s), absolute indicator(s), indicator(s) modified to account for uncertainty, model performance rating)
5. Interpret model results considering intended use (Exploratory, Planning, or Regulatory/Legal)
Table 1, 2
6. Communicate model performance by presenting key aspects of Steps 1-5 in a format and level of detail and complexity appropriate to the audience.

Fig. 1. Flowchart of steps in the methodology for hydrologic/water quality model evaluation, interpretation, and communication considering intended use.

2.1.1. Step 1a. Create graphs displaying observed and predicted values

Graphical techniques assist the user in visualizing model results relative to measured data, judging whether results make physical sense and are plausible, and determining where/how to correct model deficiencies. Graphs are needed in addition to goodness-of-fit indicators, which when used alone can lead to erroneous conclusions (e.g., see Chambers et al., 1983). Four graphical techniques are especially useful in assessing model goodness-of-fit: 1) time-series plots, 2) scatter plots, 3) percent exceedance plots (cumulative distribution plots), and 4) phase portraits (phase plots). Time-series plots help illustrate how well models reproduce observed temporal responses (e.g., in surface runoff, baseflow, constituent flux, and groundwater interactions). McCuen et al. (2006) refers to the measure of how well the time series is reproduced as time-offset bias. Scatter plots (e.g., 1:1 plots or plots of residuals) clearly show observed vs. predicted values, thus indicating potential systematic bias and obvious poor predictions. Percent exceedance (or cumulative distribution) plots are valuable in assessing the ability of models to predict typical and extreme values and to reproduce the return periods of observed values. They are also useful for evaluating how errors are propagated in time by the simulation process. In some applications, box plots are a useful alternative or supplement to percent exceedance plots, presenting less information but in a more easily digested form. Phase plots are useful in determining whether the model is reproducing observed trajectories of change and can highlight regions of poor predictions. Other plot types, including spatial representations of data and predictions where spatial patterns are of primary importance, may also be useful.

2.1.2. Step 1b. Compare summary statistics and calculate goodness-of-fit indicator values

Comparison of summary statistics (i.e., measures of central tendency and variability) of observed and predicted values is recommended as an initial assessment of overall model performance because if these summary values are not similar then a good fit along the time series is unlikely. Ideally at least one relative and at least one absolute goodness-of-fit indicator should also be calculated to assess model performance (Legates and McCabe, 1999). Taylor diagrams can be valuable for simultaneously comparing the performance of different models using multiple statistics (Taylor, 2001). This recommendation balances adequate evaluation and unnecessary complication (ASCE, 1993). Using multiple statistics and goodness-of-fit indicators does increase the likelihood of mixed interpretation of model performance; however, the benefits of a more complete model evaluation outweigh the potential disadvantages.

Two widely-used relative goodness-of-fit indicators are the Nash–Sutcliffe coefficient of efficiency (E_{NS}) and the index of agreement (d). Both indicators, and subsequent modified versions of each, are commonly applied although considerable discussion continues related to the relative merits of each (e.g., Willmott et al., 2011; Legates and McCabe, 2013).

The original E_{NS} defined by Nash and Sutcliffe (1970) is a dimensionless indicator that ranges from -∞ to 1. It 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_{NS} can be sensitive to extreme values because it squares the values of paired differences (Legates and McCabe, 1999), which may or may not be desirable. The influence of extreme values is diminished in a modified version that uses the absolute value of the deviations, which when written in its baseline adjusted form (eq. (1)) allows a better comparison to observed values, as recommended by Garrick et al. (1978), Legates and McCabe (1999), and Schaefli and Gupta (2007).

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

where $E_1{}'$ = modified E_{NS}, O_i = observed data, P_i = predicted data, and $\overline{O}'$ = baseline value of observed data against which model is to be compared.

McCuen et al. (2006) conducted an in-depth evaluation of E_{NS} and concluded that it can be a reliable indicator but emphasized the need for proper interpretation. Specifically, E_{NS} can be sensitive to sample size, outliers, and bias in magnitude and time-offset, and failure to recognize its limitations may lead to rejection of a good model solely because the E_{NS} was misapplied. McCuen et al. (2006) also provided a method for conducting hypothesis tests and computing confidence intervals with the E_{NS}. Along this line, Ritter and Muñoz-Carpena (2013) recently presented a framework for statistical interpretation of model performance designed to reduce modeler subjectivity in performance evaluation. A key component of the comprehensive procedure, which includes evaluation of the effects of bias, outliers, and repeated data, is a new method of hypothesis testing related to E_{NS} threshold values.

The original index of agreement (d) developed by Willmott (1981) is another dimensionless indicator designed not 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 sensitive to extreme values. In a manner similar to that of E_{NS}, this sensitivity is mitigated in a modified version of d_1 (eq. (2)) that uses the absolute value of the deviations instead of the squared deviations (Willmott et al., 1985, 2011). More recently, Willmott et al. (2011) refined d_1 so that it is bounded on the upper and lower ends, as shown in equation (3).

$$d_1 = 1 - \frac{\sum_{i=1}^{N}|P_i - O_i|}{\sum_{i=1}^{N}\left(|P_i - \overline{O}| + |O_i - \overline{O}|\right)} \tag{2}$$

$$\begin{aligned} d_r &= 1 - \frac{\sum_{i=1}^{N}|P_i - O_i|}{c\sum_{i=1}^{N}|O_i - \overline{O}|}, \text{ when } \sum_{i=1}^{N}|P_i - O_i| \le c\sum_{i=1}^{N}|O_i - \overline{O}| \\ d_r &= 1 - \frac{c\sum_{i=1}^{N}|O_i - \overline{O}|}{\sum_{i=1}^{N}|P_i - O_i|} - 1, \text{ when } \sum_{i=1}^{N}|P_i - O_i| > c\sum_{i=1}^{N}|O_i - \overline{O}| \end{aligned} \tag{3}$$

The root mean square error ($RMSE$, eq. (4)) and mean absolute error (MAE, eq. (5)) are widely-used absolute error goodness-of-fit indicators that describe differences in observed and predicted values in the appropriate units (Legates and McCabe, 1999). Although reporting values for at least one of these absolute indicators is recommended, a comparison is also useful because the degree to which $RMSE$ is greater than MAE indicates the extent of outliers (Legates and McCabe, 1999).

$$RMSE = \sqrt{N^{-1}\sum_{i=1}^{N}(O_i - P_i)^2} \tag{4}$$

$$MAE = N^{-1}\sum_{i=1}^{N}|O_i - P_i| \tag{5}$$

In some applications, goodness-of-fit metrics are less relevant than measures of categorical performance. Examples include prediction of whether an out-of-bank flood event will occur, whether a regulatory water quality trigger level has been exceeded, and which phytoplankton species is dominant in a water body at a given time. Bennett et al. (2013) present a range of categorical performance methods and discuss issues such as the relative weighting of positive and negative results (e.g., whether it is more important to correctly predict the chance of an exceedance or to correctly predict that the trigger will *not* be exceeded).

