Article

Not peer-reviewed version

On the Mathematical Relationship Between RMSE and NSE Across Evaluation Scenarios

Walter Chen *

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.2032.v1

Keywords: Nash–Sutcliffe efficiency (NSE); root mean square error (RMSE); sum of squared errors (SSE); hydrological model evaluation; performance metrics; dataset expansion; machine learning

 doi:10.20944/preprints202509.2032.v1

Article

On the Mathematical Relationship Between RMSE and NSE Across Evaluation Scenarios

Walter Chen

Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; waltchen@ntut.edu.tw; Tel.: +886-(2)-27712171 (ext. 2628)

Abstract

Model evaluation metrics play a crucial role in hydrology, where accurate prediction of continuous variables such as streamflow and rainfall–runoff is essential for sustainable water resources management and climate resilience. Among these metrics, the Nash–Sutcliffe efficiency (NSE) is the most widely adopted, while the Root Mean Squared Error (RMSE) is also widely used because of its clear and intuitive interpretation. Although both metrics are functions of the sum of squared errors (SSE), their different normalizations can lead to subtle differences in model ranking under certain conditions. This study systematically investigates three scenarios: (i) both models are evaluated on the same dataset, (ii) both models are re-evaluated after adding new data, and (iii) one model is evaluated on the original dataset while the other is evaluated on the expanded dataset. For each case, we derive the mathematical conditions that determine whether one model can have simultaneously lower RMSE and NSE, or vice versa. The results demonstrate that RMSE and NSE always align on the same dataset in terms of model performance, but discrepancies can emerge when dataset expansion changes the total variance and models are compared across unequal evaluation bases. These findings clarify the interpretation of NSE and RMSE in hydrological model assessment and provide guidance for their use in broader machine learning applications. An important implication is that NSE can be artificially inflated: a model may appear to perform better simply by appending new data with high variance, even if its prediction errors remain large. This insight supports more reliable evaluation of hydrological models, contributing to better water management decisions under SDG 6 (Clean Water and Sanitation) and SDG 13 (Climate Action).

Keywords: Nash–Sutcliffe efficiency (NSE); root mean square error (RMSE); sum of squared errors (SSE); hydrological model evaluation; performance metrics; dataset expansion; machine learning

1. Introduction

The evaluation of predictive models is a central concern in hydrology and related environmental sciences. Accurate assessment of model performance is essential for tasks such as streamflow forecasting, rainfall–runoff simulation, and soil erosion prediction, where decision-making often depends directly on the reliability of hydrological models. Among the many performance metrics that have been proposed, the Nash–Sutcliffe efficiency (NSE) [1] remains the most widely adopted in the hydrology community. Defined as a normalized measure of squared error relative to the variance of observations, NSE provides a convenient and interpretable scale: a value of 1 represents a perfect model, 0 corresponds to predictions no better than the observed mean, and negative values indicate performance worse than the mean benchmark.

Despite its popularity, NSE is closely related to another widely used error metric: the Root Mean Squared Error (RMSE). Both metrics are functions of the same underlying quantity, the sum of squared errors (SSE), but they differ in normalization. RMSE reports error in the same units as the observations, whereas NSE expresses model skill relative to observed variability. This duality leads to non-trivial behaviors when comparing models, particularly under dataset expansion or distributional shifts.

 doi:10.20944/preprints202509.2032.v1

Understanding these relationships is critical for fair model comparison and for interpreting results across different studies.

Beyond hydrology, the role of NSE is growing in the machine learning and deep learning literature [2–5], where it is increasingly used as a performance measure for models dealing with continuous target variables. Machine learning practitioners are drawn to NSE for its ability to contextualize model error against a baseline predictor, complementing traditional metrics such as mean squared error (MSE) and mean absolute error (MAE). As machine learning techniques are increasingly applied to hydrological problems, careful scrutiny of how NSE behaves relative to RMSE becomes especially relevant. Recently, a study of surface water velocity prediction demonstrated that a model can achieve a higher NSE while simultaneously exhibiting a higher RMSE compared to another model, highlighting an inconsistency between these two metrics [6,7].

In hydrology, numerous studies have critiqued NSE and related metrics, emphasizing their limitations and sensitivity to data characteristics. Gupta et al. [8] provided the well-known decomposition of mean squared error (MSE) into correlation, bias, and variability terms, clarifying how variance inflation can elevate NSE values without genuine error reduction. Subsequent work has shown that such dependence on local variability undermines cross-site comparisons: Williams [9] argued that NSE and the Kling–Gupta Efficiency (KGE) are unsuitable for this purpose, while Melsen [10] reviewed the widespread but problematic reliance on NSE within hydrological practice. Onyutha [11] further demonstrated that efficiency criteria can shift rankings under changing variability, bias, or outliers, underscoring that the choice of metric itself introduces calibration uncertainty. Methodological refinements, such as the probabilistic estimators proposed by Lamontagne et al. [12], aim to improve NSE and KGE performance, yet their structural sensitivities remain. Related paradoxes have also been observed for other error measures, with Chai and Draxler [13] showing that RMSE and MAE rankings can diverge depending on error distributions. Together, these critiques highlight the limitations of current practices and frame the need for a formal characterization of when and why contradictions arise between RMSE and NSE.

