Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The evaluation of the hydroclimatic factor compression for feed-forward neural discharge prediction under temporal generalization: A case study of the Tukad Badung watershed, Bali

Putu Doddy Heka Ardana

doddyhekaardana@unr.ac.id

Universitas Ngurah Rai

I Putu Sugiana

Universitas Warmadewa

I Made Dedi Suardika

Universitas Ngurah Rai

Research Article

Keywords: Discharge prediction, Exploratory factor analysis, Feed-forward neural network, Predictive sufficiency, Temporal generalization

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11684121/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The evaluation of the hydroclimatic factor compression for feed-forward neural discharge prediction under temporal generalization: A case study of the Tukad Badung watershed, Bali

Putu Doddy Heka Ardana[1,*], I Putu Sugiana[2], I Made Dedi Suardika[3]

[1]Department of Civil Engineering, Faculty of Science and Technology, Universitas Ngurah Rai, Denpasar, Bali 80238, Indonesia

[2]Department of Aquatic Resources Management, Universitas Warmadewa, Denpasar, Bali 80235, Indonesia

[3]Department of Architecture, Faculty of Science and Technology, Universitas Ngurah Rai, Denpasar, Bali 80238, Indonesia

[*]Corresponding author. E-mail: doddyhekaardana@unr.ac.id

Abstract

Dimensionality reduction is often expected to improve data-driven discharge prediction, but statistical compactness and predictive relevance are different properties. This study aimed to test whether exploratory factor analysis improves the temporal generalization of a feed-forward backpropagation artificial neural network for monthly discharge prediction in the Tukad Badung watershed, Bali, Indonesia. The monthly data for 2006–2025 were divided chronologically into training (2006–2019, n = 168), validation (2020–2024, n = 60), and a locked test period (2025, n = 12). A seven-variable baseline was compared with three factor-based scenarios, in which principal axis factoring with Promax rotation was fitted to the training period only. The selected baseline reached a training Nash–Sutcliffe efficiency of 0.834 but only 0.088 in validation and −0.018 in 2025, reflecting the overestimation of low flows and the underestimation of high flows. Two factors accounted for 53.02% of the total variance (Kaiser–Meyer–Olkin index = 0.670), and the factor-only scenario reduced the validation Nash–Sutcliffe efficiency to 0.025. Restoring wind speed (communality 0.075) to the two factors increased the validation Nash–Sutcliffe efficiency to 0.059 and reduced the 2025 root mean square error from 7.006 to 6.662 m^3/s. The 2025 Nash–Sutcliffe efficiency of

0.079 became −4.919 without the September extreme of 26.365 m³/s, and the validation values decreased to −0.833 (baseline) and −0.918 (factors plus wind speed) without the four months above the training range. These results show that factor compression did not preserve all target-relevant information and that covariance sufficiency differs from predictive sufficiency. Missing temporal information is therefore the most likely constraint to be addressed in future work.

Keywords: Discharge prediction, Exploratory factor analysis, Feed-forward neural network, Predictive sufficiency, Temporal generalization

1 Introduction

Floods are among the most damaging hydroclimatic hazards and are becoming harder to predict because urbanization, climate variability, and catchment modification are changing how rainfall is converted into runoff. Machine-learning models are currently widely used for flood and discharge forecasting but the skill depends strongly on the model, the input data, the forecast lead time, and the adoption of hybrid or decomposition strategies [1]. This is observed from the broader reviews of machine learning for flood prediction and rainfall–runoff modelling [2–4]. An example of this trend is the September 2025 floods recorded in Bali. Prawesti et al. [5] reported daily rainfall of up to 385 mm, 81 flood points in Denpasar City, and 12 in Badung Regency with the Tukad Badung watershed identified as the most affected area. The reliability of a data-driven discharge model in this type of an intensively urbanized watershed is required to be presented empirically and cannot be assumed only from the preference.

Streamflow studies in different river basins and climatic zones including tropical regions have compared neural, kernel-based, and deep-learning models [6–8]. Neural networks have also been applied in Indonesia to real-time catchment flow [9], flood susceptibility [10] and extreme water-level prediction at the Katulampa barrage in Jakarta [11]. Feed-forward backpropagation

networks have been identified to be capable of approximating the nonlinear, multivariate, and temporally dependent rainfall–runoff transformation without an explicit physical formulation and this trend explains the long-term adoption in hydrology [12–16]. However, methodological literature repeatedly warned that the success of a network did not depend only on the architecture. Maier and Dandy [17] described data division, preprocessing, input determination, architecture, training, and validation as the connected stages of one modelling process. A later review of 210 river-system studies also reported that inputs, data subsets, and model structure were often selected in an ad hoc manner [18]. Therefore, an improvement in a neural hydrological model could depend less on adding network capacity than on how the information supplied to the network was represented.

Input determination has become another independent field of study. This was observed from the development of partial mutual information, self-organizing maps combined with genetic algorithms, and a mean-shifted-histogram estimator to identify the candidate inputs carrying information about a hydrological target [19–21]. A previous comparative study showed that the outcome depended on the selection method and the stopping rule [22]. Input filtering was also reported to have changed the skill of recurrent runoff models [23]. Therefore, dimensionality reduction by factor analysis or principal components offers a different route by replacing correlated predictors with a few latent variables and promising a more parsimonious network. Principal components have been used to preprocess inputs for rainfall–runoff networks with improved accuracy in some applications and variable results in others [24, 25]. Wallis [26] had previously reported factor analysis in hydrology as a descriptive device rather than an important tool for screening predictors.

The report was associated with a precise distinction where the target was confirmed not to be included in the estimation when the common-factor model was fitted to the correlation structure among the predictors. The process showed the variables sharing the variance rather than the

78 information that improved the prediction of discharge [27]. A transformation of the predictors
79 also could not increase the information carried about the target. The best effect was to preserve
80 the information and this was only achieved completely with a sufficient representation [28, 29].
81 Jolliffe [30] analyzed regression and showed the possibility of components with small variance
82 being important predictors and the trend signaled that the ranking of latent dimensions by
83 explained variance was not a safe criterion for prediction. Therefore, covariance sufficiency is
84 used in this study as the representation that reproduces the shared covariance among the
85 predictors. Predictive sufficiency is the representation that retains the information needed to
86 predict discharge based on out-of-sample skill relative to the full predictor set. The two
87 properties are not required to coincide and the wind speed in the present dataset serves as the
88 natural test case with a communality of only 0.075.
89 The coincidence of the two properties can only be determined under a validation design that
90 imitates real applications. This is based on the ability of random partitions to allow information
91 from later years to influence the calibration of earlier periods. The trend reflects the need for
92 model development to be in line with time [17]. Large-sample experiments are also required to
93 show that the method applied in dividing a series changes the reported test performance [31,
94 32]. Moreover, aggregate metrics can provide misleading information. An example is the fact
95 that the Nash–Sutcliffe efficiency (NSE) is sensitive to outliers [33]. The zero value
96 corresponds to a mean-flow benchmark which is considered easy to beat [34, 35] and a single
97 extreme month is capable of inflating the denominator. A feed-forward network trained on
98 normalized data cannot readily reproduce a peak larger than any in the training set [36]. Recent
99 studies also reported that stand-alone long short-term memory (LSTM) networks could be
100 limited in the same manner [37] even though other experiments considered the model as
101 relatively accurate on extreme events without training [38]. Another important factor is the
102 temporal structure because lag effects persist in neural rainfall–runoff models unless the

103 temporal dependence is represented explicitly [39]. The reviewers of the field have called for
104 more objective protocols for selecting inputs and for more transparent benchmarking [40].
105 A total of three gaps are identified from the review and the first is that dimensionality reduction
106 is often analyzed through statistical compactness or aggregate accuracy with less focus on the
107 ability of retained factors to preserve target-relevant information. The second is that factor-
108 based input reduction is rarely compared with the original predictor space under a strictly
109 chronological design keeping the final evaluation period untouched. The third is the ability of
110 aggregate scores to conceal flow-dependent bias and the failure to reproduce rare high flows.
111 However, there is a lack of clarity on the capacity of hydroclimatic factor compression to
112 improve the temporal generalization of feed-forward discharge models and whether variables
113 poorly represented by latent factors have predictive value. These questions are observed not to
114 have been answered for a Bali watershed.
115 The efforts to address the gaps led to the development of an integrated exploratory factor
116 analysis (EFA) and a feed-forward backpropagation artificial neural network (FFBPANN)
117 (EFA-FFBPANN) framework that separated the estimation of hydroclimatic covariance from
118 predictive model selection. The factor structure was estimated from the training period only,
119 frozen, and compared with a seven-variable baseline of 160 networks and three factor-based
120 scenarios of 600 networks. All models were selected exclusively on a 2020–2024 validation
121 period and evaluated on a locked 2025 test period. The diagnostics exceeded aggregate error
122 to include flow-regime bias, prediction shrinkage, and the sensitivity of NSE to the September
123 2025 extreme. The objectives were to (1) quantify the predictive performance and temporal
124 generalization of the baseline model, (2) identify the latent hydroclimatic structure with a
125 training-period EFA and test whether factor-based input compression improves the network,
126 and (3) determine whether common variance is equivalent to predictive information through
127 flow-regime residual analysis and independent extreme-event evaluation.