2.1.3. Step 1c. Complete initial evaluation of model performance

Historically, modelers have stopped at this step and used some arbitrary decision criteria to support their model performance conclusion. However, recently modelers have taken this a step further and assigned "very good," "good," "satisfactory," or "unsatisfactory" model performance based on ratings developed by Moriasi et al. (2007). These ratings are typically applied based on ranges of E_{NS}, but Moriasi et al. (2007) also presents ratings based on % bias and RSR ($RMSE$ to observations standard deviation ratio).

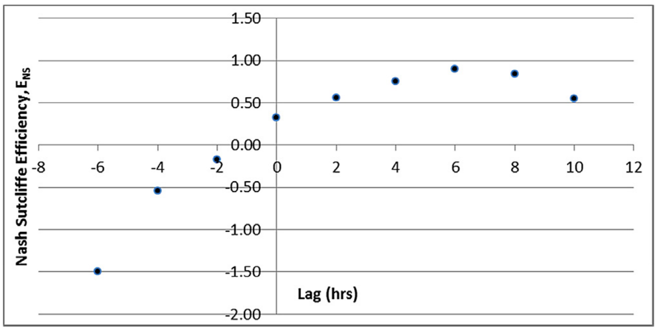


Fig. 2. Efficiogram with lag increments of 2 h from −6 to 10 h.

Alternatively, “acceptable” ranges of goodness-of-fit indicator values that represent sufficient predictive capacity in the context of the application can be useful in evaluating model performance. These should ideally be determined prior to calibration based on the modeling objectives.

Recent work by Ritter and Muñoz-Carpena (2013) provides an objective criterion for the selection of E_{NS} ranges for model performance ratings based on model prediction error represented by n_t. This indicator n_t represents the number of times that the standard deviation of observed data is greater than the model error (*RMSE*), where a good model fit is represented by a model with error that is relatively small compared to the observed data range (high n_t values). Because of the nonlinear relationship between E_{NS} and n_t, small incremental values of E_{NS} closer to 1 (perfect model fit) translate into large improvements in model prediction. As a result, in addition to the establishment of a lower threshold that defines “unsatisfactory” model performance (i.e., $E_{NS} < 0.65$, corresponding to $n_t = 0.7$), whole multipliers of n_t ($\sim 1n_t, 2n_t, 3n_t$) were established for “acceptable” ($0.65 \leq E_{NS} < 0.80$), “good” ($0.80 \leq E_{NS} < 0.90$), and “very good” ($E_{NS} \geq 0.90$) model performance ratings.

2.2. Step 2. Evaluate outliers and extremes in observed values and bias in predicted values

Whether or not the initial model evaluation (Step 1) indicates acceptable performance in terms of model accuracy, evaluation of observed values and potential bias may indicate the need for further model refinement (Fig. 1, Step 2a, 2b, 2c, 2d). Use of a frequency histogram is recommended to evaluate the presence of outliers and extreme observed values, which can make the central tendency and variability of the sample much larger or smaller than that of the population.

2.2.1. Step 2a. Evaluate potential outliers

For observed values that are much larger or smaller than the rest of the sample, it should be determined whether they are outliers (unexplainable or unrealistic values outside the assumed population) or extremes (realistic but very infrequent values). By this definition, outliers either result from mistakes (e.g., measurement or transcription error) or originate from another population; therefore, they should be removed or the populations should be analyzed separately. Removing outliers should produce more representative summary statistics, improved goodness-of-fit indicator values, and more realistic relationships with other variables. Several statistical tests such as the Dixon–Thompson, Rosner, Chi-square, Bofferoni, and Chauvenets outlier tests (McCuen, 2003) have been developed to check for outliers; however, these tests require an understanding of the probability distribution of the data. Also, even when they are applicable, discretion should be used and every effort made to identify why the observed value occurred and thus confirm its determination as an outlier or an extreme value (McCuen, 2003).

2.2.2. Step 2b. Determine whether to include extremes

Then, the modeler should identify extreme values and determine whether their inclusion is necessary to achieve project objectives based on the model's intended use. Before removing any extreme values, their importance within the hydro-climatic region should be carefully considered. In flood peak prediction projects, for example, extreme values are critical and should remain in the data set. Bulletin 17B from the United States Water Resources Council (1982) provides guidance in handling extremes when determining flood flow frequency. On the other hand, removal of extreme values may be justified in projects focused on average conditions, and their removal will produce the same benefits as removing outliers.

It should be noted that if extremes are included, their relevance to the modeling objectives can be considered and used to weight extreme values. If the extremes are not relevant to the modeling objectives, then goodness-of-fit indicators that use absolute differences or log-transformed absolute differences give less weight to the extremes. Conversely, if the extremes are important, then goodness-of-fit indicators that square or even cube the differences give more weight to extremes.

2.2.3. Step 2c. Examine potential magnitude bias in predicted values

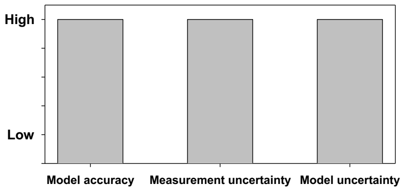
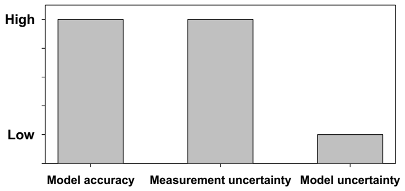
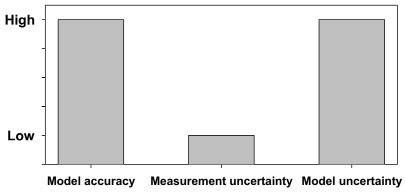
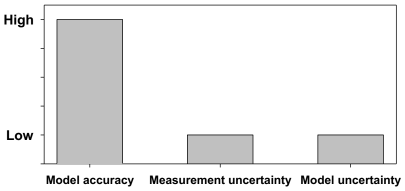
Following identification and assessment of outliers and extremes in observed values, potential bias in predicted values should be examined. There are two components of error in predicted response variables: 1) systematic error, also known as bias; and 2) nonsystematic error, also known as random error. Bias, in this sense, is a measure of the difference in magnitude of a central tendency (e.g., average, median) for predicted and observed response variables at a particular time-step. Bias ($\bar{e}$) can be reported as an average error (eq. (6)). Positive bias indicates over-prediction in the magnitude of the response variable and negative bias indicates under-prediction. The greater the value of $\bar{e}$, the greater the bias in the indicated direction.

