Fusion Journal of Engineering and Sciences

Review

Hydrological Model Evaluation Criteria Comparison

Taliah Sajid [1], Ammar Ashraf[2], Usman Pervaiz[3], Abdul Wahab[4], Adnan Akmal[1], Muhammad Waseem[1], Zeeshan Asghar[1]*

[1] Department of Civil Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23640, Pakistan; sajidtaliah@gmail.com (T.S.); adnan.akmal@giki.edu.pk (A.AK); muhammad.waseem@giki.edu.pk (M.W); zeeshan.asghar@giki.edu.pk (Z.A)

[2] Department of Civil Engineering, COMSATS University Islamabad Sahiwal Campus, Sahiwal, Pakistan., ammarmetu@gmail.com (A.AS)

[3] Department of Civil Engineering (SEN) University of Management and Technology (UMT) C-II, Johar Town, Lahore, Pakistan; usman.pervaiz@umt.edu.pk (U.P)

[4] School of Design and construction University of Southern Mississippi 118 College Dr, Hattiesburg, MS 39406, United States; abdul.wahab@usm.edu (A.W)

* Correspondence: zeeshan.asghar@giki.edu.pk

Abstract

Hydrological models are widely used to support water resources planning, flood and drought assessment, watershed management, and climate change impact analysis. The credibility of such applications depends strongly on how model performance is evaluated against observed data. Numerous statistical performance metrics have been proposed for hydrological model evaluation; however, their mathematical formulations, sensitivity to flow regimes, and interpretability differ substantially. As a result, the choice of evaluation criteria can strongly influence conclusions regarding model adequacy and comparative performance. This review synthesizes commonly used hydrological model performance metrics, including the Nash-Sutcliffe Efficiency (NSE), coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Percent Bias (PBIAS), Kling-Gupta Efficiency (KGE), RMSE-observations standard deviation ratio (RSR), and Willmott's index of agreement. For each metric, definitions, formulations, interpretation guidelines, strengths, and limitations are discussed based on classical and recent literature. Particular attention is given to the sensitivity of metrics to high-flow and low-flow conditions, systematic bias, variability, and error magnitude, as well as their suitability for different water resources applications. Comparative analysis highlights that no single metric can adequately capture all aspects of hydrological model performance, and metric-dependent ranking of models is common across studies. The review study emphasizes the importance of multi-criteria evaluation frameworks and application-specific metric selection, especially under non-stationary climatic conditions. Practical recommendations are provided to support transparent, consistent, and meaningful performance evaluation in hydrological modeling and water resources decision-making.

Keywords: Hydrological modeling; Model performance evaluation; Nash-Sutcliffe Efficiency; Kling-Gupta Efficiency; Error metrics; Water resources applications

1. Introduction

Hydrological models are essential tools for water resources planning, flood and drought forecasting, watershed management, and assessing climate change impacts. The usefulness of any model depends critically on its ability to reproduce observed hydrological behavior, including extreme events, timing, and variability. Quantitative performance metrics, also called criteria of fit, provide a standardized way to evaluate how well simulated outputs match observations. Each metric emphasizes different aspects of performance, such as bias, variability, correlation, or absolute error, so no single metric can fully capture all relevant characteristics. A multi-criteria approach is therefore recommended [1] [2]. Recent review studies indicate that, despite the widespread use of hydrological models, evaluation practices remain inconsistent across studies, with substantial variation in metric selection, threshold interpretation, and reporting standards [3] [4] [5] [6]. Similar concerns have been raised in comparative assessments of process-based and data-driven models, where differences in evaluation criteria can strongly influence perceived model superiority, particularly under non-stationary climatic conditions [7]. This review summarizes widely used performance metrics in hydrology, presents their definitions, formulas, interpretations, and limitations, highlights key findings and critiques from the literature, and offers practical recommendations for their effective use. In contrast to model-specific evaluation studies, this paper focuses on the comparative behavior, diagnostic value, and applicability of commonly used performance metrics across different hydrological conditions and water resources applications, drawing on both classical and recent literature [8] [9] [10].

2. Model Performance Evaluation in Hydrology

Hydrological models are widely used in water resources science to simulate key components of the hydrological cycle, including precipitation-runoff processes, streamflow generation, groundwater recharge, evapotranspiration, and water balance dynamics. These models support a broad range of water resources applications, such as flood forecasting, drought assessment, reservoir operation, watershed management, and evaluation of climate change impacts [1]. Advances in integrated and process-based modeling frameworks have further expanded the scope of hydrological applications, increasing the need for robust and transparent performance evaluation methodologies [11]. Because model outputs are often used to inform planning, management decisions, and policy, systematic and transparent evaluation of model performance is a fundamental step in hydrological modeling studies.

