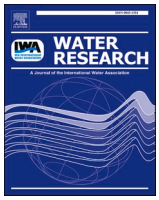
Water Research 274 (2025) 123156

Contents lists available at ScienceDirect

Water Research

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



Establishing performance criteria for evaluating watershed-scale sediment and nutrient models at fine temporal scales

Aayush Pandit [a], Sarah Hogan [a], David T. Mahoney [b], William I. Ford [c], James F. Fox [d], Christopher Wellen [e], Admin Husic [a,f,*]

[a] Department of Civil, Environmental, and Architectural Engineering, University of Kansas, United States
[b] Department of Civil and Environmental Engineering, University of Louisville, United States
[c] Department of Biosystems and Agricultural Engineering, University of Kentucky, United States
[d] Department of Civil Engineering, University of Kentucky, United States
[e] Department of Geography and Environmental Studies, Toronto Metropolitan University, Canada
[f] Department of Civil and Environmental Engineering, Virginia Tech, United States

ARTICLE INFO

Keywords:
Water quality
Watershed model
Performance measure
Evaluation criteria
Sediment
Nutrients

ABSTRACT

Watershed water quality models are mathematical tools used to simulate processes related to water, sediment, and nutrients. These models provide a framework that can be used to inform decision-making and the allocation of resources for watershed management. Therefore, it is critical to answer the question "when is a model good enough?" Established performance evaluation criteria, or thresholds for what is considered a 'good' model, provide common benchmarks against which model performance can be compared. Since the publication of prior meta-analyses on this topic, developments in the last decade necessitate further investigation, such as the advancement in high performance computing, the proliferation of aquatic sensors, and the development of machine learning algorithms. We surveyed the literature for quantitative model performance measures, including the Nash-Sutcliffe efficiency (NSE), with a particular focus on process-based models operating at fine temporal scales as their performance evaluation criteria are presently underdeveloped. The synthesis dataset was used to assess the influence of temporal resolution (sub-daily, daily, and monthly), calibration duration (< 3 years, 3 to 8 years, and > 8 years), and constituent target units (concentration, load, and yield) on model performance. The synthesis dataset includes 229 model applications, from which we use bootstrapping and personal modeling experience to establish sub-daily and daily performance evaluation criteria for flow, sediment, total nutrient, and dissolved nutrient models. For daily model evaluation, the NSE for sediment, total nutrient, and dissolved nutrient models should exceed 0.45, 0.30, and 0.35, respectively, for 'satisfactory' performance. Model performance generally improved when transitioning from short (< 3 years) to medium (3 to 8 years) calibration durations, but no additional gain was observed with longer (> 8 years) calibration. Dissolved nutrient models calibrated to load (e.g., kg/s) out-performed those calibrated to concentration (e.g., mg/L), whereas selection of target units was not significant for sediment and total nutrient models. We recommend the use of concentration rather than load as a water quality modeling target, as load may be biased by strong flow model performance whereas concentration provides a flow-independent measure of performance. Although the performance criteria developed herein are based on process-based models, they may be useful in assessing machine learning model performance. We demonstrate one such assessment on a recent deep learning model of daily nitrate prediction across the United States. The guidance presented here is intended to be used alongside, rather than to replace, the experience and modeling judgement of engineers and scientist who work to maintain our collective water resources.

* Corresponding author at: Virginia Tech, United States.
E-mail address: husic@vt.edu (A. Husic).

https://doi.org/10.1016/j.watres.2025.123156
Received 19 November 2024; Received in revised form 13 January 2025; Accepted 15 January 2025
Available online 18 January 2025

1. Introduction

Watershed models are powerful tools for simulating environmental management, estimating pollutant fate and transport, and predicting effects of climate and land use change (Fu et al., 2020; Keller et al., 2023). These models represent the transfer of matter through compartments of a watershed using a combination of empirical and theoretical equations (Fu et al., 2019). The predictions made by watershed models are most frequently compared against observations in rivers, which integrate many upland and in-stream processes into a single point (Fu et al., 2019). The relative accuracy of a model is assessed by calculating performance measures, which are derived from how well predictions fit to observations, and comparing these measures against literature-established performance evaluation criteria (Moriasi et al., 2015). In the decade(s) since Moriasi and others (2007, 2015) published their widely used benchmarks, advances in data collection and compute capabilities have pushed modelers toward simulating water quality variables (e.g., sediment) at timescales (e.g., daily) that were not considered in the original meta-analyses. Thus, there is a need to conduct a meta-analysis of the intervening literature and establish performance criteria for these variables at finer temporal resolutions.

The range of what model performance is considered 'good' or 'limited' typically varies as a function of the response variable of interest (e.g., flow or nitrate) and the temporal scale of modeling (e.g., daily or monthly). For example, a coefficient of determination (R^2) of 0.50 may be acceptable for sub-daily sediment modeling, but not for monthly flow modeling. That two models with the same performance measure (i.e., $R^2 = 0.50$) would have different qualitative interpretations indicates that modelers have more confidence, and thus higher expectations, in simulating monthly flow than sub-daily sediment. Historically, models of sediment and nutrients have been evaluated at relatively coarse temporal frequencies (e.g., monthly) (Chappell et al., 2017). This coarse evaluation was conducted in part due to limits imposed by the computational complexity of simulating fine-scale sediment and nutrient processes as well as a smaller volume of observational data. To the latter point, water quality variables have historically been collected periodically at discrete intervals, whereas flow data are broadly available and continuously monitored. As high-frequency water quality data collection efforts and computational resources improve (Bieroza et al., 2023), continual development of modeling performance criteria is needed to help guide effective model evaluation.

When selecting appropriate performance criteria for watershed water quality models, a number of factors should be considered, including the response variable sought after, temporal resolution of the model, duration of the calibration record, and the target units of the model variables. For example, modeling skill is expected to decrease for nutrients relative to flow due to the added complexity of modeling nutrient fate and transport processes (Wellen et al., 2015). Further, nutrient models adopt all of the uncertainties present in their parent hydrologic models, which control the residence time and routing of water (Husic et al., 2019). Model skill is also be expected to decrease with higher temporal resolution and longer calibration records, although this is not always the case (Moriasi et al., 2015). The expectation is that, given the same calibration data, a daily timestep model will likely encounter a greater variety of observations (variance) than a model aggregated to simulate at a monthly timestep (Chahor et al., 2014). Likewise, longer calibration records are expected to contain more diverse conditions than shorter records, thus making them more difficult to represent using a single calibration parameter set (Merz et al., 2009). Modeling performance criteria may also vary as a function of the target unit selected for evaluation (i.e., concentration, load, or yield), but this relationship is uncertain and understudied in past work.

Selection of target units for model calibration often depends on the context of a study, whether exploratory, planning, or regulatory (Harmel et al., 2014). Concentration is presented as mass per volume (e.g., mg/L), load as mass per time (e.g., kg/s), and yield as the area-normalized load (e.g., kg/ha/yr). Although load and yield encode for the same information, the use of one as opposed to the other depends on norms within a modeling community, the objectives of a modeling study, and the nature of evaluation data (e.g., channel erosion processes or reservoir sedimentation; Pandey et al., 2016). For the development of total maximum daily loads (TMDLs), a model calibrated to load (concentration times flow) may provide greater confidence in meeting specific study objectives (Amatya et al., 2010). For a study of concentration thresholds, such as how frequently a river exceeds the Environmental Protection Agency (EPA) maximum contaminant level (MCL), a concentration-based model would be more appropriate (Qi et al., 2012). The choice of the target units for calibration may change how internal processes are parameterized and how the workings of a model are interpreted or communicated to others. Given the expected variability of model performance measures as functions of response variable, temporal evaluation scale, calibration length, and target units, these factors are worthy of further investigation.

Looking forward, recent advances in data collection methods and computational approaches have the potential to shape the future of water quality modeling. Here, we focus on two such advances aquatic sensing and machine learning and indicate how a meta-analysis of the present state of watershed modeling can inform the integration of these advances into future model evaluation. First, in the last two decades, high-frequency sensors have made it possible to continuously monitor sediment and nutrient concentrations at resolutions (15-minute sampling frequency) far beyond what is feasible with traditional grab sampling (Burns et al., 2019; Rode et al., 2016). High resolution time series of water quality parameters encode information as to how watersheds function, particularly with regard to in-stream processing of nutrients (Zarnaghsh et al., 2024) and sources and flow pathways of contaminants (Mahoney et al., 2020). Second, advancements in machine learning are providing superior predictive performance in water quality modeling typically beyond what is possible with process-based models (McGill and Ford, 2024; Zhi et al., 2024). Deep learning models, a subset of machine learning models that excel at identifying patterns and dependencies in sequential data, are particularly suited for modeling water quality timeseries (Zhi et al., 2024). While we do not include machine learning models in the coming meta-analysis, their application for water quality modeling is rapidly growing and their assessments are often based on benchmarks developed for process-based models (Gorski et al., 2024; Saha et al., 2024; Song et al., 2024; Zhi et al., 2023). As the frontier for watershed modeling advances, both in terms of observational data and computational tools, it is crucial to identify the present state of water quality model evaluation so as to provide a framework for evaluating future modeling efforts.

In this paper, we carry out a meta-analysis of 229 process-based watershed model applications, extracted from 99 papers published between 2010 and 2022, with the intent to continue the development of performance evaluation criteria. We focus our meta-analysis on models that simulate sediment or nutrient dynamics, as performance criteria for sediment and nutrients is lacking relative to streamflow, particularly for finer temporal resolutions. We place emphasis on the response variable, target units, temporal resolution, and duration of calibration. The modeling results compiled from the 229 watershed models are then statistically analyzed to set up ranges for what constitutes a 'limited', 'satisfactory', 'good', and 'very good' model performance, for permutations of evaluation metric, response variable, and temporal resolution. The developed performance criteria are appended to the existing ranges reported in Moriasi and others (2015) to create a model guidance that spans a wide temporal range (sub-daily to monthly) for a diverse set of water quality variables (sediment, total nutrients, and dissolved nutrients). Lastly, we provide suggestions for the integration of sensors into modeling frameworks and we assess the performance of a recent deep learning water quality model application using the evaluation criteria established by our synthesis and meta-analysis.

