ASABE 621 JUN2017
Guidelines for Calibrating, Validating, and Evaluating Hydrologic and Water Quality (H/WQ) Models

American Society of
Agricultural and Biological Engineers

STANDARD

ASABE is a professional and technical organization, of members worldwide, who are dedicated to advancement of engineering applicable to agricultural, food, and biological systems. ASABE Standards are consensus documents developed and adopted by the American Society of Agricultural and Biological Engineers to meet standardization needs within the scope of the Society; principally agricultural field equipment, farmstead equipment, structures, soil and water resource management, turf and landscape equipment, forest engineering, food and process engineering, electric power applications, plant and animal environment, and waste management.

NOTE: ASABE Standards, Engineering Practices, and Data are informational and advisory only. Their use by anyone engaged in industry or trade is entirely voluntary. The ASABE assumes no responsibility for results attributable to the application of ASABE Standards, Engineering Practices, and Data. Conformity does not ensure compliance with applicable ordinances, laws and regulations. Prospective users are responsible for protecting themselves against liability for infringement of patents.

ASABE Standards, Engineering Practices, and Data initially approved prior to the society name change in July of 2005 are designated as "ASAE", regardless of the revision approval date. Newly developed Standards, Engineering Practices and Data approved after July of 2005 are designated as "ASABE".

Standards designated as "ANSI" are American National Standards as are all ISO adoptions published by ASABE. Adoption as an American National Standard requires verification by ANSI that the requirements for due process, consensus, and other criteria for approval have been met by ASABE.

Consensus is established when, in the judgment of the ANSI Board of Standards Review, substantial agreement has been reached by directly and materially affected interests. Substantial agreement means much more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered, and that a concerted effort be made toward their resolution.

CAUTION NOTICE: ASABE and ANSI standards may be revised or withdrawn at any time. Additionally, procedures of ASABE require that action be taken periodically to reaffirm, revise, or withdraw each standard.

ASABE, 2950 Niles Road, St. Joseph, MI 49085-9659, USA, phone 269-429-0300, fax 269-429-3852, hq@asabe.org

ASABE 621 JUN2017

Guidelines for Calibrating, Validating, and Evaluating Hydrologic and Water Quality (H/WQ) Models

Developed by the NRES-21 Hydrology committee of ASABE. Approved as an ASABE Engineering Practice June 2017.

Keywords: Hydrologic processes, Hydrology, Modeling, Model calibration, Model validation, Water quality.

1 Purpose and Scope

Hydrologic and water quality (H/WQ) models consist of interrelated assemblages of mathematical equations that represent the processes governing the movement, fate, and transport of water, sediments, nutrients, pesticides, bacteria, and other constituents in and on the surface of the soil and through groundwater, stream, river, and reservoir systems. Reliable models provide information for making sound management, policy, and regulatory decisions. While calibration, validation, and evaluation (CV&E) are essential for ensuring model accuracy and precision, there is little consistency in how model practitioners conduct and document this process and communicate the results. Therefore, this Engineering Practice was designed to contribute to good modeling practice recommendations to improve modeling methodology and application of H/WQ models through improved model performance as well as enhanced communication of the processes utilized and results to modelers, decision-makers, and other modeling stakeholders.

It builds on recommended protocols recognizing the need for formal model development, application, and communication guidelines rather than ad hoc approaches because of the often-serious human health, regulatory, ecological, socio-economic, and policy implications of modeling applications. This Engineering Practice is not meant to serve as a comprehensive guide but rather to compile and enhance guidance from foundational manuscripts such as Beck (1987), ASCE (1993), Haan et al. (1995), Refsgaard et al. (2005), Jakeman et al. (2006), Engel et al. (2007), Moriasi et al. (2007), Biondi et al. (2012), Bennett et al. (2013), Ritter and Muñoz-Carpena (2013) and Black et al. (2014), and advances described in recent compilations (e.g., Muñoz-Carpena et al., 2006; Moriasi et al., 2012, 2015a).

1.1 This Engineering Practice aims to enhance calibration, validation, and evaluation of H/WQ models through:

- establishment of consistent terminology and definitions,
- selection of model based on intended use, processes simulated, available data, and temporal and spatial scale,
- compilation and processing of both hard and soft input data and calibration data,
- model parameterization and calibration,
- evaluation of model performance,
- re-examination of input and calibration data and possible model refinement,
- re-evaluation of model results and performance considering intended use, and
- documentation of the modeling process and model results.

2 Definitions

Speaking a common language is an essential first step in improving H/WQ modeling. The following definitions were taken largely from Zeckoski et al. (2015) except where indicated.

2.1 model intended use: the purpose for which the model results will be used. Model intended use affects the required accuracy and precision of the model results (Harmel et al., 2014).

2.1.1 regulatory/legal use: model results will be used for regulatory or legal purposes or have human health implications.

2.1.2 planning use: model results will be used for planning purposes, conservation implementation, or policy formulation.

2.1.3 exploratory use: model results will be used for beta testing, model development, or for initial and approximate comparisons.

2.2 modeling objectives: the specific goals of the modeling study (e.g., determine the contributions from existing pollutant sources to a stream pollutant load or determine the impacts of climate change on drought frequency and duration).