Model performance evaluation refers to the quantitative comparison of simulated outputs with observed data to assess how well a model reproduces observed hydrological behavior. In practice, evaluation is most commonly conducted using statistical performance metrics that summarize differences between simulated and observed time series. These metrics provide objective measures of goodness-of-fit and allow comparison of model performance across different studies, catchments, and modeling approaches [1] [12]. However, multiple studies have emphasized that commonly reported performance metrics are often applied without sufficient consideration of their underlying assumptions, sensitivity to flow regimes, or suitability for specific modeling objectives [8] [9] [10] Reviews focusing on model selection and benchmarking further highlight that inconsistent metric usage complicates cross-study comparison and limits reproducibility [3] [5]. In water resources applications, performance evaluation is particularly important because errors in simulated flows can directly influence estimates of flood risk, water availability, drought severity, and long-term water balance.

A central challenge in hydrological model evaluation is that hydrological processes are complex, nonlinear, and variable across time and space. Models may perform well for certain aspects of the hydrograph (e.g., peak flows) while performing poorly for others (e.g., low flows or timing). For example, a model calibrated to reproduce flood peaks may underestimate baseflow during dry periods, while a model that captures long-term water balance may fail to reproduce short-term extremes [13]. This issue becomes more pronounced when models are evaluated across multiple catchments or climatic regimes, where differences in flow variability and seasonality can strongly influence metric behavior and interpretation. Multi-catchment and regional-scale evaluations have shown that metric rankings can change substantially across basins, underscoring the context dependency of commonly used criteria [14]. As a result, no single performance metric can fully characterize model behavior under all conditions relevant to water resources decision-making.

Performance metrics differ in what aspect of model behavior they emphasize. Some metrics focus on overall agreement between simulated and observed values relative to a benchmark (e.g., the observed mean), while others quantify absolute error magnitude, linear association, systematic bias, or variability. Error-based metrics such as RMSE and MAE describe the average magnitude of simulation errors, whereas relative efficiency measures such as the Nash-Sutcliffe Efficiency (NSE) evaluate performance relative to a simple reference model. Correlation-based metrics such as the coefficient of determination (R^2) describe how well temporal patterns are reproduced but do not capture bias or volume errors. Bias-oriented measures, such as percent bias (PBIAS), explicitly indicate systematic overestimation or underestimation [1], which is critical in water balance and resource accounting studies. Several recent studies have cautioned that over-reliance on popular metrics without diagnostic analysis can lead to misleading conclusions about model adequacy, particularly when metrics are optimized during calibration but fail to represent hydrological realism or process consistency [9] [10]. Similar critiques have emerged from comparative reviews of hydrological and hydrodynamic models, emphasizing that numerical performance alone does not guarantee physically meaningful simulations [5] [6].

Because each metric highlights different aspects of performance, reliance on a single metric can lead to incomplete or misleading conclusions. This issue is well recognized in the hydrological literature, and a multi-criteria evaluation approach is widely recommended [1] [2]. The selection of appropriate performance metrics should therefore be guided not only by convention but also by the dominant flow regime, temporal scale, and decision context of the intended application [3] [4]. Recent review papers stress that application-specific metric selection is particularly important under climate change conditions, where historical calibration performance may not translate into future reliability [4]. By combining metrics that capture goodness-of-fit, error magnitude, bias, and variability, modelers can obtain a more balanced and interpretable assessment of model performance. Such an approach is especially important in water resources applications where different management objectives may prioritize different aspects of model behavior, such as accurate flood peaks, reliable low-flow simulations, or long-term volume conservation.

In addition to numerical performance metrics, model evaluation should be interpreted in the context of the specific hydrological regime and application. For instance, flood modeling studies often emphasize metrics that are sensitive to high flows and peak timing, while drought and low-flow studies require evaluation criteria that adequately represent low-flow behavior and persistence. Similarly, climate change impact assessments often focus on long-term trends and variability rather than exact replication of individual events. Climate-focused hydrological studies increasingly emphasize robustness and consistency across scenarios rather than maximizing single-event accuracy [15][4]. Therefore, the choice and interpretation of performance metrics should be aligned with the intended water resources application.

Overall, model performance evaluation is not a purely technical exercise but a critical component of responsible hydrological modeling. Clear reporting of evaluation metrics, their definitions, limitations, and interpretation enhances transparency, reproducibility, and comparability across studies [16]. Standardized reporting practices have been repeatedly advocated to reduce ambiguity and improve synthesis across review studies [3] [6]

[9]. In the following sections, commonly used performance metrics in hydrological modeling are reviewed, with emphasis on their definitions, interpretation, limitations, and relevance for water resources applications [13].