The present study directly addresses this gap by systematically analyzing three scenarios in which RMSE and NSE can exhibit different comparative behaviors between models. We show that while both metrics are monotonic with respect to SSE on a common dataset, discrepancies may arise when models are evaluated on different data subsets or when additional variability is introduced. By deriving precise mathematical conditions for these cases, we clarify the circumstances under which NSE and RMSE rankings agree or diverge. Importantly, these insights not only advance the methodological understanding of model evaluation but also support sustainable water resources management and climate adaptation strategies, aligning with the United Nations Sustainable Development Goals, particularly SDG 6 (Clean Water and Sanitation) and SDG 13 (Climate Action).

2. Materials and Methods

This section outlines the analytical framework used to examine the relationship between RMSE and NSE. We first describe the models under consideration and how their prediction errors are represented. Next, we define the datasets and how they are expanded across cases, followed by the formal expressions of RMSE and NSE that form the basis of our derivations. Together, these elements provide the foundation for the comparative analysis presented in the Results section.

2.1. Models

Two generic predictive models are considered and denoted as Model A and Model B. The internal structures of the models are not specified because they are irrelevant to the analysis, which focuses solely on error behavior under different evaluation scenarios. The predictions of Model A and Model B at time step i are denoted by $\hat{y}_{A,i}$ and $\hat{y}_{B,i}$, respectively, while the observed value is denoted by y_i.

 doi:10.20944/preprints202509.2032.v1

2.2. Datasets

Let X denote the original dataset consisting of n observations,

$$X = \{y_1, y_2, \ldots, y_n\}.$$

A second dataset block of equal size n is introduced,

$$Z = \{y_{n+1}, y_{n+2}, \ldots, y_{2n}\},$$

and the combined dataset is denoted by

$$Y = X \cup Z.$$

The mean of X is denoted by $\bar{y}_X$, and the mean of the combined dataset Y is denoted by $\bar{y}_Y$.

2.3. Evaluation Metrics

The analysis considers two standard metrics of model performance: RMSE and NSE. For a dataset with N observations, these are defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{1}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2}$$

where $\hat{y}_i$ denotes the model prediction and $\bar{y}$ is the mean of the observed series. Both metrics are functions of the SSE, but RMSE reports error in the same units as the observations, whereas NSE normalizes error relative to the total variance of the observations.

2.4. Evaluation Scenarios

Three scenarios are investigated to examine the comparative behavior of RMSE and NSE:

1. **Case I: Same Dataset Evaluation.** Both models are evaluated on the dataset X.
2. **Case II: Expanded Dataset with Both Models Re-Evaluated.** Both models are evaluated on the combined dataset Y.
3. **Case III: Unequal Dataset Evaluation.** Model A is evaluated on X, while Model B is evaluated on Y.

These three scenarios are designed to clarify the mathematical relationship between RMSE and NSE, showing conditions where their rankings are consistent and conditions where contradictions may arise.

3. Results

The results are organized into three cases that represent different ways of evaluating and comparing models. Case I considers both models evaluated on the same dataset, Case II examines the situation where both models are re-evaluated on an expanded dataset, and Case III addresses the unequal evaluation scenario in which one model is tested only on the original dataset while the other is tested on the expanded dataset. Together, these cases make it possible to identify conditions under which RMSE and NSE provide consistent rankings and those under which discrepancies may emerge.

3.1. Case I: Same Dataset Evaluation

Definition 1. *Let* $X = \{(y_i, \hat{y}_{A,i}, \hat{y}_{B,i})\}_{i=1}^{n}$ *be the common evaluation set, with mean* $\bar{y}_X = \frac{1}{n}\sum_{i=1}^{n} y_i$ *and total sum of squares*

$$\text{SST}_X = \sum_{i=1}^{n}(y_i - \bar{y}_X)^2. \tag{3}$$

 doi:10.20944/preprints202509.2032.v1

Define the models' sums of squared errors (SSE) on X as

$$\mathrm{SSE}_A = \sum_{i=1}^{n} (y_i - \hat{y}_{A,i})^2, \qquad \mathrm{SSE}_B = \sum_{i=1}^{n} (y_i - \hat{y}_{B,i})^2. \tag{4}$$

Then

$$\mathrm{RMSE}_M = \sqrt{\frac{\mathrm{SSE}_M}{n}}, \qquad \mathrm{NSE}_M = 1 - \frac{\mathrm{SSE}_M}{\mathrm{SST}_X}, \tag{5}$$

for $M \in \{A, B\}$, with the usual caveat that $\mathrm{SST}_X > 0$ (i.e., the observed series is not constant) for NSE to be defined.