The remainder of this article is organized to include the Methods section which describes the study area, the data, and the modelling procedure. The Results section presents the baseline performance, the factor structure, and the scenario comparison. The Discussion and limitations section interprets the results while the Conclusions summarize the main outcomes.

2 Methods

2.1 Study area

The Tukad Badung watershed is approximately between 8°33′ and 8°43′ S and 115°10′ and 115°16′ E on the island of Bali, Indonesia as presented in Fig. 1. It covers an estimated area of 5495.1 ha or 54.95 km² and the main channel which is the Tukad Badung River is 19.6 km long. The watershed rises from near sea level on the coastal plain to approximately 140 m in the northern uplands that border Badung Regency. The river flows southward through the built-up core of Denpasar City before discharging into the Badung Strait. The channel passes through the older commercial district in the city where markets, housing, and roads are close to the banks. The trend shows that the exposure to high flows is concentrated along the river.

Land cover is dominated by urban surfaces as observed in the report of a Sentinel-2 classification that 63.4% of the 3944 ha mapped is built-up area, 18.6% is cropland, 13.96% is vegetation, 3.12% is bare land, and 0.92% is water bodies. The trend shows intensive urbanization through the conversion of land to residential areas, villas, and tourism facilities [5]. A spatial analysis conducted in July–August 2026 and supported by field observations provided similar information with forest and trees identified as covering 10.94% of the watershed, rice fields 16.86%, grassland and gardens 5.14%, bare land 0.61%, water bodies 2.09% while built-up land was the dominant class at 64.36%.

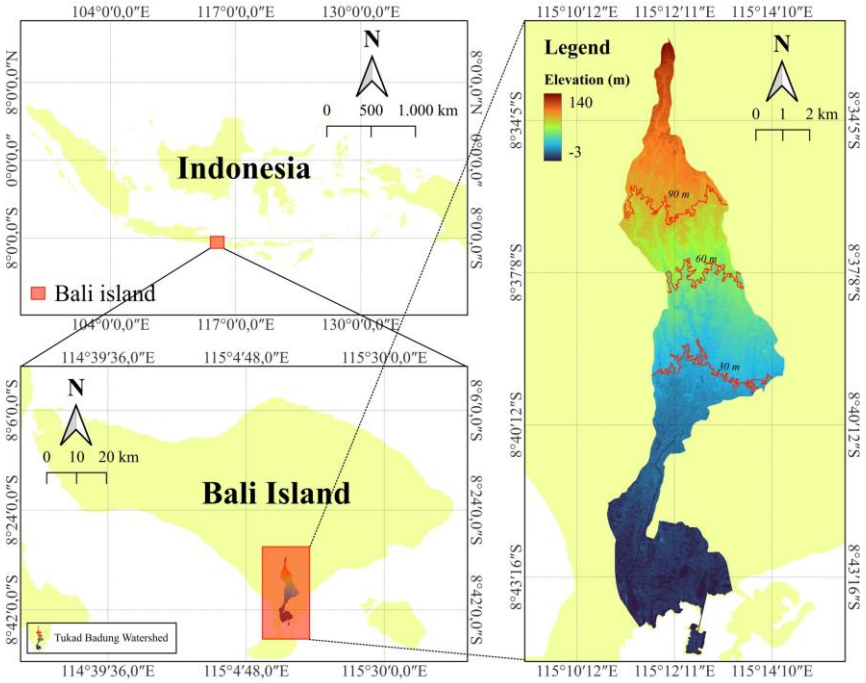

Fig. 1 Location of the Tukad Badung watershed in Indonesia and Bali.

The Tukad Badung watershed was identified as the area most affected by the September 2025 floods in Bali which was the month a discharge of 26.365 m³/s was recorded. Areal rainfall was derived from four rain gauges in and around the watershed including Sading (8°37′23″ S, 115°10′17″ E), Buagan (8°41′04″ S, 115°11′49″ E), Padang Sumbu (8°40′50″ S, 115°10′33″ E, 3 m elevation), and Suwung Kangin (8°42′39″ S, 115°13′55″ E, 3 m elevation). The Thiessen polygons covered 18.42, 13.87, 16.39, and 6.27 km² which summed to the 54.95 km² watershed area and provided the weights of 0.335, 0.252, 0.298, and 0.114 respectively. The remaining meteorological variables were recorded at the Sanglah Geophysical Station (8°40′37″ S, 115°12′36″ E). Discharge was also obtained from the Tukad Badung staff-gauge station (station no. 7-23-0-1) which was established on 21 August 1967 at 8°39′48″ S, 115°15′55″ E and 36.137 m elevation in Gerenceng, Denpasar.

The rainfall regime is strongly seasonal and the 2006–2025 areal series was used with December to February being the wettest months and August the driest. The mean monthly discharge is loosely in line with rainfall and the highest is recorded in February and December

with 5.37 m^3/s and 5.27 m^3/s respectively while the lowest is in June with 2.67 m^3/s. Some of the largest flows in the record are in the drier months with 27.7 m^3/s in July 2023 and 26.365 m^3/s in September 2025 when areal rainfall is the highest of the whole series. Rainfall and discharge of the same month are weakly correlated with $r = 0.13$ in the training period and 0.25 over the full record. The trend limits the concepts that are explainable by contemporaneous predictors. The station records showed the same seasonality where the rainfall at the Sanglah Geophysical Station in 2024 was identified as 2063.9 mm with the largest monthly data recorded in December (730.9 mm) and February (322.8 mm) and the smallest in May (1.9 mm) and August (9.7 mm) [41]. Interannual departures from the normal could be large with the Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) bulletin of July 2025 reporting a June 2025 total of 213.2 mm against a June normal of 44.1 mm (29-year mean) with a maximum daily rainfall of 108.9 mm on 25 June.

2.2 Hydroclimatic data

The dataset comprised 240 monthly observations from January 2006 to December 2025 and seven hydroclimatic predictors were used to estimate upstream discharge as the target variable in Table 1. Meteorological variables such as temperature, wind speed, humidity, and pressure were used with neural networks to represent the conditions generating runoff and floods [42, 43]. Relative humidity (KU), air temperature (TMP), wind speed (KA), sunshine duration (PM), air pressure (TU), and evaporation (PNG) were sourced from the BMKG Region III. Rainfall (CH) and discharge (DEBIT) were obtained from the Bali Penida River Basin Organization (BWS Bali Penida). Missing monthly values at a station were completed with the arithmetic mean of the available stations for the same month and rainfall was represented by the single variable CH.

Table 1 Variables used

Symbol	Variable	Unit	Role
KU	Relative humidity	%	Predictor
TMP	Air temperature	°C	Predictor
KA	Wind speed	m/s	Predictor
PM	Sunshine duration	%	Predictor
TU	Air pressure	hPa	Predictor
PNG	Evaporation	mm	Predictor
CH	Areal rainfall of the watershed	mm	Predictor
DEBIT	Upstream discharge	m^3/s	Target

2.3 Normalization and leakage control

All variables were min–max normalized (Equation 1) with the minimum and maximum of the training period only towards ensuring no information from 2020–2025 entered the scaling. Validation and test values outside the training range were retained and not clipped which was the reason normalized values exceeded one in the later periods. For example, normalized discharge reached 3.26 in validation and 3.10 in 2025. This showed that the network inputs and outputs were not rescaled again inside MATLAB R2015a and errors in m^3/s were obtained by inverse transformation.

$$X_{norm} = \frac{X - X_{min,train}}{X_{max,train} - X_{min,train}} \quad (1)$$

2.4 Chronological partition

The series was divided without shuffling and this allowed model development to consistently precede evaluation in time [17, 32] (Table 2).

Table 2 Chronological partition of the baseline and EFA-FFBPANN experiments

Subset	Period	n	Role
Training	2006–2019	168	Network fitting and EFA
Validation	2020–2024	60	Model comparison and selection
Independent test	2025	12	Final assessment, locked

2.5 Baseline FFBPANN

The baseline (scenario S1) used all seven predictors in a one-hidden-layer FFBPANN with a hyperbolic tangent sigmoid (tansig) hidden layer, a linear (purelin) output neuron, and the mean

206 squared error (MSE) as the loss. A total of four hidden-layer sizes including 7, 8, 9, and 10
207 neurons were trained with two algorithms, Levenberg–Marquardt (trainlm) [44], and Bayesian
208 regularization (trainbr) [45] with 20 independent runs per configuration which provided 4 × 2
209 × 20 = 160 networks. Both algorithms were used for runoff prediction with Levenberg–
210 Marquardt identified as the better local optimizer [16, 46] while Bayesian regularization
211 performed comparably in some studies [47] and limited over- and underfitting [48, 49].
212 Training was limited to 2000 epochs with a goal MSE of 1e−6, a minimum gradient of 1e−7,
213 and an initial μ of 0.001. Each run used a fixed seed and the internal data division of MATLAB
214 was disabled to ensure only the training period contributed to weight estimation. Furthermore,
215 the mean training-period discharge served as a naive benchmark predictor for the validation
216 and test periods.