3. Performance Metrics for Hydrological Model Evaluation

Below are eight commonly used criteria. For each, the definition, formula, interpretation, limitations, and illustrative example from the literature are given.

3.1. Nash-Sutcliffe Efficiency (NSE)

NSE measures how well simulated values follow observed values compared to using the observed mean as a predictor. It remains among the most widely used metrics in runoff and streamflow studies.

$$NSE = 1 - \frac{\sum_{t=1}^{n}(Q_{obs,t} - Q_{sim,t})^2}{\sum_{t=1}^{n}(Q_{obs,t} - \overline{Q_{obs}})^2}$$

where $Q_{obs,t}$ and $Q_{sim,t}$ are observed and simulated flows at time t, and $\overline{Q_{obs}}$ is the mean observed flow.

Interpretation

- NSE = 1 → perfect fit.
- NSE = 0 → model is as good as predicting the mean value always.
- NSE < 0 → model is worse than using the mean.
- Common (but not universal) thresholds: NSE > 0.75 (good), 0.5–0.75 (acceptable), 0.25–0.5 (fair), < 0.25 (unsatisfactory).

Limitations

- Strongly sensitive to high flows and outliers because it uses squared errors.
- Tends to emphasize peak-flow performance; may mask poor low-flow performance.
- Not normalized for variance differences; can be misleading for biased distributions.

NSE was originally introduced as a relative efficiency criterion for hydrological models [18]. It remains a foundational metric for hydrological model evaluation and is widely used, with many studies reporting NSE to assess streamflow model performance, including hydrological model intercomparisons.

3.2. Coefficient of Determination (R^2)

R^2, the square of the Pearson correlation coefficient, measures the proportion of variance in observed data explained by simulated values. It reflects the strength of the linear association between observations and simulations but does not necessarily indicate unbiased accuracy.

$$R^2 = (\frac{\sum_{t=1}^{n}(Q_{obs,t} - \overline{Q_{obs}})(Q_{sim,t} - \overline{Q_{sim}})}{\sqrt{\sum_{t=1}^{n}(Q_{obs,t} - \overline{Q_{obs}})^2 \sum_{t=1}^{n}(Q_{sim,t} - \overline{Q_{sim}})^2}}$$

Interpretation

- R^2 = 1 → perfect linear agreement.
- R^2 close to 1 → strong correlation.
- R^2 near 0 → weak correlation.
- High R^2 suggests that the model reproduces the overall trend of observed data.

Limitations

- Does not indicate bias or absolute agreement and a model could systematically over or under-estimate values but still have high R^2.
- Does not reflect absolute error magnitude.
- Sensitive to the range/variance of observed data (higher variance tends to inflate R^2).

It is reported alongside NSE and error-based metrics to show correlation or trend agreement even when bias or magnitude errors exist.

3.3. Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared difference between simulated and observed values. It is a scale-dependent, absolute measure of prediction error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Q_{sim,t} - Q_{obs,t})^2}$$

Interpretation

- RMSE = 0 → perfect fit.
- Lower RMSE → closer agreement between observed and simulated values on average.
- Because units are the same as the variable (e.g., m^3/s), RMSE provides a physically meaningful error magnitude.

Limitations

- Sensitive to large errors and outliers (peaks, extreme events).
- Because it is scale-dependent, direct comparison between basins with different flow magnitudes is problematic unless normalized.

MSE is often reported alongside relative or normalized metrics, such as RSR or NSE, to facilitate comparison across basins. RMSE or RSR is widely reported in hydrological studies to quantify absolute prediction errors [1].

3.4. Mean Absolute Error (MAE)

MAE measures the average magnitude of the absolute differences between simulated and observed values, regardless of direction. It is a scale-dependent, robust metric that is less sensitive to outliers than RMSE.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Q_{sim,t} - Q_{obs,t}\right|$$

Interpretation

- MAE = 0 → perfect match.
- Lower MAE → on average, smaller absolute errors, gives a sense of typical error magnitude

Limitations

MAE is scale-dependent, with units the same as the observed variable. It treats all errors linearly and does not reflect error variance, so a few large errors and many small errors can produce the same MAE as many moderate errors. As a result, MAE may underrepresent the importance of peaks or extreme events, which is critical in applications such as flood modeling.

This model is less sensitive to outliers and large errors unlike RMSE. Errors contribute linearly rather than quadratically, so MAE is more robust when extreme events are rare or when one cares about typical performance rather than peaks.

