

Random Cross Validation Overstates the Advantage of Machine Learning Over Stage Discharge Rating Curves in Non-Stationary Alluvial Rivers in Northwest Bangladesh

Md Abu Bakar Siddik Farhan^{1*}; Tanvir Alam²; Md Gazi Jamal Uddin³; Shilpi Mondal⁴

^{1,2,3,4}Department of Civil Engineering, Northern University of Business & Technology, Khulna, Bangladesh

Corresponding Author: Md Abu Bakar Siddik Farhan^{1*}

Publication Date: 2026/10/06

Abstract: Machine learning (ML) models are increasingly reported to estimate river discharge from stage more accurately than the conventional power-law rating curve. Most of these comparisons, however, rely on randomly split data, which mixes old and new gaugings in training and testing and ignores the fact that rating relations drift as river channels change. We compared a power-law rating curve (RC) with Random Forest, XGBoost, support vector regression and a neural network using 14,317 discharge gaugings (1969-2022) from 16 Bangladesh Water Development Board stations on 11 rivers in northwest Bangladesh. ML models were trained with stage alone and with additional hysteresis and seasonal inputs derived from daily water levels. Each model was evaluated under random five-fold cross-validation and under a chronological hold-out in which the most recent 20% of gaugings at each station were withheld. Under random cross-validation, the best ML model (Random Forest with hysteresis inputs) reached a median Nash–Sutcliffe efficiency (NSE) of 0.77 against 0.72 for the RC. Under chronological testing, this advantage disappeared: the RC achieved a median NSE of 0.60, compared with 0.49 to 0.59 for the ML models, and no ML model differed significantly from the RC after Holm correction. Skill fell significantly from random to chronological testing for the RC and five of the six ML models (Wilcoxon $p \leq 0.002$). Pettitt tests detected significant change points in rating residuals at 15 of 16 stations, confirming widespread non-stationarity. ML retained some practical value, with better low-flow accuracy and smaller bias, and it clearly outperformed the RC at three stations. We conclude that validation design influences conclusions more than algorithm choice, and that ML rating models should be tested on future gaugings before operational use.

Keywords: Rating Curve; Stage-Discharge Relationship; Machine Learning; Chronological Validation; Non-Stationarity; Bangladesh.

How to Cite: Md Abu Bakar Siddik Farhan; Tanvir Alam; Md Gazi Jamal Uddin; Shilpi Mondal (2026) Random Cross Validation Overstates the Advantage of Machine Learning Over Stage Discharge Rating Curves in Non-Stationary Alluvial Rivers in Northwest Bangladesh. *International Journal of Innovative Science and Research Technology*, 11(9), 3021-3033. <https://doi.org/10.38124/ijisrt/26sep1313>

I. INTRODUCTION

Continuous discharge records are rarely measured directly. At most hydrometric stations, discharge is gauged only occasionally, and the continuous series is computed from stage records through a stage–discharge rating curve, usually a power law of the form $Q = a(h - h_0)^b$ (Rantz, 1982). The quality of every downstream application, from flood frequency analysis to water allocation and model calibration, therefore depends on how well the rating curve represents the true stage–discharge relation. This relation is seldom fixed. Scour and deposition during floods, vegetation growth, and changes in the hydraulic control shift the curve over time, while unsteady flow

produces hysteresis in which rising and falling stages carry different discharges at the same stage (Kiang et al., 2018; Perret et al., 2022).

Hydrometric research has responded to these problems mainly through statistical and hydraulic approaches. Bayesian frameworks combine hydraulic knowledge with uncertain gaugings to estimate rating curves and their uncertainty (Le Coz et al., 2014). Mansanarez et al. (2019) extended this idea to a stage–period–discharge model that re-estimates the curve across stability periods separated by known morphological changes, and Darienzo et al. (2021) proposed a recursive segmentation procedure that detects rating shifts directly from

gaugings while accounting for observational and model uncertainty. Perret et al. (2022) developed a framework to detect hysteresis caused by flow unsteadiness and applied it across France's national hydrometric network, and Horner et al. (2022) showed that limited sensitivity of hydraulic controls is itself a source of streamflow uncertainty. A common thread in this literature is that the stage–discharge relation must be treated as time-varying.

In parallel, a large number of studies have applied ML to the same problem. He et al. (2023) reported that ensemble Random Forest and gradient-boosting models outperformed the conventional rating curve at a sluice-controlled station in China, with a coefficient of determination of 0.963. Vishwakarma et al. (2023) found that hybrid regression models outperformed the rating curve at the Gaula Barrage in India, and Kisi et al. (2024) reported a coefficient of determination of 0.9998 for a multilayer neural network on the Kizilirmak River in Turkey. Similar conclusions have been drawn with neural networks, decision trees, support vector machines and multi-point conceptual structures (Kumar et al., 2020; Alizadeh et al., 2021; Chen et al., 2023). In Bangladesh, ML has been used to forecast annual discharge at eight stations in Nilphamari district (Saleh et al., 2024) and weekly discharge of the Brahmaputra–Jamuna at Bahadurabad, where ensemble models reached NSE values of 0.85 to 0.92 (Rahim et al., 2024).

These results conflict with the hydrometric evidence that rating relations shift over time. Rozos et al. (2022) observed that ML rating studies, although motivated by the time-dependent nature of rating curves, generally handle the data as if they came from a stationary process. When gaugings from different years are randomly allocated to training and test sets, a model can learn the shifted relation of a given period from neighbouring gaugings in the same period, then be scored on gaugings it has effectively already seen. This form of temporal leakage is a recognised cause of inflated performance across ML-based science (Kapoor and Narayanan, 2023), and the choice of data split is known to change streamflow model scores (Puri et al., 2024). Yet in operation, a rating model is always fitted to past gaugings and applied to future water levels. A comparison that does not reproduce this situation may overstate the value of ML precisely at stations where channels are most unstable.

Northwest Bangladesh provides a demanding test case. The region's alluvial rivers, including the Atrai, Little Jamuna, Bangali, Mahananda, Punarbhaba and Baral, carry monsoon floods and heavy sediment loads, and the Bangladesh Water Development Board (BWDB) has maintained gauging programmes at many of them for more than four decades. Published ML applications in Bangladesh have addressed discharge forecasting at single stations or within one district (Saleh et al., 2024; Rahim et al., 2024); none has compared ML and rating-curve approaches across a regional station network, or examined how the conclusion depends on validation design. This study has three objectives: (i) to compare the accuracy of ML models and the power-law rating curve in estimating discharge from stage across a regional station network; (ii) to test whether inputs describing hysteresis and season improve ML estimates over stage alone; and (iii) to examine how the comparison changes when models are evaluated on future gaugings rather than randomly withheld ones. These objectives are addressed with 14,317 gaugings from 16 stations, seven model configurations, two validation designs, and a set of non-parametric tests that check whether the observed differences are statistically meaningful.

II. STUDY AREA AND DATA

➤ Study Area

The study area lies in northwest Bangladesh, within Rajshahi Division (approximately 24.0–25.2°N, 88.1–89.6°E; Fig. 1). The stations lie on 11 rivers: the Ganges–Padma, Mahananda, Punarbhaba, Pagla, Atrai (Karatoa–Atrai–Gur–Gumani–Hurasagar system), Little Jamuna, Tulshiganga, Bangali, Karatoa and the Baral and Fakirni Barnai distributaries. Rainfall is strongly monsoonal: about 80% of the annual total at the region's 44 BWDB rain gauges falls between June and October. Stages at most stations vary by 6 to 12 m between dry-season lows and monsoon peaks (Table 1). Channels are alluvial and mobile, which makes rating shifts likely.