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{6}$$

Model bias greatly influences E_{NS} (McCuen et al., 2006), such that bias in a predicted response variable will always reduce E_{NS} relative to unbiased prediction. McCuen et al. (2006) recommend that both bias and relative bias (R_b) be reported along with E_{NS}. Relative bias is calculated as the ratio of $\bar{e}$ to the average value of the observed random variable ($\bar{O}$). McCuen et al. (2006) state that an appropriate threshold for $|R_b|$ is 0.05 (i.e., $|R_b| > 0.05$ may be considered significant). However, a larger threshold may be warranted in certain situations, such as for multi-site calibration or when observed data are highly uncertain. Thus, when the threshold for significance increases (e.g., $|R_b| > 0.10$ or 0.20, depending on the project's intended purpose) model refinement and/or data re-examination should be considered to reduce over- or under-prediction resulting from prediction bias.

Plotting residual error as a function of P_i or O_i can reveal other patterns of bias and provide insight into possible model structural errors. Slope bias, for instance, occurs when low measured values are under-predicted while high values are over-predicted, or vice-versa. This systematic error may result in a low overall $\bar{e}$ but high *RMSE*. Alternatively, the magnitude or variance of residuals may increase or decrease with O_i. If such biases exist, the errors will not likely be normally distributed; therefore, standard methods to

Table 1
Recommendations for model interpretation and refinement for ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** uses when model accuracy is "High."[a]

 **Interpretation:** High measurement uncertainty[b] prevents definitive model accuracy conclusion, and high model uncertainty[c] further reduces confidence in predicted values. **Recommendations:** Appropriate for ***Exploratory*** uses. May be appropriate for ***Planning*** uses, although high model uncertainty and high measurement uncertainty should be reported and considered in scenario analysis. Consider collection of additional data with less uncertainty and/or refinement of model. Inappropriate for ***Regulatory/Legal*** uses. Collect additional data with less uncertainty. Determine cause of high model uncertainty, and either refine model or select a more appropriate model.	 **Interpretation:** High measurement uncertainty prevents definitive model accuracy conclusion. **Recommendations:** Appropriate for ***Exploratory*** uses. May be appropriate for ***Planning*** uses, although high measurement uncertainty should be reported and considered in scenario analysis. Consider collection of additional data with less uncertainty. Inappropriate for ***Regulatory/Legal*** uses. Collect additional data with less uncertainty.
 **Interpretation:** In spite of high model accuracy, high model uncertainty reduces confidence in predicted values. **Recommendations:** Appropriate for ***Exploratory*** uses. May be appropriate for ***Planning*** uses, although high model uncertainty should be reported and considered in scenario analysis. Consider model refinement. Inappropriate for ***Regulatory/Legal*** uses. Determine cause of high model uncertainty, and either refine model or select a more appropriate model to decrease model uncertainty.	 **Interpretation:** High precision and high accuracy provides confidence in predicted values. **Recommendations:** Appropriate for ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** uses.

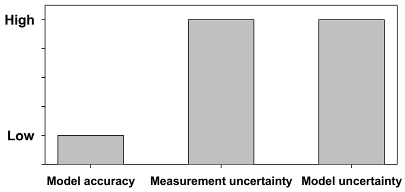
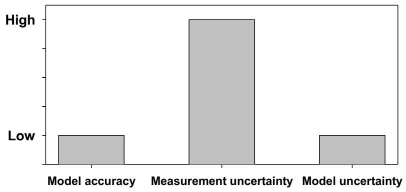
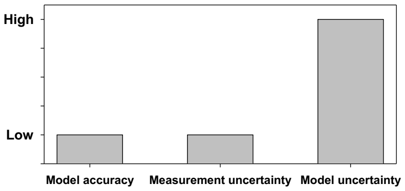
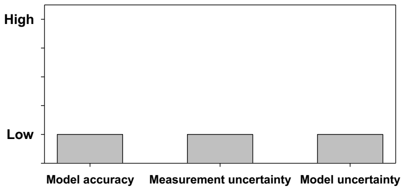
[a] A general, qualitative model performance determination based on summary statistic comparisons, goodness-of-fit indicator values, and/or graphical comparisons supplemented with model performance ratings and hypothesis testing results.
[b] Uncertainty in observed values used in model calibration and validation.
[c] Uncertainty in predicted values (in essence, the precision of predicted values).

calculate p values for indicators such as R^2 can give misleading results.

2.2.4. Step 2d. Examine potential time-offset bias in predicted values

In addition to magnitude bias, time-offset errors can occur in time-dependent models if related variables are not synchronized in time (e.g., rainfall, runoff, and groundwater response), although in some applications only the approximate timing of response needs to be predicted. Time-offset errors are essentially biases on the time scale and can affect E_{NS}; the larger the offset, the greater the decrease in E_{NS} (McCuen et al., 2006). Thus, efficiograms (plot of E_{NS} vs. time lag) can be valuable in detecting time-offset bias. Simulated time series can be lagged by some increment of Δt forward and backward in time. The resulting E_{NS} values comparing the lagged simulated time series to the observed time series are plotted against the lag times. The results show the time-offset at which E_{NS} improves compared to the no-lag condition. As an example, Fig. 2 shows the efficiogram resulting from lagging a time series in 2 h increments from −6 to 10 h. The results show an improvement in the E_{NS} from 0.32 to 0.90 when the time series is lagged by +6 h, which indicates that perhaps the model should be revised because of a consistent time-offset. In this case, the observed data should also be re-examined for potential time-offset biases due to issues

Table 2
Recommendations for model interpretation and refinement for ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** uses when model accuracy is "Low."[(a)]

