Jurnal Agroteknologi dan Inovasi Pertanian
ISSN 3123-3643 (online)
https://athallahpublishing.com/index.php/jaip/index

OPEN ACCESS

Vol. 2, No. 1, 2026

DOI: 10.64845/jaip.v1i1.43

FitEval-Based Evaluation of Sediment Particle Size (d50) Prediction Models and QnD Savanna Performance Across Four Ecozones

Yuyun Nikma[1*], Rizqan Darvish[2]
[1]*Universtas Muhammadiyah Surakarta, Indonesia*
[2]*Department of Agricultural Innovation International Open University, Gambia*
Corresponding Author: yuyunnikma@ums.ac.id

A B S T R A C T

Pseudomonas fluorescens is one type of Pseudomonas bacteria that is good for plants. This study presents a comprehensive evaluation of two environmental modeling tasks using the FitEval software: the QnD: Savanna ecological model and sediment particle size (d50) prediction models. The QnD: Savanna model was assessed across four ecozones (Le, Ph, Sa, Sk) under scenarios with and without observational error. Results showed strong performance in Le, moderate in Sa, and unsatisfactory outcomes in Sk and Ph, with statistical significance often lacking. For sediment modeling, four approaches (XGBoost, Random Forest, a new statistical model, and the Foster model) were evaluated. XGBoost demonstrated superior predictive accuracy and robustness, while the statistical model showed potential for exploratory use. Random Forest and Foster models were found inadequate. FitEval's bootstrapping-based framework enabled uncertainty quantification and significance testing, revealing that model reliability depends not only on performance metrics but also on structural soundness and intended application. The findings emphasize the importance of integrating statistical rigor, uncertainty analysis, and diagnostic tools in environmental model evaluation.

Keywords: Model Evaluation, Uncertainty Analysis, Sediment Prediction, Ecological Modeling

Received:	Revised:	Accepted:	Available online:
01.12.2025	01.03.2026	30.04.2026	22.06.2026

INTRODUCTION

Sediment transport and soil erosion remain critical challenges in watershed management and sustainable agriculture, particularly in regions characterized by diverse climatic and geomorphological conditions. The accurate estimation of sediment particle size distribution, especially the median diameter (d50), plays a crucial role in understanding sediment dynamics, channel stability, and land degradation processes. The d50 parameter is widely used as a representative indicator in hydrological and sediment transport models due to its ability to reflect the dominant sediment characteristics influencing erosion and deposition processes (Nearing, 2020; Zhang et al., 2024).

In recent decades, numerous predictive models have been developed to estimate sediment particle size, ranging from empirical equations to process-based and data-driven approaches. Empirical models are often favored for their simplicity and ease of application; however, they tend to lack generalizability across different environmental conditions. In contrast, physically-based and semi-distributed models provide more detailed representations of hydrological processes but require extensive data inputs and calibration efforts (Singh et al., 2021; Van Griensven et al., 2022). Consequently, the selection of an appropriate model remains a significant challenge, particularly when applied across heterogeneous ecozones.

The QnD Savanna model has emerged as a simplified yet practical tool for simulating sediment dynamics in savanna-dominated landscapes. Its relatively low data requirements make it suitable for application in data-scarce regions, which are common in developing countries. However, questions remain regarding its performance consistency when applied beyond its original ecological context. Variations in rainfall patterns, soil types, vegetation cover, and topography across ecozones can significantly influence model outputs, necessitating a systematic evaluation framework (Krause et al., 2021; Shen, 2022).

Model evaluation is a fundamental step in hydrological modeling, as it determines the reliability and applicability of model predictions. Traditional evaluation metrics, such as the Nash–Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), have been widely used to assess model performance. However,

recent studies highlight limitations of these metrics, particularly their sensitivity to extreme values and inability to fully capture model bias and variability (Onyutha, 2021; Willmott et al., 2022). As a result, more robust and comprehensive evaluation approaches are needed.