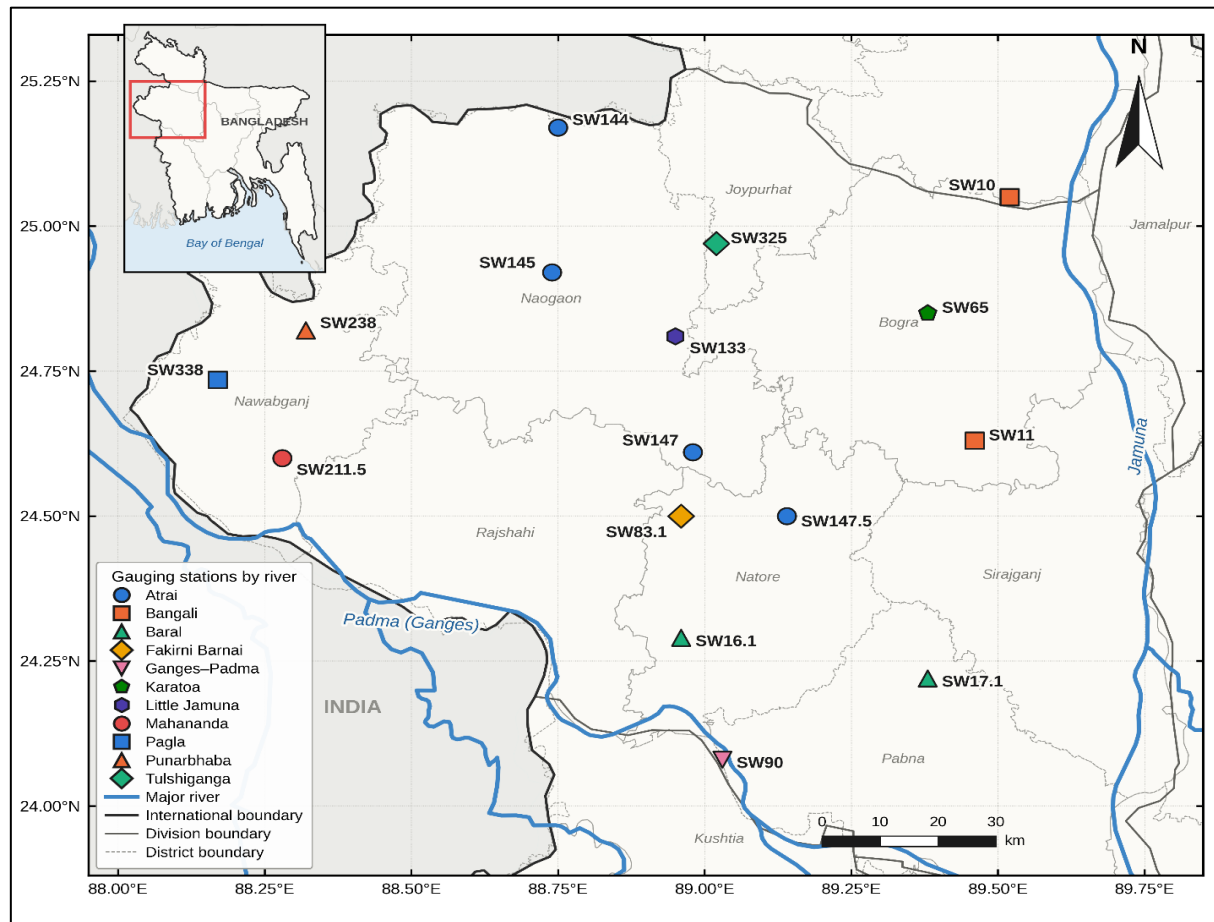


Fig 1 Locations of the 16 BWDB Gauging Stations of River.

➤ Data

Two BWDB datasets were used. The first contains field discharge gaugings, each recording date, stage (m above the PWD datum), discharge ($\text{m}^3 \text{s}^{-1}$), cross-sectional area and velocity. The second contains daily maximum, minimum and mean water levels for the same stations. The raw gauging file held 17,780 measurements at 25 stations. BWDB reports that its reference datum changed on 1 November 2022, with station-specific offsets between 0.01 and 1.13 m. Because the offset was applied inconsistently between the gauging and daily series (it appears in both at Hardinge Bridge but only in the daily

series elsewhere), all records after 31 October 2022 were excluded to avoid introducing an artificial shift.

After quality control (Section 3.1), 20 stations retained at least 200 valid gaugings. Four of these had no overlapping daily water level record, so the final modelling set comprised 16 stations and 14,317 gaugings (Table 1). Record lengths range from 13 years (Cansart) to 54 years (Hardinge Bridge). Hardinge Bridge, on the Ganges, contributes 6,270 gaugings and discharges up to $76,000 \text{ m}^3 \text{s}^{-1}$; the other stations are smaller rivers with 211 to 888 gaugings and maximum discharges between 152 and $3,660 \text{ m}^3 \text{s}^{-1}$.

Table 1 Stations and Gauging Records Used in the Analysis (after Quality Control, 1969–2022).

ID	Station	River	N	Period	Stage range (m)	Median Q (m^3/s)	Max Q (m^3/s)
SW10	Simulbari	Bangali	679	1983–2022	11.84–17.90	91.5	762
SW11	Khanpur	Bangali	673	1980–2022	5.23–14.65	66.8	936
SW133	Naogaon	Little Jamuna	532	1981–2022	5.73–15.37	68.6	547
SW144	Chak Hariharpur	Atrai	562	1989–2022	15.08–22.39	58.6	1,453
SW145	Mohadebbpur	Atrai	660	1986–2022	10.78–18.67	50.9	1,530
SW147	Atrai Rly. Bridge	Atrai	599	1980–2022	6.25–13.84	126.3	904
SW147.5	Singra	Atrai	532	1981–2022	5.07–13.50	144.4	629
SW16.1	Malanchi	Baral	804	1978–2022	7.58–14.62	17.3	812
SW17.1	Baral Rly. Bridge	Baral	522	1994–2022	2.26–11.94	60.9	338
SW211.5	Chapai Nawabganj	Mahananda	888	1981–2022	10.27–22.00	143.5	3,660

SW238	Rohanpur	Punarbhaba	365	1994–2022	10.89–23.49	98.5	748
SW325	Sonaimukhi	Tulshiganga	358	1994–2022	7.66–17.03	9.5	274
SW338	Cansart	Pagla	211	2010–2022	15.03–21.72	20.8	180
SW65	Bogra	Karatoa	394	1994–2022	10.20–15.55	7.7	152
SW83.1	Naldanga	Fakirni Barnai	268	1994–2022	5.34–13.59	69.0	252
SW90	Hardinge Bridge	Ganges–Padma	6,270	1969–2022	3.27–14.68	1,534	76,000

N = Number of Gaugings Retained. Stage is Referred to the PWD Datum.