 **Interpretation:** High model uncertainty[(c)] and low accuracy (even with high measurement uncertainty[(b)]) provide little confidence in predicted values. **Recommendations:** Inappropriate for ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** uses. Collect additional data with less uncertainty. Determine cause of high model uncertainty and low model accuracy, and either refine model or select a more appropriate model.	 **Interpretation:** Although model uncertainty is low, low model accuracy (even with high measurement uncertainty) reduces confidence in predicted values. **Recommendations:** May be appropriate for ***Exploratory*** uses; however, application of a model with low accuracy should be clearly presented and well justified. Inappropriate for ***Planning*** and ***Regulatory/Legal*** uses. Collect additional data with less uncertainty. Determine cause of low model accuracy, and either refine model or select a more appropriate model.
 **Interpretation:** High model uncertainty and low accuracy provide little confidence in predicted values. **Recommendations:** Likely inappropriate for ***Exploratory*** uses; however, may be useful to highlight range of possible system behaviors and develop hypotheses. Determine cause of high model uncertainty and low model accuracy, and either refine model or select a more appropriate model. Inappropriate for ***Planning*** and ***Regulatory/Legal*** uses. Determine cause of high model uncertainty and low model accuracy, and either refine model or select a more appropriate model.	 **Interpretation:** Although model uncertainty is low, low model accuracy reduces confidence in predicted values. **Recommendations:** May be appropriate for ***Exploratory*** uses; however, application of a model with low accuracy should be clearly presented and well justified. Inappropriate for ***Planning*** and ***Regulatory/Legal*** uses. Determine cause of low model accuracy, and either refine model or select a more appropriate model.

[(a)] A general, qualitative model performance determination based on summary statistic comparisons, goodness-of-fit indicator values, and/or graphical comparisons supplemented with model performance ratings and hypothesis testing results.
[(b)] Uncertainty in observed values used in model calibration and validation.
[(c)] Uncertainty in predicted values (in essence, the precision of predicted values).

with sampling methodology. If the time-offset is not consistent, it may be appropriate to compare measured and predicted daily maximums, for example, rather than values at a particular point in time, especially in applications in which only the *approximate* timing of events is relevant. To complicate the analysis further, a single predictive model can have both magnitude and time-offset biases, both of which can contribute to low accuracy; however, a single E_{NS} value cannot identify which factor is the principal source of the bias (McCuen et al., 2006).

2.3. Step 3. Estimate uncertainty in observed data and predicted values

The value of uncertainty estimates for observed data used for calibration/validation and in predicted values and communicating that uncertainty to scientific, regulatory, policy, and public interests is increasingly emphasized (e.g., Beck, 1987; Kavetski et al., 2002; Reckhow, 2003; Beven, 2006; Muñoz-Carpena et al., 2006; Shirmohammadi et al., 2006; Van Steenbergen et al., 2012). (Fig. 1, Step 3a, 3b, 3c). Presenting uncertainty estimates for model predictions and observed data enhances the ability of modelers and decision-makers to assess and quantify confidence in the observed and predicted values (Harmel et al., 2010).

2.3.1. Step 3a. Estimate uncertainty in observed values used for calibration/validation

Estimating and reporting the uncertainty in observed data used in calibration and validation is recommended because of its impact on the evaluation and interpretation of model results. An uncertainty estimation framework specifically for measured water

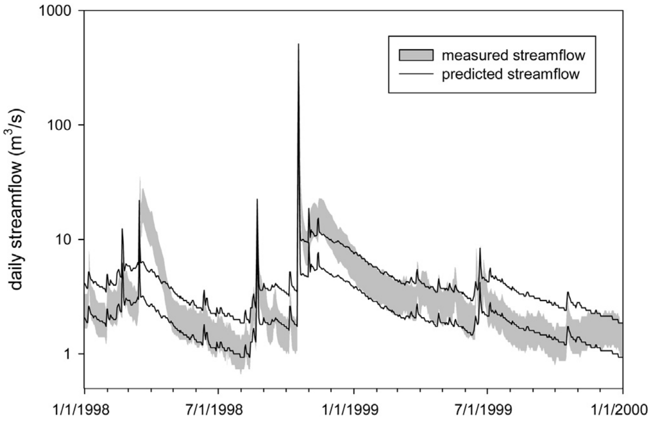


Fig. 3. Measured and predicted daily streamflow example (adapted from Harmel et al., 2010). The uncertainty boundaries (Cv = 0.192) are presented as the shaded area for measured streamflow and as the upper and lower boundary lines (Cv = 0.192) for predicted streamflow.

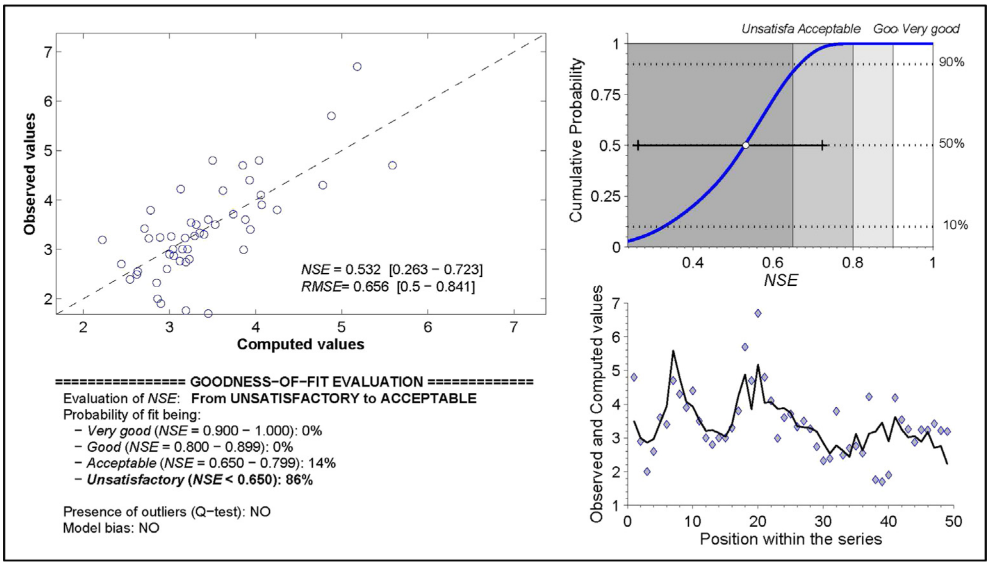
quality data was recently established (Harmel et al., 2006) and modified into a more user-friendly software format (Harmel et al., 2009). Both the software and its framework-basis use *RMSE* propagation to estimate uncertainty for data collection procedural categories (discharge measurement, sample collection, sample preservation/storage, laboratory analysis, and data processing and management) and for resulting discharge, concentration, and load values. The only known similar work is that of Montgomery and Sanders (1986). Research by Harmel et al. (2006, 2009) and others (e.g., Ramsey, 1998) has indicated that substantial error can be introduced by each procedural category, especially sample collection, although most standard methods focus on sample preservation/storage and laboratory analysis.