2. Materials and methods

2.1. Selection of a large sample of watershed water quality model applications

A number of modeling selection attributes were considered when choosing studies to include in the meta-analysis. First, we used 2010 as a lower constraint to our model selection given that it is soon after Moriasi and others (2007) published statistical benchmarks, which have since become standard for model evaluation. Further, the review by Wellen et al. (2015) included papers up until the year 2010, thus assessing water quality model evaluation in the subsequent years is necessary. Many post-2015 studies also reference the refined suggestions found in Moriasi and others (2015). Second, we chose models that were process-orientated in nature, as opposed to purely empirical or data-driven. Distributed, semi-lumped, and lumped modeling frameworks fall under our acceptable classification. The third selection criteria was that at least one of the forms of sediment, nitrogen (dissolved or total), or phosphorus (dissolved or total) was simulated. Total nutrients represent the sum of the particulate and dissolved forms of nitrogen and phosphorus, respectively, and are frequently used to monitor ecological condition (Chambers et al., 2011). The final criteria was that at least one statistical metric was used to evaluate the model's predictive ability compared to observed data.

We used two types of journal database searches to find papers for inclusion in the meta-analysis. First, given that studies by Moriasi et al. (2007, 2015) have now become the standard for benchmarking watershed-scale water quality models, we utilized the "Cited by" feature in Google Scholar to search for and identify papers that evaluated the ability of a watershed model to predict sediment or nutrient transport using a benchmark identified by Moriasi et al (2007, 2015). We narrowed the "Cited by" list to contain at least one of the following keywords, including: "watershed modeling" in combination with "sediment (s)," "nutrient(s)," "nitrogen," "nitrate," "phosphorus," or "phosphate." Second, in a separate search of Google Scholar, we specifically sought out studies that utilized high temporal resolution water quality data, such as in situ sensors or automated samplers. The second cited reference search contained the previous keywords plus one of the following: "sensor(s)," "high-resolution," "sub-daily," "hourly," or "event." The 99 papers we synthesized included 229 water quality model applications using 37 unique model types in 27 different countries (Fig. 1). The total number of water quality model applications in our synthesis for sediment, nitrogen, and phosphorus is similar to that of Moriasi et al. (2015), providing us confidence in proceeding with the meta-analysis.

With the studies aggregated, we seek to answer the following questions:

1. Does the temporal resolution (sub-daily, daily, or monthly) of model evaluation systematically impact performance metrics?
2. Does the duration (< 3 years, 3 to 8 years, or > 8 years) of the calibration record systematically impact performance metrics?
3. Does the selection of modeled units (concentration, load, or yield) systematically impact performance metrics?
4. With the proliferation of high-frequency sensors and rise of machine learning algorithms in the 2010s, what should the benchmarks of water quality modeling be for the next decade(s)?

2.2. Extraction of data from watershed water quality model applications

For each study identified in Section 2.1, a common spreadsheet was completed to standardize data extraction and facilitate the meta-analysis. The spreadsheet included areas to enter metadata, such as the article title, author list, and publication year. The watershed model used and the number of basins modeled were extracted. When readily available, the watershed area, dominant land use, mean annual precipitation, and mean annual temperature were recorded. For each basin, the primary values of interest for our meta-analysis were organized as follows:

- **Response variable**: streamflow, sediment, total nutrients, and dissolved nutrients. Total nutrients include total nitrogen (TN) and total phosphorus (TP). Dissolved nutrients include nitrate (NO_3^-) and orthophosphate (PO_4^{3-}).
- **Temporal resolution**: sub-daily (SD), daily (D), and monthly (M).
- **Target units**: concentration, load, and yield.
- **Duration of calibration**: recorded as a number of years and grouped into three categories (< 3 years, 3 to 8 years, and > 8 years). The duration of validation (when performed) was also recorded.
- **Evaluation metric**: Nash-Sutcliffe efficiency (NSE), coefficient of determination (R^2), and percent bias (pBIAS).

An example entry is the following: a model simulates flow and

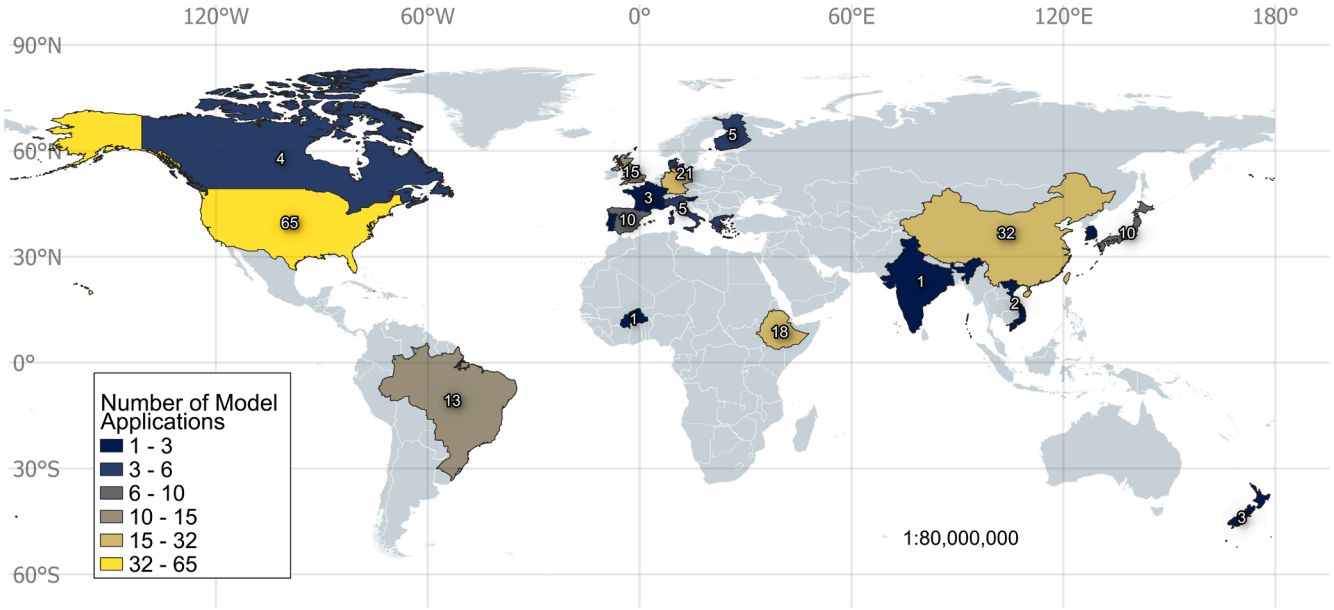


Fig. 1. Spatial coverage of watershed model applications (n = 229) obtained from papers meeting study search criteria.

sediment concentration at a daily timestep for a 5-year calibration period and is evaluated using NSE and R^2 metrics. The individual entries for each model application were compiled into a machine-readable format that serves as the dataset for this meta-analysis (see Supplemental Data Set 1).

Regarding performance measures, a large variety of statistical metrics were reported in the literature. We focus our analysis on NSE, R^2, and pBIAS. These performance measures are commonly applied, in part due to their simplicity of use, wide recognition of these metrics in the community, and availability of benchmarks from published works, such as Moriasi et al. (2007, 2015). The NSE metric is a measure of how well the temporal patterns of observed data are captured by the model (Nash and Sutcliffe, 1970). An NSE value of 1 indicates model explains all temporal patterns properly whereas a value of 0 indicates the model only performs as good as the mean of observation. NSE is bound by negative infinity (-∞). The metric can be calculated as follows:

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(Q_{obs} - Q_{sim})^2}{\sum_{i=1}^{n}(Q_{obs} - \overline{Q_{obs}})^2}\right]$$

where, Q_{obs} is the observed streamflow and Q_{sim} is the simulated streamflow. A bar above either variable denotes averaging. The NSE has been noted to emphasize the peaks of the hydrograph or pollutograph, and de-emphasize the lower values (Moriasi et al., 2015).

The R^2 metric measures the proportion of variability in the observed data that can be explained by the model (Pearson, 1901). A value of 1 implies model can explain all variability while a 0 implies the model explains none of the variability in the observed data. The metric can be calculated as follows:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Q_{obs} - \overline{Q_{obs}})(Q_{sim} - \overline{Q_{sim}})}{\sum_{i=1}^{n}(Q_{obs} - \overline{Q_{obs}})^2\sum_{i=1}^{n}(Q_{sim} - \overline{Q_{sim}})^2}\right]^2$$

The pBIAS metric measures average magnitude and direction of bias error of model (Sorooshian et al., 1993). A positive pBIAS indicates model overpredicts while a negative pBIAS indicates model suffers from underprediction bias. This can give inflated results when model converges to mean of observed data and therefore will be insensitive to model variability. The metric can be caclulated as follows:

$$pBIAS = 100 \times \left[\frac{\sum_{i=1}^{n}(Q_{obs} - Q_{sim})}{\sum_{i=1}^{n}(Q_{obs})}\right]$$

2.3. Meta-analysis of modeling evaluation metrics

The dataset generated in Section 2.2 was assembled and stratified under several conditions, which were used as the groupings for generating statistics and conducting inter-group comparison. One stratification condition is model temporal resolution (SD, D, and M). A second stratification condition is duration of calibration data (< 3 years, 3 to 8 years, and > 8 years). A third stratification condition is the units of the target variable (concentration, load, and yield). For each condition, the NSE, R^2, and pBIAS values during calibration (and validation) were aggregated for flow, sediment, total nutrient, and dissolved nutrient models, respectively.

Box-and-whisker plots were generated for each permutation within the three stratification conditions. The plots indicate the general distribution of literature-reported values with a central mark that indicates the median (50th percentile), box edges that indicate the 25th and 75th percentiles, and whiskers that extend to 1.5 times the interquartile range from the box edges. The box-and-whisker plots were used for visual assessment of the distribution of evaluation metrics.