III. METHODS

➤ Methodological Approach

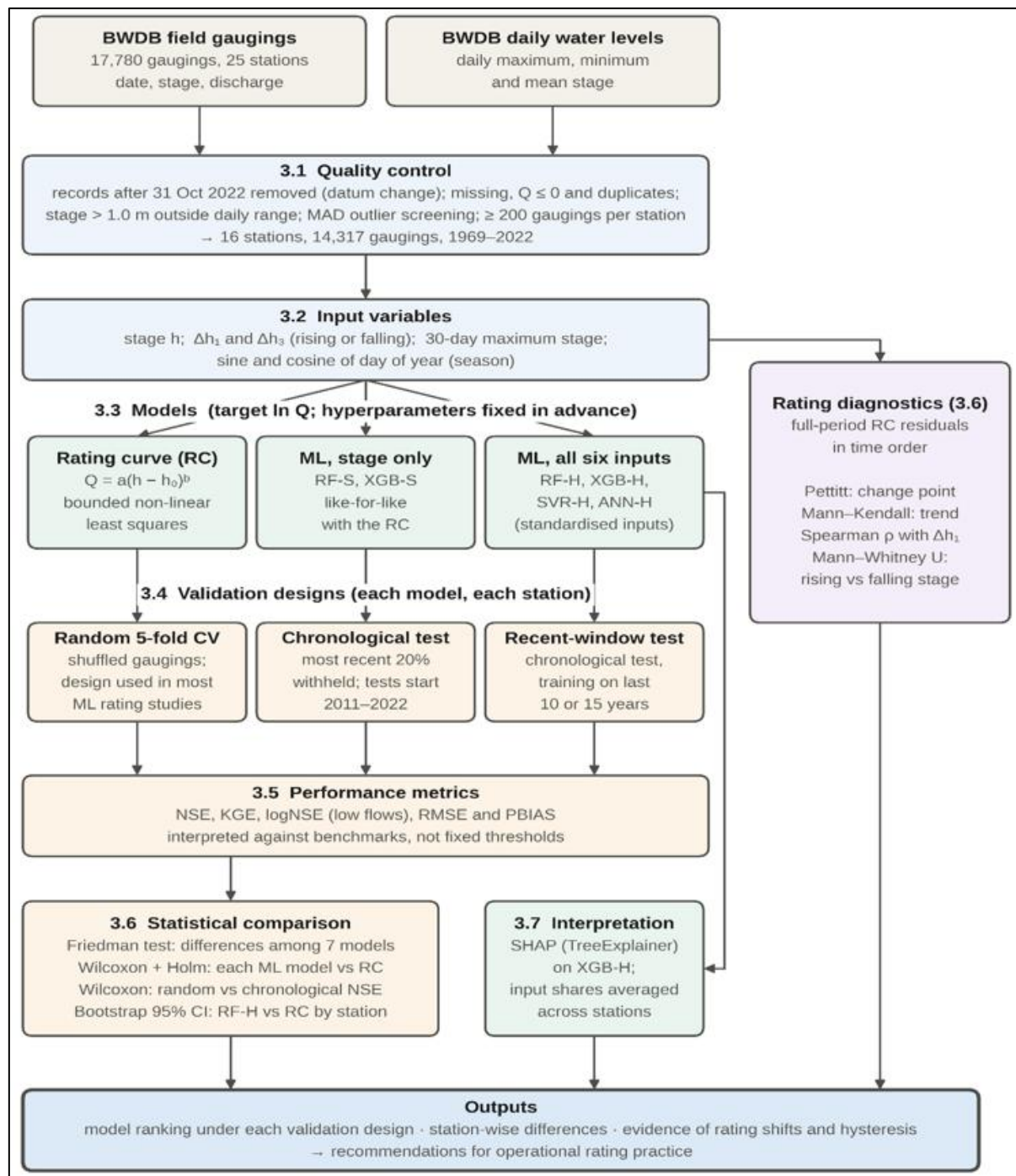


Fig 2 Methodological Framework of the Study, from Data Quality Control and Model Development to Validation, Statistical Comparison and Interpretation.

➤ Quality Control

Gaugings with missing stage or discharge, non-positive discharge, or exact duplication were removed. The gauging file also contained typing errors in stage (for example, 18,435.04 m at Chak Hariharpur and 120.17 m at Singra), so any gauging whose stage fell more than 1 m outside the station's recorded daily water level range was removed (44 gaugings). Remaining outliers were flagged by binning each station's gaugings into 15 stage classes and removing points whose log-discharge departed from the class median by more than 3.5 times the scaled median absolute deviation (306 gaugings). Finally, stations with fewer than 200 valid gaugings were excluded.

➤ Input Variables

For each gauging, five variables were derived from the daily mean water level series in addition to stage h : the one-day stage change Δh_1 (m day^{-1}) and three-day stage change Δh_3 , which indicate whether the river was rising or falling; the maximum stage over the preceding 30 days, which represents recent flood history; and the sine and cosine of day of year, which represent season. Stage recorded at the gauging matched the daily mean stage with a median absolute difference of 0.00 m (95th percentile m), indicating that the two series share the same gauging record.

➤ Models

The benchmark was a single power-law rating curve fitted separately to each station:

$$Q = a(h - h_0)^b \quad (1)$$

Where h_0 is the stage of zero flow, estimated jointly with a and b by bounded non-linear least squares on log-discharge (h_0 constrained below the minimum observed stage, $0.3 \leq b \leq 5$). Six ML configurations were compared with it. Random Forest (RF; Breiman, 2001) and XGBoost (XGB; Chen and Guestrin, 2016) were each trained with stage alone (RF-S, XGB-S) for a like-for-like comparison with the rating curve, and with all six inputs (RF-H, XGB-H). Support vector regression (SVR-H) and a two-layer multilayer perceptron (ANN-H; 16 and 8 neurons, L2 penalty 0.1) were trained with all six inputs after standardisation. All models predicted $\ln Q$, and predictions were back-transformed by exponentiation. Hyperparameters were fixed in advance at conventional values (RF: 500 trees, minimum leaf size 3; XGB: 400 trees, learning rate 0.03, maximum depth 3 or 4, subsampling 0.8) rather than tuned on test data, so that the comparison reflects typical practice.

➤ Validation Designs

Each model was evaluated at each station under two designs. In random five-fold cross-validation, gaugings were shuffled and split into five folds, each fold predicted by a model trained on the other four. This mirrors the design used in most published ML rating studies. In the chronological hold-out, gaugings were sorted by date, and the most recent 20% were withheld for testing. Test periods began between 2011 and 2022 depending on the station. A third experiment repeated the chronological test with training restricted to the 10 or 15 years immediately preceding the test period, to examine whether recent data alone can track rating shifts.

➤ Performance Metrics

Performance was measured with the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), the Kling–Gupta efficiency (KGE; Gupta et al., 2009), the NSE of log-transformed discharge (logNSE), which gives more weight to low flows, the root mean square error (RMSE), and percent bias, $\text{PBIAS} = 100 \times \Sigma(Q_{\text{sim}} - Q_{\text{obs}}) / \Sigma Q_{\text{obs}}$, for which negative values indicate underestimation. Following Knoben et al. (2019), we interpret NSE and KGE values relative to benchmarks rather than against fixed thresholds, and we report both because they respond differently to bias and variability errors.

➤ Statistical Analysis

Because performance scores across stations are not normally distributed and the same stations are used by every model, non-parametric paired tests were applied, following the recommendations of Demšar (2006) for comparing several models across multiple datasets. The Friedman test (Friedman, 1937) assessed whether the seven models differed in NSE within each design. Each ML model was then compared with the rating curve using the Wilcoxon signed-rank test (Wilcoxon, 1945), with p-values adjusted by the Holm (1979) procedure for six comparisons. The Wilcoxon test was also used to compare each model's NSE between the random and chronological designs.