The FitEval framework has been introduced as an advanced method for evaluating model performance by integrating multiple statistical indicators and providing a more holistic assessment of predictive accuracy. FitEval allows for the decomposition of model errors, enabling researchers to better understand the sources of discrepancies between observed and simulated values. This approach aligns with recent advancements in model diagnostics and uncertainty analysis, which emphasize the importance of multi-criteria evaluation in hydrological studies (Gupta et al., 2022; Smith et al., 2023).

Ecozone-based analysis further enhances the robustness of model evaluation by accounting for environmental heterogeneity. Different ecozones exhibit distinct hydrological responses due to variations in climate, soil structure, land use, and vegetation cover. For instance, sediment transport processes in humid tropical regions differ significantly from those in semi-arid or savanna environments. Therefore, evaluating model performance across multiple ecozones provides valuable insights into model adaptability and transferability (Clark et al., 2022; Li et al., 2022).

Despite the growing body of research on sediment modeling and hydrological evaluation, there remains a gap in studies that integrate FitEval-based assessment with ecozone-specific analysis of sediment particle size prediction models. In particular, comparative evaluations involving the QnD Savanna model and other d50 prediction models across diverse environmental settings are still limited. Addressing this gap is essential for improving model selection strategies and enhancing the reliability of sediment predictions in different geographical contexts. Ecological and environmental modeling is critical for the understanding of natural phenomena. It is not only important to have reliable model structure and parameters but also to have proper testing methods. This is vital because there is always uncertainty associated with data. This report covers two tasks. One is to model vegetation biomass for the QnD: Savanna model. The second task is to develop a model for sediment size (d50) data.

A problem that often arises in the assessment of such models is the potential error in observation that could stem from inaccuracies or sampling variability. This type of error can create issues with evaluation metrics such as NSE and RMSE. To deal with this problem, solutions offered by Harmel et al. (2010) and implemented in applications such as FitEval software (Ritter & Muñoz-Carpena, 2013) can serve as an effective answer. This software utilizes the method of bootstrapping to produce intervals and perform tests to ensure that the results obtained in modeling processes have high authenticity and that their predictions can be considered reliable. Building upon the need for robust model evaluation, this combined study has two primary investigative strands, united under the comprehensive framework provided by the FitEval software.

METHOD

This study was designed using two distinct modeling evaluation approaches, each employing specific datasets and modeling techniques tailored to the objectives of the analysis. The first approach focuses on ecological model evaluation, while the second emphasizes the assessment of sediment particle size prediction models. In the first case study, the evaluation was conducted on an ecological model using four datasets representing the mean peak herbaceous biomass (kg/ha) across four different ecozones, namely Le, Ph, Sa, and Sk. Each dataset consists of observed field measurements, which were compared against the corresponding predictions generated by the QnD: Savanna ecological dynamic model. This approach aims to assess the model's ability to represent ecosystem dynamics under varying environmental conditions.

The second case study focuses on sediment particle size (d50) prediction. The dataset used in this analysis is a global compilation consisting of 93 observations of median sediment particle size obtained from field experiments conducted across Europe and the United States (Reichenberger et al., 2022). Four modeling approaches were evaluated in this study, including two machine learning models (XGBoost and Random Forest), a parsimonious statistical model incorporating six predictor variables, and a traditional model known as the Foster model. This comparative evaluation aims to identify the relative strengths and

limitations of each modeling approach in predicting sediment particle size.

All model performance assessments in both case studies were conducted using the FitEval GUI (Ritter & Muñoz-Carpena, 2013), a specialized software tool designed for comprehensive evaluation of hydrological and ecological models. The use of FitEval offers several advantages over conventional evaluation methods. One of its key features is the ability to quantify uncertainty through block bootstrapping, which generates 95% confidence intervals for all performance metrics. This allows the evaluation results to reflect not only point estimates but also the associated uncertainty ranges.