2.3 parameter: quantifiable characteristic of a feature or process represented in a model (Malone et al., 2015).

2.3.1 parameterization: the process of assigning values to model parameters. Parameterization includes calibration and validation, which both require model performance evaluation.

2.3.2 calibration: the process of adjusting selected input parameter values and initial conditions to obtain simulated values that match measured observations with the desired accuracy.

2.3.2.1 manual calibration: the sequence of parameter adjustments is left entirely to the modeler.

2.3.2.2 automatic calibration: a search algorithm is used to decide what parameters are adjusted and by what amount based on the value of an objective function.

2.3.2.3 calibrated model: a model applied to a particular physical setting through appropriate parameterization and calibrated to measured data.

2.3.3 validation: the process of verifying that a calibrated model reproduces measured observations for conditions different than were used for the model calibration.

2.3.4 evaluation: the process of using graphical, quantitative, and/or statistical techniques, along with performance ratings and model intended use to judge the quality of model predictions (Harmel et al., 2014).

2.3.5 verification: the process of determining that the model code is correctly implemented and reflects the conceptual model.

2.4 state variable: time series or function used as input for the model or calculated by the model.

2.5 hard data: data measured within the study area (Arnold et al., 2015).

2.6 soft data: information on individual processes that may not be directly measured in the study area, including temporal or spatial averages, estimated quantities using other models, or qualitative knowledge from experimentalists (Arnold et al., 2015).

2.7 initial conditions: values taken by variables (e.g. soil moisture, snow depth, stage of vegetation) that describe the study area at the beginning of the simulation.

2.8 warm-up period: a simulation period that precedes the time period of interest and is long enough for variables such as soil moisture, perennial vegetation, or reservoir levels to reach values that are independent from initial conditions.

2.9 sensitivity: the relative change of a model to a change in a parameter or input variable.

2.10 sensitivity analysis: the process of determining how sensitive the model output is to selected input parameters.

2.11 uncertainty analysis: the process of estimating the effect of input data and model structure uncertainty on model output variables.

2.12 model performance measures: tools and techniques that evaluate how well simulated values represent measured observations over a specified time period.

3 Consider the intended model use

Upon initiation of modeling projects, the model's intended use (e.g., Exploratory, Planning and Regulatory/Legal) should be taken into consideration (Harmel et al., 2014). Although these intended use categories are not mutually exclusive and may not cover the entire spectrum of modeling applications, they represent general categories that warrant different expectations related to model CV&E.

3.1 Determine the model intended use (e.g., legal/regulatory, planning, or exploratory).

3.2 Determine the modeling objectives.

3.3 Given the intended uses and modeling objectives, determine:

3.3.1 Processes to be simulated.

3.3.2 Spatial and temporal scale of the simulation.

3.3.3 Expectations for accuracy and precision of model predictions.

4 Select model

Select the model that best meets the intended use, modeling objectives, and performance expectations. Moriasi et al. (2012) summarized more than 20 commonly applied H/WQ models; however, that collection is not meant to be exhaustive.

4.1 Select a model that is appropriate given:

- processes that need to be simulated,
- required outputs,
- required spatial and temporal resolution,
- available input and calibration data,
- modeling expertise, and
- computing resources.

4.2 If multiple spatial scales must be considered, the following options should be taken into account (Baffaut et al., 2015):

4.2.1 Simplify by breaking the project into smaller questions using appropriate scales and models. Link two models when addressing interactions of processes that operate over differing spatial scales.

4.2.2 If multiple spatial extents are considered within the same model, perform calibration and validation for smaller areas first, increasing progressively to larger areas. Alternatively, multiple scales can be considered simultaneously.

4.2.3 In the absence of calibration and validation data at the relevant spatial or temporal scale, calibrated parameter values from a model calibrated at a different temporal or spatial scale may not be representative and should only be used with caution. A model calibrated or validated using data of a different scale should be flagged with an explanation and appropriate caveats.

4.2.4 Similarly, model results should be interpreted with caution when they are applied to locations, spatial or temporal scales, or processes that are different from those used in calibration (e.g., be cautious when using a model calibrated with only watershed outlet data to evaluate edge-of-field management effects).

5 Process input data and calibration/validation/evaluation data

Modeling typically requires input data including boundary and initial condition to drive the model (e.g., weather, management) and to parameterize the model (e.g., land use, soil, topography). Measured values for comparison with model predictions are also necessary for model performance evaluation, but these CV&E data need to be reviewed for accuracy and consistency with model objectives.

5.1 Identify available input data.

5.1.1 Review input data and remove obvious errors.

5.1.2 Evaluate input data to ensure that they are appropriate for the modeling objectives. The examples below provide an initial list, but all input data should be reviewed:

- Verify that soil properties are consistent with soil texture.
- Verify that the soil profile is accurately described (e.g., cracking potential, restrictive layer, nutrient stratification).
- Evaluate soil data collection methods, including sample timing, frequency, compositing methods, processing, handling/storage, and analysis as well as the data uncertainty.
- Ensure that land management practices are representative of the area simulated.
- Specify land management practices with a degree of accuracy and precision consistent with modeling objectives.
- Review weather data collection, processing, and Quality Assurance and Quality Control (QA/QC) methods.
- Strive to utilize the densest and most accurate precipitation data, as these factors can contribute substantial prediction uncertainty.