3.5. Percent Bias (PBIAS)

PBIAS measures the average tendency of simulated flows to systematically over or underestimate observed values, expressed as a percentage.

$$PBIAS = 100 \times \frac{\sum_{t=1}^{n}(Q_{sim,t} - Q_{obs,t})}{\sum_{t=1}^{n} Q_{obs,t}}$$

Interpretation

- PBIAS = 0 → no bias, ideal.
- Positive PBIAS → model overestimates flows.
- Negative PBIAS → model underestimates flows.
- Common thresholds: ±10% = excellent, ±10–15% = good, ±15–25% = satisfactory, > ±25% = unsatisfactory.

Limitations

- Shows only average directional bias, not error distribution.
- Large positive and negative errors can cancel out, masking poor performance.

Reporting PBIAS for bias detection is widely recommended for bias detection and provide practical thresholds for model acceptability [1].

3.6. Kling-Gupta Efficiency (KGE)

KGE provides a balanced performance metric by decomposing errors into correlation, bias, and variability components.

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$$

where:

r is the linear correlation coefficient between simulated and observed, β = bias ratio, γ = ratio of coefficients of variation

Interpretation

- KGE = 1 (perfect agreement)
- Values closer to 1 indicate better overall performance.
- Component analysis allows diagnostic insight into whether poor performance is due to bias, variability, or correlation.

Limitations

- Different KGE formulations exist; specifying which version is used is important.
- Aggregates multiple aspects into a single number; component analysis is required for detailed diagnostics.

Comparative studies have shown that KGE can be more diagnostically informative than NSE when component terms are examined [13]. Because KGE explicitly shows which component (r, bias, or variability) is responsible for poor performance, it is more diagnostic than NSE.

3.7. RSR (RMSE-observations standard deviation Ratio)

RSR standardizes RMSE by the standard deviation of observed data, allowing comparison across basins or datasets.

$$RSR = \frac{RMSE}{SD\ \ (Q_{obs})} = \frac{\sqrt{\frac{1}{n}\sum(Q_{sim} - Q_{obs})^2}}{\sqrt{\frac{1}{n-1}\sum(Q_{obs} - \overline{Q_{obs}})^2}}$$

Interpretation

- Lower RSR → better fit; ideal = 0.
- Common thresholds: ≤0.5 = very good, 0.5–0.6 = good, etc.

Limitations

- Because it divides by observation variability, RSR can be influenced by low variability in observed series.
- Like RMSE, it still penalizes large errors; RSR is a normalized but still scale-sensitive metric.

RSR is promoted as a normalized error metric suitable for inter-basin comparison [1].

3.8. Willmott Index of Agreement (d) / refined index (d_r)

Willmott's index quantifies agreement between simulated and observed flows, bounded between 0 and 1. The refined index (d_r) improves statistical interpretability.

$$d = 1 - \frac{\sum_{t=1}^{n}(Q_{sim,t} - Q_{obs,t})^2}{\sum_{t=1}^{n}\left(\left|Q_{sim,t} - \overline{Q_{obs}}\right| + \left|Q_{obs,t} - \overline{Q_{obs}}\right|\right)^2}$$

Interpretation

- d = 1 indicates perfect agreement; d = 0 indicates no agreement.
- Because d is bounded, it is often easier to communicate than unbounded MSE or RMSE.

Limitations

- d can be overly sensitive to the sample distribution and may inflate apparent skill when variability is low.
- The original d has known statistical issues, use Willmott's refined index (d_r) where appropriate and explain which version is used.

The refined index d_r was introduced to improve statistical interpretation of agreement metrics [17], the index is used in climate and hydrology model comparisons when a bounded measure is desired.

4. Comparative Analysis of Metrics

Table 1 summarizes commonly used hydrological performance metrics, highlighting their primary focus, strengths, and limitations as reported in the literature.

Table 1: Commonly Used Hydrological Model Performance Metrics: Properties and Limitations

Metric	What it measures	Sensitive to	Ideal value	Main limitation
NSE	Fit relative to mean (goodness-of-fit)	High flows, outliers	1 (perfect)	Overemphasizes peaks; can hide low-flow errors
R^2	Proportion of variance explained (correlation)	Range/variance of data	1	Doesn't show bias or absolute error
RMSE	Average magnitude of error (squared)	Large errors/outliers	0	Scale-dependent; penalizes large errors
MAE	Average absolute error	All errors equally	0	Scale-dependent; underweights large errors
PBIAS	Average bias (%)	Systematic over/underestimation	0%	Cancelling errors can mask problems
KGE	Combined correlation, bias, variability	Different components (r, bias, cv)	1	Multiple formulations; must examine components
RSR	RMSE normalized by obs SD	Observation variability	0	Influenced by low obs variability
Willmott d/d_r	Bounded agreement measure	Distribution of observations	1	Original d has statistical issues; use d_r when possible