To check the physical assumptions behind the analysis, a single full-period rating curve was fitted at each station and its log residuals were examined in time order. The Pettitt (1979) test identified the most likely change point, and the Mann–Kendall test (Mann, 1945; Kendall, 1975) tested for monotonic trend. Hysteresis was examined by correlating residuals with Δh_1 (Spearman) and by comparing residuals on rising ($\Delta h_1 > 0.02 \text{ m day}^{-1}$) and falling ($\Delta h_1 < -0.02 \text{ m day}^{-1}$) stages with the Mann–Whitney U test (Mann and Whitney, 1947). Finally, station-wise 95% confidence intervals for the difference in chronological NSE between RF-H and the rating curve were obtained from 1,000 bootstrap resamples of the test gaugings (Efron and Tibshirani, 1993). A significance level of 0.05 was used throughout.

➤ Model Interpretation

The contribution of each input to the XGB-H estimates was quantified with SHAP values computed by the TreeExplainer algorithm (Lundberg et al., 2020). For each station, mean absolute SHAP values were normalised to sum to one, and shares were then averaged across stations. All analyses were carried out in Python (scikit-learn, XGBoost, SHAP and SciPy), and the code is available from the corresponding author.

IV. RESULTS

➤ Non-Stationarity of the Rating Relations

Visual inspection shows that the stage–discharge relation has moved at many stations (Fig. 3). At Baral Rly. Bridge and Naogaon, and at Baghabari and Ullapara (shown in Fig. 2 but excluded from modelling because they lack daily water level records), gaugings from the 2010s and 2020s plot in a different part of the stage–discharge space from those of the 1980s and

1990s, which is consistent with bed aggradation or degradation. The statistical tests confirm this impression (Table 2). The Pettitt test found a significant change point in rating residuals at 15 of the 16 stations, and the Mann–Kendall test found a significant monotonic trend at 12.

Most change points fall between 1992 and 2016. The only station without a significant change point, Singra, still showed a clear hysteresis signal.

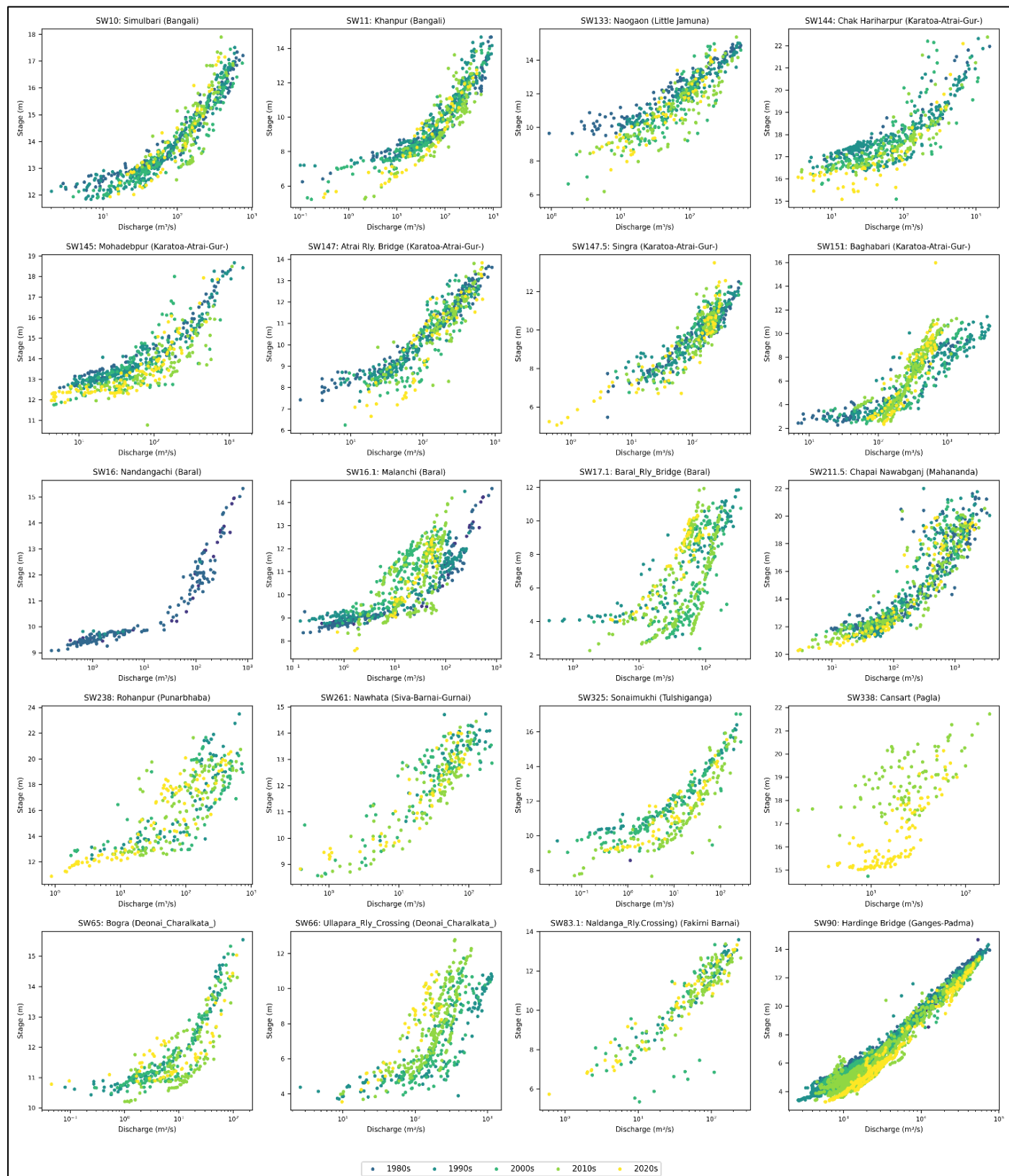


Fig 3 Stage-Discharge Gaugings at the 20 Quality-Controlled Stations, Coloured by Decade; Stations Marked * were Not Modelled.

Evidence for hysteresis at the daily scale was weaker. Residuals on rising stages exceeded those on falling stages significantly at only three stations (Simulbari, Singra and

Naldanga), and at Chapai Nawabganj the difference ran in the opposite direction. At most stations, the effect of stage change was small compared with the long-term drift.

Table 2 Tests of Non-Stationarity and Hysteresis in Log Residuals of a Single Full-Period Rating Curve.

Station	Pettitt change year	Pettitt p	Mann–Kendall τ	Median residual rising / falling	Mann–Whitney p
SW10 Simulbari	1992	<0.001	0.08*	0.047 / –0.022	0.038
SW11 Khanpur	2006	<0.001	0.17*	0.074 / 0.002	0.157
SW133 Naogaon	1997	<0.001	0.24*	0.138 / 0.019	0.145
SW144 Chak Hariharpur	1997	<0.001	0.22*	0.281 / 0.290	0.811
SW145 Mohadebpur	2004	<0.001	0.28*	0.258 / 0.158	0.523
SW147 Atrai Rly. Bridge	2006	<0.001	0.20*	0.084 / 0.001	0.161
SW147.5 Singra	1995	0.067	–0.03	0.127 / –0.019	0.021
SW16.1 Malanchi	2010	<0.001	–0.01	0.258 / 0.127	0.527
SW17.1 Baral Rly. Bridge	2015	<0.001	–0.14*	0.063 / –0.004	0.538
SW211.5 Chapai Nawabganj	1985	<0.001	0.04	–0.022 / 0.165	<0.001
SW238 Rohanpur	2016	<0.001	–0.11*	–0.091 / 0.168	0.072
SW325 Sonaimukhi	2008	<0.001	0.29*	0.050 / 0.086	0.936
SW338 Cansart	2021	<0.001	0.20*	0.246 / 0.197	0.577
SW65 Bogra	2006	<0.001	0.24*	–0.043 / 0.026	0.243
SW83.1 Naldanga	2006	0.002	–0.01	0.109 / –0.059	<0.001
SW90 Hardinge Bridge	2011	<0.001	0.42*	0.002 / –0.020	0.468

* Mann–Kendall Trend Significant at $p < 0.05$. Residuals ($\ln Q_{\text{obs}} - \ln Q_{\text{fit}}$) are in Natural-Log Units Rising and Falling Stages Defined by $\Delta h_1 > 0.02$ and < -0.02 m Day⁻¹.