5.1.3 Carefully examine spatial and temporal resolution of input data to guide decisions on model spatial discretization (Guzman et al., 2015). Consider the model processes and note any data limitations relative to the modeling objectives.

5.2 Identify and compile CV&E data that are compatible with the chosen spatial and temporal resolution. Use site-specific measured data when available. Clearly note if other types of data are used to evaluate model predictions, such as regression-based estimates (e.g., sediment based on discharge or total phosphorus based on sediment), spatial or temporal interpolation (e.g., soil moisture), aggregation/disaggregation (e.g., large scale ≥1 km resolution evapotranspiration estimates), or re-scaling estimates of model outputs.

5.2.1 Identify and compile available hard data (Daggupati et al., 2015).

- Determine if there are sufficient data for model calibration and validation. If not, consider a proxy site approach.
- Evaluate if data collection (i.e., timing, frequency, compositing), handling/storage, processing, analysis, and QA/QC methods are suitable for intended purpose and modeling objectives. If not, consider collecting or utilizing additional data.
- Assess data collection methods to determine the uncertainty of the measured data (Harmel et al., 2006; Shirmohammadi et al., 2006; Guzman et al., 2015).
- Determine if the quality of the measured data is sufficient given the intended model use. If not, consider collecting or utilizing additional data with less uncertainty.
- Assess consistency among data sources (if multiple sources).
- Identify potential outliers and investigate whether they should be included or excluded from the data set.

5.2.2 Identify and compile relevant soft data (Arnold et al., 2015).

- Ensure soft data are representative of the site.

- Consider soft data sources such as: evapotranspiration estimated from remote sensing data, crop yield estimated from annual agricultural statistics, flood stages, results from prior modeling studies, and data in databases (e.g., Measured Annual Nutrient loads from AGricultural Environments, MANAGE, Harmel et al., 2016).

5.3 Separate the datasets into calibration and validation data sets.

5.3.1 The separation can be temporal (i.e., different periods) or spatial (i.e., different site locations), or both.

5.3.2 For large areas or watersheds, multi-site calibration should be used if data are available.

5.3.3 If there are not sufficient data for model calibration and evaluation, consider switching to a model with a strong physical/mechanistic basis as opposed to an empirical basis, which if properly parameterized is likely to better represent the system when uncalibrated.

6 Determine model performance measures to use in model evaluation

Model performance measures are critical for evaluating how well simulated values represent measured data. Selection of these measures should occur *a priori* to avoid a potential perception of biased evaluation. Although utilization of multiple performance measures, including graphical techniques and quantitative indicators, requires additional effort, it produces a more comprehensive evaluation of model performance and is recommended (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, Ritter and Muñoz-Carpena, 2013).

6.1 Determine the performance measures to be used for model performance evaluation.

6.1.1 Graphical measures include:

- scatter plots between measured and simulated values with the regression slope and intercept displayed,
- time series of simulated and measured values,
- cumulative distributions,
- duration curves,
- maps for field-scale and watershed-scale result comparison.
- In addition to standard graphical measures, also depict uncertainty in measured and/or predicted values (e.g., Figure 1).

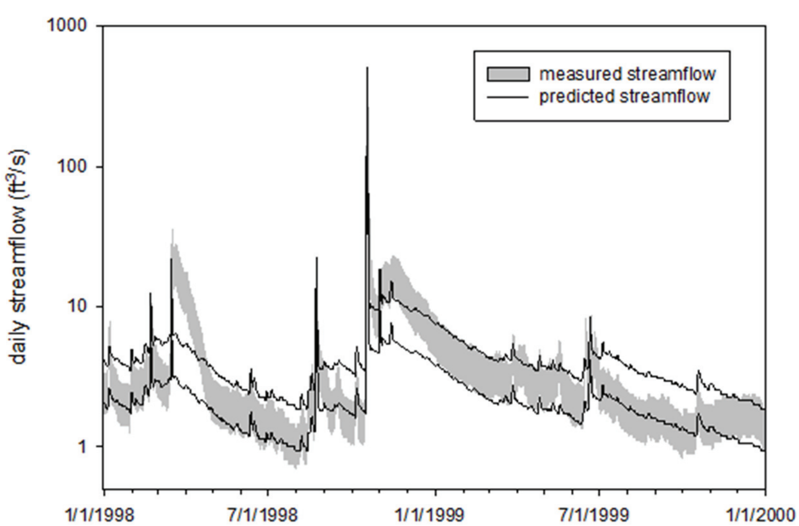


Figure 1 – Measured and simulated daily streamflow hydrograph for the Medina River (from Harmel et al., 2010). The uncertainty boundaries are presented as the shaded area for measured streamflow and as the upper and lower boundary lines for simulated streamflow.