Proposition 1. *If $\mathrm{RMSE}_A < \mathrm{RMSE}_B$ on the same dataset X, then $\mathrm{NSE}_A > \mathrm{NSE}_B$.*

Proof. Because the square root is strictly increasing and n is identical for both models,

$$\begin{aligned} \mathrm{RMSE}_A < \mathrm{RMSE}_B &\iff \sqrt{\tfrac{\mathrm{SSE}_A}{n}} < \sqrt{\tfrac{\mathrm{SSE}_B}{n}} \\ &\iff \tfrac{\mathrm{SSE}_A}{n} < \tfrac{\mathrm{SSE}_B}{n} \\ &\iff \mathrm{SSE}_A < \mathrm{SSE}_B. \end{aligned}$$

Now consider the NSE difference:

$$\mathrm{NSE}_A - \mathrm{NSE}_B = \left(1 - \tfrac{\mathrm{SSE}_A}{\mathrm{SST}_X}\right) - \left(1 - \tfrac{\mathrm{SSE}_B}{\mathrm{SST}_X}\right) = \tfrac{\mathrm{SSE}_B - \mathrm{SSE}_A}{\mathrm{SST}_X}. \tag{6}$$

Since $\mathrm{SST}_X > 0$, the sign of $\mathrm{NSE}_A - \mathrm{NSE}_B$ matches the sign of $\mathrm{SSE}_B - \mathrm{SSE}_A$. From $\mathrm{SSE}_A < \mathrm{SSE}_B$ we obtain $\mathrm{SSE}_B - \mathrm{SSE}_A > 0$, hence

$$\mathrm{NSE}_A - \mathrm{NSE}_B = \tfrac{\mathrm{SSE}_B - \mathrm{SSE}_A}{\mathrm{SST}_X} > 0 \implies \mathrm{NSE}_A > \mathrm{NSE}_B.$$

This proves the claim. □

Proposition 2 (Equivalent linear relation)**.** *On a fixed dataset X, NSE is a strictly decreasing linear function of* RMSE^2.

Proof. Since $\mathrm{MSE} = \mathrm{RMSE}^2 = \mathrm{SSE}/n$, one can write

$$\mathrm{NSE} = 1 - \frac{n\,\mathrm{RMSE}^2}{\mathrm{SST}_X}. \tag{7}$$

Therefore,

$$\mathrm{RMSE}_A < \mathrm{RMSE}_B \iff \mathrm{RMSE}_A^2 < \mathrm{RMSE}_B^2 \iff \mathrm{NSE}_A > \mathrm{NSE}_B.$$

□

Interpretation. On the same dataset, RMSE and NSE always produce consistent rankings: whichever model achieves lower RMSE necessarily achieves higher NSE. Ties occur only if $\mathrm{SSE}_A = \mathrm{SSE}_B$, while NSE becomes undefined if $\mathrm{SST}_X = 0$. This confirms that for hydrological model evaluation under fixed datasets, RMSE and NSE cannot yield contradictory conclusions.

3.2. Case II: Expanded Dataset with Both Models Re-Evaluated

Definition 2. *Let the original dataset X contain n observations, with Model A and Model B errors measured by*

$$\mathrm{SSE}_{A1} = \sum_{i=1}^{n} (y_i - \hat{y}_{A,i})^2, \qquad \mathrm{SSE}_{B1} = \sum_{i=1}^{n} (y_i - \hat{y}_{B,i})^2. \tag{8}$$

 doi:10.20944/preprints202509.2032.v1

Assume that $\mathrm{SSE}_{A1} < \mathrm{SSE}_{B1}$*, so Model A initially has lower RMSE and higher NSE. Define the deficit of Model B as*

$$\Delta_1 = \mathrm{SSE}_{B1} - \mathrm{SSE}_{A1} > 0. \tag{9}$$

A new block of n observations, denoted Z, is added, producing the expanded dataset $Y = X \cup Z$ *with* $2n$ *points. Model errors on Z are*

$$\mathrm{SSE}_{A2} = \sum_{i=n+1}^{2n} (y_i - \hat{y}_{A,i})^2, \qquad \mathrm{SSE}_{B2} = \sum_{i=n+1}^{2n} (y_i - \hat{y}_{B,i})^2. \tag{10}$$

Thus, on Y the total squared errors are

$$\mathrm{SSE}_A = \mathrm{SSE}_{A1} + \mathrm{SSE}_{A2}, \qquad \mathrm{SSE}_B = \mathrm{SSE}_{B1} + \mathrm{SSE}_{B2}. \tag{11}$$

Proposition 3 (Consistency on expanded dataset)**.** *When both models are evaluated on the expanded dataset Y, RMSE and NSE always produce consistent rankings:*

$$\mathrm{RMSE}_A < \mathrm{RMSE}_B \iff \mathrm{NSE}_A > \mathrm{NSE}_B. \tag{12}$$

Proof. Both RMSE and NSE are monotonic functions of SSE when evaluated on the same dataset. Therefore, comparing RMSE_A and RMSE_B is equivalent to comparing SSE_A and SSE_B, which in turn determines the sign of $\mathrm{NSE}_A - \mathrm{NSE}_B$. Hence, contradictory outcomes such as $\mathrm{RMSE}_A < \mathrm{RMSE}_B$ together with $\mathrm{NSE}_A < \mathrm{NSE}_B$ are impossible. □

Proposition 4 (Condition for reversal)**.** *For Model B to surpass Model A on Y, the requirement is*

$$\mathrm{SSE}_B < \mathrm{SSE}_A. \tag{13}$$

Equivalently,

$$\mathrm{SSE}_{B1} + \mathrm{SSE}_{B2} < \mathrm{SSE}_{A1} + \mathrm{SSE}_{A2} \iff \mathrm{SSE}_{A2} - \mathrm{SSE}_{B2} > \Delta_1. \tag{14}$$