217 **2.6 Training-only EFA**

218 EFA was applied to the seven predictors using only the 168 training observations while
219 sampling adequacy was checked with the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's
220 test of sphericity. Factors were extracted by principal axis factoring (PAF) which modeled
221 common variance and were rotated with the oblique Promax method without forcing
222 independence between latent dimensions [27]. The number of factors followed the eigenvalue
223 criterion and interpretability with loadings below 0.30 treated as negligible. Regression-based
224 factor scores (Equation 2) were computed for all 240 months from z-scores standardized with
225 training-period means and standard deviations. The coefficients were frozen after the training-
226 period solution to ensure the factor structure was never re-estimated on validation or test data
227 [50]. The KMO measure, Bartlett's test, factor extraction, rotation, and factor scores were
228 computed in IBM SPSS Statistics 22 while MATLAB was used only for the FFBPANN.

229 $$F_{k,t} = \sum_{j=1}^{7} w_{kj} Z_{j,t} \, , Z_{j,t} = \frac{X_{j,t} - \bar{X}_{j,train}}{s_{j,train}} \qquad (2)$$

2.7 Input scenarios

The four input scenarios defined were S1 as the baseline in addition to S2, S3, and S4 as factor-based as presented in Table 3.

Table 3 Input scenarios of the baseline and EFA-FFBPANN experiments

Scenario	Inputs	Number of inputs	Purpose
S1 (baseline)	KU, TMP, KA, PM, TU, PNG, CH	7	Full predictor space
S2	F1, F2	2	Are only the latent factors sufficient?
S3	F1, F2, KA	3	Does a low-communality variable retain predictive value?
S4	F1, F2, KA, PNG	4	Does evaporation add information?

2.8 Experimental design and model selection

Each factor-based scenario was trained with five hidden-layer sizes including 5, 7, 8, 9, and 10 neurons, two algorithms in the form of trainlm and trainbr, as well as 20 runs to provide $3 \times 5 \times 2 \times 20 = 600$ networks all of which converged. The seeds were fixed and sequential with the best network within each scenario selected only based on the maximum validation NSE while the minimum validation RMSE served as the tie-breaker. The 2025 period was a locked test set that was not considered in training, scenario comparison, or model selection. Small differences in the scenarios were required to be interpreted with caution because every reported best network was the maximum of 200 candidates with 160 for the baseline. This was achieved by selecting the best of several candidate biases based on the apparent upward performance [51, 52].

2.9 Performance metrics and diagnostics

Performance was assessed at four levels including aggregate statistics, hydrographs, residuals by flow regime, and extreme-event behavior. The primary criterion was the NSE [53] (Equation 3) where Q_o and Q_p were the observed and predicted discharge respectively while $\bar{Q}_o$ was the mean observed discharge.

$$NSE = 1 - \frac{\sum_t \left(Q_{o,t} - Q_{p,t}\right)^2}{\sum_t \left(Q_{o,t} - \bar{Q}_o\right)^2} \quad (3)$$

The root mean square error (RMSE), mean absolute error (MAE), and bias were calculated in m^3/s after inverse transformation. The Pearson correlation as well as the slope and intercept of the regression of predicted or observed discharge were also used to quantify prediction shrinkage. Moreover, the coefficient of determination reported for the training period was computed as 1 − SSE/SST and was numerically identical to the in-sample NSE. The benchmark NSE of the training mean was reported for each period because zero was adopted for a constant mean-flow predictor [34, 35]. Residuals were grouped into low, medium, and high flow classes defined as terciles of the observed discharge over the full 240-month record. The classes included low at ≤ 2.48 m^3/s, medium at 2.49–4.46 m^3/s, and high at ≥ 4.48 m^3/s with 80 months each and were used only to describe errors and not to fit any model.

2.10 Extreme-event sensitivity

The 2025 NSE was calculated twice for all 12 months and after excluding September 2025 where discharge reached 26.365 m^3/s considering the ability of one month to dominate the denominator [33]. Peak capture was also defined as the ratio of predicted to observed discharge in a month because neural streamflow models tended to underestimate the largest flows [54].

2.11 Software and reproducibility

Networks were trained in MATLAB R2015a with the normalized input workbook, the scripts, and result tables made available from the corresponding author upon reasonable request.

3 Results and discussion

3.1 Baseline performance and the generalization gap

The best baseline network among 160 candidates which was selected through validation NSE was the 7-7-1 architecture trained with trainlm in run 3. The training NSE was recorded as

0.834 but the validation NSE reduced to 0.088 and the locked 2025 test to −0.018 as presented in Table 4. The RMSE in discharge units increased from 0.901 m³/s in training to 4.493 m³/s in validation and 7.006 m³/s in 2025. The MAE also increased from 0.714 to 2.846 and 4.527 m³/s respectively. The mean training discharge which was used as a naive benchmark provided validation and test NSE of −0.016 and −0.006 respectively. This showed that the selected network only exceeded the benchmark in validation by a small margin. Fig. 2 compares observed and predicted discharge for 2006–2025. The repeated-run design also showed that nominally successful networks could differ substantially in validation skill. This was observed in the trainlm runs that converged but generalized very poorly. The focus of the current report is only on the best candidate and the interpretation of the selection is presented in the Discussion section.

Table 4 Performance of the selected baseline FFBPANN 7-7-1 model

Period		n	MSE (normalized)	RMSE (normalized)	RMSE (m³/s)	MAE (m³/s)	NSE
Training	2006–2019	168	0.0117	0.1084	0.901	0.714	0.834
Validation	2020–2024	60	0.2923	0.5407	4.493	2.846	0.088
Independent	2025	12	0.7107	0.8430	7.006	4.527	−0.018

The model was trained with trainlm (run 3) with the MSE and normalized RMSE presented in the normalized target space and the errors in m³/s were back-transformed. The partition is provided in Table 2.

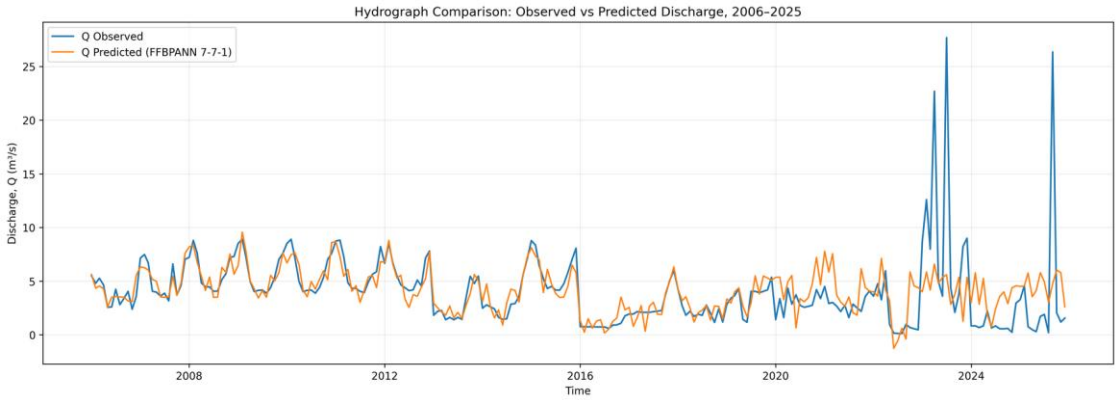

Fig. 2 Observed and predicted discharge of the selected baseline FFBPANN 7-7-1 model, 2006–2025.

3.2 Prediction shrinkage and flow-regime errors

The hydrograph and scatter diagnostics showed that the network was in line with the seasonal and interannual structure under ordinary conditions but compressed the amplitude of discharge toward the central range as presented in Figs. 2 and 3. The slope of predicted discharge on observed value was recorded as 0.317 over all 240 months and was within approximately 0.131 in the validation period. High flows are pulled down and low flows are pulled up to show that the error is systematic and not random as presented in Fig. 4.

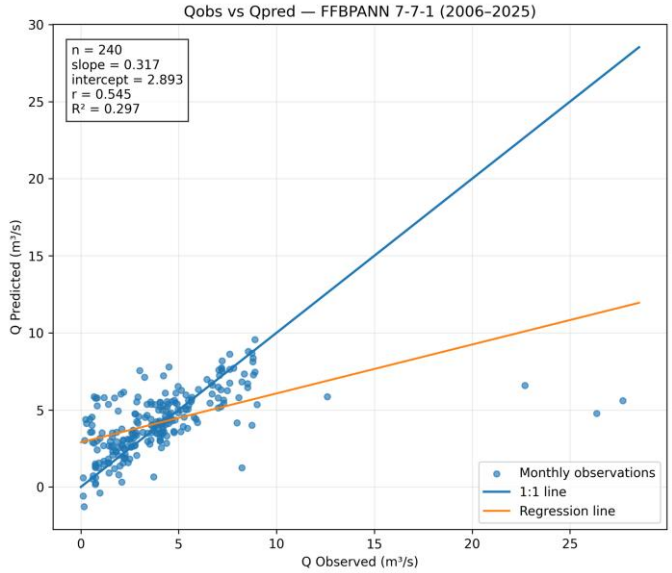

Fig. 3 Observed versus predicted discharge for the selected baseline FFBPANN 7-7-1 model.