6.1.2 Statistical measures include relative and absolute performance measures:

- Relative: Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970), index of agreement (d, d-index, Willmott, 1981; Willmott et al., 1985, 2011), percent bias (PBIAS, Gupta et al., 1999; Moriasi et al., 2007), ratio of RMSE to observations standard deviation (RSR, Moriasi et al., 2007).
- Absolute: root mean square error (RMSE), mean absolute error (Legates and McCabe, 1999).
- In addition to standard statistical measures and their possible variations to target a specific range of output values (Legates and McCabe, 1999), include modified measures that consider uncertainty in measured and/or predicted values (Harmel and Smith, 2007; Harmel et al., 2010).

7 Parameterize and calibrate the model

Parameterization is the process of determining a set of parameter values deemed suitable for model use in a specific study area. Calibration can be part of the parameterization process and consists of adjusting input parameter values and initial or boundary conditions within reasonable ranges until the simulated results closely match measured data. Calibration is generally reserved for those parameters that are not easily measurable, or are intrinsically heterogeneous or uncertain. The following guidelines, described in more detail in Daggupati et al. (2015) and Malone et al. (2015), are recommended good practices for parameterizing and calibrating H/WQ models.

7.1 Sensitivity analysis is highly recommended to determine which parameters most impact model outputs and should be prioritized for calibration.

7.1.1 Use informal approaches (e.g., experience with model, knowledge of study area, literature review) to identify parameters to be included in the sensitivity analysis. It is important to carefully and transparently describe the sources and assumptions used to establish the variability of the parameters. Be aware that changes made to model code can affect its parameter sensitivity.

7.1.2 Constrain parameter values within ranges appropriate to the study area within which to conduct sensitivity analysis.

7.1.3 Select an appropriate sensitivity analysis method (Yuan et al., 2015). A sensitivity analysis should be conducted for any new site to which the model is applied.

7.2 Select calibration parameters, minimizing the number of parameters.

7.2.1 Determine the variability of the sensitive parameters objectively based on published literature values, prior model calibration in a similar setting, or expert judgment.

7.2.2 For each sensitive parameter, determine whether it should be set using: (a) published values, (b) knowledge of the site, or (c) a model calibration procedure (Daggupati et al., 2015).

7.2.3 For each parameter selected for optimization, revise the range of parameter values based on the sensitivity analysis.

7.2.4 Use caution when optimizing correlated parameters. Ideally parameters selected for optimization should not be correlated.

7.3 Select the optimization objective function(s). Objective functions based on multiple model output variables are strongly recommended.

7.4 Develop the calibration strategy for the parameters that will be calibrated. The calibration strategy will vary with the number, location, and spatial scale of calibration/validation data sets, and the interaction between them.

7.4.1 Simple strategy: Applies for a single site with one or more variables calibrated independently. The outcome is one optimal parameter set.

7.4.2 Complex strategy: Applies for any combination of variables and sites that cannot be calibrated independently. Complex calibration strategies produce several “optimal” input parameter sets for each variable, site, and spatial and temporal scale, which must be reduced to a smaller number of unified “calibrated” parameter sets. Conditions where a complex strategy is needed include: one variable, multiple sites; multiple variables, single site; one variable, multiple spatial scales; and multi-variable, multi-site, multi-temporal scale.

7.5 Select one of the following three systematic calibration approaches. For any approach, consider manual calibration, automatic calibration, or a combination of both, depending on the number of calibration parameters and stages.

7.5.1 Single stage, the Pareto optimal approach. This approach is suitable for one single objective function, which can include one or multiple output variables. It is suited for calibrating a single or multiple input parameters. Combinations of parameter values are given as model inputs until an optimal combination is identified. It is possible to arrive at a range of parameter values for one single parameter or a set of parameter value combinations for multiple parameters that all produce equally optimal results.

7.5.2 Stepwise, single-pass. This approach is suitable when input parameters or sets of input parameters are optimized sequentially. Each optimization can use the same or different objective functions. Once a parameter or a set of input parameters has been calibrated, their values are not adjusted during subsequent optimizations.

7.5.3 Stepwise, iterative strategy. This approach is similar to the previous one in that input parameters (or sets of input parameters) are optimized sequentially but after each optimization, the previous set of parameters is re-adjusted. This is suited when there are significant interactions between simulated processes. As soon as the number of input parameters gets larger than 3 or 4, the process becomes tedious and automated calibration procedures are recommended.

7.6 Utilize a model warm-up period to reduce model dependence on estimates of state variables’ initial conditions. Choose a period that is appropriate to the model and the output variable. If this is unknown, it can be determined through a sensitivity analysis on the initial conditions.