Proof. Substituting the blockwise decompositions $\mathrm{SSE}_A = \mathrm{SSE}_{A1} + \mathrm{SSE}_{A2}$ and $\mathrm{SSE}_B = \mathrm{SSE}_{B1} + \mathrm{SSE}_{B2}$, the inequality $\mathrm{SSE}_B < \mathrm{SSE}_A$ rearranges to (14), where $\Delta_1 = \mathrm{SSE}_{B1} - \mathrm{SSE}_{A1}$. This condition means that Model B must outperform Model A on the added block Z by more than its initial deficit on X. □

Interpretation. On the expanded dataset Y, RMSE and NSE remain consistent in their ranking of models. Model B can only overtake Model A if its relative improvement on Z outweighs its earlier disadvantage Δ_1 from X. Otherwise, Model A continues to be superior on both metrics. Thus, when both models are assessed on the same expanded dataset, RMSE and NSE cannot yield conflicting conclusions.

3.3. Case III: Unequal Dataset Evaluation

Definition 3. *Let the original dataset X contain n observations, with sums of squared errors* SSE_{A1} *and* SSE_{B1} *for Model A and Model B, respectively. Assume* $\mathrm{SSE}_{A1} < \mathrm{SSE}_{B1}$ *so that Model A initially outperforms Model B. Model A is evaluated only on X, while Model B is evaluated on the expanded dataset* $Y = X \cup Z$*, where Z is a new block of n observations. For Model A,*

$$\mathrm{RMSE}_A = \sqrt{\frac{\mathrm{SSE}_{A1}}{n}}, \qquad \mathrm{NSE}_A = 1 - \frac{\mathrm{SSE}_{A1}}{\mathrm{SST}_X}, \qquad \mathrm{SST}_X = \sum_{i=1}^{n} (y_i - \bar{y}_X)^2, \tag{15}$$

while for Model B on Y,

$$\mathrm{SSE}_B = \mathrm{SSE}_{B1} + \mathrm{SSE}_{B2}, \qquad \mathrm{RMSE}_B^{(Y)} = \sqrt{\frac{\mathrm{SSE}_{B1} + \mathrm{SSE}_{B2}}{2n}}, \qquad \mathrm{NSE}_B^{(Y)} = 1 - \frac{\mathrm{SSE}_{B1} + \mathrm{SSE}_{B2}}{\mathrm{SST}_Y}, \tag{16}$$

 doi:10.20944/preprints202509.2032.v1

where SST_Y *is the total sum of squares of the combined dataset.*

Proposition 5 (RMSE condition)**.** *For Model B to have larger RMSE than Model A on Y, it is necessary and sufficient that*

$$\mathrm{SSE}_{B1} + \mathrm{SSE}_{B2} > 2\,\mathrm{SSE}_{A1}. \tag{17}$$

Proof. The inequality $\mathrm{RMSE}_B^{(Y)} > \mathrm{RMSE}_A$ expands to $\sqrt{\frac{\mathrm{SSE}_{B1}+\mathrm{SSE}_{B2}}{2n}} > \sqrt{\frac{\mathrm{SSE}_{A1}}{n}}$. Squaring both sides and multiplying through by $2n$ gives the stated condition. □

Proposition 6 (NSE condition)**.** *For Model B to have larger NSE than Model A, it is necessary and sufficient that*

$$\mathrm{SSE}_{B1} + \mathrm{SSE}_{B2} < \frac{\mathrm{SST}_Y}{\mathrm{SST}_X}\,\mathrm{SSE}_{A1}. \tag{18}$$

Proof. The inequality $\mathrm{NSE}_B^{(Y)} > \mathrm{NSE}_A$ expands to $1 - \frac{\mathrm{SSE}_{B1}+\mathrm{SSE}_{B2}}{\mathrm{SST}_Y} > 1 - \frac{\mathrm{SSE}_{A1}}{\mathrm{SST}_X}$. Rearranging yields the stated condition. □

Proposition 7 (Combined conditions)**.** *Both conditions hold simultaneously if and only if*

$$2\,\mathrm{SSE}_{A1} < \mathrm{SSE}_{B1} + \mathrm{SSE}_{B2} < \frac{\mathrm{SST}_Y}{\mathrm{SST}_X}\,\mathrm{SSE}_{A1}. \tag{19}$$

Equivalently, the feasible interval for SSE_{B2} *is*

$$\max\left(0,\ 2\,\mathrm{SSE}_{A1} - \mathrm{SSE}_{B1}\right) < \mathrm{SSE}_{B2} < \frac{\mathrm{SST}_Y}{\mathrm{SST}_X}\,\mathrm{SSE}_{A1} - \mathrm{SSE}_{B1}. \tag{20}$$

Proof. Combining the RMSE and NSE inequalities (17) and (18) gives the stated two-sided bound. The equivalent expression follows by solving for SSE_{B2}. □

Proposition 8 (Variance decomposition)**.** *For equal block sizes, the variance ratio satisfies*

$$\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} = 1 + \frac{\mathrm{SST}_Z}{\mathrm{SST}_X} + \frac{n}{2\,\mathrm{SST}_X}(\bar{y}_X - \bar{y}_Z)^2, \tag{21}$$

where $\mathrm{SST}_Z = \sum_{i=n+1}^{2n}(y_i - \bar{y}_Z)^2$ *is the within-block variance of Z and* $\bar{y}_Z$ *is its mean. Hence the feasible interval is nonempty if and only if* $\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} > 2$.