2.3.2. Step 3b. Estimate uncertainty in predicted values

Model uncertainty (excluding components related to measurement uncertainty) may be attributed to parameterization (parameter uncertainty), algorithm selection and ability to represent natural processes, and natural process variability based on temporal and spatial scales (Vicens et al., 1975; Beven, 1989; Haan, 1989). The methodology presented here does not address alternatives for model uncertainty estimation [e.g., first-order approximation (Haan, 2002), Monte Carlo simulation (Haan et al., 1995), Bayesian methods (e.g., Jin et al., 2010) and generalized likelihood uncertainty estimation (GLUE, Beven and Freer, 2001)] but instead focuses on the importance of estimating model uncertainty because of its policy, regulatory, and management implications (Shirmohammadi et al., 2006). When these techniques are not viable for a specific application, a qualitative assessment of model uncertainty is likely appropriate. For example, the modeler can reasonably assume the model uncertainty is low if the model algorithms reasonably represent natural processes, if the input and calibration/validation data sets are extensive and have low measurement uncertainty, and if parameter values are realistic.

The model interpretation and refinement recommendations for various model uses rely on qualitative groupings of model uncertainty (Tables 1 and 2). It can be argued that these qualitative groupings are preferable to quantitative thresholds because of the inability of current techniques to adequately consider numerous dependent model processes and thus make definitive uncertainty estimates for complex, distributed models. However, whether quantitative or qualitative techniques are used, the process of estimating uncertainty in predicted values produces an estimation of model precision, which should be considered in determination of model appropriateness for a specific application.

For certain ***Exploratory*** and ***Planning*** applications in which scenarios take the model beyond the bounds of the calibration and validation data, statistical measures may not be sufficient to characterize model uncertainty. A common example is model application to simulate hydrologic impacts in long-term climate change scenarios, in which temperatures and weather events are more extreme than those in the calibration/validation data sets. For such extrapolative prediction, greater weight should be given to observational data or time-periods that more closely resemble expected scenarios. In addition, the model should be investigated for structural relevance: Does the model rely on submodels that may not be valid in the extrapolated conditions? For example, were temperature and evaporation relationships established for a limited temperature and humidity range? Were sediment yield relationships developed assuming continuation of current land use, cropping patterns, and/or storm intensities?



```
NSE = 0.532  [0.263 − 0.723]
RMSE= 0.656  [0.5 − 0.841]
```

```
================ GOODNESS−OF−FIT EVALUATION ============
 Evaluation of NSE:   From UNSATISFACTORY to ACCEPTABLE
 Probability of fit being:
   − Very good (NSE = 0.900 − 1.000): 0%
   − Good (NSE = 0.800 − 0.899): 0%
   − Acceptable (NSE = 0.650 − 0.799): 14%
   − Unsatisfactory (NSE < 0.650): 86%

 Presence of outliers (Q−test): NO
 Model bias: NO
```

Fig. 4. Example FITEVAL output.

2.3.3. Step 3c. Report estimated uncertainty in observed and predicted values

For both predicted and observed data sets, the uncertainty should be estimated for individual values (although it is possible that the same uncertainty may apply to each value in the set). These estimates are needed in Step 4 to modify goodness-of-fit indicator values considering both measurement and model uncertainty by the method of Harmel et al. (2010), which uses the coefficient of variation (Cv) to report uncertainty as did Haan et al. (1995). In contrast, Harmel et al. (2006, 2009) present uncertainty estimates for individual values as ±% uncertainty, while others have used the standard deviation.

The example in Fig. 3 illustrates the benefits of presenting uncertainty estimates for corresponding calibration/validation data and predicted values. As is clearly evident in Fig. 3, the model does a poor job of simulating hydrograph recession following high peak flow events. By estimating the uncertainty for measured and predicted values and presenting them graphically, the regions of good and poor model fit and the influence of uncertainty can be observed more clearly than in simple time series comparisons.

2.4. Step 4. Re-evaluate model performance considering accuracy, precision, and hypothesis testing

Complete model evaluation requires both operational examination of the accuracy (Step 1) and precision (Step 3b) of predicted values and examination of whether the model is scientifically and conceptually valid (Willmott et al., 1985; Loague and Green, 1991) and whether it uses currently accepted modeling practices such that the conceptual and algorithmic basis of the model is correct. A model is a good representation of reality only if it can be used to predict, within a calibrated and validated range, an observable phenomenon with acceptable accuracy and precision (Loague and Green, 1991). Thus, model performance should be re-evaluated (Fig. 1, Step 4) once outliers and extreme values in observed data used for calibration/validation are considered (Step 2), potential bias in predicted values is addressed (Step 2), and uncertainty in observed and predicted values is estimated (Step 3).

The same graphical techniques, summary statistics, and goodness-of-fit indicators described in Step 1 can be used. In addition, the method of Harmel et al. (2010) can be used to produce correction factors for the traditional error term ($O_i - P_i$) in goodness-of-fit calculations and thus modify indicator values accounting for measurement and model uncertainty. This method is based on Haan et al. (1995), who state that the degree of overlap between corresponding probability density functions for observed and predicted values is indicative of model predictive ability.

Cibin et al. (2011) proposed another goodness-of-fit evaluation method, similar to that of Harmel and Smith (2007), which can be used to modify E_{NS} considering prediction uncertainty. This method also applies when “measured” values are estimated or assumed, which is relatively common when no continuous or daily water quality data are available because of periodic weekly or monthly sampling. In this case regression models, such as LOADEST (Runkel et al., 2004), can be used to convert periodic measurements into more frequent or continuous data for calibration and validation; however, the uncertainty introduced should be carefully considered and clearly reported.