Significant inter-group differences identified by the Kruskal-Wallis test (Kruskal and Wallis, 1952) were investigated using the two-sided Wilcoxon rank sum test (Wilcoxon, 1945). This test is nonparametric and evaluates the null hypothesis that two random variables have equal medians ($\alpha = 0.05$). Rejecting the null hypothesis indicates that the medians of the random variables are significantly different. Prior studies have shown significant differences in the values of evaluation metrics for flow versus water quality variables (Wellen et al., 2015) and calibration versus validation periods (Moriasi et al., 2015). Thus, we do not repeat these analyses. In this study, we focus on within-variable (e.g., sediment) differences in evaluation metrics (NSE, R^2, pBIAS) during calibration based on the three stratification conditions: temporal resolution (SD, D, M), calibration duration (< 3 years, 3 to 8 years, > 8 years), and target units (concentration, load, yield).

2.4. Developing performance evaluation criteria at fine temporal scales

The intent of this study was not to replace the metrics established in Moriasi et al. (2015). Rather, we use their established metrics, together with our meta-analysis, to refine their suggested performance thresholds, particularly where finer temporal resolutions benchmarks were lacking. Thus, we begin with the Moriasi et al. (2015) metric suggestions as the default values and refine them if our results show robust evidence toward a new performance evaluation criteria. Performance metrics during calibration and validation (within the same permutation) were combined to increase sample size for generating recommendations. Recommendations were not provided when data were insufficient (i.e., small sample size, n < 10) for a given pairing of a water quality variable and temporal scale. The collective modeling experience and judgement of this study's authors were used to create performance criteria that serve as a resource to modelers for use in assessing the accuracy of their models.

To develop performance metric recommendations for flow, sediment, total nutrient, and dissolved nutrient variables, we assessed the distributions of reported NSE, R^2, and pBIAS values. First, it must be acknowledged that not every model that is published can be considered 'satisfactory' or even 'acceptable'. For example, our aggregated dataset includes the reported NSE calibration values for 103 models of sediment, regardless of model timestep (SD, D, M) or target unit (concentration, load, yield). Seven of these model applications have negative NSE values, which are defined as having a worse predictive capability than the mean of the observation data. By all accounts, these models would be considered 'unacceptable'. While these models may not be acceptable, their publication and evaluation can nonetheless be informative. Thus, for this reason, direct use of published statistics, such as through percentiles of the raw reported values, may not constitute a recommendation that reflects satisfactory model performance. To mitigate this issue, we performed bootstrapping with replacement to generate robust statistics (Efron, 1979). These statistics formed the basis of our initial performance thresholds. A concept diagram of our approach is shown in Figure S1. For each bootstrap sampling of a metric-variable-timestep pairing (e.g., NSE for daily nitrate), the resampled median was calculated. This process was repeated 1000 times to generate a distribution of resampled medians. Thereafter, the 2.5th, 50th, and 97.5th percentile values of the resampled medians were used as initial guideposts for identifying thresholds of 'limited', 'satisfactory', 'good', and 'very good' model performance, respectively. This approach yielded reasonably similar performance evaluation criteria for the lower resolution timesteps evaluated in Moriasi et al. (2015), providing us confidence in the general approach.

The modeling results compiled from the 229 watershed models were statistically analyzed to set up initial performance thresholds for each permutation of response variable, temporal resolution, and evaluation metric. A 'limited' model is one where the evaluation metric is below the 2.5th percentile of resampled medians. A 'satisfactory' model exceeds the 2.5th percentile, a 'good' model exceeds the 50th percentile, and a 'very good' model exceeds 97.5th percentile of resampled medians. As in Moriasi et al. (2015), these approximate initial recommendations are further refined using the synthesis of performance criteria from the existing literature as well as the modeling experience of the authors. In

general, we tend to keep in place the Moriasi et al. (2015) recommendations for permutations assessed in their study, such as for their robust streamflow assessment and their lower-resolution temporal evaluation of water quality variables. For simplicity, thresholds were rounded conservatively to a lower increment of 0.05 (for NSE and R^2). The minimum width for each NSE and R^2 performance evaluation criteria bin was set to 0.05. If multiple temporal resolutions (SD, D, or M) had similar distributions for a given metric and water quality variable, they were aggregated into a single recommendation. The final step in determining guidelines was that the performance criteria for models evaluated at a finer timescale (e.g., SD) should not be more strict than those at a coarser timescale (e.g., M). In instances where this occurs, the two (or more) timescales were aggregated into a single performance evaluation criteria, as in Moriasi et al. (2015).

3. Results

A total of 229 model applications, from 99 research articles, were synthesized for this analysis of water quality model performance (Fig. 1). Ninety-seven of the 229 model applications (41.8%) were conducted either in the United States or China. The Soil Water and Assessment Tool (SWAT) was by far the most frequently used water quality model, accounting for 57% of all model applications (Fig. 2), followed by the Hydrological Predictions for the Environment (HYPE), MIKE Systeme Hydrologique Europeen (MIKE SHE), and Hydrological Simulation Program (HSPF) models at around 5% each. Thirty-two other modeling variants constitute the remaining 29% of model applications. Models were most frequently evaluated at a monthly timestep (42%), followed by daily (40%), and then sub-daily (18%). When a dominant land use was reported in a paper, it was most frequently agriculture, followed by forest, and then grassland. Models were applied across a wide range of basin areas, most frequently between 10 km^2 and 10,000 km^2. The mean (± one standard deviation) annual precipitation and air temperature of all basins was 1086 ± 490 mm/year and 13.5 ± 6.1 °C, respectively. Flow model performance was evaluated in 67% of model applications (Fig. 3), followed by sediment in 50%, NO_3^- in 38%, TP in 28%, TN in 21%, and PO_4^{3-} in 9%. Below, we focus the main text presentation on NSE as the performance measure of interest but note results for R^2 and pBIAS as well.

3.1. Effects of temporal resolution

Performance measures for sub-daily (SD), daily (D), and monthly (M) evaluation timesteps are shown in Fig. 4. The sample size of couplings of water quality variable and evaluation timestep ranged from n = 2 for sub-daily total nutrient evaluation to n = 66 for daily flow evaluation. Models of flow generally performed the best, followed by sediment, dissolved nutrient, and then total nutrient models. The low performance of total nutrient models may be a result of the compounding uncertainties of simulating both particulate and dissolved forms in one model. For most couplings, median values of NSE were generally greater during calibration than validation, although the difference in values was rarely significant when assessed with the Wilcoxon test. Hereafter, all results are in reference to the calibration period unless noted otherwise, as recent work has pointed to reconsidering the split-sample approach and forgoing validation altogether (Shen et al., 2022).

For the flow variable, models evaluated at sub-daily timesteps had the greatest median NSE (0.79) (Fig. 4). The median NSE for daily evaluation (0.69) was significantly lower than both sub-daily ($p = 0.04$) and monthly ($p = 0.01$) evaluation. The sub-daily and monthly median NSE values were not significantly different ($p = 0.66$). The lowest NSE whisker boundary for a flow model (0.32) occurs at the daily timestep. Similar to flow model evaluation, the median NSE for sediment, total nutrient, and dissolved nutrient models decreases from sub-daily to daily before increasing from daily to monthly. However, trends for these variables were not significant. The variability (spread) in reported water quality model performance is much greater than for flow models for NSE, R^2, and pBIAS. Several sediment, total nutrient, and dissolved

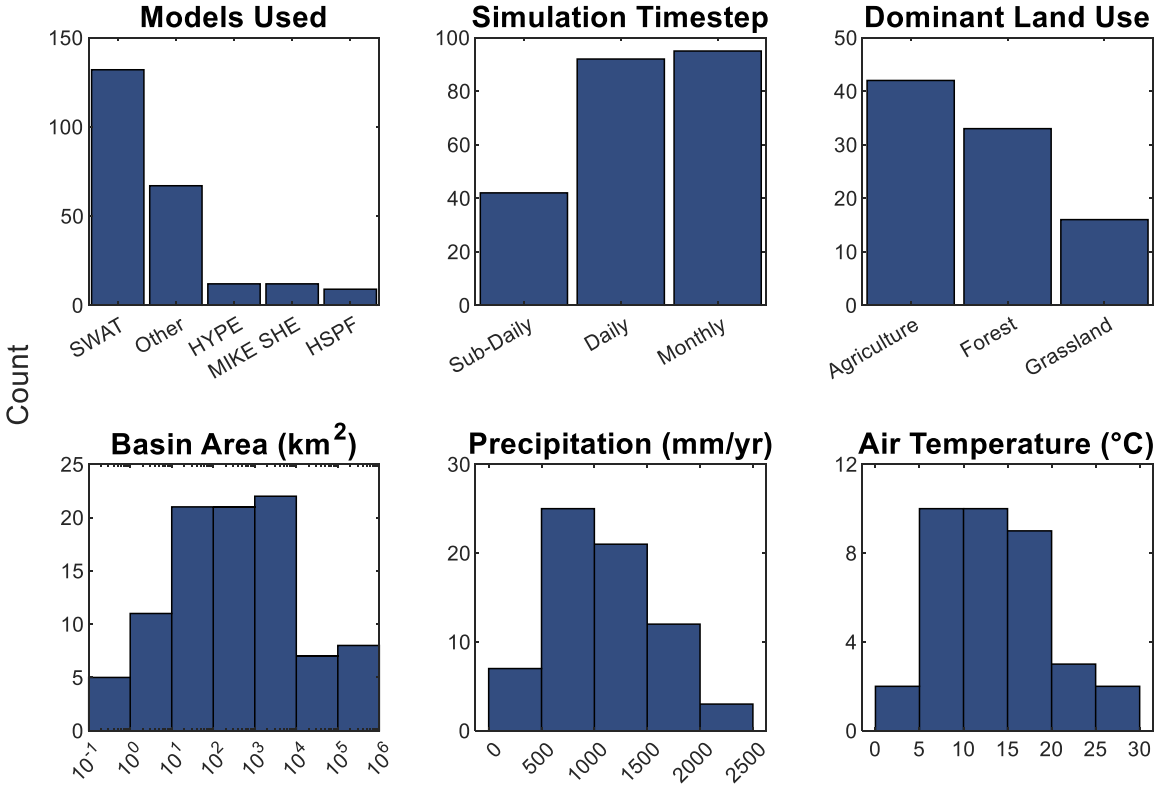


Fig. 2. Histograms of models used, simulation timestep, dominant land use, basin area, precipitation, and air temperature for basins in the meta-analysis. Not all studies reported values for each variable, thus counts may differ between plots.