Proof. Let $\bar{y}_X = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{y}_Z = \frac{1}{n}\sum_{i=n+1}^{2n} y_i$, and $\bar{y}_Y = \frac{1}{2}(\bar{y}_X + \bar{y}_Z)$. By definition,

$$\mathrm{SST}_Y = \sum_{i=1}^{2n}(y_i - \bar{y}_Y)^2 = \sum_{i=1}^{n}(y_i - \bar{y}_Y)^2 + \sum_{i=n+1}^{2n}(y_i - \bar{y}_Y)^2.$$

For block X, write $y_i - \bar{y}_Y = (y_i - \bar{y}_X) + (\bar{y}_X - \bar{y}_Y)$ and expand:

$$\sum_{i=1}^{n}(y_i - \bar{y}_Y)^2 = \underbrace{\sum_{i=1}^{n}(y_i - \bar{y}_X)^2}_{\mathrm{SST}_X} + n(\bar{y}_X - \bar{y}_Y)^2,$$

where the cross term vanishes because $\sum_{i=1}^{n}(y_i - \bar{y}_X) = 0$. For block Z, similarly,

$$\sum_{i=n+1}^{2n}(y_i - \bar{y}_Y)^2 = \mathrm{SST}_Z + n(\bar{y}_Z - \bar{y}_Y)^2.$$

 doi:10.20944/preprints202509.2032.v1

Adding both parts gives

$$\mathrm{SST}_Y = \mathrm{SST}_X + \mathrm{SST}_Z + n(\bar{y}_X - \bar{y}_Y)^2 + n(\bar{y}_Z - \bar{y}_Y)^2.$$

Since $\bar{y}_Y = \frac{1}{2}(\bar{y}_X + \bar{y}_Z)$, we obtain

$$n(\bar{y}_X - \bar{y}_Y)^2 + n(\bar{y}_Z - \bar{y}_Y)^2 = \tfrac{n}{2}(\bar{y}_X - \bar{y}_Z)^2.$$

Hence

$$\mathrm{SST}_Y = \mathrm{SST}_X + \mathrm{SST}_Z + \tfrac{n}{2}(\bar{y}_X - \bar{y}_Z)^2,$$

which yields (21) after dividing by SST_X. □

Interpretation. In the unequal dataset case, contradictions between RMSE and NSE become possible. Specifically, Model B may have larger RMSE yet simultaneously larger NSE than Model A, provided that the variance of the combined dataset more than doubles relative to X. This inflation of SST_Y arises when Z exhibits high within-block variability, a substantial mean shift from X, or both. In such cases, the enlarged denominator in the NSE formula reduces the relative penalty for Model B's higher errors, allowing NSE to rank it above Model A even though RMSE confirms that Model A remains superior in absolute error.

4. Discussion

The preceding mathematical analysis considered three distinct scenarios for comparing models under the RMSE and NSE metrics. While the Results section established precise conditions under which the two metrics agree or diverge, it is equally important to interpret these findings in the broader context of hydrological model evaluation and machine learning practice. This section discusses the implications of each case, emphasizing when the metrics remain consistent and when apparent contradictions can arise.

4.1. Case I and Case II

When both models are evaluated on the same dataset, either the original dataset X (Case I) or the expanded dataset Y (Case II), the ranking of models by RMSE and NSE is always consistent. Since both metrics are monotonic functions of the sum of squared errors (SSE) relative to the same variance baseline, we obtain the equivalence

$$\mathrm{RMSE}_A < \mathrm{RMSE}_B \iff \mathrm{NSE}_A > \mathrm{NSE}_B. \tag{22}$$

Therefore, no contradictions can arise: whichever model achieves a lower RMSE necessarily achieves a higher NSE. This alignment ensures that, in practice, evaluations on the same dataset leave no ambiguity about which model is superior.

4.2. Case III

In contrast, when the two models are evaluated on unequal datasets, it is possible for the rankings to diverge. Specifically, Model A may be evaluated only on X, while Model B is evaluated on the expanded dataset $Y = X \cup Z$. In this situation, Model B's NSE is normalized by SST_Y, the variance of the combined dataset. If SST_Y is much larger than SST_X, Model B's relative error ratio can shrink even if its absolute error (and thus its RMSE) remains larger. The necessary and sufficient condition for this outcome is

$$2\,\mathrm{SSE}_{A1} < \mathrm{SSE}_{B1} + \mathrm{SSE}_{B2} < \frac{\mathrm{SST}_Y}{\mathrm{SST}_X}\,\mathrm{SSE}_{A1}, \tag{23}$$

with feasibility requiring

$$\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} > 2. \tag{24}$$

 doi:10.20944/preprints202509.2032.v1

Using the pooled variance decomposition,

$$\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} = 1 + \frac{\mathrm{SST}_Z}{\mathrm{SST}_X} + \frac{n}{2\,\mathrm{SST}_X}(\bar{y}_X - \bar{y}_Z)^2, \tag{25}$$

this condition can be satisfied if the added block Z has either very large internal variability (large SST_Z), a substantial mean shift relative to X (large $|\bar{y}_X - \bar{y}_Z|$), or both. Under such conditions, it becomes possible for Model A to achieve a lower RMSE yet a lower NSE than Model B, thus producing a genuine contradiction between the two metrics.