Interpretation of E_{NS} values is often subjective and may be biased by the magnitude and number of observed data points as well as outliers and repeated data. Thus, Ritter and Muñoz-Carpena (2013) proposed hypothesis testing of model goodness-of-fit indicators and developed a statistically-based framework to accept or reject model performance. The procedure assesses the significance of indicator values based on approximated probability distributions for two common indicators (E_{NS} and $RMSE$) and does not assume normality. The distributions are derived with bootstrapping (block bootstrapping when dealing with time series), which is a method of approximating unknown underlying distributions of any shape for the goodness-of-fit indicators, followed by bias corrected and accelerated calculation of confidence intervals. Once the approximate probability distribution is known, the statistical significance of model goodness-of-fit can be evaluated. For example, the null hypothesis (H0) represents that the median E_{NS} is less than the threshold E_{NS} value below which the goodness of fit is not acceptable ($E_{NS} < 0.65$) and the alternative hypothesis (H1) when it is acceptable ($E_{NS} \geq 0.65$). The null hypothesis is rejected, and the result (i.e., acceptable goodness-of-fit) is statistically significant when the p-value is less than the significance level α. The p-value represents the probability of wrongly accepting the fit ($E_{NS} \geq 0.65$) when it should be rejected (i.e., when H0 is true). The choice of α should be based on the modeling project and its intended purpose (i.e., how strong the evidence needs to be for accepting or rejecting H0). As a starting point, Ritter and Muñoz-Carpena (2013) suggest adopting the least restrictive significance level of $\alpha = 0.10$; however, it might be more appropriate to lower the significance level to $\alpha = 0.05$ or 0.01 when substantial uncertainty is present in observed values or for model applications with ***Planning*** and certainly ***Regulatory/Legal*** purposes. This hypothesis testing procedure together with calculation of graphical and goodness-of-fit indicator values, outliers, and bias can be easily performed using the software FITEVAL (see Fig. 4) (Ritter and Muñoz-Carpena, 2013).

These modified indicator values, along with model performance ratings (Moriasi et al., 2007), provide valuable supplemental information designed to be used in conjunction with (not instead of) traditionally-applied indicators. All of these methods applied and considered together provide a detailed overview of model performance.

2.5. Step 5. Interpret model results considering intended use

As stated by Loague and Green (1991), the intended use of the model should determine the strictness of criteria used to determine acceptance during model performance evaluation (Fig. 1, Step 5). Thus, model interpretation and recommendation guidelines were developed for three categories of intended use: ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** (Tables 1 and 2). At this stage in the methodology, uncertainty in observed data used in calibration/validation and in model predictions will have been determined (Step 3) and model accuracy will have been evaluated (Step 4), so these factors can guide model interpretation and refinement for each of the intended uses. In Tables 1 and 2, recommendations are given for each intended model use for each of the eight combinations of High/Low model accuracy, High/Low measurement uncertainty, and High/Low model uncertainty. *Model accuracy* is a general, qualitative model performance determination based on summary statistic comparisons, goodness-of-fit indicator values, and/or graphical comparisons supplemented with model performance ratings (Moriasi et al., 2007) and hypothesis testing results (Ritter and Muñoz-Carpena, 2013). *Measurement uncertainty* is the uncertainty in observed values used in model calibration and validation. *Model uncertainty* is the uncertainty in predicted values (in essence, the precision of predicted values).

2.6. Step 6. Communicate model performance

After completing Steps 1–5, the key aspects of these steps need to be presented in a format appropriate to the project's intended audience, and at a level of detail and complexity that the audience can comprehend (Fig. 1, Step 6). It is not practical to reproduce every detail from each of these steps; thus, the information presented depends to some degree on the audience, the findings, and the intended use of the H/WQ model. The objective of communicating the results of Steps 1–5 should be to clearly demonstrate the rationale and approach used to evaluate and refine the model.

2.6.1. *Step 1. Communication (Evaluate initial model performance)*

The need for and benefits of extensive reporting and communication of results in this step are limited in many modeling applications because it serves as an initial evaluation of model performance; however, there are circumstances when these initial findings might be of interest. For example, graphs identifying a clear pattern, such as under-prediction of flow peaks, would be important in flood prediction projects. Once outliers and extremes in observed values and bias in predicted values are evaluated and possibly removed (Step 2) and uncertainty in observed data used for calibration/validation and in predicted values is estimated (Step 3), model performance will be re-evaluated in Step 4.

2.6.2. *Step 2. Communication (Evaluate outliers and extremes in observed values and bias in predicted values)*

Results of this step should be clearly communicated to provide the audience insight into the observed calibration and validation data and model behavior, which is needed for understanding and interpreting results. For example, plots or R_b values showing substantial bias in model results inform the audience and provide direction and rationale for model refinement. In addition, data or model modifications to reduce or remove the influences of outliers and extreme values in observed data and magnitude and time-offset bias in predicted values should be clearly reported. Specifically, the assumptions used to systematically account for these influences should be defined and presented.

2.6.3. *Step 3. Communication (Estimate uncertainty in observed data and predicted values)*

This step focuses on uncertainty analyses of observed data and predicted values, which must be clearly communicated to the audience because modeling results are often used to make strategic decisions with economic, ecological, and health and safety implications. Kloprogge et al. (2007) contains detailed guidance on the communication of uncertainty to non-technical audiences, and therefore, is a valuable reference. It provides a general introduction to the issue of communicating uncertainty information; it assists writers in meeting the audience's information needs and in reflecting on anticipated uses and possible impacts of the uncertainty information; and it contains practical information on how to communicate uncertainties.

Uncertainty should be explained with careful consideration of the audience's level of understanding, but an audience unfamiliar with uncertainty should not be used as a reason not to communicate uncertainty (Pappenberger and Beven, 2006). In most situations, presenting numerous equations is not beneficial. Instead, presenting the range of probable predicted outcomes and the inherent uncertainty in observed data is recommended to enhance communication of model results (e.g., Fig. 3). Clear presentation of the uncertainty associated with modeling projects will facilitate more appropriate model applications, better understanding and interpretation of model results, and enhanced implementation of actions and programs based on model results (Van Steenbergen et al., 2012).

Communication of uncertainty has received considerable attention in the last decade in climate modeling (e.g., Patt and Dessai, 2005; Budescu et al., 2012; Lorenz et al., 2013) and clinical health research (e.g., Politi et al., 2011; Han, 2013); H/WQ modelers can benefit from this work. For example, Budescu et al., 2012 found that redundant presentation of uncertainty information (presentation of the same information in both descriptive and numerical terms) can facilitate understanding. Additionally, modelers should be aware that qualitative descriptions of uncertainty (e.g., "high", "medium" and "low") and even standard probability statistics will be interpreted very differently by different stakeholders, and this will affect (and be affected by) interpretation of and trust in modelers and model results (Budescu et al., 2012).