➤ Random Versus Chronological Validation

Under random five-fold cross-validation, every tree-based ML model scored a higher median NSE than the rating curve (Table 3, Fig. 4). RF-H performed best, with a median NSE of 0.771 against 0.717 for the rating curve, and it also had the best mean Friedman rank (2.00 of 7). The Friedman test indicated significant differences among models ($\chi^2 = 30.00$, $p < 0.001$). This ranking matches published comparisons (He et al., 2023; Kisi et al., 2024).

Under chronological testing, the rating curve achieved the highest median NSE (0.598), and the ML models ranged from 0.490 (SVR-H) to 0.586 (XGB-S). Differences among models remained significant (Friedman $\chi^2 = 15.19$, $p = 0.019$), but after Holm correction no ML model differed significantly from the rating curve in either design (all adjusted $p \geq 0.27$; Table 4). Median KGE was less decisive: the rating curve (0.656) sat within the ML range (0.614 to 0.704), with XGB-H and ANN-H slightly higher.

Table 3 Median Performance Across the 16 Stations Under the Two Validation Designs.

Model	NSE random	NSE chrono	KGE random	KGE chrono	logNSE random	logNSE chrono	PBIAS chrono (%)
RC	0.717	0.598	0.752	0.656	0.708	0.567	–8.7
RF-S	0.742	0.506	0.797	0.653	0.714	0.593	–5.7
XGB-S	0.756	0.586	0.781	0.646	0.721	0.601	–4.9
RF-H	0.771	0.515	0.797	0.646	0.817	0.691	2.0
XGB-H	0.744	0.521	0.766	0.671	0.811	0.671	3.1
SVR-H	0.729	0.490	0.738	0.614	0.805	0.690	2.6
ANN-H	0.692	0.519	0.795	0.704	0.806	0.672	7.1

Bold marks the best value in each column (for PBIAS, the value closest to zero). -S = stage only; -H = stage plus hysteresis and seasonal inputs.

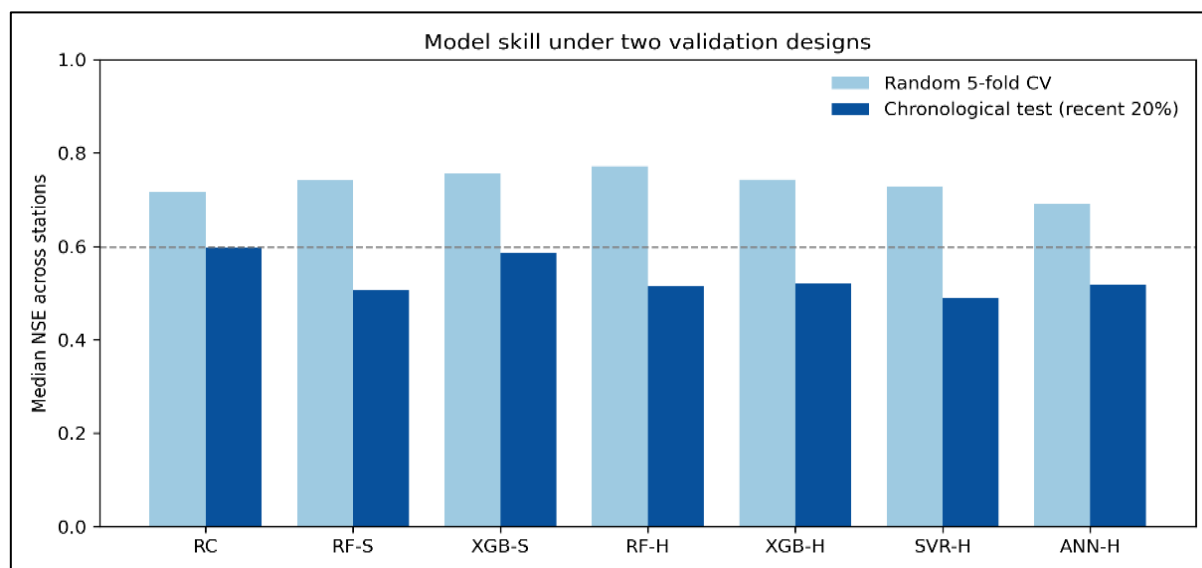


Fig 4 Median NSE Across Stations for Each Model Under Random Five-Fold Cross-Validation and the Chronological hold-out. The Dashed Line Marks the Rating Curve's Chronological Score.

Table 4 Wilcoxon Tests of each ML Model against the Rating Curve (NSE, Holm-Adjusted) and Effect of Validation Design.

Model	Median Δ NSE vs RC (random)	p_Holm (random)	Median Δ NSE vs RC (chrono)	p_Holm (chrono)	Median NSE drop, random chrono	p (design)
RC	–	–	–	–	0.080	<0.001
RF-S	–0.007	1.00	–0.026	1.00	0.176	0.001
XGB-S	0.007	1.00	–0.007	1.00	0.169	0.002
RF-H	0.015	0.47	–0.023	1.00	0.185	<0.001
XGB-H	–0.001	1.00	–0.042	1.00	0.191	<0.001
SVR-H	–0.015	1.00	–0.074	0.56	0.286	<0.001
ANN-H	–0.039	0.27	–0.098	0.56	0.149	0.376

Δ NSE is the paired station-wise difference (model minus RC); positive values favour the ML model. p (design) is the Wilcoxon test of random versus chronological NSE for the same model. A median paired difference can be negative even where the model's median NSE is higher, because the median of differences need not equal the difference of medians.

Skill declined for every model when moving from random to chronological testing, and the decline was significant for the rating curve and five of the six ML models (Wilcoxon $p \leq 0.002$; Table 4). The rating curve lost a median of 0.080 in NSE, while the ML models lost between 0.149 and 0.286. The difference in decline between each ML model and the rating curve was not itself significant ($p = 0.23$ to 0.82), so the evidence supports a general optimism of random validation rather than a demonstrably larger penalty for ML. The ranking of methods nevertheless changed with the design, and it is this ranking that decides which method is used in practice.

➤ Station-Wise Performance and Bootstrap Comparison

Station-wise results varied widely (Fig. 5). Hardinge Bridge was estimated almost perfectly by all methods (NSE 0.93 to 0.96), and the Atrai Rly. Bridge, Bogra and Naldanga stations reached NSE values of 0.7 to 0.8. At the other extreme, every method failed at Baral Rly. Bridge (NSE between -2.9 and -5.7), where the gaugings form two separate branches (Fig. 3), and all ML models failed at Cansart, which has the shortest record. The bootstrap comparison (Supplementary Table S1) found that RF-H was significantly better than the rating curve at three stations (Simulbari, Khanpur and Hardinge Bridge), significantly worse at five (Malanchi, Baral Rly. Bridge, Chapai Nawabganj, Rohanpur and Cansart), and not distinguishable at the remaining eight.