Performance metrics used in hydrological modeling differ substantially in their mathematical formulation, interpretation, and sensitivity to specific characteristics of the simulated time series. As a result, different metrics may lead to different conclusions about model performance when applied to the same dataset. Understanding the comparative strengths and limitations of commonly used evaluation criteria is therefore essential for meaningful model assessment, particularly in water resources applications where modeling outcomes inform decision-making. Large-scale reviews and benchmarking exercises confirm that metric-dependent ranking of models is common, especially when contrasting conceptual, process-based, and data-driven approaches [3][5][7].

One of the most widely used metrics in hydrology is the Nash-Sutcliffe Efficiency (NSE) [18], which evaluates model performance relative to the mean of observed data. NSE is sensitive to large errors and places greater emphasis on high flows due to the squared error formulation. Consequently, NSE is often well suited for applications focused on peak flows, such as flood modeling and flood frequency analysis. However, NSE has well-documented limitations, including reduced sensitivity to low flows and a tendency to penalize models that perform reasonably well during dry periods but fail to reproduce extreme events. As a result, reliance on NSE alone may lead to biased assessments of model performance, particularly in drought studies or baseflow-dominated catchments.

The coefficient of determination (R^2) is commonly used alongside NSE to assess the strength of the linear relationship between observed and simulated values [18]. While R^2 provides useful information about how well temporal patterns and variability are captured, it does not account for systematic bias or differences in magnitude. A model may exhibit a high R^2 while consistently overestimating or underestimating streamflow volumes. For this reason, R^2 is insufficient as a standalone performance metric and should be interpreted in combination with error- and bias-based measures.

Error-based metrics such as the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) quantify the average magnitude of simulation errors [1]. RMSE is more sensitive to large errors due to the squaring of residuals, making it particularly responsive to peak flow mismatches. MAE, by contrast, treats all errors equally and provides a more robust measure of overall error magnitude. In comparative terms, RMSE is often preferred in flood-related studies where peak errors are critical, while MAE may be more appropriate for general water balance assessments or long-term simulations where extreme values should not dominate evaluation.

Percent bias (PBIAS) explicitly measures the average tendency of a model to overestimate or underestimate observed values [1]. This metric is especially relevant in water resources planning and management, where systematic bias can lead to incorrect estimates of water availability, reservoir inflows, or consumptive use. However, PBIAS does not provide information about the timing or variability of simulated flows and may indicate acceptable performance even when temporal dynamics are poorly represented. As such, PBIAS is most informative when used in conjunction with metrics that capture variability and correlation.

More recently, composite metrics such as the Kling-Gupta Efficiency (KGE) have been proposed to address some of the limitations of NSE. KGE decomposes model performance into correlation, bias, and variability components, allowing a more balanced assessment of model behavior [14]. This decomposition facilitates diagnostic evaluation by identifying whether poor performance arises from errors in timing, systematic bias, or misrepresentation of variability. As a result, KGE has gained increasing attention in hydrological modeling studies, particularly in comparative and multi-objective evaluation frameworks.

The Ratio of the Root Mean Square Error to the Standard Deviation of Observations (RSR) provides a standardized measure of error relative to observed variability [1]. Lower RSR values indicate better model performance, and the metric is often used in conjunction with NSE and PBIAS in model evaluation guidelines. RSR is useful for comparing performance across different watersheds or datasets with varying levels of variability, although it shares sensitivity to large errors similar to RMSE.

Willmott's index of agreement aims to overcome some limitations of correlation-based metrics by providing a bounded measure of agreement between simulated and observed values [17]. These indices are sensitive to both systematic and random errors and have been applied in various hydrological and environmental modeling studies. However, their interpretation is less intuitive compared to more widely used metrics, which may limit their adoption in applied water resources studies.

Comparative analyses across the literature consistently indicate that no single metric is sufficient to fully characterize hydrological model performance. Metrics may yield conflicting evaluations depending on the dominant flow regime, temporal scale, and modeling objective. For example, a model may achieve high NSE and low RMSE for flood events while exhibiting substantial bias and poor low-flow performance, leading to misleading conclusions if evaluation focuses only on peak flows.

Therefore, a multi-metric evaluation framework is widely recommended, particularly for water resources applications that involve diverse hydrological conditions and management objectives. Combining metrics that capture efficiency, error magnitude, bias, correlation, and variability provides a more comprehensive and defensible assessment of model performance. Such an approach enhances transparency and ensures that model evaluation results are aligned with the specific requirements of flood management, drought assessment, climate change impact analysis, and watershed-scale water resources planning.