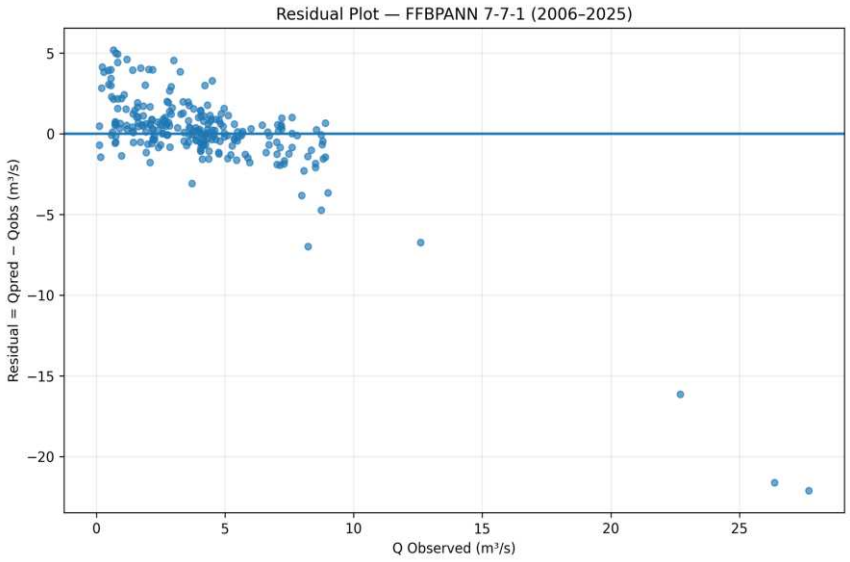

Fig. 4 Residual diagnostic for the baseline FFBPANN 7-7-1 model.

The pattern was further confirmed through the flow-class analysis with terciles of observed discharge including low at ≤ 2.48, medium at 2.49–4.46, and high at ≥ 4.48 m³/s as presented in Table 5. The baseline averagely overestimated the low flows by 1.26 m³/s and underestimated high flows by 1.37 m³/s over the full record. The bias increased to +2.17 m³/s for low flows and −5.81 m³/s for high flows in validation. The change of sign from low to high flows was the signature of regression toward the mean. This showed that an increase in the hidden layer was not a remedy within the range tested because the 160-network experiment included 7, 8, 9, and 10 hidden neurons and the smallest was used in the best validation network.

Table 5 Flow-regime dependent error of the baseline FFBPANN model

Period	Flow class	n	Observed mean (m³/s)	Predicted mean (m³/s)	Bias (m³/s)	MAE (m³/s)	RMSE (m³/s)
All 240	Low	80	1.346	2.602	+1.256	1.555	2.125
All 240	Medium	80	3.617	4.078	+0.461	0.947	1.298
All 240	High	80	7.251	5.876	−1.375	1.812	4.243
Validation	Low	25	0.953	3.125	+2.172	2.481	2.889
Validation	Medium	24	3.269	4.348	+1.079	1.593	1.986
Validation	High	11	10.649	4.837	−5.812	6.409	9.085

3.3 Independent 2025 evaluation and the September 2025 extreme

The locked 2025 period was the clearest stress test with the observed discharge reaching 26.365 m³/s in September. The baseline predicted 4.771 m³/s which was 18.1% of the observed peak as presented in Table 6 and Fig. 5. The predicted series was within a narrow range throughout the year while the observations contained significantly larger excursions.

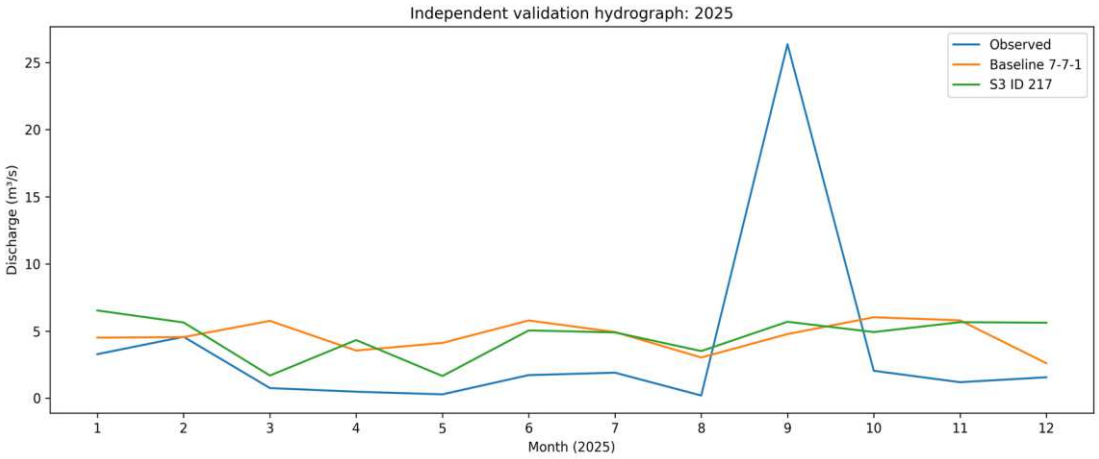

Fig. 5 Independent 2025 hydrograph comparison between the baseline and S3 models. The trend shows S3 as the best validation-selected EFA-FFBPANN model.

Table 6 Monthly independent-test predictions for 2025

2025 month	Observed Q (m³/s)	Baseline Q_{pred} (m³/s)	Baseline residual (m³/s)	S3 Q_{pred} (m³/s)	S3 residual (m³/s)
January	3.276	4.508	+1.232	6.531	+3.255
February	4.570	4.548	−0.022	5.637	+1.067
March	0.752	5.757	+5.005	1.683	+0.932
April	0.480	3.546	+3.066	4.328	+3.848
May	0.284	4.114	+3.829	1.649	+1.365
June	1.717	5.786	+4.069	5.044	+3.327
July	1.898	4.926	+3.028	4.897	+2.999
August	0.197	3.029	+2.832	3.507	+3.311
September	26.365	4.771	−21.594	5.690	−20.675
October	2.039	6.027	+3.988	4.922	+2.884
November	1.188	5.795	+4.607	5.659	+4.471
December	1.556	2.603	+1.047	5.619	+4.063

The September event dominates the total squared error and is not evidence of general predictive skill. Residuals are predicted minus observed discharge.

The S3 model raised the all-month 2025 NSE from −0.018 to 0.079 but the gain depended on one month. The removal of September changed the NSE of S3 to −4.919 and the value for the baseline to −5.904 because the extreme value inflated the variance in the denominator. S3 underpredicted the September peak by 20.675 m³/s and captured 21.6% against 18.1% for the baseline. The exclusion of September also changed the RMSE to 3.091 m³/s for S3 and 3.339 m³/s for the baseline, the MAE was 2.866 and 2.975 m³/s, r^2 was 0.451 and 0.041 while the Spearman correlation was 0.769 and 0.357 respectively. This allowed S3 to rank the non-

extreme months better but both models were inadequate for the low flows and for the flood peak.

3.4 Factor structure of the hydroclimatic predictors

The training-period factorability tests supported EFA in Table 7. This led to the retention of two factors with the initial eigenvalues recorded to be 3.008 and 1.476 which explained 42.97% and 21.09% of the standardized variance and a cumulative of 64.06%. The sums of squared loadings after principal axis extraction were 37.92% and 15.10% which showed the two factors accounted for 53.02% of the total variance as common. Promax rotation also allowed correlated factors but the final factor correlation was −0.005 which practically reflected a lack of correlation.

Table 7 Training-period EFA diagnostics for the seven hydroclimatic predictors

Diagnostic	Result	Interpretation
Sample size	168	Training period only
KMO	0.670	Adequate factorability
Bartlett χ^2 (df = 21)	441.334, $p < 0.001$	Correlation matrix differs from identity
Extraction/rotation	Principal axis factoring / Promax	Common variance; oblique rotation
Retained factors	2	Eigenvalues above 1 and interpretability
Initial cumulative variance	64.06%	Two eigenvalues above 1
Extracted common variance	53.02%	Share of total variance represented by the two factors
Factor correlation	−0.005	Effectively uncorrelated

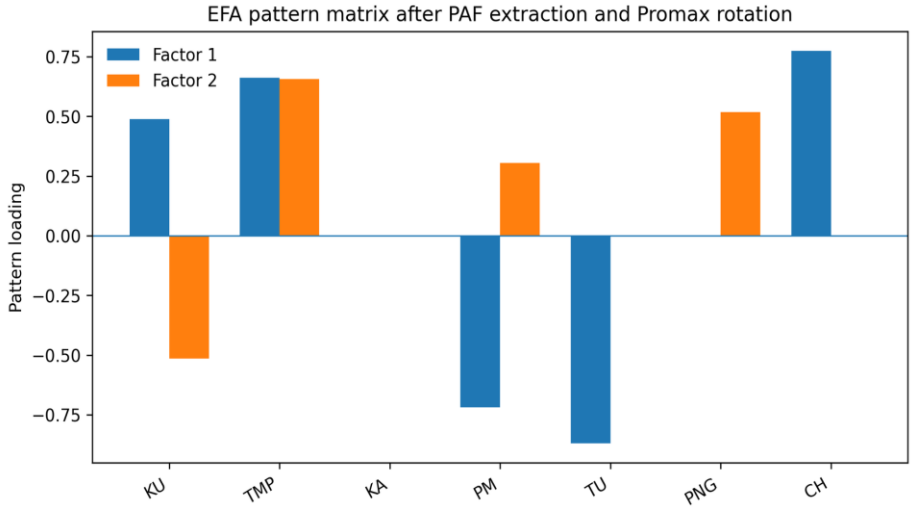

Fig. 6 Pattern loadings of the two-factor PAF-Promax solution. The solution was fitted to the 2006–2019 training period.