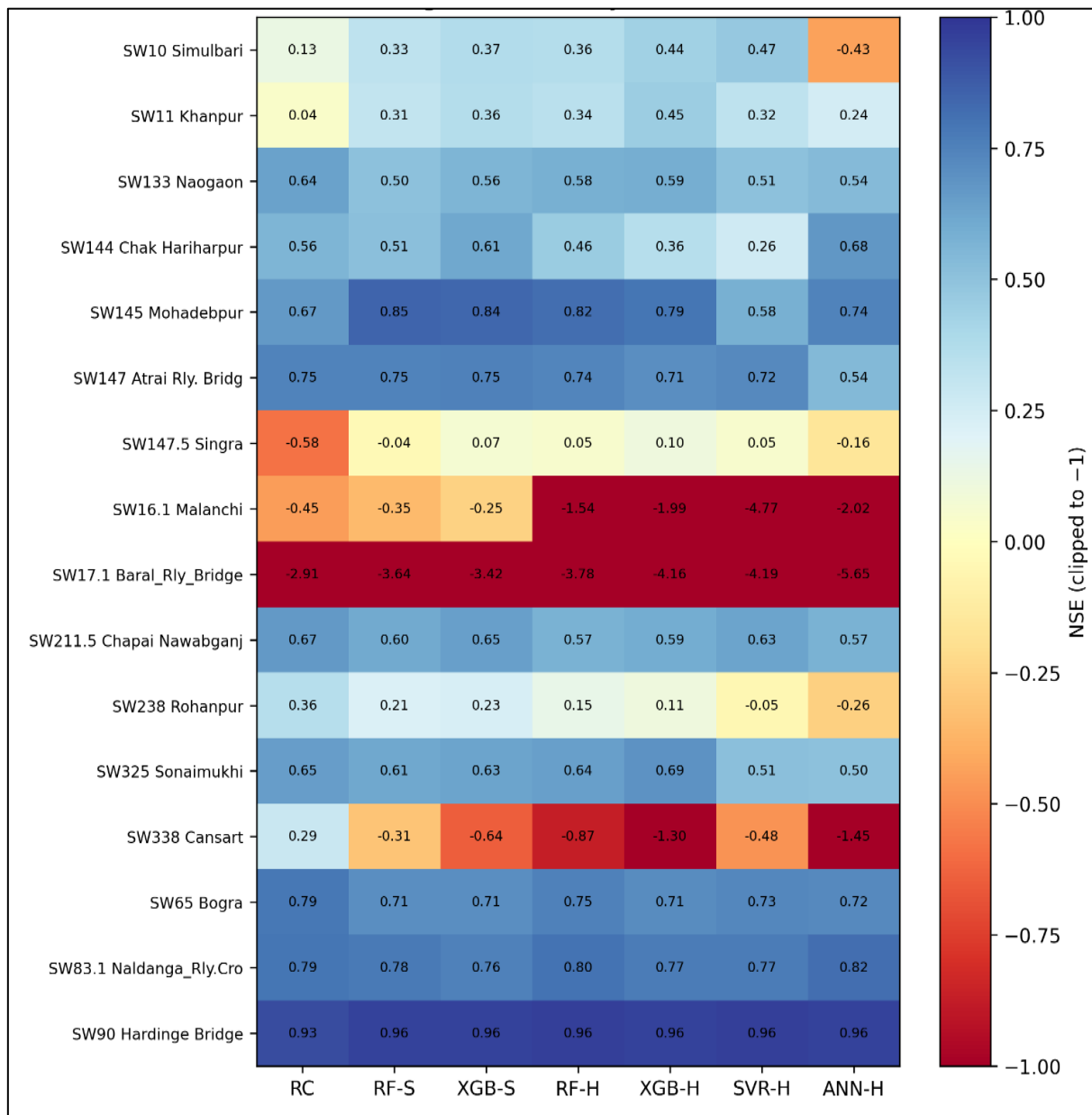


Fig 5 Station-Wise NSE for Each Model Under the Chronological hold-out. Values Below -1 are Shown in the Darkest Color.

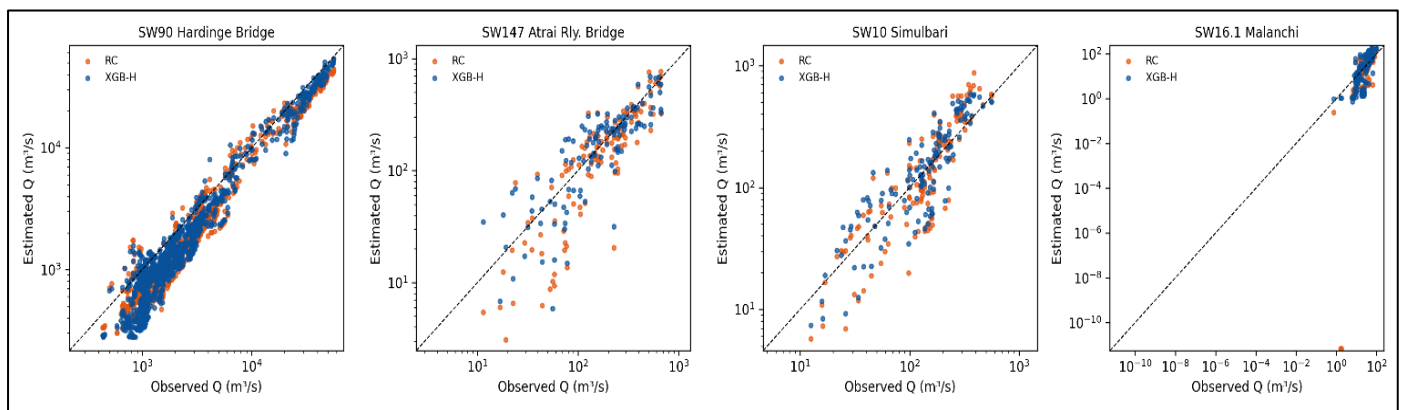


Fig 6 Observed Versus Estimated Discharge (Rating Curve and XGB-H) in the Chronological Test at Four Stations: Hardinge Bridge (Large River, Good fit), Atrai Rly. Bridge (Small River, Good Fit), Simulbari (ML Advantage) and Malanchi (Shifted Rating, all Methods Poor).

➤ Effect of Training on Recent Years Only

Restricting training to recent years did not rescue either approach consistently. With 15 years of training data, the rating curve's median chronological NSE fell slightly to 0.566, and with 10 years it fell to 0.464, largely because shorter records sampled fewer high flows. RF-H improved to 0.567 with 10 years, and XGB-H stayed close to 0.52 regardless of window length. The neural network, by contrast, degraded sharply with shorter windows (median NSE -0.234 with 10 years), which reflects its larger data requirement. At individual stations with strong recent shifts, such as Singra and Malanchi, a short window improved the rating curve considerably (from -0.58 to 0.47 at Singra and from -0.45 to 0.44 at Malanchi with 10 years), and RF-H improved even more at Singra (from 0.05 to 0.78). The same window, however, reduced the rating curve at Mohadebpur from 0.67 to -1.35, because the shorter record

contained few high flows. Rating updating based on recent gaugings is therefore an effective remedy at shifted stations, provided the update window still covers the full range of stages.