4.3. Implications and Potential for Inflated NSE

The results highlight that contradictions between RMSE and NSE rankings only arise when models are evaluated on different datasets. In hydrological and machine learning practice, such situations can occur when models are trained or tested on unequal time periods or spatial domains. In machine learning in particular, it is common for models to be benchmarked on entirely different datasets. In these settings, direct comparison of NSE values across studies or regions is not meaningful: a higher NSE in one study does not necessarily indicate a better model than one with a lower NSE in another, because the underlying datasets may have very different variance structures, mean levels, or event distributions. Without controlling for dataset characteristics, NSE alone is unsuitable for comparing models across independent experiments. Case III further emphasizes that NSE can be artificially inflated by adding new data with sufficiently high variance or shifted means. A weak model may therefore appear superior under NSE simply because the denominator in the NSE formula grows faster than its errors.

Beyond variance inflation, other potential forms of NSE manipulation include: (i) selectively extending the dataset with extreme events that dominate the variance without proportionally increasing the error; (ii) rescaling the observed series (e.g., through unit changes or aggregation choices) so that the variance baseline increases; and (iii) cherry-picking evaluation periods with naturally high variability (such as wet seasons in rainfall–runoff studies). Each of these strategies can make a weak model appear competitive under NSE, while its absolute accuracy as measured by RMSE remains poor. Importantly, RMSE is not subject to these distortions, since it directly reflects the magnitude of prediction errors without reference to the variance of the observations. Unlike NSE, RMSE cannot be inflated by dataset characteristics such as variance shifts or mean differences. As a result, RMSE provides a more stable and accurate measure of absolute model performance, making it a valuable complement to NSE when comparing models.

These findings suggest that NSE should always be interpreted with caution, especially in comparative studies involving datasets of different scales or variability. For hydrology and related sustainability fields, the broader implication is that rigorous model assessment requires transparency in dataset selection and metric reporting, ensuring that apparent performance gains are not merely artifacts of evaluation design.

5. Algorithmic Demonstration of NSE Inflation

The preceding analysis shows that contradictions between RMSE and NSE rankings arise only when models are evaluated on unequal datasets. In particular, increasing the variance of an expanded evaluation set $Y = X \cup Z$ can raise the NSE even when absolute errors remain large. This section provides an explicit constructive algorithm which, for a given dataset $X = \{(y_i, \hat{y}_i)\}_{i=1}^n$ and desired margin $\delta > 0$, produces a block Z such that

$$\mathrm{NSE}_Y = \mathrm{NSE}_X + \delta. \tag{26}$$

 doi:10.20944/preprints202509.2032.v1

5.1. Algorithm Outline

Let

$$SST_X = \sum_{i=1}^{n}(y_i - \bar{y}_X)^2, \quad SSE_X = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \mathrm{NSE}_X = 1 - \frac{SSE_X}{SST_X}.$$

We target (26), which is equivalent to the required squared error of the added block Z:

$$SSE_Z = \left(\frac{SSE_X}{SST_X} - \delta\right) SST_Y - SSE_X. \tag{27}$$

Write the variance ratio $k_{\mathrm{var}} := SST_Y/SST_X > 1$. Then (27) becomes

$$SSE_Z = \left[(1 - \mathrm{NSE}_X - \delta)\, k_{\mathrm{var}} - (1 - \mathrm{NSE}_X)\right] SST_X. \tag{28}$$

To keep $SSE_Z \geq 0$, the minimal variance inflation factor must satisfy

$$k_{\mathrm{var}} \;\geq\; \frac{1 - \mathrm{NSE}_X}{1 - \mathrm{NSE}_X - \delta}. \tag{29}$$

For equal block sizes $|Z| = |X| = n$, the pooled-variance identity is

$$\frac{SST_Y}{SST_X} = 1 + \frac{SST_Z}{SST_X} + \frac{n}{2SST_X}(\bar{y}_X - \bar{y}_Z)^2. \tag{30}$$

Hence any desired k_{var} can be realized by choosing a within-block spread SST_Z and/or a mean shift $|\bar{y}_Z - \bar{y}_X|$. Two canonical constructions are provided below: (i) *All Spread* (no mean shift), and (ii) *All Shift* (minimal spread). A short "Mixed" version is also included.

Proposition 9 (Guarantee). *Fix X, choose δ with $0 < \delta < 1 - \mathrm{NSE}_X$, and select k_{var} satisfying (29). If Z is constructed to realize $SST_Y = k_{\mathrm{var}}\, SST_X$ via (30), and its prediction errors satisfy (27), then $\mathrm{NSE}_Y = \mathrm{NSE}_X + \delta$. Moreover, if SSE_Z is strictly smaller than (27), then $\mathrm{NSE}_Y > \mathrm{NSE}_X + \delta$.*

5.2. Practical Notes and Constraints

- Valid range: $0 < \delta < 1 - \mathrm{NSE}_X$; otherwise NSE_Y would exceed 1 or feasibility fails.
- Larger δ demands larger k_{var} per (29), achievable through SST_Z and/or mean shift $|\bar{y}_Z - \bar{y}_X|$ in (30).
- If the right-hand side of (27) is negative, the chosen k_{var} cannot deliver the target δ; increase k_{var} or reduce δ.
- Error assignment on Z: construct a vector $e \in \mathbb{R}^n$ with $\|e\|_2^2 = SSE_Z$, then set $\hat{y}_j = y_j - e_j$ for $j = n+1, \ldots, 2n$. This can be done deterministically (equal-magnitude entries with alternating signs) or stochastically (i.i.d. random draws rescaled to the exact norm).