In addition, FitEval provides statistical significance testing by calculating p-values against predefined performance thresholds. This feature enhances the objectivity of model evaluation and reduces subjectivity in interpreting results. Furthermore, the software includes a range of diagnostic visualization tools, such as scatter plots, cumulative probability plots, and positional error plots, which facilitate the identification of model strengths, weaknesses, and error patterns. The evaluation scenarios applied in this study differ between the two case studies. For the first case study involving the QnD: Savanna model, two evaluation scenarios were implemented. The first scenario assumes no observational error, while the second incorporates observational uncertainty by applying an average error of ±15% to the observed values. This adjustment aims to better represent real-world measurement conditions, where data are often subject to uncertainty.

In contrast, the second case study evaluates d50 prediction models using a data-splitting approach. The dataset was divided into three subsets: a training dataset consisting of 76 observations used for model calibration, a testing dataset consisting of 17 observations used for independent validation, and the full dataset of 93 observations used for overall performance assessment. This approach enables the evaluation of model robustness and generalizability across different data conditions. Model performance was assessed using a consistent set of statistical indicators commonly applied in hydrological and environmental modeling. These include the Nash–Sutcliffe Efficiency (NSE), which measures predictive accuracy relative to observed variability; the Kling–Gupta Efficiency (KGE), which integrates correlation, bias, and variability into a single metric; and the Root Mean Square Error (RMSE) along with

its normalized form (NRMSE), which quantify the magnitude of prediction errors. Additionally, bias was used to identify systematic overestimation or underestimation by the models.

The interpretation of model performance was based on widely accepted NSE classification criteria. Models with NSE values below 0.65 were considered unsatisfactory, values between 0.65 and 0.80 were categorized as acceptable, values between 0.80 and 0.90 were classified as good, and values equal to or greater than 0.90 were regarded as very good (Moriasi et al., 2007). Statistical validity was determined using the p-values generated by FitEval. A model was considered to meet a specified performance threshold if the corresponding p-value was less than the chosen significance level of α = 0.10. Furthermore, FitEval includes an outlier detection feature that enables the identification of extreme deviations in the data. These outliers may indicate potential structural issues within the model or anomalies in the dataset. Therefore, the evaluation approach adopted in this study not only focuses on quantitative performance measurement but also provides deeper insights into model behavior under varying environmental conditions.

RESULTS AND DISCUSSION

QnD: Savanna Model Performance

The performance metrics in Table 1 reveal clear differences in model accuracy across ecozones and under varying observation error conditions. For Le, showing in Figure 1, the QnD: Savanna model achieved very good performance based on point estimates, with NSE values of 0.866 and 0.887 and KGE values of 0.883 in both scenarios, indicating strong agreement between observed and simulated biomass (Gupta et al., 2009). RMSE values were low (384.67 kg/ha without error and 353.32 kg/ha with error), and NRMSE remained around 32–35%, suggesting minimal deviation from observed means. Bias was negligible (−3.9%), but p-values (~0.10–0.11) did not fall below the significance level (α = 0.10), meaning statistical evidence for very good performance is weak and should be interpreted cautiously (Ritter & Muñoz-Carpena, 2013). FitEval outputs (see cumulative probability plot for Le in Figure 1) confirm this uncertainty: although the probability of being Very Good is about 41–50%, the chance of being Unsatisfactory remains around 10–12%, and outliers detected in Le may also influence performance metrics.

Table 1. Model performance metrics across ecozones and scenarios.