➤ Input Importance in the ML Models

SHAP analysis showed that stage dominated the XGB-H estimates, accounting for an average of 47% of mean absolute SHAP values (Fig. 7). The 30-day maximum stage was second (21%), followed by the two seasonal terms (together 23%). The one-day and three-day stage changes, which were included to represent hysteresis, contributed only 4% and 5%. This agrees with the residual analysis in Section 4.1: at a daily time step, the rate of stage change carries little information beyond stage itself, whereas recent flood history and season, which are plausibly linked to seasonal bed adjustment, carry more.

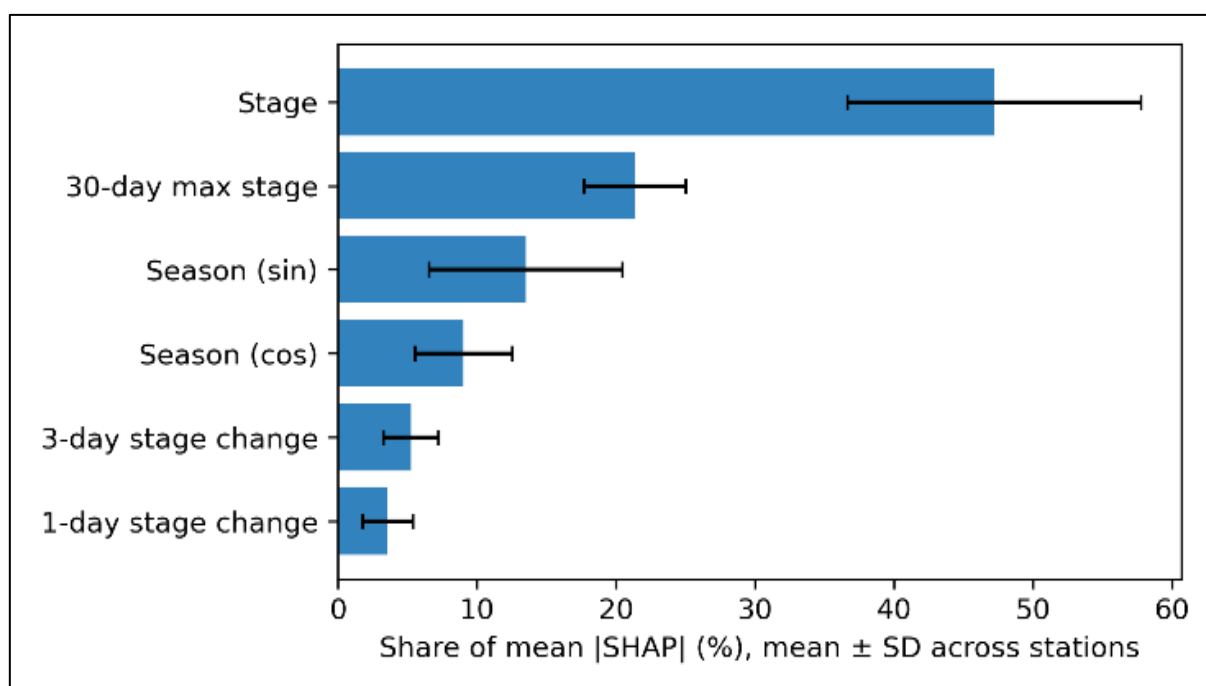


Fig 7 Share of Mean Absolute SHAP Value for Each Input of the XGB-H Model, Averaged Across the 16 Stations (Error Bars Show One Standard Deviation).

V. DISCUSSION

➤ Loss of the ML Advantage Under Chronological Testing

The apparent superiority of ML over the rating curve depends on how the comparison is made. With random cross-validation, our findings match those of He et al. (2023), Vishwakarma et al. (2023) and Kisi et al. (2024): flexible models fit the data better than a three-parameter curve. With chronological testing, which reflects how a rating model is actually used, the simpler curve is at least as good. The explanation lies in the non-stationarity documented in Table 2. When rating shifts are present, random splitting places gaugings from the same stability period in both training and test sets, so a flexible model can memorise the local relation for that period. The rating curve cannot exploit this in the same way because it fits one relation for the whole record, which is why its score changes less between designs. This is a specific case of the temporal leakage problem described by Kapoor and Narayanan (2023), and it echoes the observation of Rozos et al.

(2022) that ML rating models are usually built as if the data were stationary.

The reported accuracy of some recent ML rating studies should therefore be read with care. Coefficients of determination above 0.99 are possible for smooth, stable sections, but they say little about how a model will perform after the next morphogenic flood. Our results indicate that ML rating models need time-ordered evaluation before any claim of superiority is made, a point that Nearing et al. (2021) raised more broadly for ML in hydrology.

➤ Where ML Still Helps

ML models gave lower bias and better low-flow performance than the rating curve. Under random testing, RF-H improved median logNSE from 0.708 to 0.817, a difference that was significant (Wilcoxon $p = 0.001$). Under chronological testing, the improvement (0.567 to 0.691) was smaller and not significant ($p = 0.14$), and median PBIAS moved from -8.7%

for the rating curve to +2.0% for RF-H. The RC's tendency to underestimate in recent years is consistent with the positive residual trends found at most stations (Table 2), because a curve fitted to the whole record lags behind a channel that now conveys more water at a given stage. ML also clearly outperformed the rating curve at Simulbari, Khanpur and Hardinge Bridge, where the relation is smoother or the record is long. For low-flow applications such as drought monitoring and dry-season allocation, a tree-based model with seasonal inputs may therefore be worth using alongside the rating curve, provided it is tested chronologically.

➤ *Hysteresis at the Daily Scale*

We expected hysteresis inputs to be the main source of ML improvement, as suggested by earlier studies (Kumar et al., 2020; Alizadeh et al., 2021). This was not the case. Significant rising–falling differences appeared at only three stations, and stage-change inputs contributed less than 10% of SHAP importance. Two factors probably explain this. First, daily mean water levels smooth out the sub-daily stage changes that drive unsteady-flow hysteresis, which is better detected with continuous or hourly data (Perret et al., 2022). Second, in these rivers the slow drift of the rating relation over years is larger than the loop within a single flood. The 30-day maximum stage and seasonal terms were more informative, possibly because they act as indirect indicators of the seasonal state of the channel bed.

➤ *Practical Implications*

For BWDB and similar agencies, three recommendations follow. First, rating curves at these stations should be updated regularly with recent gaugings; our window experiment showed large gains at shifted stations from this step alone, and formal shift-detection methods such as those of Mansanarez et al. (2019) and Darienzo et al. (2021) could make the process systematic. Second, any ML rating model proposed for operational use should be validated on gaugings collected after its training period, and its skill should be compared with an updated rating curve rather than a single historical one. Third, stations where all methods fail, such as Baral Rly. Bridge, need field investigation of the gauge datum, the control section and possible relocations before any modelling is attempted.

➤ *Limitations*

Several limitations should be noted. The benchmark was a single power-law curve per station, whereas agencies often use segmented curves or periodic shifts; a comparison with a dynamically updated rating curve would be a stricter test and is planned for future work. Hyperparameters were fixed rather than tuned, which may disadvantage the ML models slightly, although tuning on chronologically ordered folds would not remove the underlying shift problem. Records after October 2022 were excluded because of the datum change, and hysteresis could only be represented with daily data. Gauging uncertainty was not modelled explicitly (Kiang et al., 2018). Predictions of $\ln Q$ were back-transformed without a retransformation-bias correction (Duan, 1983), which may bias all methods slightly low. The random and chronological designs were also scored on different sets of test gaugings, so part of the difference between them may reflect the flow range of the recent period rather than temporal leakage alone. Finally,

the network is regional, and the results may differ for rivers with more stable controls.