7.7 Uncertainty analysis is highly recommended (Shirmohammadi et al., 2006; Guzman et al., 2015) to:

7.7.1 Guide potential model refinement following preliminary evaluation of model performance.

7.7.2 Evaluate model performance considering modeling objectives and the intended model use (Section 9).

7.7.3 Interpret and communicate the level of confidence in model results (Section 10).

7.8 Complete preliminary evaluation of model performance.

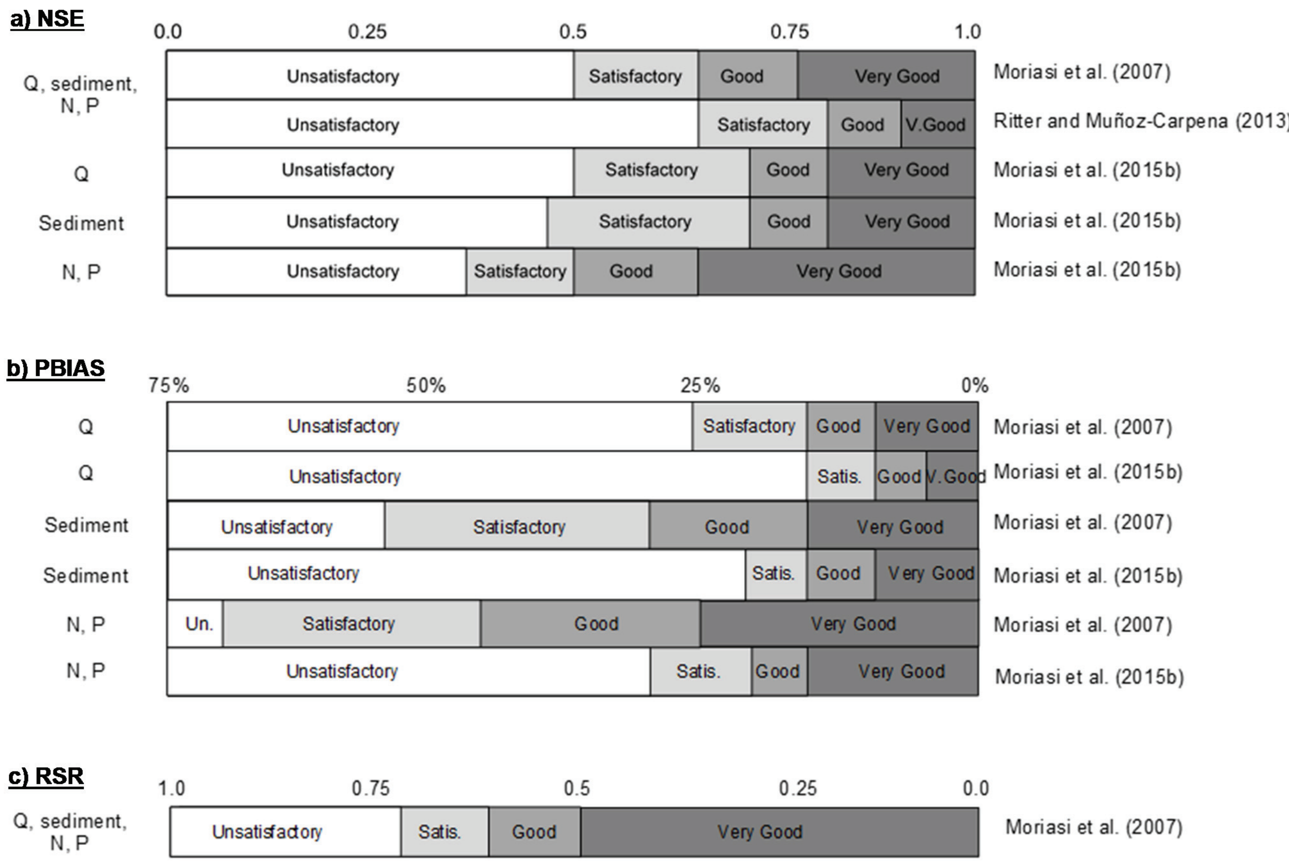
7.8.1 Examine graphical performance measures.

7.8.2 Calculate at least one relative and one absolute performance measure.

7.8.3 Determine performance rating (Figure 2).

7.8.3.1 The criteria developed by Moriasi et al. (2007, 2015b) apply to continuous, long-term simulations for the listed output responses, medium to large spatial scale studies, using measured data for calibration and validation obtained under the typical data quality scenario (Figure 2). The modified NSE and RSR values (Clause 6.1.2) can be used to account for uncertainty in CV&E data and model uncertainty. The rating guidelines can be adjusted to be stricter for Regulatory/Legal intended model uses and relaxed for Exploratory intended uses.

7.8.3.2 Evaluation of the statistical significance of the model’s goodness-of-fit is recommended. Ritter and Muñoz-Carpena (2013) provide an objective criterion for the selection of NSE ranges for model performance ratings based on a model prediction error indicator relative to the RMSE (Figure 2). This method, which is available in the FITEVAL software (Ritter and Muñoz-Carpena, 2013), evaluates the probability of the goodness-of-fit and can take into account the uncertainty of observed data and model predictions. The choice of a confidence level to reject the null hypothesis of unsatisfactory model performance should be based on the model intended use (i.e., how strong the evidence needs to be for accepting or rejecting H_0).



- For the Moriasi et al. (2007) ratings, ***all criteria*** must be met to achieve higher rating. Adjust strictness of criteria based on several factors including model intended use.
- The Ritter and Muñoz-Carpena (2013) ratings were designed to determine significance of model fit based on hypothesis test results. Utilize less restrictive significance level (e.g., α=0.10) for Exploratory uses; lower to α=0.05 or 0.01 with highly uncertain measured data or for Planning uses; lower to α=0.01 for Regulatory/Legal uses.
- For the Moriasi et al. (2015b) ratings, ***all criteria*** must be met to achieve higher rating. See Moriasi et al. (2015b) for applicable temporal scales for each statistical measure and output. Adjust strictness of criteria based on several factors including model intended use.
- Q - includes streamflow, surface runoff, base flow, and tile flow.