5. Implications for Water Resources Applications

The choice of hydrological model performance metrics has direct implications for water resources planning, management, and decision-making. Different water resources applications prioritize different aspects of model behavior, such as peak flows, low flows, long-term water balance, or variability. Consequently, the selection of evaluation criteria should be aligned with the specific objectives of the study, rather than relying on a single, generic metric.

5.1. Flood Modeling and Flood Risk Management

Under climate change scenarios, flood-focused performance evaluation becomes even more sensitive to bias and variability metrics, as changes in extreme-event frequency and magnitude can distort traditional efficiency scores [15]. In flood-related applications, accurately simulating peak flows, timing, and rising limbs of hydrographs is critical for infrastructure design, early warning systems, and risk assessment. Metrics such as the Nash-Sutcliffe Efficiency (NSE) and RMSE are frequently used in flood modeling because they emphasize high-flow conditions and penalize large errors. However, NSE's sensitivity to extreme values can lead to acceptable scores even when flood volumes or timing are poorly represented.

Bias-oriented metrics such as PBIAS are therefore essential companions in flood studies, as systematic over- or underestimation of peak discharge can significantly affect flood hazard mapping and structural design decisions. In this context, Kling-Gupta Efficiency (KGE) offers advantages by explicitly accounting for correlation, bias, and variability, allowing modelers to diagnose whether deficiencies arise from incorrect flood magnitude, timing, or variability. For flood-focused water resources applications, a combination of NSE or KGE with RMSE (or RSR) and PBIAS provides a more reliable assessment than any single metric alone.

5.2. Drought Analysis and Low-Flow Assessment

Drought studies, environmental flow assessments, and water supply planning require accurate simulation of low flows and flow persistence rather than peak events. Traditional metrics such as NSE and RMSE often perform poorly in these contexts because they are dominated by high-flow periods and squared errors. As a result, models that reproduce flood peaks well may still misrepresent low-flow behavior while achieving acceptable NSE values.

For low-flow and drought applications, absolute error measures such as MAE, bias indicators such as PBIAS, and normalized metrics like RSR are often more informative. Additionally, specialized efficiency measures or modified NSE formulations that emphasize low-flow conditions have been recommended in the literature. The implications for water resources management are significant: misrepresentation of low flows can lead to incorrect assessments of water availability, ecosystem stress, and drought severity. Therefore, metric selection should explicitly reflect the importance of low-flow performance in drought-related studies.

5.3. Climate Change Impact Assessment

Climate change impact studies focus on changes in hydrological regimes, including shifts in mean flow, variability, and extremes over long time horizons. In this context, evaluation metrics must be capable of capturing not only short-term accuracy but also long-term bias and variability. Metrics such as PBIAS are particularly important for assessing systematic errors in simulated water balance, while KGE provides insight into whether discrepancies arise from changes in variability, correlation, or mean flow. Recent climate-impact modeling studies emphasize that failure to evaluate long-term bias can propagate substantial uncertainty into adaptation and infrastructure planning decisions [15][4].

R^2 is often reported in climate impact studies to demonstrate trend agreement between observed and simulated series; however, it should not be interpreted as evidence of accurate magnitude or volume reproduction. Poor metric selection in climate change assessments can propagate uncertainty into water resources planning decisions, such as reservoir operation, irrigation demand estimation, and adaptation strategies. A multi-metric framework is therefore essential to ensure robust conclusions under changing climatic conditions.

5.4. Watershed-Scale Water Balance and Resource Planning

For watershed-scale water balance studies, the accurate representation of long-term volumes and seasonal patterns is often more important than individual event simulation. Metrics that explicitly quantify bias, such as PBIAS, play a central role in evaluating whether models conserve water mass over time. Absolute error measures (RMSE or MAE) provide information on typical deviations, while normalized metrics such as RSR facilitate comparison across watersheds with differing hydrological regimes.

In integrated water resources management, model outputs inform decisions related to allocation, storage, and sustainability. Inadequate evaluation, such as relying solely on NSE, can mask systematic biases that lead to overestimation of available water resources or underestimation of deficits. Consequently, transparent reporting of multiple complementary metrics is critical for ensuring that hydrological models provide credible support for water resources planning and policy.