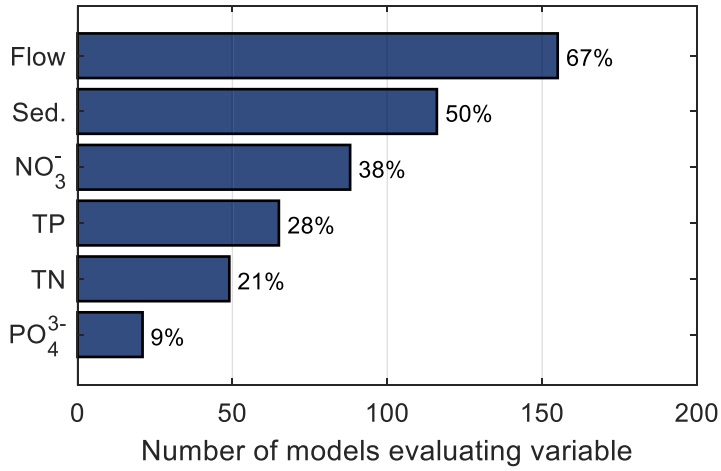


Fig. 3. Number (and percentage) of model applications that evaluate a given variable. Sed. = sediment, TN = total nitrogen, TP = total phosphorus, NO_3^- = nitrate, and PO_4^{3-} = orthophosphate.

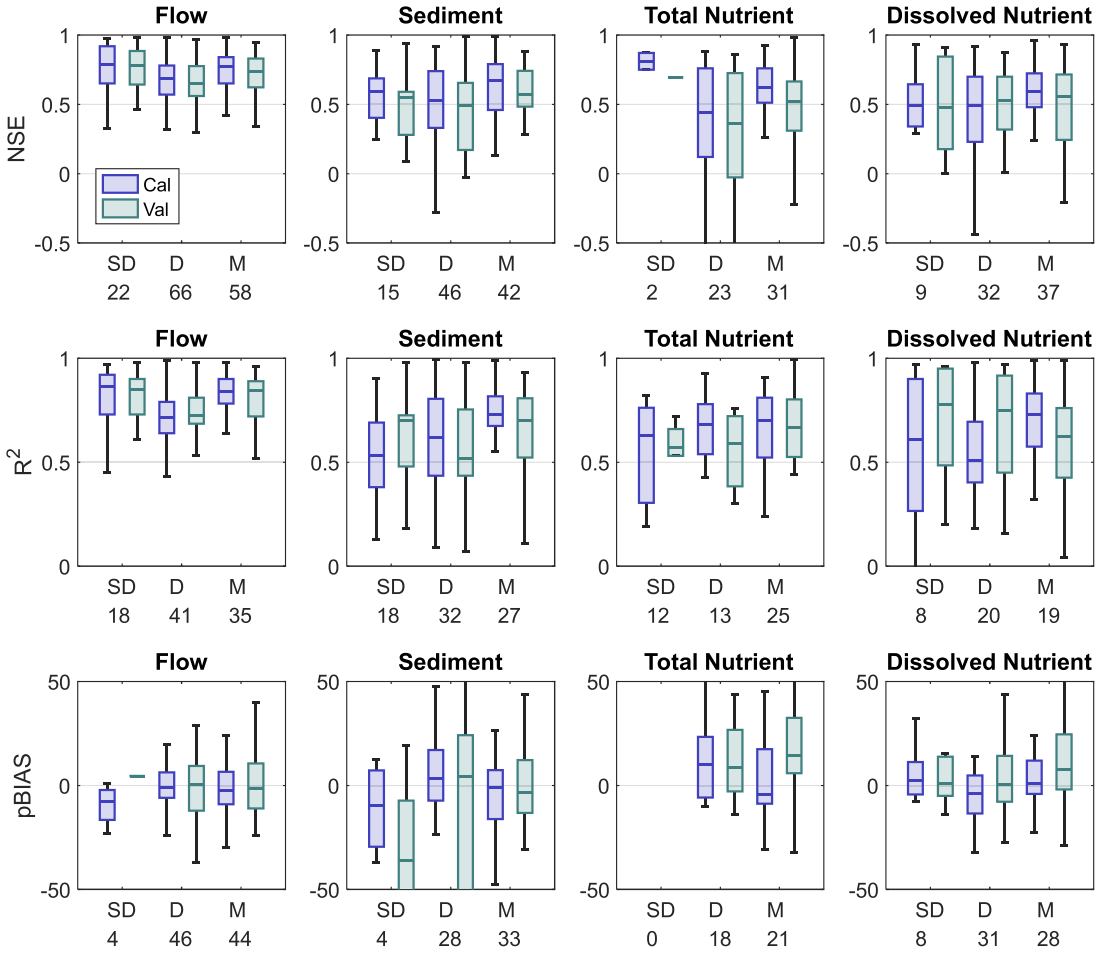


Fig. 4. Nash-Sutcliffe efficiency (NSE), coefficient of determination (R^2), percent bias (pBIAS) performance for flow, sediment, total nutrients (TN + TP), and dissolved nutrients ($NO_3^- + PO_4^{3-}$) prediction as a function of the model temporal resolution: sub-daily (SD), daily (D), or monthly (M). The number of basins included in a group are listed below the x-tick label. Boxplots show the median, interquartile range, and whiskers for each data grouping (outliers not shown). Cal = calibration. Val = validation.

nutrient model applications have NSE values lower than 0. Negative NSE values indicate that the model predictions are less accurate than using the observed mean value as the prediction.

3.2. *Effects of calibration duration*

Performance evaluation metrics for different durations of model calibration (< 3 years, 3 to 8 years, or > 8 years) are shown in Fig. 5. For all variables, the median NSE increased when models were calibrated on

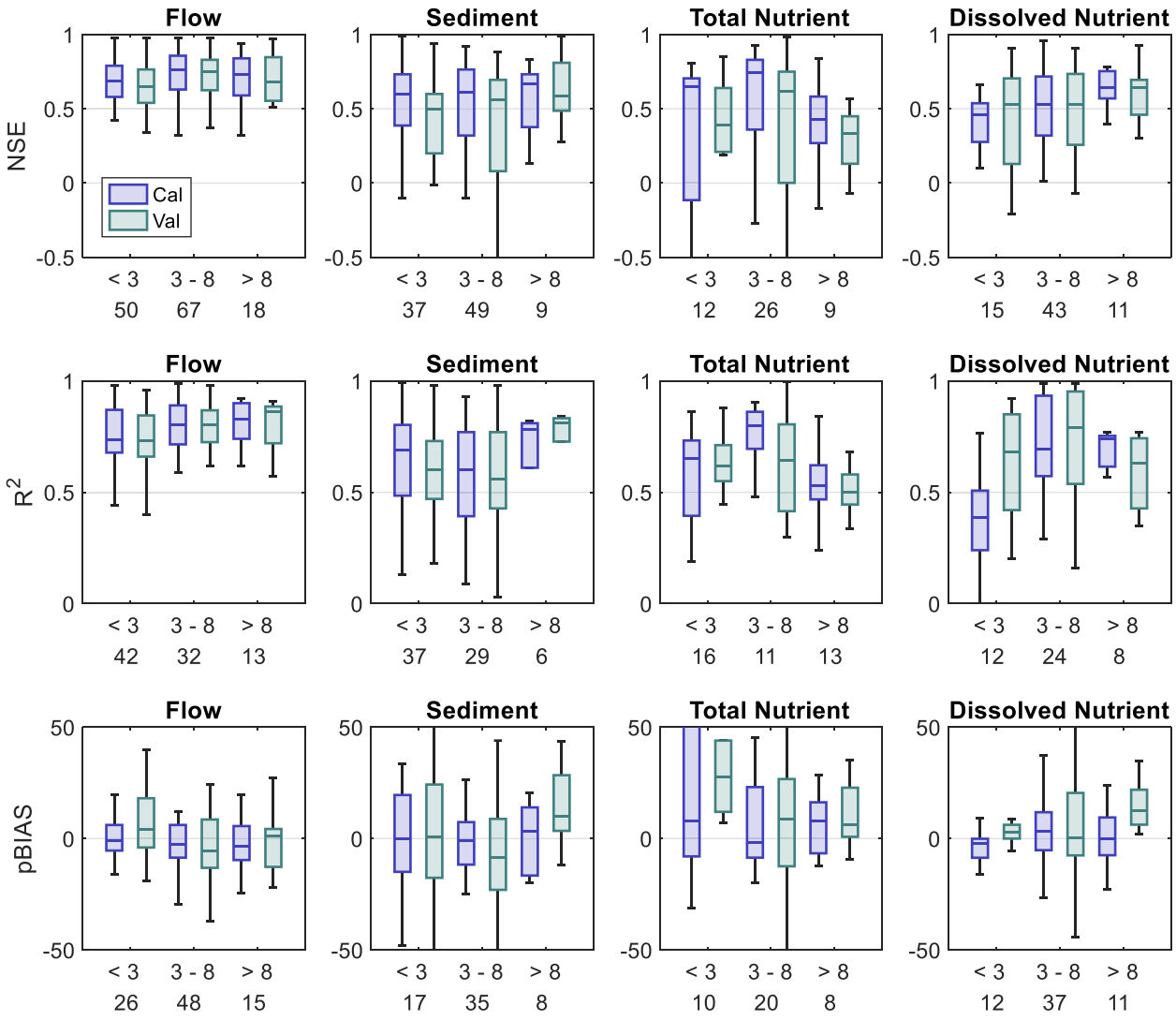


Fig. 5. Nash-Sutcliffe efficiency (NSE), coefficient of determination (R^2), percent bias (pBIAS) performance for flow, sediment, total nutrients (TN + TP), and dissolved nutrients ($NO_3^- + PO_4^{3-}$) simulation as a function of calibration duration: < 3 years, 3 to 8 years, or > 8 years. The number of basins included in a group are listed below the x-tick label. Boxplots show the median, interquartile range, and whiskers for each data grouping (outliers not shown). Cal = calibration. Val = validation.

a longer duration of observations, at least initially. This increase in median NSE was consistent when transitioning from short (< 3 years) to medium (3 to 8 years) calibration records for all variables but was only significant for flow ($p = 0.04$). From medium (3 to 8 years) to long (> 8 years) records, no significant increases in median NSE for any variables. Similar patterns for initial (short-to-medium duration) improvement in performance with more calibration data were also noted for R^2 for flow, total nutrient, and dissolved nutrient models, but not sediment. As calibration duration increased, the width of the interquartile range generally reduced for most evaluation metrics and variables.