Fig. 6 and Table 8 show that factor 1 is dominated by air pressure (TU −0.869), areal rainfall (CH 0.774), sunshine duration (PM −0.719), air temperature (TMP 0.662), and relative humidity (KU 0.489). This reflects a wet-condition pattern of high rainfall with low pressure and short sunshine. Factor 2 is defined by air temperature (TMP 0.657), evaporation (PNG 0.518), relative humidity (KU −0.515), more weakly sunshine duration (PM 0.305), as well as a thermal–evaporative pattern of warm, dry, and evaporative conditions. The trend showed that both labels were identified to be interpretive. Moreover, Wind speed (KA) had a communality of only 0.075 with no loadings above 0.30 on either factor and was barely represented by the common-factor structure. The factor-score coefficients frozen for 2020–2025 were KU (0.147, −0.216), TMP (0.330, 0.819), KA (−0.026, 0.041), PM (−0.203, 0.195), TU (−0.323, 0.197), PNG (−0.021, 0.160), and CH (0.228, −0.093).

Table 8 Final PAF-Promax pattern loadings and communalities

Variable	F1 loading	F2 loading	Communality
KU	0.489	−0.515	0.506
TMP	0.662	0.657	0.865
KA	< 0.30	< 0.30	0.075
PM	−0.719	0.305	0.611
TU	−0.869	< 0.30	0.756
PNG	< 0.30	0.518	0.273
CH	0.774	< 0.30	0.624

3.5 Performance of the EFA-FFBPANN scenarios

The results showed that all 600 factor-based networks converged. The best network within each scenario which was selected through only validation NSE adopted five hidden neurons and trainlm as presented in Table 9. None of the three scenarios exceeded the baseline validation NSE of 0.088. This was observed from the fact that factor-only scenario S2 reached 0.025 with a training NSE of 0.256. The restoring wind speed in S3 raised the training NSE to 0.441, the validation NSE to 0.059, and the 2025 NSE to 0.079 as well as provided the lowest validation RMSE for the three at 4.566 m³/s. The addition of evaporation in S4 lifted the training NSE to 0.545 but maintained the validation NSE at 0.029 and the 2025 NSE at −0.082 as presented in Fig. 7.

Table 9 Best validation-selected network of each scenario compared with the baseline

Scenario	Inputs	Hidden neurons	Algorithm	Run	Training NSE	Validation NSE	Validation RMSE (m³/s)	2025 NSE	2025 RMSE (m³/s)
S2	F1 + F2	5	trainlm	5	0.256	0.025	4.646	−0.017	7.003
S3	F1 + F2 + KA	5	trainlm	17	0.441	0.059	4.566	0.079	6.662
S4	F1 + F2 + KA + PNG	5	trainlm	16	0.545	0.029	4.636	−0.082	7.221
Baseline (S1)	7 original variables	7	trainlm	3	0.834	0.088	4.493	−0.018	7.006

RMSE is in discharge units after inverse transformation.

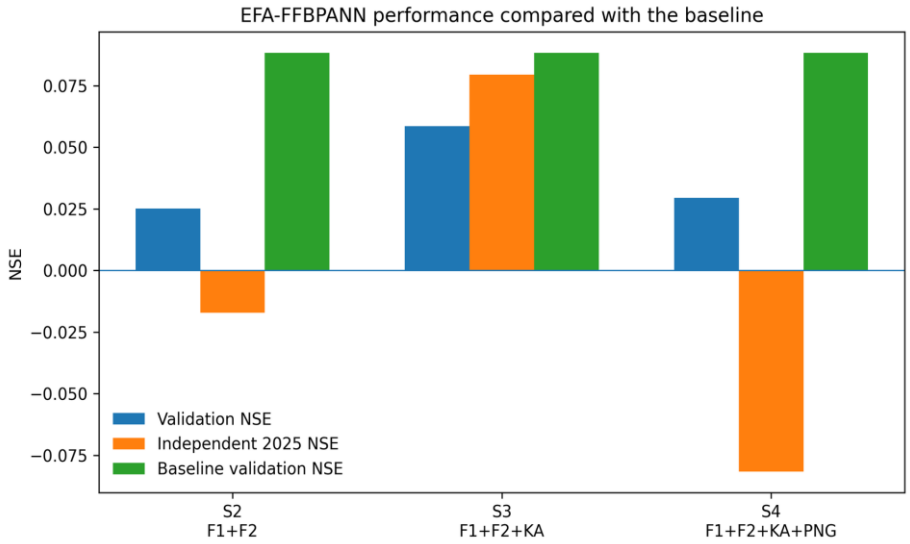

Fig. 7 Validation and independent-test NSE of the best EFA-FFBPANN models. The baseline validation NSE is presented for comparison.

3.6 Residual comparison between the baseline and S3

S3 was identified as the best factor-based scenario on validation and was comprehensively compared with the baseline in Figs. 8 and 9. The observed–predicted slope was estimated at 0.087 in validation against 0.131 for the baseline and approximately 0.441 in training compared to 0.834. Therefore, the predictions were even more compressed than the trend observed in the baseline. S3 generally showed stronger low-flow overprediction (+2.138 against +1.256 m^3/s) and stronger high-flow underprediction (−2.082 against −1.375 m^3/s) for the whole 240 months as presented in Table 10.

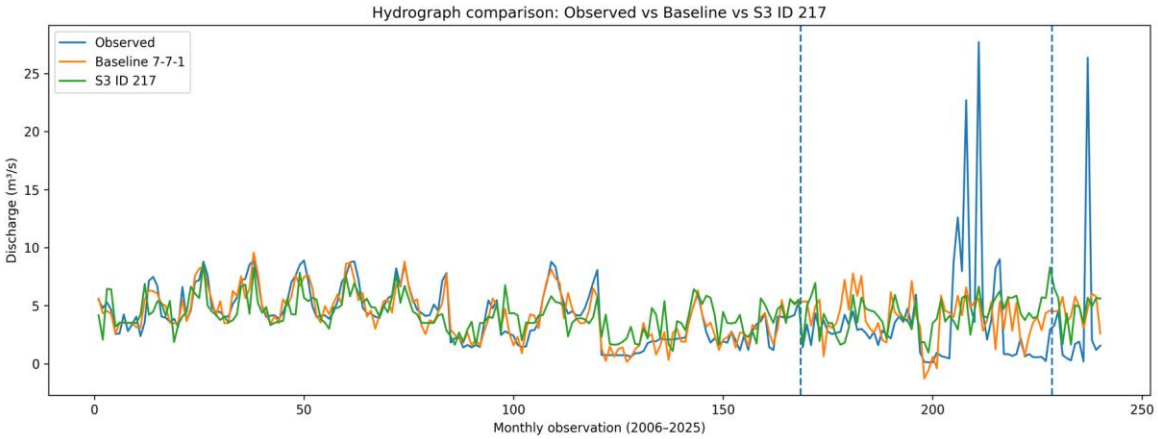

Fig. 8 Observed discharge, baseline FFBPANN, and best S3 EFA-FFBPANN model covering 2006–2025.

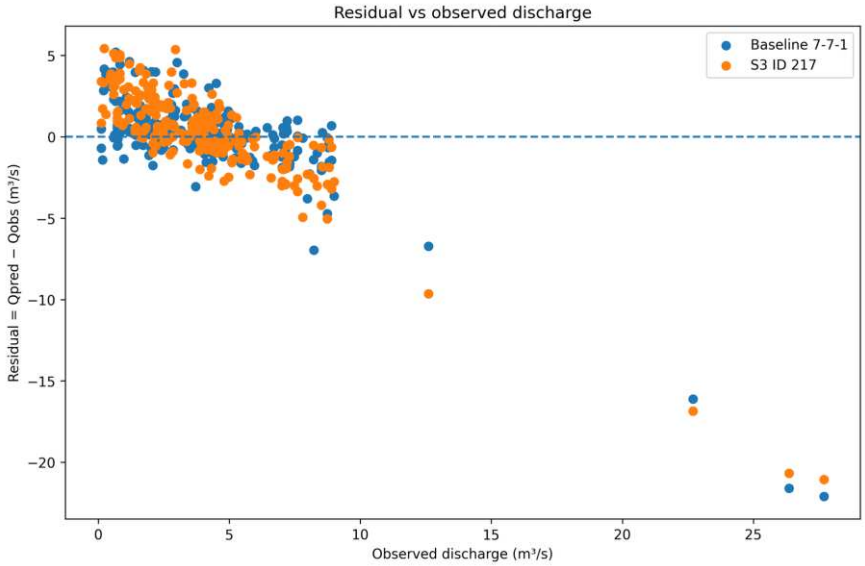

Fig. 9 Residuals versus observed discharge for the baseline and S3 models.