Figure 2 – Performance rating guidelines for a) NSE, b) PBIAS, and c) RSR of flow components at any resolution and monthly water quality loads. It is critical in applying these performance ratings to acknowledge that they are guidelines for establishing various levels of acceptability of model performance. They *do not* establish strict standards that can be applied in isolation to justify or reject model performance conclusions.

8 Refine the model and/or re-examine input and calibration data

Following the preliminary evaluation of model performance, additional examination of the model and input and calibration data is encouraged following the process outlined in Harmel et al. (2014).

8.1 Evaluate potential outliers. Ritter and Muñoz-Carpena (2013) provide a straightforward method for outlier detection. Evaluate potential outliers identified earlier in measured data used for calibration. Remove confirmed outliers. Annotate data points where the metadata suggest incorrect or questionable data. When the metadata do not flag the outliers, annotate those in the graphs and provide in the accompanying text a potential explanation of their cause and effect, and potential limits of the model.

8.2 Determine whether inclusion of extreme values 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. 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.

8.3 Examine potential magnitude bias or systematic error in predicted values. Bias 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. McCuen et al. (2006) demonstrated how bias can affect other performance measures.

8.4 Examine the time-offset bias, which 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. Efficiograms (plot of NSE vs. time lag) can be valuable in detecting and correcting time-offset bias. When time-offset bias is present, measured data should be re-examined as the problem can be caused by sampling or recording errors. 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.

8.5 Repeat uncertainty analysis if the model was refined or if input and calibration data were edited.

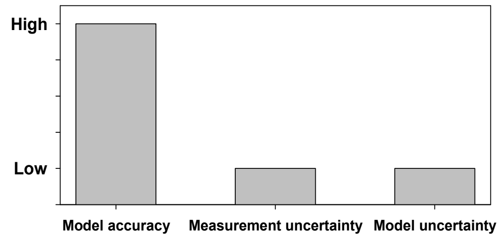
8.6 Use care when interpreting the results of the calibration or uncertainty analysis. Uncertainty estimates for predicted values for a given set of inputs do not necessarily translate to a different set of inputs or impute similar uncertainty characteristics to other predicted values.

9 Re-evaluate model performance

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). The model's intended use should determine the strictness of criteria used to determine acceptable model performance (Loague and Green, 1991) and be considered when interpreting model results and presenting relevant limitations (Wagener et al., 2001; Refsgaard et al., 2005; Jakeman et al., 2006).

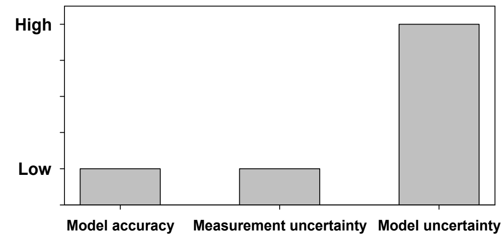
9.1 Re-apply the same graphical and statistical measures used in Clause 7.8.1 and 7.8.2. These measures along with model performance ratings (Moriasi et al., 2007, 2015b; Ritter and Muñoz-Carpena, 2013; Figure 2) considered together provide a comprehensive overview of model performance.

9.2 Interpret model results considering intended use, data uncertainty, and model accuracy and precision (Figure 3) based on recommendations of Harmel et al. (2014).



Interpretation: High precision and high accuracy provides confidence in predicted values.

Recommendations:
Appropriate for ***Exploratory***, ***Planning***, and ***Regulatory/Legal*** uses.



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.

Figure 3 – Model interpretation and refinement guidelines for intended model uses (from Harmel et al., 2014). Two examples from the eight combinations of High/Low model accuracy, High/Low measurement uncertainty, and High/Low model uncertainty (precision) are presented.

10 Document the process and report results

The increasing use of H/WQ models for technical, policy, and legal decision-making calls for greater transparency in communicating the methods used and decisions made. Many of the following recommendations described in detail in Saraswat et al. (2015) were developed to increase model reproducibility and facilitate appropriate interpretation of published modeling studies.

10.1 Report model intended use (e.g., exploratory, planning, and/or regulatory/legal) and describe how the intended model use affects model performance ratings.

10.2 Report descriptive information about the model.

- Model name, version, and release date (to facilitate reproducibility). Indicate how to obtain the model.
- Model type (e.g., mechanistic, quasi-mechanistic or empirical), spatio-temporal scales (e.g., point to watershed; sub-hourly to yearly), and model structure and components (e.g., hydrology, nutrients, sediment, chemicals, carbon, vegetation).

10.3 Describe the study area using maps where appropriate.

10.3.1 Location including scale, latitude and longitude, and a map showing the study area with the perimeter. Include the identifying watershed code(s), e.g., the Hydrologic Unit Code (HUC) for US watersheds.

10.3.2 Topography, including a general description (e.g., floodplain, rolling hills, mountainous), range of elevation, and slopes (average, maximum, minimum).