3.3. Effects of target units

Performance evaluation metrics for different units of the constituent variable (concentration, load, and yield) are shown in Fig. 6. Load was the most common calibration target for sediment and total nutrients. For dissolved nutrients, concentration was the most common unit used for calibration. Units of yield were common for sediment modeling, but not total or dissolved nutrient modeling, likely due to norms in the sediment modeling community (Pandey et al., 2016). Sediment models calibrated to units of concentration (NSE = 0.64) had similar performance ($p = 0.25$) to models calibrated on load (NSE = 0.61). Total nutrient models calibrated to load (NSE = 0.66) were not significantly different ($p = 0.84$) from those calibrated to concentration (NSE = 0.59). Dissolved nutrient models calibrated to load (NSE = 0.60) significantly outperformed ($p = 0.01$) those calibrated to concentration (NSE = 0.45). The sediment and total nutrient models calibrated to yield observations had the lowest median NSE and the widest ranges of NSE values. Models calibrated to yield were less common and generally had the lowest sample size. The performance of yield-based models of total nutrients (n = 9) were poor (median NSE = -0.27) as nearly half of the model applications (n = 4) were from a small-scale study that yielded negative values (Ramirez-Avila et al., 2017).

3.4. Sensor data use in model evaluation

We noted water quality model applications that used aquatic sensors, which allow for daily or sub-daily model evaluation. Of the 229 model applications, 47 were informed by some type of sub-daily measurement system, often in situ sensors, for sediment (as turbidity) or dissolved nutrients (nitrate and orthophosphate). Sensors for total nutrients are not presently available, although turbidity was occasionally used as a

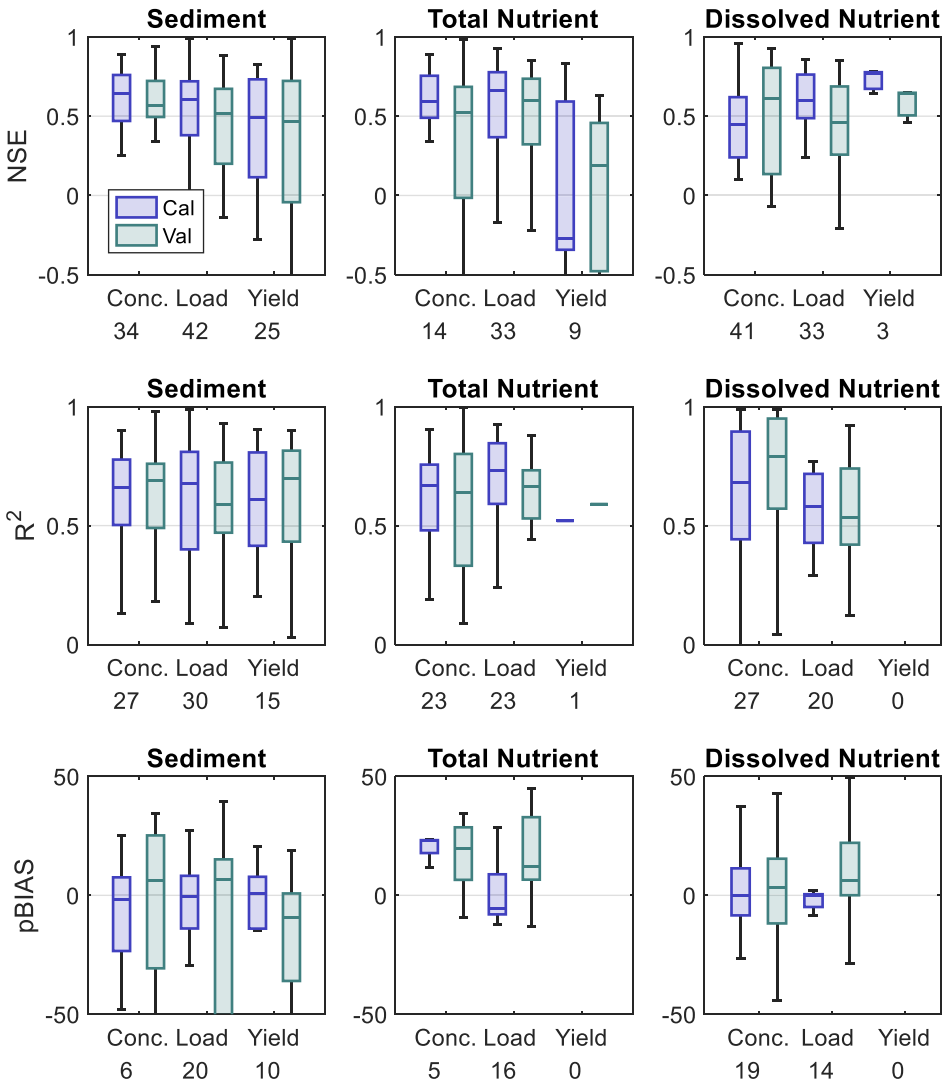


Fig. 6. Nash-Sutcliffe efficiency (NSE), coefficient of determination (R^2), percent bias (pBIAS) performance for sediment, total nutrients (TN + TP), and dissolved nutrients ($NO_3^- + PO_4^{3-}$) models as a function of the units of the target variable: concentration, load, or yield. The number of basins included in a group are listed below the x-tick label. Boxplots show the median, interquartile range, and whiskers for each data grouping (outliers not shown). Cal = calibration. Val = validation.

proxy for TN and TP. For sediment, common sensors were the YSI 6-Series and YSI EXO2 multiparameter sondes (Xylem Inc., USA) and the Compact-CLW and Infinity-CLW turbidity sensors (JFE Advantech Co., Ltd., Japan). For nitrate, common sensors were the S::CAN spectralyser sensor (Messtechnik GmbH, Austria), the OPUS UV sensor (TriOS GmbH, Germany), the Nitratax UV sensor (Hach Company, USA), and the SUNA V2 sensor (Sea-Bird Scientific, USA). For orthophosphate, common sensors were the HydroCycle-PO_4 wet chemical sensor (Sea-Bird Scientific, USA) and the Phosphax Sigma analyzer (Hach Company, USA). As these sensors indirectly measure the variable of interest, sensor estimates were often related to laboratory measurements derived from either grab samples or automated samplers (e.g., ISCO 6712; Teledyne, Inc., USA).

3.6. Bootstrapping performance metric benchmarks

Boot strapping statistics are shown for flow (Table S1), sediment (Table S2), total nutrient (Table S3), and dissolved nutriens (Table S4) variables. Bootstrapping estimates were created to generate a distribution of resampled medians from recorded performance measures (Figure S1). For flow, the median resampled NSE for SD, D, and M evaluation were 0.76, 0.65, and 0.72 respectively (Table S1). The confidence intervals (2.5th and 97.5th percentiles) for NSE were [0.69, 0.84] for SD evaluation, [0.63, 0.71] for D evaluation, and [0.72 0.80] for M evaluation. For sediment, dissolved nutrients, and total nutrients, distributions were generated for both temporal scale of evaluation (SD, D, M) and the constituent units (concentration, load, yield). For sediment, the median resampled NSE for SD, D, and M evaluation of a concentration-based model were 0.55, 0.54, and 0.68, respectively (Table S2). For the load-based sediment model, the median resampled NSE for D and M (no SD estimates were made due to small sample size) were 0.50 and 0.67, respectively. For total nutrients (Table S3) and dissolved nutrients (Table S4), the confidence intervals were wider than for sediment and flow given greater variability in reported model performances and the smaller sample sizes for nutrient-based models. For dissolved nutrients, the sampled median NSE for SD, D, and M evaluation of a concentration-based model were 0.42, 0.52, and 0.55, respectively (Table S4). The uncertainty bounds for the concentration-based dissolved nutrient model were [0.29, 0.75] for the SD evaluation, [0.35, 0.66] for D evaluation, and [0.44, 0.65] for M evaluation. These bootstrapped results were used in conjunction with benchmarks established in Moriasi et al. (2015) to update performance evaluation criteria, particular at finer temporal scales (SD and D).

4. Discussion

In this discussion, we recommend performance criteria for water quality model evaluation at fine temporal scales. Thereafter, we discuss modeling considerations and practices, including whether to calibrate a water quality model to concentration or load. Lastly, we look ahead to the next generation of water quality models where big data and machine learning will challenge the present approach to how models are constructed, evaluated, and operationalized.

4.1. Recommended performance evaluation criteria

A major goal of this work was to recommend performance evaluation criteria for water quality models, particularly those evaluated at fine temporal resolutions. The results of this analysis are presented in Table 1. Our analysis builds upon the work of Moriasi et al. (2015) in several key ways. The finest temporal resolution of performance metric (for NSE and R^2) evaluation established in their study was daily for flow and monthly for sediment, total nutrients, and dissolved nutrients. We extend their analysis and present daily performance criteria for NSE and R^2 for all water quality constituents and sub-daily for a subset of constituents having sufficient data. For pBIAS, Moriasi and others (2015) did provide daily performance criteria for all constituents, and we add sub-daily metrics where data were sufficient. Because NSE was the most commonly applied metric, we focused the discussion in the following text on NSE performance criteria. Recommendations for R^2 and pBIAS follow a similar logic.

Here, we note the minimum thresholds for a model to be *at least* satisfactory at sub-daily and daily timescales (Table 1). The performance evaluation criteria we use for streamflow models largely matches that of Moriasi and others (2015), as their analysis for flow was robust across timescales (daily, monthly, and annually). We now add a sub-daily recommendation for flow evaluation equal to the daily criteria (NSE > 0.50) as sub-daily flow models had the highest performance (Fig. 4). For water quality variables, we propose common criteria regardless of target units as significant differences were typically uncommon between concentration and load performance measures (Fig. 4). For sediment, we recommend sub-daily and daily model performance to exceed a lower threshold NSE of 0.45. For total nutrients (TN and TP), a lower NSE threshold of 0.30 should be exceeded for daily model evaluation. This threshold is quite low as reported performance measures for concentration-based total nutrient models were generally low. For dissolved nutrients (NO_3^- and PO_4^{3-}), a model evaluated at sub-daily resolution should exceed an NSE of 0.30 while a daily model should exceed an NSE 0.35.