VI. CONCLUSIONS

Using 14,317 gaugings from 16 BWDB stations in northwest Bangladesh, we compared a power-law rating curve with six ML configurations under random and chronological validation. The main conclusions are as follows.

- Under random five-fold cross-validation, tree-based ML models achieved higher median NSE than the rating curve (best 0.77 versus 0.72), reproducing the pattern common in the literature.
- Under chronological testing, the rating curve achieved the highest median NSE (0.60 versus 0.49 to 0.59), and no ML model differed significantly from it after Holm correction.
- Random validation was significantly more optimistic than chronological testing for the rating curve and five of the six ML models.
- Rating relations were non-stationary at almost every station, with significant change points at 15 of 16 stations, while daily-scale hysteresis was weak.
- ML offered smaller bias and better low-flow skill and outperformed the rating curve at three stations, but its advantage was station-specific rather than general.

The choice of validation design changed the conclusion of this comparison more than the choice of algorithm did. We recommend that future ML rating studies report chronological or period-based validation as the primary test, and that agencies prioritise regular rating updates at unstable stations.

DECLARATIONS

- Data availability. The discharge and water level data were obtained from the Bangladesh Water Development Board (BWDB) and are available from BWDB on request, subject to its data-sharing conditions. The analysis code is available from the corresponding author on reasonable request.

ACKNOWLEDGEMENTS

The authors thank the Hydroinformatics and Flood Forecasting Circle of the Bangladesh Water Development Board for providing the hydrological data.

- Funding. This study received no specific grant from any funding agency in the public, commercial, or not for profit sectors.
- Competing interests. The authors declare no competing interests.
- Author contributions: All authors read and approved the final manuscript.
- Declaration of generative AI use. During the preparation of this work, the authors used Claude (Anthropic) to assist with drafting and language editing, checking the consistency of reported values, verifying reference details and writing code for figures. The authors reviewed and edited all content and take full responsibility for the content

of the publication. Every author has read, comprehended, and cooperated with the Instructions for Authors' "Ethical responsibilities of Authors" declaration.

REFERENCES

- [1]. Alizadeh, F., Faregh Gharamaleki, A., & Jalilzadeh, R. (2021). A two-stage multiple-point conceptual model to predict river stage-discharge process using machine learning approaches. *Journal of Water and Climate Change*, 12(1), 278–295. <https://doi.org/10.2166/wcc.2020.006>
- [2]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [3]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- [4]. Chen, Y. C., Yang, H. C., Lee, S. P., & Ho, C. H. (2023). Machine learning in establishing the stage–discharge rating curve of an irrigation canal. *Paddy and Water Environment*, 21(2), 181–191.
- [5]. Darienzo, M., Le Coz, J., Renard, B., & Lang, M. (2021). Detection of stage-discharge rating shifts using gaugings: A recursive segmentation procedure accounting for observational and model uncertainties. *Water Resources Research*, 57(4), e2020WR028607. <https://doi.org/10.1029/2020WR028607>
- [6]. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
- [7]. Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383), 605–610. <https://doi.org/10.1080/01621459.1983.10478017>
- [8]. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- [9]. Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
- [10]. Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- [11]. He, S., Niu, G., Sang, X., Sun, X., Yin, J., & Chen, H. (2023). Machine learning framework with feature importance interpretation for discharge estimation: A case study in Huitanggou sluice hydrological station, China. *Water*, 15(10), 1923. <https://doi.org/10.3390/w15101923>
- [12]. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- [13]. Horner, I., Le Coz, J., Renard, B., Branger, F., & Lagouy, M. (2022). Streamflow uncertainty due to the limited sensitivity of controls at hydrometric stations. *Hydrological Processes*, 36(2), e14497. <https://doi.org/10.1002/hyp.14497>
- [14]. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [15]. Kendall, M. G. (1975). *Rank correlation methods* (4th ed.). Charles Griffin.
- [16]. Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., et al. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, 54, 7149–7176. <https://doi.org/10.1029/2018WR022708>
- [17]. Kisi, O., Azamathulla, H. M., Cevat, F., Kulls, C., Kuhdaragh, M., & Fuladipannah, M. (2024). Enhancing river flow predictions: Comparative analysis of machine learning approaches in modeling stage-discharge relationship. *Results in Engineering*, 22, 102017. <https://doi.org/10.1016/j.rineng.2024.102017>
- [18]. Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>
- [19]. Kumar, M., Kumari, A., Kushwaha, D., Kumar, P., Malik, A., Ali, R., & Kuriqi, A. (2020). Estimation of daily stage–discharge relationship by using data-driven techniques of a perennial river, India. *Sustainability*, 12(19), 7877. <https://doi.org/10.3390/su12197877>
- [20]. Le Coz, J., Renard, B., Bonnifait, L., Branger, F., & Le Boursicaud, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology*, 509, 573–587. <https://doi.org/10.1016/j.jhydrol.2013.11.016>
- [21]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- [22]. Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3), 245–259.
- [23]. Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50–60.
- [24]. Mansanarez, V., Renard, B., Le Coz, J., Lang, M., & Darienzo, M. (2019). Shift happens! Adjusting stage-discharge rating curves to morphological changes at known times. *Water Resources Research*, 55, 2876–2899. <https://doi.org/10.1029/2018WR023389>
- [25]. Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I: A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- [26]. Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources*

- Research, 57(3), e2020WR028091.
<https://doi.org/10.1029/2020WR028091>
- [27]. Perret, E., Lang, M., & Le Coz, J. (2022). A framework for detecting stage-discharge hysteresis due to flow unsteadiness: Application to France's national hydrometry network. *Journal of Hydrology*, 608, 127567. <https://doi.org/10.1016/j.jhydrol.2022.127567>
- [28]. Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2), 126–135.
- [29]. Puri, D., Sihag, P., Thakur, M. S., Jameel, M., Chadee, A. A., & Azamathulla, H. M. (2024). Analysis of data splitting on streamflow prediction using random forest. *AIMS Environmental Science*, 11(4), 593–609. <https://doi.org/10.3934/environsci.2024029>
- [30]. Rahim, M. A., Liu, S., Hu, K., Li, H., Abedin, M. A., & Akter, F. (2024). Multiple-model based prediction of weekly discharge of the Brahmaputra-Jamuna by assimilating antecedent hydrological regime. *Geocarto International*, 39(1), 2413551. <https://doi.org/10.1080/10106049.2024.2413551>
- [31]. Rantz, S. E. (1982). *Measurement and computation of streamflow: Volume 2, Computation of discharge* (Water-Supply Paper 2175). U.S. Geological Survey.
- [32]. Rozos, E., Leandro, J., & Koutsoyiannis, D. (2022). Development of rating curves: Machine learning vs. statistical methods. *Hydrology*, 9(10), 166. <https://doi.org/10.3390/hydrology9100166>
- [33]. Saleh, M. A., Rasel, H. M., & Ray, B. (2024). From data to decisions: Leveraging ML for improved river discharge forecasting in Bangladesh. *Watershed Ecology and the Environment*, 6, 209–226. <https://doi.org/10.1016/j.wsee.2024.09.004>
- [34]. Vishwakarma, D. K., Kuriqi, A., Abed, S. A., Kishore, G., Al-Ansari, N., Pandey, K., Kumar, P., Kushwaha, N. L., & Jewel, A. (2023). Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. *Heliyon*, 9(5), e16290. <https://doi.org/10.1016/j.heliyon.2023.e16290>
- [35]. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.