10.3.3 Soil and geology characteristics, especially those that affect hydrology and biogeochemistry, extent and distribution of major soil groups or types (depending on scale, with soil taxonomical name, textural class, and/or hydrologic soil group information) and relevant geological formations and materials.

10.3.4 Land use and land cover, including summary information on the types, extent, and distribution of land use (e.g., crop type, percent of watershed area, region(s) having the greatest density), any changes during the study period, and typical management practices (e.g., tillage, agrochemical application, irrigation, drainage, and/or fire management descriptions and timing).

10.3.5 Water resource structures (e.g., ditches or other constructed waterways; drainage structures, terraces, water detention ponds, or reservoirs) and other constructed or natural watershed features (e.g., canals; wetlands; vegetative filter strips or riparian buffers; connected floodplains).

10.3.6 Climate, including long-term average temperatures, precipitation, and the comparison of weather during the modeling period to the longer-term climate record.

10.3.7 Location and description of point sources, such as wastewater treatment plants, including constituent concentrations and total discharges into stream reaches (for watershed models).

10.3.8 Specific features of interest that help readers understand key aspects of the watershed or location.

10.4 Describe the CV&E data.

- Location, station identifier, and description of monitoring station(s) including stream bed and channel characteristics.
- Frequency and timing of all measurements.
- QA/QC protocols for all measurements.
- Estimated uncertainty of all data.
- For water quality constituent samples (e.g., sediment, nutrients, chemicals), describe method of sample collection, storage, preservation, and analysis, including citation to standard methods, if appropriate.
- For flow measurement, describe stage and/or velocity sensor, equipment model, manufacturer's reported accuracy and precision, presence of a weir or flume, method to determine flow (e.g., stage discharge relationship, velocity and cross-section measurement, Manning's equation), and equipment maintenance and calibration procedures.
- For in-situ sensors, describe equipment model, manufacturer's reported accuracy and precision, location and orientation in cross-section and in the water column, and equipment maintenance and calibration procedures.
- For automated water quality samplers, describe equipment model, location and orientation of the intake in cross-section and in the water column, sample type (discrete or composite), sampling interval (time or flow), and equipment maintenance and calibration procedures.
- Method used to derive constituent mass loads from streamflow and measured concentrations.
- Report temporal or spatial discontinuity between data sources. For example, many 24-hour precipitation values are recorded for 7:00 a.m. to 7:00 a.m. (or some other standard time), whereas the 24-hour USGS streamflow values are reported for 12:00 midnight to 12:00 midnight.

10.5 Describe input data used in the model.

10.5.1 If published data are used, cite the data source, including the metadata. Published datasets often specify a preferred citation, which should be used. When input data are not from a published source, basic information should be included, and the full dataset should be uploaded to an electronic repository if possible.

10.5.2 Report dates for which each source is used if more than one data source is combined (e.g., to fill incomplete weather records).

10.5.3 Report modifications, simplifications, or "data cleaning" procedures used in preparing input data, including assumptions made to acquire or process site-specific management data to make it compatible with the model.

10.5.4 For geospatial input data, note and report the resolution of the vector/raster dataset. For derived geospatial data, such as thematic map (e.g., land use raster), the acquisition date for the original imagery is essential for proper temporal representation of study-area features. In addition, note if the original resolution has been modified using a data layer with a different spatial resolution.

10.6 Report the approach used to select calibration parameters (e.g., sensitivity analysis, experience with model, literature review, expert opinion).

10.6.1 Include a table with the following information for each important model parameter.

- For default parameters (determined by model developers without user modification), cite reference(s).
- For baseline parameters (determined by model user based on a priori knowledge), provide justification and/or citation to support each value.
- For calibration parameters (modified after comparison to a response variable), provide the range of each one tested during calibration, a justification for that range, and the resolution of the parameter (e.g., ± 0.1 unit, ± 5%). This information should be provided even if the parameter retains its default value after calibration.

10.7 Report the following information related to the CV&E strategy.

- Locations and model processes for which model was calibrated.
- Systematic calibration approach used to adjust parameters and assess output.
- Allocation of the spatio-temporally distributed data (i.e., data splitting) for either calibration or validation.
- Description of automated calibration processes or algorithms used including: algorithm or software name and revision or version number, objective function, sampling method, number of simulations, any assumptions related to statistical distribution of the parameters, likelihood function, prior, and posterior distributions for Bayesian-type approaches.

10.8 Report the following information related to performance evaluation.

- Results of model uncertainty analysis
- Methods used to assess model performance for each output variable (graphical and statistical measures).
- Objective function(s) used to assess model performance either in the form of equation(s) and/or with citation(s) to standard references. If multiple objective functions are used, the sequence or combination of their use should be described. If more than one site is calibrated, describe the method used to combine them.
- Criteria used to determine when CV&E is deemed successful, considering the model intended use.
- Model calibration and validation results for each output variable, including comparison of observed and modeled values, model performance statistics, using an appropriate combination of text, tables, and graphics.