There is a philosophical question as to whether criteria from past studies, which constitute the status quo, should necessarily be acceptable criteria for future work. Ultimately, the acceptance of a model and its results should rely on the specific problem at hand and the experience of the modeler, which carries a degree of subjectivity. Nevertheless, modeling performance criteria based on statistical metrics cited in prior studies provides a common benchmark that can be used as one of many tools in the model evaluation process. As an example, the performance evaluation criteria we present from our global sample of modeling studies may require regional refinement. Recent studies suggest that lower benchmarks of model performance are not only model-dependent but are also dependent on regionally dominant hydrologic processes. In the Pacific Northwest (USA) and Rocky Mountains (USA), baseline (uncalibrated) models simulating discharge achieved NSE values greater than 0.70 in many catchments (Seibert et al., 2018), exceeding our proposed benchmark for 'good' daily flow simulation. These randomly parameterized, uncalibrated models were not guided by modeler expertise or calibration algorithms and yet were broadly successful. Therefore, modelers should be aware of the regional variability in baseline model performance when considering lower benchmarks for process-based models, adjusting lower benchmarks as needed, and aiming for upper benchmarks of what is possible with a given dataset.

Several additional considerations should be made prior to the use of the suggested performance criteria in Table 1. First, we echo the many sentiments in Moriasi et al. (2015) that encourage modelers to employ multiple evaluation criteria, to report multiple statistics, and to be reasonably flexible in their use of performance criteria. We also reinforce recommendations by Wellen et al. (2015) and Mai (2023) that modelers conduct sensitivity analyses, apply parameter optimization algorithms, and employ uncertainty analyses. Further, the performance criteria provided may not be suitable in all situations. For example, the majority of model applications considered in this study (57%) used SWAT, thus the performance criteria are most representative of this

Table 1
Recommended model performance metrics for hydrologic and water quality models for sub-daily (SD), daily (D), and monthly (M) evaluation frequencies. SD, D, and M metric thresholds are reported when sufficient data were available to generate reliable distributions. If two or more temporal scales had similar distributions, they were aggregated. Recommendations for sediment, total nutrient, and dissolved nutrient models are the same for different target units, such as concentration or load. Total nutrient models include total nitrogen and total phosphorus. Dissolved nutrient models include nitrate and orthophosphate. NSE = Nash-Sutcliffe efficiency. R^2 = coefficient of determination. pBIAS = percent bias.

Metric	Variable	Temporal Scale	Limited	Satisfactory	Good	Very Good
NSE	Flow	SD-D-M	< 0.50	$0.50 \leq NSE < 0.70$	$0.70 \leq NSE < 0.80$	≥ 0.80
	Sediment	SD-D	< 0.45	$0.45 \leq NSE < 0.55$	$0.55 \leq NSE < 0.65$	≥ 0.65
		M	< 0.50	$0.50 \leq NSE < 0.60$	$0.60 \leq NSE < 0.70$	≥ 0.70
	Total Nutrient	D	< 0.30	$0.30 \leq NSE < 0.45$	$0.45 \leq NSE < 0.60$	≥ 0.60
		M	< 0.45	$0.45 \leq NSE < 0.60$	$0.60 \leq NSE < 0.70$	≥ 0.70
	Dissolved Nutrient	SD	< 0.30	$0.30 \leq NSE < 0.45$	$0.45 \leq NSE < 0.60$	≥ 0.60
		D	< 0.35	$0.35 \leq NSE < 0.50$	$0.50 \leq NSE < 0.65$	≥ 0.65
		M	< 0.45	$0.45 \leq NSE < 0.55$	$0.55 \leq NSE < 0.65$	≥ 0.65
R^2	Flow	SD-D-M	< 0.60	$0.60 \leq R^2 < 0.75$	$0.75 \leq R^2 < 0.85$	≥ 0.85
	Sediment	SD-D	< 0.45	$0.45 \leq R^2 < 0.55$	$0.55 \leq R^2 < 0.65$	≥ 0.65
		M	< 0.60	$0.60 \leq R^2 < 0.70$	$0.70 \leq R^2 < 0.80$	≥ 0.80
	Total Nutrient	SD	<0.30	$0.30 \leq R^2 < 0.40$	$0.40 \leq R^2 < 0.60$	≥ 0.60
		D	< 0.40	$0.40 \leq R^2 < 0.50$	$0.50 \leq R^2 < 0.60$	≥ 0.60
		M	< 0.50	$0.50 \leq R^2 < 0.60$	$0.60 \leq R^2 < 0.70$	≥ 0.70
	Dissolved Nutrient	SD-D	< 0.45	$0.45 \leq R^2 < 0.60$	$0.60 \leq R^2 < 0.75$	≥ 0.75
		M	< 0.50	$0.50 \leq R^2 < 0.65$	$0.65 \leq R^2 < 0.80$	≥ 0.80
pBIAS	Flow	D-M	$> \pm 15$	$\pm 10 < pBIAS \leq \pm 15$	$\pm 5 < pBIAS \leq \pm 10$	$\leq \pm 5$
	Sediment	D	$> \pm 20$	$\pm 10 < pBIAS \leq \pm 20$	$\pm 5 < pBIAS \leq \pm 10$	$\leq \pm 5$
		M	$> \pm 15$	$\pm 10 < pBIAS \leq \pm 15$	$\pm 5 < pBIAS \leq \pm 10$	$\leq \pm 5$
	Total Nutrient	D	$> \pm 40$	$\pm 20 < pBIAS \leq \pm 40$	$\pm 10 < pBIAS \leq \pm 20$	$\leq \pm 10$
		M	$> \pm 20$	$\pm 10 < pBIAS \leq \pm 20$	$\pm 5 < pBIAS \leq \pm 10$	$\leq \pm 5$
	Dissolved Nutrient	D-M	$> \pm 15$	$\pm 10 < pBIAS \leq \pm 15$	$\pm 5 < pBIAS \leq \pm 10$	$\leq \pm 5$

modeling framework. The use of SWAT for semi-distributed watershed water quality modeling is common due in part to the large body of literature related to the model that serves as reference material and its accessibility through graphical user interfaces (Keller et al., 2023). Although modern computational resources have made fully distributed modeling more accessible than ever, our results indicate that modelers are trading physical realism for models that are either more parsimonious or familiar (Sadayappan et al., 2024). We do not discuss any further the pros and cons of various modeling packages as that has been well-covered in previous studies (Wellen et al., 2015). Lastly, while the NSE, R^2, and pBIAS performance measures are useful, they are not without their limitations in the hydrologic and water quality sciences (Knoben et al., 2019; Mizukami et al., 2019). Alternative metrics are emerging, such as the Kling-Gupta Efficiency (Gupta et al., 2009), and performance evaluation criteria for these metrics will likely be established as the number of studies that use them grows.

4.2. Considerations in water quality modeling

We posed several questions in the Methods (Section 2.1) that anchored this analysis. Is the performance of water quality models systematically impacted by (1) temporal resolution of model evaluation, (2) duration of the calibration data record, or (3) the target units of the modeled variable? Below, we provide considerations and recommendations to each of these points.

The temporal resolution of model evaluation is an influential variable to consider when assessing performance (Fig. 4). An interesting phenomena we observed was that sub-daily model performance (for median NSE) was equal to or better than daily model performance across all variables (Fig. 4). This goes against the typical "rule of thumb" that models perform worse when evaluated at finer temporal resolutions due to their presumed inability to capture fine-scale patterns (Massmann, 2020). In their survey over 494 watershed model applications, Wellen and others (2015) likewise noticed a gradual decline in model performance from annual to monthly to daily evaluation but with a similar increase in performance at sub-daily evaluation. That the greatest model performance occurs for sub-daily evaluation could reflect that modelers apply high temporal resolution models to systems where they have greater confidence in their understanding and implementation of key processes. It could also represent that the sub-daily data used to evaluate the model are capturing dynamic trends that may be missed with coarser sampling, allowing models to become more representative. Thus, matching the spatial representation of modeled processes to the resolution of input and evaluation data is crucial for accurate depiction of water quality processes.

The duration of calibration data was influential in improving model performance, up to a point (Fig. 5). We saw consistent, but not always significant, increases in NSE for all variables when models were calibrated with 3 to 8 years of data versus less than 3 years of data. Merz and others (2009) also pointed out improvement in model accuracy is largest from 1 to 5 years, with a longer dataset ensuring hydrological processes are properly incorporated. However, no clear pattern in the median NSE was observed when additional data (more than 8 years) was included in model calibration. This "upper ceiling" on performance could be the result of epistemic uncertainty (Gupta et al., 2012). Because our knowledge of the physical world is incomplete, our model representations of the influential processes will be inadequate no matter the amount of data considered. Another important note is that for sediment, total nutrient, and dissolved nutrient models, longer duration datasets do not necessarily correlate to a greater density of data. For example, a nitrate model with discrete samples collected over the duration of a decade could have fewer data points than a model with one year of daily nitrate sensor data. Despite no clear trend in the change of median NSE with more calibration data, we do observe that shorter calibration periods have wider bounds on reported values than longer periods (Fig. 5). That is, shorter calibration periods have a mix of high performing and low performing models, whereas the spread in performance metrics decreases the longer the calibration period becomes. Therefore, models calibrated to longer duration datasets are expected to be less sensitive to data length and provide steady estimates of model performance and parameters (Li et al., 2010; Yapo et al., 1996). Further, even if they do not directly improve model performance, longer calibration periods can enhance the internal representation of hydrologic conditions under changing circumstances, such as prolonged droughts, which are crucial for ensuring reliability in future scenario simulations (Fu et al., 2020).