2.6.4. *Step 4. Communication (Re-evaluate model performance considering accuracy, precision, and hypothesis testing)*

This step updates the initial evaluation of model performance (Step 1) and generally produces a large number of graphs and statistics for modelers to interpret results and evaluate model performance. Stakeholders will likely not be interested in seeing all of these graphs along with detailed discussion with numerous goodness-of-fit indicators but will instead benefit from summary information and example plots related to model performance.

Ideally, evaluation of model performance should include summary or example graphs, summary statistic comparisons, at least one relative goodness-of-fit indicator (e.g., E_1' and/or d_1 or d_r), at least one absolute goodness-of-fit indicator (e.g., *RSME* and/or *MAE*), and one goodness-of-fit indicator modified to account for uncertainty (e.g., Harmel et al., 2010). Presenting goodness-of-fit indicator values for both calibration and validation periods along with the simulation duration (e.g., event, continuous), simulation time-step (e.g., daily, monthly), spatial scale (e.g. field, small watershed, basin), and data type (e.g., flow volume, sediment load) along with a brief discussion of any data or model abnormalities is recommended. In conjunction with graphs and indicator values, model performance ratings such as developed by Moriasi et al. (2007) are useful to qualify performance as "good," "satisfactory," "unsatisfactory," etc. Lastly, the hypothesis testing methodology of Ritter and Muñoz-Carpena (2013) provides an indication of the statistical significance of model predictions and provides a probability of the following model performance ratings: "unsatisfactory" ($E_{NS} < 0.65$), "acceptable" ($0.65 \leq E_{NS} < 0.80$), "good" ($0.80 \leq E_{NS} < 0.90$), and "very good" ($E_{NS} \geq 0.90$).

2.6.5. *Step 5. Communication (Interpret model results considering intended use)*

In addition to graphs and indicator values, which have historically been used to assign some arbitrary judgment related to model performance, other important aspects of model performance (e.g., model performance ratings, hypothesis testing, and estimates of uncertainty in observed data used for calibration/validation and model output) are now recommended as good modeling practices. In addition, according to Wagener et al. (2001), Refsgaard et al. (2005), and Jakeman et al. (2006), the intended model use should be considered when interpreting model results and presenting relevant limitations. Tables 1 and 2 were developed by expanding the model evaluation matrix in Harmel et al. (2010) to facilitate consideration of intended model use (i.e., ***Regulatory/Legal***, ***Planning***, and ***Exploratory***).

In addition, the spatial variability, spatial scale, system complexity, available data, and parameter uncertainty should be considered when interpreting and communicating model performance (Loague and Green, 1991). In all steps, but especially in this final one, accurate, straight-forward communication of model performance avoiding technical jargon is critical but too often not achieved.

3. Summary and conclusions

The methodology for evaluation, interpretation, and communication for H/WQ model performance was formulated to assist modelers in more effectively evaluating and interpreting model performance and more accurately communicating that performance to stakeholders and decision-makers while considering the model's intended use. The methodology focuses on interpretation and communication of model results, not on model development or initial calibration and validation. Rather the methodology applies to model application following initial calibration and addresses model refinement, evaluation, interpretation, and communication. The methodology includes steps for evaluating initial model performance; evaluating outliers and extremes in observed values and bias in predicted values; estimating uncertainty in observed data and predicted values; re-evaluating model performance considering accuracy, precision, and hypothesis testing; interpreting model results considering intended model use; and communicating model performance.

In addition, a flowchart and user-friendly tables were developed to guide model interpretation, refinement, and proper application considering intended model uses (i.e., ***Regulatory/Legal*** includes modeling projects with regulatory, legal, and/or human health implications; ***Planning*** includes modeling for planning purposes, conservation implementation, and policy formulation; and ***Exploratory*** includes modeling projects in which initial or approximate comparisons or beta model development is desired). These various intended uses necessitate different levels of confidence in model results; thus, intended model use should be considered when determining the standard by which models are judged. For instance, when high measurement uncertainty prevents a definitive model accuracy conclusion and high model uncertainty further reduces confidence in predicted values, modeling results are likely inappropriate for ***Regulatory/Legal*** purposes but may be appropriate for ***Planning*** and ***Exploratory*** purposes. In contrast, when high model precision and accuracy along with low measurement uncertainty provides considerable confidence in predicted values, modeling results are likely appropriate for ***Regulatory/Legal***, ***Planning***, and ***Exploratory*** purposes. This manuscript provides substantive interpretation and communication guidance that considers the intended use for H/WQ modeling, which to date has received limited attention in the literature. The methodology was designed to serve as recommended guidance and contribute to "good modeling practices." It is not meant to be a definitive standard or required methodology but to contribute to enhanced model application emphasizing the importance of the model's intended use. The goal of continued "good modeling practice" development is to improve modeling methodology and application of H/WQ models through enhanced evaluation and interpretation of model performance as well as enhanced communication of that performance to decision-makers and other modeling stakeholders.

Acknowledgments

To address the broad need for model calibration and validation guidelines, the Soil and Water Division of the American Society of Agricultural and Biological Engineers (ASABE) formed a Model Calibration and Validation Workgroup in 2010. The "Process" subcommittee has since developed a compilation of model-specific papers on H/WQ model use, calibration, and validation (Moriasi et al., 2012) and is developing a series of manuscripts on specific steps and relevant aspects of model calibration and validation. The "Communications" subcommittee developed the recommended methodology for evaluation, interpretation and communication of H/WQ model results presented in this manuscript. It should be kept in mind that ASABE Standards can range from Engineering Guidelines to Recommended Practices to actual Standards, and considerable debate has occurred and is expected to continue on whether a standard(s) should be developed for model calibration and validation and, if so, how the resulting standard(s) will be developed and applied.

To the modelers before us that built the foundation for the hydrologic/water quality models used today and along the way mentored and guided us. To them, we express our sincere appreciation: Jeff Arnold, Keith Beven, Bernie Engel, Tom Haan, Walt Knisel, John Parsons, Ken Reckhow, Wayne Skaggs, Dan Storm, Jimmy Williams, and many more, and to numerous colleagues, who have provided valuable contributions and unwavering support. Thoughtful reviews and comments by Mike White and Evelyn Steglich are also greatly appreciated.