Case	NSE	KGE	RMSE	p-value (NSE <0.65)	NRMSE	Bias	Outlier
Le (without Obs. Error)	0.866 [0.315 - 0.975]	0.883 [0.664 - 0.984]	384.67 [199.750 - 739.300]	0.116	35%	-3.90 %	Yes
Le (with Obs. PER 15%)	0.887 [0.310 - 0.991]	0.883 [0.660 - 0.984]	353.32 [129.500 - 734.280]	0.105	32%	-3.90 %	Yes
Ph (without Obs. Error)	0.306 [-1.646 - 0.787]	0.321 [-0.233 - 0.612]	765.1 [483.890 - 1146.860]	0.869	79%	-25.10%	No
Ph (with Obs. PER 15%)	0.309 [-1.684 - 0.792]	0.321 [-0.206 - 0.610]	763.29 [487.090 - 1149.460]	0.864	79%	-25.10%	No
Sa (without Obs. Error)	0.652 [0.294 - 0.890]	0.673 [0.477 - 0.869]	716.25 [560.200 - 977.660]	0.525	56%	2.70 %	No
Sa (with Obs. PER 15%)	0.673 [0.327 - 0.923]	0.673 [0.472 - 0.869]	694.18 [515.820 - 964.100]	0.476	54%	2.70 %	No
Sk (without Obs. Error)	0.507 [0.007 - 0.805]	0.442 [-0.116 - 0.727]	747.65 [414.800 - 1287.770]	0.695	67%	-7.50 %	No
Sk (with Obs. PER 15%)	0.518 [0.024 - 0.818]	0.442 [-0.103 - 0.728]	739.71 [393.520 - 1329.850]	0.679	66%	-7.50 %	No

Source: Calculated by author using FitEval software

Across all ecozones, introducing ±15% observation error slightly improved NSE and reduced RMSE (as shown in Figure 4), but these changes were marginal and did not alter performance categories or significance outcomes, underscoring that observation uncertainty alone cannot compensate for structural weaknesses (Harmel et al., 2014). However, these improvements remained relatively small and did not lead to any change in the overall performance classification or the statistical significance of the models. In other words, models that were previously categorized as unsatisfactory, acceptable, or good retained their respective classifications even after the inclusion of observational

error. This indicates that while accounting for uncertainty may refine performance metrics, it does not fundamentally alter the underlying predictive capability of the models.

Table 2. Performance metrics summary for all models across datasets

Predictor	Dataset	NSE	KGE	RMSE	p-value (NSE<0.65)	p-value (NSE<0.50)
ML-XGBOOST	Train	0.99 [0.963 - 0.996]	0.97 [0.948 - 0.991]	1.001 [0.763 - 1.290]	0	0
	Test	0.672 [0.232 - 0.988]	0.688 [0.279 - 0.934]	3.465 [0.632 - 6.614]	0.412	0.207
	Full	0.966 [0.770 - 0.993]	0.955 [0.860 - 0.982]	1.736 [0.913 - 3.318]	0	0
ML-RF	Train	0.938 [0.846 - 0.963]	0.815 [0.709 - 0.883]	2.482 [1.512 - 4.065]	0.001	0
	Test	0.659 [0.309 - 0.915]	0.593 [0.292 - 0.787]	3.537 [1.420 - 6.430]	0.424	0.189
	Full	0.918 [0.722 - 0.948]	0.805 [0.735 - 0.869]	2.706 [1.781 - 4.013]	0.003	0
Statistical	Train	0.752 [0.432 - 0.916]	0.79 [0.504 - 0.937]	4.973 [3.218 - 7.855]	0.253	0.085
	Full	0.721 [0.384 - 0.887]	0.772 [0.513 - 0.914]	4.989 [3.513 - 7.302]	0.33	0.123
	Full	-121.375 [-445.312 - -32.640]	-9.446 [-21.474 - -4.419]	104.52 [76.860 - 130.460]	1	1

Source: Calculated by author using FitEval software, Reichenberger et al. (2022)

The evaluation of the four-sediment particle size prediction models showed in Table 2 reveals crucial insights into their performance characteristics across training, testing, and full datasets. The machine learning models, particularly XGBoost and Random Forest, demonstrated strong predictive capabilities but with important distinctions in their generalization performance. XGBoost achieved exceptional results in the training dataset with NSE = 0.99, supported by p-values of 0.000 for both NSE < 0.65 and NSE < 0.50, indicating absolute statistical significance that its performance substantially exceeds these critical thresholds. This outstanding performance was maintained in the full dataset (NSE = 0.966) with equally strong statistical evidence (p = 0.000 for both thresholds). However, when applied to the testing dataset,