The target units of the modeled water quality variables (concentration, load, or yield) were an important consideration in evaluating model performance (Fig. 6). Most models of sediment and total nutrients were calibrated to load while dissolved nutrients were more frequently calibrated to concentration. Because flow models tend to outperform sediment and nutrient models (Fig. 4), we originally hypothesized that load-based models would outperform concentration-based models as the load-based estimates are the product of flow and concentration and thus benefit from the superior flow models. That is, a good flow model could mask deficiencies in an underperforming sediment or nutrient concentration model. While we observed this for the dissolved nutrient variable, it was not the case for the sediment and total nutrient variables. Load-based measures may be required as part of a regulatory activities, such as developing total maximum daily loads (Amatya et al., 2010). For general, non-regulatory model applications, we recommend the use of concentration rather than load as a calibration goal. Concentration provides a flow-independent measure of the water quality variables. Thus, if the goal of a project is to seek a mechanistic understanding of water quality variation, a concentration-based calibration can provide greater insight.

4.3. Frontiers for the next generation of water quality models

The next generation of water quality models will have access to greater levels of computing and a larger variety of big data than presently available. We recommend that these new resources be leveraged to overcome existing limitations. For example, in past decades, the cost of computing made it prohibitive to run and calibrate models at high temporal resolutions. However, coarse resolutions fail to capture the fine-scale processes within a watershed, especially during high-frequency events like storms, which contribute the majority of sediment and nutrient loads in short periods (Loperfido, 2013). Integrating aquatic sensors into hydrological modeling offers promising advancements in water quality assessment. Continuous data from sensors can improve model accuracy by capturing variability and fine-scale processes that are often missed by grab sampling (Husic et al., 2023). Further, hydrologic and water quality signatures that can be readily determined with sensors, such as flushing, hysteresis, diel patterns, and loading curves, can provide an additional evaluation target to improve model realism (Mahoney et al., 2020; Zhang et al., 2023). Beyond evaluation data, advancements in the temporal resolution and accuracy of model inputs, such as climatic forcings, point source inputs, and management activities, are providing reductions to epistemic uncertainties (Possantti et al., 2023). Collectively, these factors are pushing the standards higher for hydrologic and water quality modeling performance. With the sub-daily and daily performance evaluation criteria we propose in this study (Table 1), modelers will be in a position to critically evaluate their high temporal resolution models.

Next-generation models may leverage machine learning and artificial intelligence to process large datasets, enabling better prediction capabilities, and adaptability to diverse hydrological conditions (Höge et al., 2022; Kratzert et al., 2019b). As an example, we show the results of a recent long-short term memory (LSTM) network – a purely data-driven deep learning model – that was trained on daily flow, nitrate concentration, and nitrate yield observation at 95 sites that had high-frequency nitrate sensor data across the United States (Pandit, 2024; Fig. 7). The model makes robust and accurate predictions and can

capture differing hydrologic and nitrate regimes. Through global model training, the LSTM network was able to leverage information at all sites, thus exchanging information from one watershed to another when beneficial (Kratzert et al., 2024). Regionalized approaches to calibration (for process-based models, Abbaspour et al., 2015; Phillips et al., 2024) or training (for machine learning models, Kratzert et al., 2019a) lead to models that have a greater transferability of insights and accuracy when applied to ungauged basins.

Due to their nascent rise, no water quality performance evaluation metrics presently exist that are specific to machine learning models. Nonetheless, studies that utilize machine learning models typically apply process-based benchmarks to evaluate their performance (e.g., Gorski et al., 2024; Saha et al., 2024). We suggest that the performance evaluation criteria established in this study (see Table 1) can serve to inform the assessment of machine learning models, particularly those that operate at fine temporal resolutions. To demonstrate, we evaluate the performance of the deep learning model of daily nitrate concentration from Pandit (2024). The model successfully captures mobilization of nitrate during storms at Kankakee River (NSE = 0.63) and the dilution of nitrate during storms at Difficult Run River (NSE = 0.55). Based on the NSE recommendations for daily nitrate evaluation in Table 1, the model performance at these two sites would be considered 'good'. For overall model performance at the 95 sites, the median NSE values (not plotted) for flow, nitrate concentration, and nitrate yield were 0.68, 0.46, and 0.49, respectively. Based on Table 1 recommendations, the median site for this model had 'satisfactory' flow performance and 'good' nitrate performance. Here, a single machine learning model was able to capture dynamic behavior and generate robust, satisfactory predictions. Although the machine learning model is a black box, whose internal workings are largely a mystery, explainable artificial intelligence methods are being developed to provide intuitive interpretations to data driven predictions (Zhi et al., 2024). Further, the integration of machine learning into process-based water quality models, such as through differentiable modeling (Shen et al., 2023), can deliver simulations that leverage both interpretability and performance to support more effective water management practices.

5. Conclusions

In this study, we assessed the performance of 229 model applications published in 99 articles between 2010 and 2022. Models of streamflow were the most accurate, followed by sediment, then dissolved nutrients, and lastly total nutrients. Regarding temporal resolution of model evaluation, monthly and daily evaluation were equally common, with sub-daily evaluation occurring in fewer than 20% of studies. We found that model performance tended to improve when the duration of calibration data increased from low (< 3 years) to medium (3 to 8 years). However, additional calibration data beyond 8 years did not consistently improve median model performance, although it did reduce the spread in reported evaluation metrics. In situ aquatic sensors facilitated much of the sub-daily and daily model evaluation for water quality constituents (sediment, nitrate, and phosphate). We recommend concentration (mass per volume), rather than load (mass per time), as a calibration goal for water quality models as it provides a flow-independent measure of sediment, nitrogen, or phosphorus variation.

A major contribution of our work is the establishment of fine temporal scale (sub-daily and daily) performance criteria for water quality model evaluation (Table 1). We build upon existing guidelines, which were made at coarser temporal resolutions (Moriasi et al., 2015), to provide a resource for modelers to rapidly evaluate the performance of their models. As computational capabilities increase, models become more accessible, and new types of frameworks (e.g., deep learning) become standard in water quality modeling, Table 1 can serve as a guide for evaluating the accuracy of models that will grow in their predictive abilities. The guidance presented here is intended to be used alongside, rather than to replace, the experience and modeling judgement of engineers and scientist who work to maintain our collective water resources.

CRediT authorship contribution statement

Aayush Pandit: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Sarah Hogan:** Data curation.

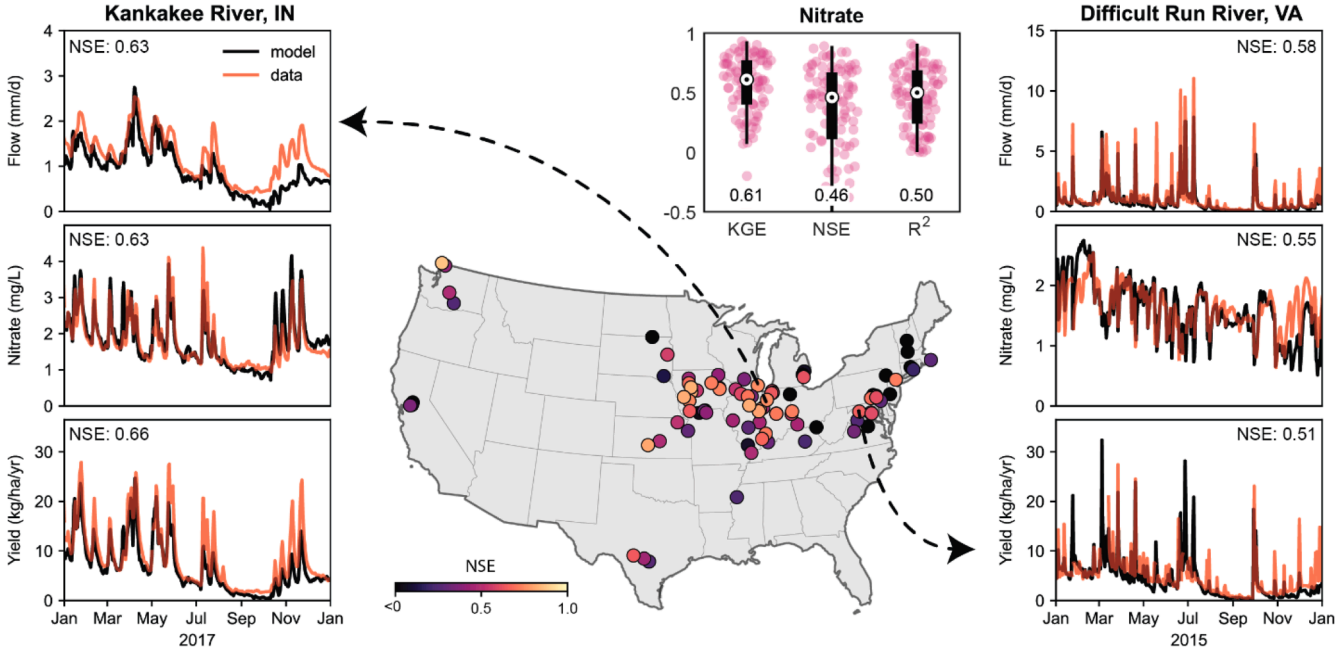


Fig. 7. The results of a globally trained deep learning model of daily nitrate in the contiguous US (n = 95 sites; adapted from Pandit, 2024). Summary statistics for the nitrate concentration model are shown in a boxplot. A text inset in the boxplot figure denotes the median KGE, NSE, and R^2 values. Timeseries of modeled flow, concentration, and yield are shown for two sites: Kankakee River (IN) and Difficult Run River (VA). The model successfully captures differing nitrate responses to storm events: mobilization occurs at the Kankakee River (left) while dilution occurs at Difficult Run River (right). KGE = Kling-Gupta efficiency, NSE = Nash-Sutcliffe efficiency, R^2 = coefficient of determination.

David T. Mahoney: Writing – review & editing, Conceptualization. **William I. Ford:** Writing – review & editing. **James F. Fox:** Writing – review & editing. **Christopher Wellen:** Writing – review & editing. **Admin Husic:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the associate editor and two anonymous reviewers for their comments, which have greatly improved the quality of this manuscript. The authors report no conflicts of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2025.123156.