Table 10 Flow-dependent residual comparison between the baseline and S3 models

Period	Flow class	Baseline bias (m³/s)	S3 bias (m³/s)	Baseline RMSE (m³/s)	S3 RMSE (m³/s)
All 240	Low	+1.256	+2.138	2.125	2.610
All 240	Medium	+0.461	+0.529	1.298	1.432
All 240	High	−1.375	−2.082	4.243	4.377
2020–2024	Low	+2.172	+3.187	2.889	3.487
2020–2024	Medium	+1.079	+0.798	1.986	1.731
2020–2024	High	−5.812	−5.719	9.085	8.918
2025	Low	+3.497	+3.022	3.668	3.221
2025	High	−10.808	−9.804	15.270	14.639

S3 was not uniformly worse because the bias and RMSE were slightly smaller than the trend recorded for baseline in the validation high-flow class. Both the low-flow bias and the high-

flow error were also reduced in 2025. The improvements were significantly small to offset the lower training and validation skill.

3.7 Sensitivity to extreme months and variability between candidate networks

A total of three checks were used to qualify the validation ranking as presented in Table 11. First, the training discharge ranged from 0.59 to 8.90 m^3/s with a standard deviation of 2.22 m^3/s while the validation period contained four values above 8.90 m^3/s with a maximum of 27.7 m^3/s in 2023 and eight values under 0.59 m^3/s. The 2025 period had one value above and three under the training range. The standard deviations of discharge were also 4.75 m^3/s in validation and 7.25 m^3/s in 2025. The absence of the four validation values above the training range (n = 56) reduced the validation NSE to −0.833 for the baseline and −0.918 for S3. Second, only four of the 200 S2 networks, 18 of the 200 S3 networks, and one of the 200 S4 networks had a positive validation NSE with medians of −0.017, −0.005 and −0.045. Third, the order S3 > S2 > S4 was consistent but weak. The probability of a random S3 network surpassing a random S2 network was 0.62 and a random S2 network leading a random S4 network was 0.73. This was observed from the Mann–Whitney ($p < 0.001$) reported descriptively because the runs shared the same data and were not independent. The positive 2025 NSE was reached by 63% of the S2, 77% of the S3, and 60% of the S4 networks. This led to the confirmation that a positive 2025 NSE was not evidence of skill.

Table 11 Sensitivity of the validation ranking to extreme months and network spread

Check	Result
Training discharge range	0.59–8.90 m^3/s (SD 2.22)
Validation values outside the training range	4 above, 8 below (maximum 27.7 m^3/s, 2023); SD 4.75
2025 values outside the training range	1 above, 3 below; SD 7.25
Validation NSE without the 4 values above the training range (n = 56)	Baseline −0.833; S3 −0.918
Networks with positive validation NSE (of 200)	S2 4; S3 18; S4 1
Median validation NSE	S2 −0.017; S3 −0.005; S4 −0.045
Networks with positive 2025 NSE	S2 63%; S3 77%; S4 60%
Probability S3 > S2; S2 > S4 (Mann–Whitney)	0.62; 0.73 ($p < 0.001$)

406 SD, standard deviation; NSE, Nash–Sutcliffe efficiency; S2, S3, and S4, factor-based input scenarios.
407 The probabilities are descriptive because the runs share the same data and are not independent.

408 **3.8 Discussion and limitations**

409 ***3.8.1 The generalization gap, non-stationarity, and temporal information***

410 The reduction in NSE from 0.834 in training to 0.088 in validation and −0.018 in 2025 showed
411 that the network learned a relationship not transferred across time. This was the concern raised
412 in the neural-network hydrology literature where strong calibration statistics could coexist with
413 limited predictive reliability when the validation sample contained conditions poorly
414 represented in calibration [17, 18] and where a genuine out-of-sample test was needed [12, 55].
415 The training record spanned 0.59–8.90 m^3/s while later years contained significantly higher
416 values and non-stationary streamflow was known to degrade data-driven prediction [56]. This
417 showed the need to interpret the validation NSE of 0.088 as the best of the tested configurations
418 rather than evidence of a robust model. The trend was supported by a few extreme months as
419 presented in the sensitivity analysis in Table 11. The final period was also maintained at locked
420 because the method applied in dividing the series changed the reported test performance [32].
421 The possible explanation for the poor transfer which was not tested in this study was the
422 absence of temporal memory. The network received and mapped contemporaneous monthly
423 variables directly to discharge while the runoff also depended on soil moisture, groundwater
424 and channel storage, antecedent rainfall, and previous flow which were not analyzed as
425 presented in Fig. 10. Lag effects persisted in neural rainfall–runoff models except for where
426 temporal dependence was represented explicitly [39]. The reviews of flood forecasting also
427 reported that sequence-aware and hybrid methods often outperformed stand-alone feed-
428 forward networks [1, 3]. Furthermore, larger hidden layers were not supportive within the range
429 of 7–10 neurons tested for the baseline in line with a limit set by the information supplied rather
430 than the network width. The weak same-month correlation between rainfall and discharge (r =

0.13 in training) showed similar trend even though the measurement and regulation effects at the gauge possibly contributed as discussed in the Limitations section.

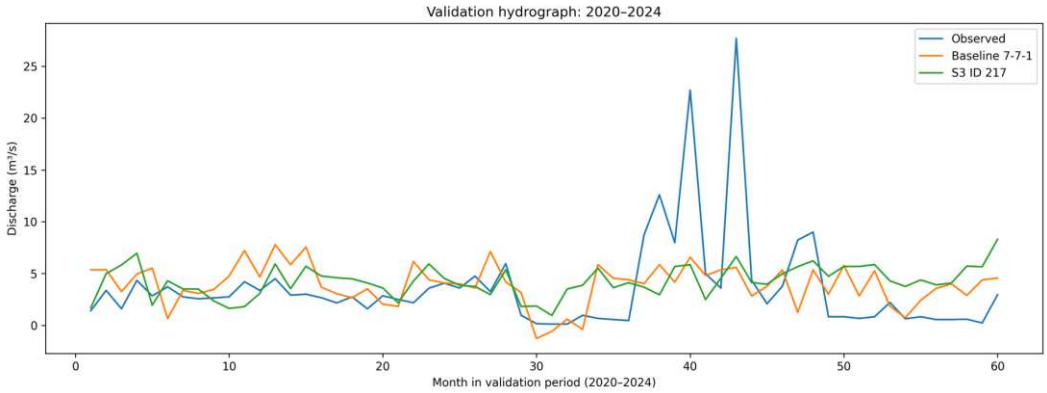

Fig. 10 Validation-period hydrograph of the baseline and S3 models for 2020–2024.

3.8.2 Covariance sufficiency is not predictive sufficiency

The two factors effectively summarized the covariance of the predictors. Factor 1 described wet conditions and factor 2 focused thermal–evaporative conditions. Both were able to contribute 53.02% of the total variance as common with the KMO and Bartlett tests supporting factorability. The factor-only network S2 with a validation NSE of 0.025 and training NSE of 0.256 was weaker than the baseline. This was consistent with the compression removing information that the network needed even though the difference was small relative to the spread between the runs as subsequently discussed. The representation that reproduces shared covariance does not need to retain the factors predicting discharge [27, 30] and Wallis [26] has also previously described factor analysis in hydrology as a descriptive device rather than a predictor-screening tool.

Wind speed was identified as the instructive case with a communality of 0.075 which led to the treatment within the factor model as nearly unique variance. The restoration in S3 increased the training NSE from 0.256 to 0.441, the validation NSE from 0.025 to 0.059, and the 2025 NSE from −0.017 to 0.079. This showed the possibility of a variable weakly shared with the other predictors to carry target-relevant information. The same reasoning explained the reason

the optimal input set identified by information-based methods was not required to coincide with the set that explained the most covariance [19–21]. The addition of evaporation to S3 did not provide any assistance and the compression could not be repaired simply by restoring the variables. The trends identified were in agreement with principal-component studies that reported gains in some set-ups and variable results in others [24, 25]. The scale of the effect was also important because the differences between scenarios were small and tended not to be useful in practice.

3.8.3 Selection bias and the weight of small differences

Every reported best network was the maximum of 200 candidates and 160 for the baseline. The maximum of several noisy estimates overstated the true performance of the winner [51, 52]. Table 11 shows how little separates the scenarios with only four, 18, and one networks recording a positive validation NSE, the medians were negative, and the probability that S3 surpassed S2 was 0.62. The trend reflected that the S3 > S2 > S4 ordering was only indicative. A positive 2025 NSE was identified in 60–77% of networks and this showed the single positive S3 value of 0.079 in 2025 did not have weight to serve as evidence of improvement. The phenomenon was the reason for the more importance associated with the locked test period and the second assessment without September 2025 compared to the headline scores. The stochastic variability between nominally successful networks was also part of the uncertainty of neural flood forecasts [48, 57].