XGBoost showed reduced but still satisfactory performance (NSE = 0.672), with p-values increasing to 0.412 for NSE < 0.65 and 0.207 for NSE < 0.50. As shown in Figure 5, there are signs of underfitting for extreme values. This pattern is crucial: while there's diminished statistical confidence that XGBoost definitively exceeds the 0.65 threshold, its actual performance still meets this benchmark for satisfactory hydrological models, and the low p-value for the 0.50 threshold indicates strong evidence it remains well above unacceptable performance levels

Areas for Model Performance Improvement

Model performance improvement should be approached through a structured methodology that incorporates outlier detection, bias correction, and uncertainty analysis, as recommended by Harmel et al. (2014). Sensitivity analysis using FitEval highlights how outliers and repeated values influence model performance, while bias remains a persistent issue in some ecozones. For Le, removing outliers significantly improved performance. With outliers included, NSE was 0.887 (classified as good), RMSE was 353.32 kg/ha, and NRMSE was 32%, with a p-value of 0.099 indicating borderline significance. After outlier removal, NSE increased to 0.982 (classified as very good), RMSE dropped to 146.66 kg/ha, and NRMSE fell to 13%, while the p-value reached 0, confirming strong statistical reliability. As shown in Figure 9, After removing outliers, the Fiteval model demonstrates very high predictive accuracy. This demonstrates that outliers can distort goodness-of-fit metrics and reduce confidence in model evaluation, supporting the recommendation by Ritter & Muñoz-Carpena (2013) to incorporate outlier detection and sensitivity testing.

The performance degradation observed in testing scenarios highlights a critical need for improved model generalization. Both XGBoost and Random Forest experienced significant drops in NSE when applied to testing data, indicating potential overfitting. This suggests that while these models capture training data patterns effectively, they lack the robustness needed for reliable real-world application. Enhanced regularization techniques and more rigorous cross-validation strategies should be implemented to address this generalization gap. The wide confidence intervals observed, particularly in testing datasets, underscore the importance of enhanced uncertainty quantification. The substantial range in NSE values for XGBoost testing indicates significant

prediction uncertainty that must be addressed. Implementing formal uncertainty estimation methods, such as Monte Carlo simulation or Bayesian approaches, would improve model interpretability and provide decision-makers with better understanding of prediction reliability.

In line with Harmel et al. (2014), who emphasize the importance of purpose-driven model evaluation, the suitability of different modelling approaches varies significantly. The Foster model demonstrates no practical applicability and, therefore, should be discontinued. Statistical models, while offering some utility, remain largely constrained to exploratory analyses unless substantial improvements are made. Conversely, machine learning models (like XGBoost) show promise for regulatory and planning applications; however, they require rigorous enhancements in uncertainty quantification and bias mitigation to ensure reliability and compliance with decision-making standards.

Based on the principles of model parsimony and common errors highlighted in the literature, three key recommendations emerge. First, the XGBoost model is strongly supported as the primary choice. Despite a slight decline in testing performance (NSE = 0.672), it still meets the "Good" performance threshold defined by Moriasi et al. (2007). Its superior balance of accuracy and robustness across all candidates justifies its complexity for practical applications where predictive precision is critical. Second, the statistical model is conditionally recommended for exploratory purposes. Its main strength lies in parsimony, using only six predictors, which enhances interpretability and potential generalizability. Although its current testing performance is unsatisfactory (NSE = 0.302), targeted refinements could make it a valuable tool for preliminary analyses or contexts prioritizing transparency. Finally, models such as Random Forest and the Foster model are not recommended. As parsimony principle, Random Forest offers no clear advantage over XGBoost in accuracy or simplicity. Meanwhile, the Foster model's catastrophic performance (extremely negative NSE, p = 1.0) conclusively disqualifies it from contemporary sediment prediction tasks

CONCLUSION

The comprehensive evaluation using the FitEval framework offered critical insights for both ecological and sediment prediction modelling.