5.3. Minimal Worked Algorithms

All Spread (no mean shift)

This construction achieves the required variance inflation entirely by enlarging the within-block spread of Z while keeping its mean identical to that of X. It demonstrates how NSE can be raised through variance inflation alone.

 doi:10.20944/preprints202509.2032.v1

Algorithm 1 Variance-Only Construction (All Spread).

Input: $X = \{(y_i, \hat{y}_i)\}_{i=1}^{n}$, target margin $\delta \in (0, 1 - \mathrm{NSE}_X)$
Compute SST_X, SSE_X, NSE_X.
Set $k_{\mathrm{var}} \leftarrow \dfrac{1 - \mathrm{NSE}_X}{1 - \mathrm{NSE}_X - \delta}$.
Set $\bar{y}_Z \leftarrow \bar{y}_X$. ▷ No mean shift
Set $SST_Z \leftarrow (k_{\mathrm{var}} - 1)\, SST_X$.
Construct $y_{n+1:2n}$ with mean $\bar{y}_X$ and variance SST_Z.
Compute $SST_Y \leftarrow k_{\mathrm{var}}\, SST_X$.
Compute SSE_Z from (27).
Build error vector e with $\|e\|_2^2 = SSE_Z$; set $\hat{y}_j \leftarrow y_j - e_j$.
Output: $Y = X \cup Z$ with $\mathrm{NSE}_Y = \mathrm{NSE}_X + \delta$.

All Shift (minimal spread)

This construction achieves the required variance inflation almost entirely by shifting the mean of Z, while keeping its internal variance negligible. It illustrates that even a simple mean displacement can artificially inflate NSE.

Algorithm 2 Mean-Shift-Only Construction (All Shift).

Input: $X = \{(y_i, \hat{y}_i)\}_{i=1}^{n}$, target margin $\delta \in (0, 1 - \mathrm{NSE}_X)$
Compute SST_X, SSE_X, NSE_X.
Set $k_{\mathrm{var}} \leftarrow \dfrac{1 - \mathrm{NSE}_X}{1 - \mathrm{NSE}_X - \delta}$.
Set $SST_Z \approx 0$. ▷ Minimal variance
Set $\Delta_\mu \leftarrow \sqrt{\frac{2SST_X}{n}(k_{\mathrm{var}} - 1)}$.
Set $\bar{y}_Z \leftarrow \bar{y}_X + \Delta_\mu$; construct $y_{n+1:2n}$ with mean $\bar{y}_Z$ and negligible variance.
Compute $SST_Y \leftarrow k_{\mathrm{var}}\, SST_X$.
Compute SSE_Z from (27); build e with $\|e\|_2^2 = SSE_Z$.
Output: $Y = X \cup Z$ with $\mathrm{NSE}_Y = \mathrm{NSE}_X + \delta$.

Mixed Spread+Shift (optional)

Choose any nonnegative pair (SST_Z, Δ_μ) satisfying

$$1 + \frac{SST_Z}{SST_X} + \frac{n}{2SST_X}\,\Delta_\mu^2 \;=\; k_{\mathrm{var}},$$

then apply (27). For example, fix Δ_μ to control the mean of Z and solve for $SST_Z = (k_{\mathrm{var}} - 1)SST_X - \frac{n}{2}\Delta_\mu^2$.

These constructions serve as stress tests, highlighting NSE's sensitivity to dataset design under unequal evaluation bases. By contrast, in Cases I and II, where the dataset is fixed, RMSE and NSE rankings remain strictly consistent.

5.4. Ensuring Larger RMSE Together with Higher NSE

The constructions above guarantee that the NSE of the expanded dataset $Y = X \cup Z$ can be increased by any desired margin δ. However, this alone does not ensure that RMSE_Y also exceeds RMSE_X. To investigate this, consider

$$\mathrm{RMSE}_X = \sqrt{\frac{SSE_X}{n}}, \qquad \mathrm{RMSE}_Y = \sqrt{\frac{SSE_X + SSE_Z}{2n}}.$$

We have $\mathrm{RMSE}_Y > \mathrm{RMSE}_X$ if and only if $SSE_Z > SSE_X$.

Substituting (28) and using $SSE_X = (1 - \mathrm{NSE}_X)SST_X$, this condition becomes

$$k_{\mathrm{var}} > \frac{2(1 - \mathrm{NSE}_X)}{1 - \mathrm{NSE}_X - \delta}, \tag{31}$$

 doi:10.20944/preprints202509.2032.v1

which is a stricter requirement than (29).

Interpretation. Inequality (29) establishes the minimum variance inflation needed to raise NSE by δ, but it does not constrain RMSE. The stronger bound (31) guarantees that $\mathrm{NSE}_Y > \mathrm{NSE}_X$ *and* $\mathrm{RMSE}_Y > \mathrm{RMSE}_X$. Importantly, when $\delta = 0$, (31) reduces to $\frac{SST_Y}{SST_X} > 2$, exactly the threshold identified earlier for contradictions to be possible. Thus, the simple "greater than two" rule is revealed as a special case of this general framework.