3.8.4 Extremes and implications for flood prediction

The September 2025 peak of 26.365 m³/s was predicted as 4.771 m³/s by the baseline and 5.690 m³/s by S3. A feed-forward network trained on normalized data could not readily exceed the largest output in the training set [36] and the maximum was only 8.90 m³/s. The trend showed that part of the failure was related to data-range problem. The same limit was identified for stand-alone LSTM networks [37] but other experiments reported the possibility of the models

476 to remain relatively accurate on extremes without training [38]. This suggested that the cause
477 was in architecture and information rather the fixed property of neural models. The
478 underestimation of the largest flows was also reported for other neural streamflow forecasts
479 [54]. Moreover, the training with the MSE was discovered to support predictions near the
480 conditional mean when extremes were rare [58]. A model with this peak attenuation should not
481 be used independently for flood warning and could serve as a statistical reference for normal
482 flows or as a component of a hybrid framework.
483 The temporal holdout protected against a false conclusion because the positive S3 NSE could
484 have been read as an improvement had 2025 been used in the tuning process. The locked period
485 showed that the all-month gain depended on one extreme observation and the remaining
486 months remained poorly reproduced. This was despite the fact that S3 ranked the months fairly
487 better than the baseline (Spearman 0.769 against 0.357, n = 11). The trend agreed with the
488 perception that validation should reflect the intended use [18].

489 ***3.8.5 Contribution relative to earlier studies***

490 The results agreed with the established literature that feed-forward networks learned nonlinear
491 hydroclimatic relationships but were sensitive to input selection, data division, preprocessing,
492 architecture, training, and validation [12, 17, 18]. The trends identified also supported the
493 emphasis on input determination and temporal information [19, 39]. Previous studies that
494 reported only the best aggregate statistic were qualified. Moreover, repeated stochastic runs
495 exposed large variability, 2025 was never used for tuning, and the residual analysis showed
496 that a near-zero NSE could coexist with severe flow-dependent bias.
497 The main contribution is to show that covariance reduction and predictive sufficiency are
498 different properties under leakage-controlled factor estimation and a locked temporal test rather
499 than claiming EFA improves FFBPANN performance. It also reflects that retaining a variable
500 poorly represented by the factors can partly restore predictive information. The workflow offers

a simple checklist for practitioners working in data-scarce watersheds to evaluate input reduction by estimating every transformation on the training period only, keeping a final period locked, reporting the spread among repeated runs, and examining errors based on flow regime as well as with and without the dominant extreme.

3.8.6 Limitations and future studies

Several limitations were identified which included the adoption of monthly data when flood peaks were recorded daily or at shorter scales. The few extreme discharges in the calibration record also caused difficulty in determining the upper tail with an unweighted squared-error objective. The predictors did not contain antecedent storage or lagged discharge and the EFA was developed to determine shared structure between the predictors rather than maximize target information. The 2025 test also focused on 12 months which were considered stringent but small. The discharge was obtained from a staff-gauge station and the foundational stage–discharge conversion for the monthly values was not evaluated.

The large dry-season value of 27.7 m^3/s in July 2023 as well as the weak same-month correlation between rainfall and discharge suggested that measurement and regulation effects could contribute to the unexplained variance. Moreover, the candidate pools differed between the baseline with 160 networks and 7–10 hidden neurons and the factor-based scenarios with 200 networks each and 5–10 hidden neurons. Linear, seasonal-climatology, or other machine-learning benchmarks that were needed to establish the level of skill any of the tested networks added were not included.

The next stage is expected to add temporal structure before network depth in the form of lagged configurations of t, t − 1, t – 2, and possibly t − 3 months with the same chronological division as well as all normalization and factor parameters estimated from 2006–2019 only. The loss function and target transformation also need to be tested for sensitivity to high flows such as with a high-flow-weighted or multi-objective formulation [58]. Moreover, there is a need to

evaluate improvement based on peak magnitude and timing, flow-regime bias, and skill after removing the dominant extreme event rather than only NSE. Finer temporal resolution, additional watersheds, and comparison with other input-reduction methods are also required to test the generalization of the results.

4 Conclusions

In conclusion, this study aimed to determine the ability of compressing hydroclimatic predictors with EFA to improve the temporal generalization of a feed-forward neural network for monthly discharge prediction in the Tukad Badung watershed. The results showed that the framework did not cause any improvement under a chronological design with a locked 2025 test period. The selected 7-7-1 baseline which was the best of 160 networks fitted the training period effectively with an NSE of 0.834 but reached only 0.088 in 2020–2024 and −0.018 in 2025 with RMSE increasing from 0.901 to 4.493 and 7.006 m^3/s. A validation NSE of 0.088 was only modestly above the mean-flow benchmark of −0.016 which led to the identification of the network as the best configuration tested and not a robust predictor. The associated errors were observed to be systematic because low flows were overestimated, and high flows underestimated while the 2020–2024 high-flow bias reached −5.812 m^3/s.

The training-only EFA extracted two factors which represented 53.02% of the total variance (KMO = 0.670) while the factor-only scenario S2 lowered validation NSE to 0.025. The trend showed that two latent dimensions discarded the information needed by the network. Restoring wind speed in S3 which was a variable with a communality of only 0.075 also increased validation NSE to 0.059 and reduced the 2025 RMSE from 7.006 to 6.662 m^3/s while the addition of evaporation in S4 did not have any effect as observed with the validation NSE of 0.029.

The differences identified were indicative and in conclusive because each reported network was the best of 160 or 200 candidates. The results showed that only 18 of the 200 S3 networks

had a positive validation NSE. The exclusion of four validation months above the training range of 0.59–8.90 m^3/s reduced the validation NSE to −0.833 for the baseline and −0.918 for S3. The positive 2025 NSE of 0.079 for S3 also depended on one month by reducing to −4.919 when September 2025 was excluded against −5.904 for the baseline. Both models captured only about a fifth of the 26.365 m^3/s peak as observed in the 21.6% for S3 and 18.1% for the baseline.

The central contribution of this study was both conceptual and empirical. For example, the reduction of hydroclimatic covariance did not preserve the information required to predict discharge. This supported treating covariance and predictive sufficiencies as distinct properties in neural hydrological modelling. Therefore, factor analysis was considered best for the determination of the structure among predictors. Variables that were weakly represented by the factors should not be discarded without a predictive test. The methodological contribution was that the combination of training-only factor extraction, chronological validation, a locked test period, repeated stochastic runs, and flow-regime diagnostics exposed limitations a single aggregate score could have hidden. The results did not support operational flood forecasting with the present monthly formulation. Future work should maintain the same chronological design while adding lagged hydroclimatic predictors, antecedent rainfall and antecedent discharge where operationally appropriate. The loss functions providing significant weight to high flows should also be examined.

Declarations

Availability of data and materials

The datasets used and analyzed during this study, including the normalized input workbook, the MATLAB scripts, and the result tables, are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through the State University Operational Assistance Program and the Implementation Contract between the Higher Education Service Institution Region VIII and Ngurah Rai University under the Regional Fundamental Research (PFR) scheme for fiscal year 2026. Contract numbers: LLDIKTI 135/C3/DT.05.00/PL-REGULER LANJUTAN/2026; 695/DST/LL8/DT.05.00/2026; and 0323/1.02.01/1.09/2026.

Authors' contributions

PDHA contributed to conceptualization, methodology, investigation, analysis, and writing including review and editing. IMDS conducted data curation, validation, and supervision. IPS contributed to data collection, survey and investigation, supervision and validation. All authors read and approved the final manuscript.

Declaration of AI

The authors acknowledge the use of artificial intelligence (AI)-assisted tools solely for language editing and improvement of manuscript clarity. These tools were utilized to refine the English expression and organization of the text. All scientific content, data analysis, interpretation, and conclusions presented in this study were developed and verified by the authors. The authors take full responsibility for the integrity and accuracy of the work.

Acknowledgements

The authors appreciate the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for funding this study. Ngurah Rai University, the project team, and all other parties who supported the completion are also acknowledged.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Asif M, Kuglitsch MM, Pelivan I, Albano R (2025) Review and intercomparison of machine learning applications for short-term flood forecasting. *Water Resour Manag* 39:1971–1991. https://doi.org/10.1007/s11269-025-04093-x

2. Jehanzaib M, Ajmal M, Achite M, Kim T-W (2022) Comprehensive review: advancements in rainfall–runoff modelling for flood mitigation. *Climate* 10(10):147. https://doi.org/10.3390/cli10100147

3. Mosavi A, Ozturk P, Chau K-W (2018) Flood prediction using machine learning models: literature review. *Water* 10(11):1536. https://doi.org/10.3390/w10111536

4. Tosan M, Nourani V, Kisi O, Dąbrowska D (2025) Evolution of ensemble machine learning approaches in water resources management: a review. *Earth Sci Inform* 18:416. https://doi.org/10.1007/s12145-025-01911-z

5. Prawesti I, Sari DN, Udrekh U, Priyana Y (2026) Flood inundation analysis on land cover in the Tukad Badung watershed, Bali Province: a case study of the September 2025 Bali floods. *E3S Web Conf* 738:02005. https://doi.org/10.1051/e3sconf/202673802005

6. Adikari KE, Shrestha S, Ratnayake DT, Budhathoki A, Mohanasundaram S, Dailey MN (2021) Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions. *Environ Model Softw* 144:105136. https://doi.org/10.1016/j.envsoft.2021.105136