Data availability

The data are attached as a Supplemental File

References

Abbaspour, K.C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., Kløve, B., 2015. A continental-scale hydrology and water quality model for Europe: calibration and uncertainty of a high-resolution large-scale SWAT model. J. Hydrol. 524, 733–752.

Amatya, D.M., Jha, M., Edwards, A., Williams, T., Hitchcock, D., 2010. Application of SWAT model on chapel branch creek watershed with a karst topography, SC. In: 2010 ASABE Annual International Meeting. Pittsburgh, Pennsylvania, pp. 5595–5614.

Bieroza, M., Acharya, S., Benisch, J., ter Borg, R.N., Hallberg, L., Negri, C., Pruitt, A., Pucher, M., Saavedra, F., Staniszewska, K., van't Veen, S.G.M., Vincent, A., Winter, C., Basu, N.B., Jarvie, H.P., Kirchner, J.W., 2023. Advances in catchment science, hydrochemistry, and aquatic ecology enabled by high-frequency water quality measurements. Environ. Sci. Technol.

Burns, D.A., Pellerin, B.A., Miller, M.P., Capel, P.D., Tesoriero, A.J., Duncan, J.M., 2019. Monitoring the riverine pulse: applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. Wiley Interdiscip. Rev. Water 6, e1348.

Chahor, Y., Casalí, J., Giménez, R., Bingner, R.L., Campo, M.A., Goñi, M., 2014. Evaluation of the AnnAGNPS model for predicting runoff and sediment yield in a small Mediterranean agricultural watershed in Navarre (Spain). Agric. Water Manag. 134, 24–37.

Chambers, P.A., Benoy, G.A., Brua, R.B., Culp, J.M., 2011. Application of nitrogen and phosphorus criteria for streams in agricultural landscapes. Water Sci. Technol. 64, 2185–2191.

Chappell, N.A., Jones, T.D., Tych, W., 2017. Sampling frequency for water quality variables in streams: systems analysis to quantify minimum monitoring rates. Water Res. 123, 49–57.

Efron, B., 1979. Bootstrap methods: another look at the Jackknife. Ann. Stat. 7, 1–26.

Fu, B., Horsburgh, J.S., Jakeman, A.J., Gualtieri, C., Arnold, T., Marshall, L., Green, T.R., Quinn, N.W.T.T., Volk, M., Hunt, R.J., Vezzaro, L., Croke, B.F.W.W., Jakeman, J.D., Snow, V., Rashleigh, B., 2020. Modeling water quality in watersheds: from here to the next generation. Water Resour. Res. 56, e2020WR027721.

Fu, B., Merritt, W.S., Croke, B.F.W., Weber, T.R., Jakeman, A.J., 2019. A review of catchment-scale water quality and erosion models and a synthesis of future prospects. Environ. Model. Softw. 114, 75–97.

Gorski, G., Larsen, L., Wingenroth, J., Zhang, L., Bellugi, D., Appling, A.P., 2024. Stream nitrate dynamics driven primarily by discharge and watershed physical and soil characteristics at intensively monitored sites: insights from deep learning. Water Resour. Res. 60.

Gupta, H.V., Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48, 1–16.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.

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

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., Fenicia, F., 2022. Improving hydrologic models for predictions and process understanding using neural ODEs. Hydrol. Earth Syst. Sci. 26, 5085–5102.

Husic, A., Fox, J., Adams, E., Ford, W., Agouridis, C., Currens, J., Backus, J., 2019. Nitrate pathways, processes, and timing in an agricultural karst system: development and application of a numerical model. Water Resour. Res. 55, 2079–2103.

Husic, A., Fox, J.F., Clare, E., Mahoney, T., Zarnaghsh, A., 2023. Nitrate hysteresis as a tool for revealing storm-event dynamics and improving water quality model performance. Water Resour. Res. 59, e2022WR033180.

Keller, A.A., Garner, K., Rao, N., Knipping, E., Thomas, J., 2023. Hydrological models for climate-based assessments at the watershed scale: a critical review of existing hydrologic and water quality models. Sci. Total Environ. 867, 161209.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23, 4323–4331.

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS opinions: never train an LSTM on a single basin. Hydrol. Earth Syst. Sci. Discuss. 1–19.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019a. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55, 11344–11354.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621.

Li, C.Z., Wang, H., Liu, J., Yan, D.H., Yu, F.L., Zhang, L., 2010. Effect of calibration data series length on performance and optimal parameters of hydrological model. Water Sci. Eng. 3, 378–393.

Loperfido, J.J.V., 2013. Surface water quality in streams and rivers: scaling and climate change. In: Ahuja, S.B.T.-C.W.Q, P (Eds.), Comprehensive Water Quality and Purification. Elsevier, Waltham, pp. 87–105.

Mahoney, D.T., Fox, J., Al-Aamery, N., Clare, E., 2020. Integrating connectivity theory within watershed modelling part II: application and evaluating structural and functional connectivity. Sci. Total Environ. 740, 140386.

Mai, J., 2023. Ten strategies towards successful calibration of environmental models. J. Hydrol. 620, 129414.

Massmann, C., 2020. Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. Hydrol. Process. 34, 4–20.

McGill, T., Ford, W.I., 2024. Extreme learning machine predicts high-frequency stream flow and nitrate-N concentrations in a Karst Agricultural Watershed. J. ASABE 67, 73–87.

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res. 45, 1–15.

Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H.V., Kumar, R., 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. Hydrol. Earth Syst. Sci. 23, 2601–2614.

Moriasi, D.N., Arnold, J.G., Liew, M.W.Van, Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. Am. Soc. Agric. Biol. Eng. 50, 885–900.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: performance measures and evaluation criteria. Trans. ASABE 58, 1763–1785.

Nash, J.E., Sutcliffe, J.V, 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290.

Pandey, A., Himanshu, S.K., Mishra, S.K., Singh, V.P., 2016. Physically based soil erosion and sediment yield models revisited. Catena 147, 595–620.

Pandit, A., 2024. Deep Learning and Aquatic Sensing Reveal Nitrate Dynamics in Mississippi River Basin. University of Kansas.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. London, Edinburgh. Dublin Philos. Mag. J. Sci. 2, 559–572.

Phillips, A.K., Mandal, S., Mohamed, M., Sorichetti, R.J., Ross, C.A., Thomas, J.L., Wellen, C.C., 2024. Is comprehensive event sampling necessary for constraining process models of water quality? A comparison of high and low frequency phosphorus sampling programs for constraining the HYPE water quality model. J. Hydrol. 639, 131502.

Possantti, I., Barbedo, R., Kronbauer, M., Collischonn, W., Marques, G., 2023. A comprehensive strategy for modeling watershed restoration priority areas under epistemic uncertainty: a case study in the Atlantic Forest, Brazil. J. Hydrol. 617, 129003.

Qi, Z., Ma, L., Helmers, M.J., Ahuja, L.R., Malone, R.W., 2012. Simulating nitrate-nitrogen concentration from a subsurface drainage system in response to nitrogen application rates using RZWQM2. J. Environ. Qual. 41, 289.

Ramirez-Avila, J.J., Radcliffe, D.E., Osmond, D., Bolster, C., Sharpley, A., Ortega-Achury, S.L., Forsberg, A., Oldham, J.L., 2017. Evaluation of the APEX model to simulate runoff quality from agricultural fields in the Southern Region of the United States. J. Environ. Qual. 46, 1357–1364.

Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan, P., Kronvang, B., Halliday, S.J., Skeffington, R.A., Rozemeijer, J.C., Aubert, A.H., Rinke, K., Jomaa, S., 2016. Sensors in the stream: the high-frequency wave of the present. Environ. Sci. Technol. 50, 10297–10307.

Sadayappan, K., Stewart, B., Kerins, D., Vierbicher, A., Zhi, W., Smykalov, V.D., Shi, Y., Vis, M., Seibert, J., Li, L., 2024. BioRT-HBV 1.0: a biogeochemical reactive transport model at the watershed scale. J. Adv. Model. Earth Syst. 16, 315–333.
Saha, G., Shen, C., Duncan, J., Cibin, R., 2024. Performance evaluation of deep learning based stream nitrate concentration prediction model to fill stream nitrate data gaps at low-frequency nitrate monitoring basins. J. Environ. Manage. 357, 120721.
Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in hydrological modelling. Hydrol. Process. 32, 1120–1125.
Shen, C., Appling, A.P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C.J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., Rahmani, F., Song, Y., Beck, H.E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., Lawson, K., 2023. Differentiable modelling to unify machine learning and physical models for geosciences. Nat. Rev. Earth Environ. 4, 552–567.
Shen, H., Tolson, B.A., Mai, J., 2022. Time to update the split-sample approach in hydrological model calibration. Water Resour. Res. 58, 1–26.
Song, Y., Chaemchuen, P., Rahmani, F., Zhi, W., Li, L., Liu, X., Boyer, E., Bindas, T., Lawson, K., Shen, C., 2024. Deep learning insights into suspended sediment concentrations across the conterminous United States: strengths and limitations. J. Hydrol. 639.
Sorooshian, S., Duan, Q., Gupta, V.K., 1993. Calibration of rainfall-runoff models: application of global optimization to the Sacramento Soil Moisture Accounting Model. Water Resour. Res. 29, 1185–1194.
Wellen, C., Kamran-Disfani, A.-R., Arhonditsis, G.B., 2015. Evaluation of the current state of distributed watershed nutrient water quality modeling. Environ. Sci. Technol. 49, 3278–3290.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bull. 1, 80–83.
Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. J. Hydrol. 181, 23–48.
Zarnaghsh, A., Kelly, M., Burgin, A., Husic, A., 2024. Revealing nitrate uptake and dispersion dynamics using high-frequency sensors and two-dimensional modeling in a large river system. Adv. Water Resour., 104693
Zhang, X., Yang, X., Hensley, R., Lorke, A., Rode, M., 2023. Disentangling in-stream nitrate uptake pathways based on two-station high-frequency monitoring in high-order streams. Water Resour. Res. 59, 1–17.
Zhi, W., Appling, A.P., Golden, H.E., Podgorski, J., Li, L., 2024. Deep learning for water quality. Nat. Water.
Zhi, W., Ouyang, W., Shen, C., Li, L., 2023. Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. Nat. Water 1, 249–260.