This work was presented at the 2012 ASABE Annual International Meeting, Dallas, TX, and at the 2013 International Conference on Water Resources and Environmental Research: Water and Environmental Dynamics, Koblenz, Germany.

USDA is an equal opportunity provider and employer.

References

ASCE, 1993. Criteria for evaluation of watershed models. J. Irrig. Drain. Eng. 119 (3), 429–442.

Augusiak, J., Van den Brink, P.J., Grimm, V., 2014. Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. Ecol. Model. http://dx.doi.org/10.1016/j.ecolmodel.2013.11.009 (in press).

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

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Beven, K., 2006. On undermining the science? Hydrol. Process. 20 (14), 3141–3146.

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

Beven, K., Freer, J., 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.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montanari, A., 2012. Validation of hydrological models: conceptual basis, methodological approaches and a proposal for a code of practice. Phys. Chem. Earth 42–44, 70–76.

Black, D.C., Wallbrink, P.J., Jordan, P.W., 2014. Towards best practice implementation and application of models for analysis of water resources management scenarios. Environ. Model. Softw. 52, 136–148.

Budescu, D.V., Por, H.-H., Broomell, S.B., 2012. Effective communication of uncertainty in the IPCC reports. Clim. Change 113 (2), 181–200. http://dx.doi.org/10.1007/s10584-011-0330-3.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. PWS-Kent Publishing Co., Boston, MA, p. 395.

Cibin, R., Chaubey, I., Engel, B., 2011. Simulated watershed scale impacts of corn stover removal for biofuels on hydrology and water quality. Hydrol. Process.. http://dx.doi.org/10.1002/hyp.8280.

Engel, B., Storm, D., White, M., Arnold, J., Arabi, M., 2007. A hydrologic/water quality model application protocol. J. Am. Water Resour. Assoc. 43 (5), 1223–1236.

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

Haan, C.T., 2002. Statistical Methods in Hydrology, second ed. Iowa State Press, Ames, IA.

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.

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

Han, P.K., 2013. Conceptual, methodological, and ethical problems in communicating uncertainty in clinical evidence. Med. Care Res. Rev. 70, 14S–36S.

Harmel, R.D., Smith, P.K., Migliaccio, K.L., 2010. Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Trans. ASABE 53 (1), 55–63.

Harmel, R.D., Smith, D.R., King, K.W., Slade, R.M., 2009. Estimating storm discharge and water quality data uncertainty: a software tool for monitoring and modeling applications. Environ. Model. Softw. 24 (7), 832–842.

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.

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

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

Jakeman, A.J., Lester, R.A., Norton, J.P., 2006. Ten interative steps in development and evaluation of environmental models. Environ. Model. Softw. 21, 602–614.

Jin, X., Xu, C.-Y., Zhang, Q., Singh, V.P., 2010. Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. J. Hydrol. 383 (3–4), 147–155.

Kloprogge, P., Van der Sluijs, J., Wardekker, J., 2007. Uncertainty Communication: Issues and Good Practice. Report NWS-E-2007-199, Copernicus Institute for Sustainable Development and Innovation. University of Utrecht, p. 60.

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

Legates, D.R., McCabe Jr., G.J., 2013. A refined index of model performance: a rejoinder. Int. J. Climatol. 33, 1053–1056. http://dx.doi.org/10.1002/joc.3487.

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.

Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. J. Contam. Hydrol. 7, 51–73.

Lorenz, S., Dessai, S., Paavola, J., Forster, P.M., 2013. The communication of physical science uncertainty in European National Adaptation Strategies. Clim. Change. http://dx.doi.org/10.1007/s10584-013-0809-1.

McCuen, R.H., 2003. Modeling Hydrologic Change. Lewis Publishers, Boca Raton, FL.

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

Montgomery, R.H., Sanders, T.G., 1986. Uncertainty in water quality data. Dev. Water Sci. 270, 17–29.

Moriasi, D.N., Wilson, B.N., Douglas-Mankin, K.R., Arnold, J.G., Gowda, P.H., 2012. Hydrologic and water quality models: use, calibration, and validation. Trans. ASABE 55 (4), 1241–1247.

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

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

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: 7 reasons not to use uncertainty analysis. Water Resour. Res. 42 (5). http://dx.doi.org/10.1029/2005W05302.

Patt, A., Dessai, S., 2005. Communicating uncertainty: lessons learned and suggestions for climate change assessment. C. R. Geosci. 337, 425–441. http://dx.doi.org/10.1016/j.crte.2004.10.004.

Politi, M.C., Clark, M.A., Ombao, H., Dizon, D., Elwyn, G., 2011. Communicating uncertainty can lead to less decision satisfaction: a necessary cost of involving patients in shared decision making? Health Expect. 14 (1), 84–91 http://dx.doi.org/10.1111/j.1369-7625.2010.00626.x.

Ramsey, M.H., 1998. Sampling as a source of measurement uncertainty: techniques for quantification and comparison with analytical sources. J. Anal. At. Spectrom. 13, 97–104.

Reckhow, K.J., 2003. On the need for uncertainty assessment in TMDL modeling and implementation. J. Water Resour. Plan. Manage. 129 (4), 245–246.

Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholte, H., Kassahun, A., 2005. Quality assurance in model based water management – review of existing practice and outline of new approaches. Environ. Model. Softw. 20, 1201–1215.

Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45.

Runkel, R.L., Crawford, C.G., Cohn, T.A., 2004. Load Estimator (LOADEST): a FORTRAN Program for Estimating Constituent Loads in Streams and Rivers. U.S. Geological Survey Techniques and Methods Book 4 Chapter A5. http://pubs.usgs.gov/tm/2005/tm4A5/.

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080.

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.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. (1984–2012) 106 (D7), 7183–7192.

United States Water Resources Council, 1982. Guidelines for Determining Flood Flow Frequency. Bulletin 17B. U.S. Department of Interior, Geological Survey, Office of Water Data Coordination, Reston, VA.

Van Steenbergen, N., Ronsyn, J., Willems, P., 2012. A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication. Environ. Model. Softw. 33, 92–105.

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.

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5 (1), 13–26. http://dx.doi.org/10.5194/hess-5-13-2001.

Willmott, C.J., Robeson, S.R., Matsuura, K., 2011. A refined index of model performance. Int. J. Climatol.. http://dx.doi.org/10.1002/joc.2419.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Connell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90 (C5), 8995–9005.

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