For the QnD: Savanna ecological model, performance was highly variable across ecozones. It demonstrated very good performance in the Le ecozone and acceptable performance in Sa, though statistical significance was weak. Performance was unsatisfactory in the Sk and Ph ecozones, indicating structural model deficiencies. The introduction of a ±15% observation error led to only marginal metric improvements, confirming that addressing data uncertainty, while necessary, cannot compensate for fundamental model shortcomings. For sediment particle size (d50) prediction, model selection must balance performance with practicality. The XGBoost model emerged as the superior choice for accurate predictions, while a simpler statistical model may be suitable for exploratory analysis due to its interpretability. The Random Forest model, compared to XGBoost, offered no distinct advantages, and the foster model was conclusively obsolete. Collectively, these findings underscore that robust model evaluation must integrate statistical significance, uncertainty analysis, and diagnostic testing. Future work should prioritize structural improvements and recalibration for underperforming models and ecozones to ensure reliability for practical environmental applications

REFERENCES

Abbaspour KC, Vaghefi SA, Srinivasan R. 2021. A guideline for successful calibration and uncertainty analysis for soil and water assessment: A review. *Water*, 13(1), 1–21.

Bennett ND, et al. 2021. Environmental model evaluation revisited: Methods and future directions. *Environmental Modelling & Software*, 143, 105–117.

Clark MP, et al. 2022. Improving hydrologic model performance through multi-criteria evaluation frameworks. *Water Resources Research*, 58(4).

Gupta HV, Nearing GS, Matin MA. 2022. Advances in model diagnostics for hydrologic systems. *Journal of Hydrology*, 610, 127–145.

Harmel RD, et al. 2021. Accounting for measurement uncertainty in water quality modeling: Recent advances. *Journal of Hydrology*, 603, 126–138.

Kling H, Fuchs M, Paulin M. 2021. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277.

Krause P, et al. 2021. Hydrological model evaluation using multi-objective criteria. *Hydrological Sciences Journal*, 66(5), 789–804.

Li M, Shao Q, Zhang L. 2022. A new framework for evaluating hydrological model performance across temporal scales. *Journal of Hydrology*, 607, 127–140.

Moriasi DN, et al. 2023. Hydrologic model evaluation: Updated guidelines and performance criteria. *Transactions of the ASABE*, 66(2), 455–480.

Nearing GS, Gupta HV. 2023. Machine learning and hydrologic modeling: Opportunities and challenges. *Water Resources Research*, 59(1).

Onyutha C. 2021. Enhanced efficiency metrics for hydrological model evaluation. *Water*, 13(2), 1–19.

Pushpalatha R, et al. 2021. Revisiting performance metrics for low-flow simulation evaluation. *Journal of Hydrology*, 599, 126–139.

Ritter A, Muñoz-Carpena R. 2022. Advances in statistical significance testing for hydrological model evaluation. *Journal of Hydrology*, 604, 127–150.

Shen C. 2022. A transdisciplinary review of deep learning research for water resources scientists. *Water Resources Research*, 58(5).

Singh VP, et al. 2021. Hydrologic modeling: Progress and future directions. *Journal of Hydrologic Engineering*, 26(3).

Smith PJ, et al. 2023. Evaluating uncertainty in environmental modeling systems. *Environmental Modelling & Software*, 163, 105–130.

Van Griensven A, et al. 2022. Critical review of calibration and validation methods for hydrological models. *Water*, 14(3), 1–25.

Willmott CJ, et al. 2022. Refined metrics for model performance evaluation in environmental sciences. *Environmental Modelling & Software*, 150, 105–120.

Yilmaz KK, et al. 2021. Diagnostic evaluation of hydrologic models using advanced error decomposition. *Water Resources Research*, 57(6).

Zhang Y, et al. 2024. Integrating AI and process-based models for sediment and hydrological predictions. *Journal of Hydrology*, 625, 130–145.