This result demonstrates that it is possible to construct a dataset Z such that the combined dataset $Y = X \cup Z$ yields higher NSE but also higher RMSE compared with X. In other words, an inferior model with larger errors can nevertheless appear superior under NSE once the evaluation dataset is artificially expanded, underscoring the vulnerability of NSE as a performance metric in unequal evaluation settings.

6. Conclusions

This study examined the relationship between NSE and RMSE across three evaluation scenarios. The key findings can be summarized as follows:

1. When models are evaluated on the same dataset (Cases I and II), the rankings by RMSE and NSE are always consistent. A lower RMSE necessarily implies a higher NSE, and no contradictory outcomes are possible.
2. When models are evaluated on unequal datasets (Case III), contradictions may arise. In this setting, it is possible for one model to have a lower RMSE but simultaneously a lower NSE. The necessary and sufficient condition for this outcome is that the total variance of the expanded dataset more than doubles that of the original dataset, i.e.,

$$\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} > 2.$$

   This situation may occur if the new data block has very large variability, a substantial mean shift, or both.
3. A strengthened bound was derived showing that, for a targeted increase of δ in NSE, one requires

$$k_{\mathrm{var}} > \frac{2(1 - \mathrm{NSE}_X)}{1 - \mathrm{NSE}_X - \delta},$$

   which guarantees that the combined RMSE is also larger than the original RMSE. This result generalizes the > 2 rule: when $\delta = 0$, the strengthened bound reduces exactly to $\frac{\mathrm{SST}_Y}{\mathrm{SST}_X} > 2$. The implication is that an inferior model, already worse in RMSE, can nevertheless appear superior under NSE once the dataset is artificially expanded. In other words, the model remains less accurate in absolute terms yet appears better in relative efficiency, exposing a structural vulnerability of NSE.

In summary, the analysis demonstrates that RMSE and NSE provide fully consistent guidance when applied to a common dataset, but contradictions emerge once models are compared across different evaluation bases. The inequalities derived here formalize the exact conditions under which such paradoxes occur, clarifying the mechanisms that drive apparent improvements in NSE despite deteriorating RMSE. For hydrological and machine learning applications, this emphasizes the critical importance of consistent evaluation datasets and cautions against overinterpreting NSE values in cross-dataset comparisons, where a model may seem improved by NSE even though its RMSE—and thus its absolute accuracy—has worsened. By making these conditions explicit, the study contributes a rigorous theoretical foundation for interpreting efficiency metrics and highlights the need for transparency in evaluation design.

 doi:10.20944/preprints202509.2032.v1

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, and funding acquisition, W.C. The author has read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the National Science and Technology Council (Taiwan) under Research Project Grant Numbers NSTC 114-2121-M-027-001 and NSTC 113-2121-M-008-004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT 5 (OpenAI) for assistance in editing and polishing the writing. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290.
2. Liu, C.-Y.; Ku, C.-Y.; Wu, T.-Y.; Chiu, Y.-J.; Chang, C.-W. Liquefaction susceptibility mapping using artificial neural network for offshore wind farms in Taiwan. *Eng. Geol.* **2025**, *351*, 108013.
3. Zhang, Q.; Miao, C.; Gou, J.; Zheng, H. Spatiotemporal characteristics and forecasting of short-term meteorological drought in China. *J. Hydrol.* **2023**, *624*, 129924.
4. Hu, J.; Miao, C.; Zhang, X.; Kong, D. Retrieval of suspended sediment concentrations using remote sensing and machine learning methods: A case study of the lower Yellow River. *J. Hydrol.* **2023**, *627*, 130369.
5. Sahour, H.; Gholami, V.; Vazifedan, M.; Saeedi, S. Machine learning applications for water-induced soil erosion modeling and mapping. *Soil Tillage Res.* **2021**, *211*, 105032.
6. Chen, W.; Nguyen, K.A.; Lin, B.-S. Rethinking Evaluation Metrics in Hydrological Deep Learning: Insights from Torrent Flow Velocity Prediction. *Sustainability* **2025**, under review.
7. Chen, W.; Nguyen, K.A.; Lin, B.-S. Deep Learning and Optical Flow for River Velocity Estimation: Insights from a Field Case Study. *Sustainability* **2025**, *17*, 8181.
8. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91.
9. Williams, G.P. Friends Don't Let Friends Use Nash–Sutcliffe Efficiency (NSE) or KGE for Hydrologic Model Accuracy Evaluation: A Rant with Data and Suggestions for Better Practice. *Environ. Model. Softw.* **2025**, *106*, 106665.
10. Melsen, L.A.; Puy, A.; Torfs, P.J.J.F.; Saltelli, A. The Rise of the Nash–Sutcliffe Efficiency in Hydrology. *Hydrol. Sci. J.* **2025**, 1–12.
11. Onyutha, C. Pros and Cons of Various Efficiency Criteria for Hydrological Model Performance Evaluation. *Proc. IAHS* **2024**, *385*, 181–187.
12. Lamontagne, J.R.; Barber, C.A.; Vogel, R.M. Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data. *Water Resour. Res.* **2020**, *56*, e2020WR027101.
13. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments Against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250.