6. Key Recommendations for Practice and Research

The No single performance metric fully captures hydrological model behavior across all flow regimes and applications. Commonly used metrics such as NSE and RMSE are intuitive and widely reported, but they emphasize different aspects of model performance. NSE evaluates predictive skill relative to the observed mean and tends to emphasize peak flows, while RMSE quantifies absolute error magnitude in physical units. The coefficient of determination (R^2) describes the strength of linear association between observed and simulated values but provides no information about bias or error magnitude. In contrast, PBIAS explicitly quantifies systematic over- or under-estimation and is therefore essential when volume conservation or water balance accuracy is important.

Kling-Gupta Efficiency (KGE) addresses several limitations of NSE by decomposing model performance into correlation, bias, and variability components, making it particularly useful for diagnostic evaluation. Normalized or bounded metrics, such as RSR and Willmott's refined index of agreement (d_r), further support comparison across catchments or studies with differing flow magnitudes. Consequently, reliance on a single metric can lead to misleading conclusions, especially when models perform unevenly across high-flow and low-flow conditions.

In practice, multi-criteria evaluation is strongly recommended. Reporting a combination of complementary metrics, such as NSE or KGE (overall goodness-of-fit), RMSE or

RSR (error magnitude), and PBIAS (bias), provides a concise yet comprehensive assessment of model performance. For applications focused on low-flow behavior or drought analysis, additional metrics sensitive to low flows (e.g., modified NSE formulations, percentile-based metrics, or flow-regime-specific indicators) should be included [13].

Specific recommendations include:

- Use a suite of metrics: At minimum, report one relative goodness-of-fit metric (NSE or KGE), one absolute or normalized error metric (RMSE, MAE, or RSR), and one bias indicator (PBIAS).
- NSE: Appropriate for assessing overall streamflow performance and peak flows; should always be paired with bias and low-flow-sensitive metrics.
- KGE: Preferred when diagnostic insight is needed, as it reveals whether deficiencies arise from bias, variability, or correlation.
- PBIAS: Essential when volume conservation, water balance, or systematic over- or under-estimation is critical.
- RMSE/MAE/RSR: RMSE provides intuitive error magnitudes; MAE offers robustness to outliers; RSR facilitates inter-basin comparison.
- R^2: Useful for contextualizing correlation but should never be used alone to claim model adequacy.
- Willmott's d/d_r: Suitable when a bounded metric is desired for communication; the refined index (d_r) is preferred due to improved statistical behavior.

Finally, numerical metrics should always be complemented with visual diagnostics, such as hydrographs, scatter plots, flow-duration curves, and Q-Q plots, as aggregated statistics can mask important differences in timing, magnitude, or flow-regime behavior. Statistical diagnostics drawn from hydrological statistics literature further support this recommendation, particularly for identifying non-normality, heteroscedasticity, and regime-dependent errors [19]. Future research should aim to develop clearer, application-specific guidelines for metric selection and to integrate flow-regime-focused and diagnostic evaluation frameworks into standard hydrological modeling practice.

7. Conclusion

Hydrological model performance metrics are complementary rather than interchangeable tools. Robust model evaluation requires combining goodness-of-fit, error-magnitude, and bias measures, such as NSE or KGE together with RMSE or RSR and PBIAS, alongside graphical diagnostics. Explicit reporting of metric definitions, versions (e.g., KGE formulation), thresholds, and component terms enhances transparency and interpretability. This need for clarity has been repeatedly highlighted in recent methodological critiques and synthesis studies. Thoughtful selection and combined use of evaluation metrics are therefore essential for reliable model assessment and for supporting informed decision-making in water resources applications.

Author Contributions: Conceptualization, T.S., A.AS., and M.W.; methodology, A.AS.; literature review, T.S., U.P., and A.W.; writing—original draft preparation, T.S.; writing—review and editing, A.AK., A.AS., U.P., A.W., M.W., and Z.A.; supervision, Z.A.; project administration, Z.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, “Model evaluation guidelines for systematic quantification of accuracy in watershed simulations,” Transactions of the ASABE, vol. 50, no. 3, pp. 885–900, 2007, doi: 10.13031/2013.23153.

[2] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, “Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling,” Journal of Hydrology, vol. 377, no. 1–2, pp. 80–91, 2009, doi: 10.1016/j.jhydrol.2009.08.003.

[3] M. Nesru, “Hydrological model selection and performance evaluation: A review,” Arabian Journal of Geosciences, vol. 16, no. 102, 2023, doi: 10.1007/s12517-023-11194-7.

[4] Y. T. Bihon, T. K. Lohani, A. T. Ayalew, B. G. Neka, A. K. Mohammed, G. B. Geremew, and E. G. Ayele, “Performance evaluation of various hydrological models with respect to hydrological responses under climate change scenario: a review,” Cogent Engineering, vol. 11, no. 1, 2360007, 2024, doi: 10.1080/23311916.2024.2360007.