Annex A
(Informative)

Bibliography

Arnold, J.G., Youssef, M.A., Yen, H., White, M.J., Sheshukov, A.Y., Sadeghi, A.M, Moriasi, D.M., Steiner, J.L., Amatya, D.M., Skaggs, R.W., Haney, E.B., Jeong, J., Arabi, M., and Gowda, P.H. (2015). Hydrological processes and model representation: Implications of soft data on calibration. *Trans. ASABE* 58(6), 1637-1660.

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

Baffaut, C., Dabney, S., Smolen, M., Youssef, M.A., Bonta, J.V., Chu, M.L., Guzman, J.A, Shedekar, V.S, Jha M.K., and Arnold, J.G. (2015). Hydrologic and water quality modeling: Spatial and temporal considerations. *Trans. ASABE* 58(6), 1661-1680.

Beck, M.B. (1987). Water quality modeling: A review of the analysis of uncertainty. *Water Resources 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., and Andreassian, V. (2013). Characterising performance of environmental models. *Environ. Modelling Software* 40, 1-20.

Biondi, D., G. Freni, V., Iacobellis, G., Mascaro, and A., Montanari. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics Chem. Earth* 42-44, 70-76.

Black, D.C., Wallbrink, P.J., and Jordan, P.W. (2014). Towards best practice implementation and application of models for analysis of water resources management scenarios. *Environ. Modelling Software* 52, 136-148.

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K.R., Zeckoski, R., Jeong, J., Parajuli, P.B., Saraswat, D., and Youssef, M.A. (2015). Calibration and validation strategies for hydrologic and water quality models. *Trans. ASABE* 58(6), 1705-1719.

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

Gupta, H.V., Sorooshian, S., and Yapo, P.O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrologic Eng.* 4(2), 135-143.

Guzman, J.A., Shirmohammadi, A., Sadeghi, A., Wang, X., Chu, M.L., Jha, M.K., Parajuli, P.B., Harmel, R.D., Khare, Y.P., and Hernandez, J. (2015). Uncertainty considerations in calibration and validation of hydrological and water quality models. *Trans. ASABE* 58(6), 1745-1762.

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

Harmel, R.D., Christianson, L.E., Smith, D.R., McBroom, M.W., and Higgs, K.D. (2016). Expansion of the MANAGE Database with Forest and Drainage Studies. *J. American Water Resources Assoc.* doi:10.1111/1752-1688.12438.

Harmel, R.D., Smith, P.K., Migliaccio, K.W., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Muñoz-Carpena, R., and Robson, B.J. (2014). Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. *Environ. Modelling Software* 57, 40-51.

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

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

Jain, S.K., and 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., and, Norton J.P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environ. Modelling Software* 21, 602-614.

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

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

Malone, R.W., Yagow, G., Baffaut, C., Gitau, M., Qi, Z., Amatya, D., Parajuli, P.B., Bonta, J.V., and Green, T. (2015). Parameterization guidelines and considerations for hydrologic models. *Trans. ASABE* 58(6), 1681-1703.

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

Moriasi, D.N., Zeckoski, R.W., Arnold, J.G., Baffaut, C., Malone, R.W., Daggupati, P., Guzman, J.A., Saraswat, D., Yuan, Y., Wilson, B.N., Shirmohammadi, A., and Douglas-Mankin, K.R. (2015a). Hydrologic and water quality models: Key Calibration and Validation Topics. *Trans. ASABE* 58(6), 1609-1618.

Moriasi, D.N., Gitau, M.W., Pai, N., and Daggupati, P. (2015b). Hydrologic and water quality models: Performance measures and criteria. *Trans. ASABE* 58(6), 1763-1785.

Moriasi, D.N., Wilson, B.N., Douglas-Mankin, K.R, Arnold, J.G., and 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., and 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., and Wallender, W. (2006). Evaluation of modeling tools for TMDL development and implementation. *Trans. ASABE* 49(4), 961-965.

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

Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholte, H., and Kassahun, A. (2005). Quality assurance in model based water management — Review of existing practice and outline of new approaches. *Environ. Modelling Software* 20, 1201-1215.

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

Saraswat, D., Frankenberger, J.R., Pai, N., Ale, S., Daggupati, P., Douglas-Mankin K.R., and Youssef M.A. (2015). Hydrologic and water quality models: Documentation and reporting procedures for calibration, validation, and use. *Trans. ASABE* 58(6), 1787-1797.

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., and Sohrabi, T.M. (2006). Uncertainty in TMDL models. *Trans. ASABE* 49(4), 1033-1049.

Yuan, Y., Khare, Y., Wang, X., Parajuli, P.B., Kisekka, I., and Finsterle, S. (2015). Hydrologic and water quality models: Sensitivity. *Trans. ASABE* 58(6), 1721-1744.

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., and Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrol. Earth Syst. Sci.* 5(1), 13-26.

Willmott, C.J., Robeson, S.R., and Matsuura, K. (2011). A refined index of model performance. *Intl. J. Climatology.* doi: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., and 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. *Physical Geographer* 2(2), 184-194.

Zeckoski, R. W., Smolen, M., Moriasi, D., Frankenberger, J., and Feyereisen, G. (2015). Hydrologic and water quality terminology as applied to modeling. *Trans. ASABE* 58(6), 1619-1635.