[5] C. Kant, R. S. Meena, and S. K. Singh, “A critical appraisal on various hydrological and hydrodynamic models,” Water Conservation Science and Engineering, vol. 10, article 24, 2025, doi: 10.1007/s41101-024-00328-x.

[6] M. Dewangan and P. Vishvakarma, “Hydrological modeling and its applications: A review,” International Journal of Engineering in Computer Science and Electronics, vol. 12, no. 2, pp. 966–971, 2024, doi: 10.48047/intjecse/v12i2.201169.

[7] Y. Zhang, A. Ye, K. Hsu, S. Sorooshian, J. Li, P. Nguyen, and B. Analui, “Hydrological models vs machine learning models: Multi-model weighting ensembles improve global streamflow simulations,” Journal of Hydrology, vol. 655, 132904, 2025, doi: 10.1016/j.jhydrol.2025.132904.

[8] D. R. Legates and G. J. McCabe, “Evaluating the use of ‘goodness-of-fit’ measures in hydrologic and hydroclimatic model validation,” Water Resources Research, vol. 35, no. 1, pp. 233–241, 1999, doi: 10.1029/1998WR900018.

[9] M. P. Clark, R. M. Vogel, J. R. Lamontagne, N. Mizukami, W. J. M. Knoben, G. Tang, S. Gharari, J. Freer, P. H. Whitfield, K. Shook, and S. M. Papalexiou, “The abuse of popular performance metrics in hydrologic modeling,” Water Resources Research, vol. 57, no. 9, e2020WR029001, 2021, doi: 10.1029/2020WR029001.

[10] G. Cinkus, N. Mazzilli, H. Jourde, A. Wunsch, T. Liesch, N. Ravbar, Z. Chen, and N. Goldscheider, “When best is the enemy of good – critical evaluation of performance criteria in hydrological models,” Hydrology and Earth System Sciences, vol. 27, pp. 2397–2411, 2023, doi: 10.5194/hess-27-2397-2023.

[11] M. P. Clark, B. Nijssen, J. D. Lundquist, D. Kavetski, R. Rupp, R. Woods, J. Freer, E. D. Gutmann, A. Wood, L. Brekke, J. Arnold, D. Gochis, and R. M. Rasmussen, “A unified approach for process-based hydrologic modeling: 1. Modeling framework,” Water Resources Research, vol. 51, no. 4, pp. 2498–2514, 2015, doi: 10.1002/2015WR017198.

[12] G. Golmohammadi, S. Prasher, A. Madani, and R. Rudra, “Evaluating three hydrological distributed watershed models: MIKE-SHE, APEX, SWAT,” Hydrology, vol. 1, no. 1, pp. 20–39, 2014, doi: 10.3390/hydrology1010020.

[13] R. Pushpalatha, C. Perrin, N. Le Moine, and V. Andréassian, “A review of efficiency criteria suitable for evaluating low-flow simulations,” Journal of Hydrology, vol. 420–421, pp. 171–182, 2012, doi: 10.1016/j.jhydrol.2011.11.055.

[14] J. Rasmussen, H. Madsen, K. H. Jensen, and J. C. Refsgaard, “Evaluation of hydrological model performance across multiple catchments,” Hydrology Research, vol. 46, no. 6, pp. 887–903, 2015, doi: 10.2166/nh.2014.200.

[15] B. K. Mishra, K. Kobayashi, A. Murata, S. Fukui, and K. Suzuki, “Hydrologic modeling and flood-frequency analysis under climate change scenario,” Modeling Earth Systems and Environment, vol. 10, pp. 5621–5633, 2024, doi: 10.1007/s40808-024-02082-4.

[16] W. J. M. Knoben, J. E. Freer, and R. A. Woods, “Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores,” Hydrology and Earth System Sciences, vol. 23, pp. 4323–4331, 2019, doi: 10.5194/hess-23-4323-2019.

[17] C. J. Willmott, S. M. Robeson, and K. Matsuura, “A refined index of model performance,” International Journal of Climatology, vol. 32, no. 13, pp. 2088–2094, 2012, doi: 10.1002/joc.2419.

[18] J. E. Nash and J. V. Sutcliffe, “River flow forecasting through conceptual models. Part I: A discussion of principles,” Journal of Hydrology, vol. 10, no. 3, pp. 282–290, 1970, doi: 10.1016/0022-1694(70)90255-6.

[19] D. R. Helsel, R. M. Hirsch, K. R. Ryberg, S. A. Archfield, and E. J. Gilroy, Statistical methods in water resources. U.S. Geological Survey Techniques and Methods, Book 4, 2020, doi: 10.3133/tm4A3.