622 7. Essam Y, Huang YF, Ng JL, Birima AH, Ahmed AN, El-Shafie A (2022) Predicting
623 streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms.
624 *Sci Rep* 12:3883. https://doi.org/10.1038/s41598-022-07693-4
625 8. Parisouj P, Mohebzadeh H, Lee T (2020) Employing machine learning algorithms for
626 streamflow prediction: a case study of four river basins with different climatic zones in the
627 United States. *Water Resour Manag* 34(13):4113–4131. https://doi.org/10.1007/s11269-020-
628 02659-5
629 9. Aqil M, Kita I, Yano A, Nishiyama S (2007) Neural networks for real time catchment flow
630 modeling and prediction. *Water Resour Manag* 21:1781–1796.
631 https://doi.org/10.1007/s11269-006-9127-y
632 10. Priscillia S, Schillaci C, Lipani A (2021) Flood susceptibility assessment using artificial
633 neural networks in Indonesia. *Artif Intell Geosci* 2:215–222.
634 https://doi.org/10.1016/j.aiig.2022.03.002
635 11. Kardhana H, Valerian JR, Rohmat FIW, Kusuma MSB (2022) Improving Jakarta's
636 Katulampa Barrage extreme water level prediction using satellite-based long short-term
637 memory (LSTM) neural networks. *Water* 14(9):1469. https://doi.org/10.3390/w14091469
638 12. Dawson CW, Wilby RL (2001) Hydrological modelling using artificial neural networks.
639 *Prog Phys Geogr* 25(1):80–108. https://doi.org/10.1177/030913330102500104
640 13. Sahoo GB, Ray C (2006) Flow forecasting for a Hawaii stream using rating curves and
641 neural networks. *J Hydrol* 317:63–80. https://doi.org/10.1016/j.jhydrol.2005.05.008
642 14. Wu JS, Han J, Annambhotla S, Bryant S (2005) Artificial neural networks for forecasting
643 watershed runoff and stream flows. *J Hydrol Eng* 10(3):216–222.
644 https://doi.org/10.1061/(ASCE)1084-0699(2005)10:3(216)

15. Yaseen ZM, Sulaiman SO, Deo RC, Chau K-W (2019) An enhanced extreme learning machine model for river flow forecasting. *J Hydrol* 569:387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069

16. Zounemat-Kermani M, Kisi O, Rajaee T (2013) Performance of radial basis and LM-feed forward artificial neural networks for predicting daily watershed runoff. *Appl Soft Comput* 13(12):4633–4644. https://doi.org/10.1016/j.asoc.2013.07.007

17. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables. *Environ Model Softw* 15(1):101–124. https://doi.org/10.1016/S1364-8152(99)00007-9

18. Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems. *Environ Model Softw* 25(8):891–909. https://doi.org/10.1016/j.envsoft.2010.02.003

19. Bowden GJ, Dandy GC, Maier HR (2005) Input determination for neural network models in water resources applications. Part 1. *J Hydrol* 301:75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021

20. Fernando TMKG, Maier HR, Dandy GC (2009) Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach. *J Hydrol* 367:165–176. https://doi.org/10.1016/j.jhydrol.2008.10.019

21. Leahy P, Kiely G, Corcoran G (2008) Structural optimisation and input selection of an artificial neural network for river level prediction. *J Hydrol* 355:192–201. https://doi.org/10.1016/j.jhydrol.2008.03.017

22. Snieder E, Abogadil K, Khan UT (2020) A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *J Hydrol* 583:124299. https://doi.org/10.1016/j.jhydrol.2019.124299

669 23. Wang Q, Liu Y, Yue Q, Zheng Y, Yao X, Yu J (2020) Impact of input filtering and
670 architecture selection strategies on GRU runoff forecasting: a case study in the Wei River
671 Basin, Shaanxi, China. *Water* 12(12):3532. https://doi.org/10.3390/w12123532
672 24. Hu T, Wu F, Zhang X (2007) Rainfall–runoff modeling using principal component analysis
673 and neural network. *Hydrol Res* 38(3):235–248. https://doi.org/10.2166/nh.2007.010
674 25. Zhang J, Chen X, Khan A, Zhang Y-K, Kuang X, Liang X, Taccari ML, Nuttall J (2021)
675 Daily runoff forecasting by deep recursive neural network. *J Hydrol* 596:126067.
676 https://doi.org/10.1016/j.jhydrol.2021.126067
677 26. Wallis JR (1968) Factor analysis in hydrology: an agnostic view. *Water Resour Res*
678 4(3):521–527. https://doi.org/10.1029/WR004i003p00521
679 27. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of
680 exploratory factor analysis in psychological research. *Psychol Methods* 4(3):272–299.
681 https://doi.org/10.1037/1082-989X.4.3.272
682 28. Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken.
683 https://doi.org/10.1002/047174882X
684 29. Fisher RA (1922) On the mathematical foundations of theoretical statistics. *Philos Trans R*
685 *Soc Lond A* 222:309–368. https://doi.org/10.1098/rsta.1922.0009
686 30. Jolliffe IT (1982) A note on the use of principal components in regression. *J R Stat Soc Ser*
687 *C Appl Stat* 31(3):300–303. https://doi.org/10.2307/2348005
688 31. Chen J, Zheng F, May R, Guo D, Gupta H, Maier HR (2022) Improved data splitting
689 methods for data-driven hydrological model development based on a large number of
690 catchment samples. *J Hydrol* 613:128340. https://doi.org/10.1016/j.jhydrol.2022.128340
691 32. Shen H, Tolson BA, Mai J (2022) Time to update the split-sample approach in hydrological
692 model calibration. *Water Resour Res* 58(3). https://doi.org/10.1029/2021WR031523

33. McCuen RH, Knight Z, Cutter AG (2006) Evaluation of the Nash–Sutcliffe efficiency index. *J Hydrol Eng* 11(6):597–602. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597)

34. Knoben WJM, Freer JE, Woods RA (2019) Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol Earth Syst Sci* 23:4323–4331. https://doi.org/10.5194/hess-23-4323-2019

35. Seibert J, Vis MJP, Lewis E, van Meerveld HJ (2018) Upper and lower benchmarks in hydrological modelling. *Hydrol Process* 32:1120–1125. https://doi.org/10.1002/hyp.11476

36. Hettiarachchi P, Hall MJ, Minns AW (2005) The extrapolation of artificial neural networks for the modelling of rainfall–runoff relationships. *J Hydroinform* 7(4):291–296. https://doi.org/10.2166/hydro.2005.0025

37. Baste S, Klotz D, Acuña Espinoza E, Bardossy A, Loritz R (2025) Unveiling the limits of deep learning models in hydrological extrapolation tasks. *Hydrol Earth Syst Sci* 29:5871. https://doi.org/10.5194/hess-29-5871-2025

38. Frame JM, Kratzert F, Klotz D, Gauch M, Shalev G, Gilon O, Qualls LM, Gupta HV, Nearing GS (2022) Deep learning rainfall–runoff predictions of extreme events. *Hydrol Earth Syst Sci* 26:3377. https://doi.org/10.5194/hess-26-3377-2022

39. Wu CL, Chau KW (2011) Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *J Hydrol* 399:394–409. https://doi.org/10.1016/j.jhydrol.2011.01.017

40. Abrahart RJ, Anctil F, Coulibaly P, Dawson CW, Mount NJ, See LM, Shamseldin AY, Solomatine DP, Toth E, Wilby RL (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Prog Phys Geogr* 36(4):480–513. https://doi.org/10.1177/0309133312444943

41. BPS Kota Denpasar (2025) Kota Denpasar dalam angka 2025 [Denpasar Municipality in figures 2025], vol 33. Badan Pusat Statistik Kota Denpasar, Denpasar

42. Kassem Y, Gökçekuş H, Hussein AS (2025) Relating urban runoff with soil moisture, temperature, wind speed and vapor pressure using artificial neural network: a case study of Mogadishu in Somalia. In: 16th International Conference on Applications of Fuzzy Systems, Soft Computing and Artificial Intelligence Tools (ICAFS-2023). Springer, Cham. https://doi.org/10.1007/978-3-031-72506-7_10

43. Sankaranarayanan S, Prabhakar M, Satish S, Jain P, Ramprasad A, Krishnan A (2020) Flood prediction based on weather parameters using deep learning. *J Water Clim Change* 11(4):1766–1783. https://doi.org/10.2166/wcc.2019.321

44. Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. *IEEE Trans Neural Netw* 5(6):989–993. https://doi.org/10.1109/72.329697

45. MacKay DJC (1992) Bayesian interpolation. *Neural Comput* 4(3):415–447. https://doi.org/10.1162/neco.1992.4.3.415

46. Máca P, Pech P, Pavlásek J (2014) Comparing the selected transfer functions and local optimization methods for neural network flood runoff forecast. *Math Probl Eng* 2014:782351. https://doi.org/10.1155/2014/782351

47. Bor A, Okan M (2025) Comparison of Levenberg–Marquardt and Bayesian regularization learning algorithms for daily runoff forecasting. *J Innov Sci Eng*. https://doi.org/10.38088/jise.1375510

48. Khan MS, Coulibaly P (2006) Bayesian neural network for rainfall–runoff modeling. *Water Resour Res* 42(7). https://doi.org/10.1029/2005WR003971

49. Kingston GB, Lambert MF, Maier HR (2005) Bayesian training of artificial neural networks used for water resources modeling. *Water Resour Res* 41(12). https://doi.org/10.1029/2005WR004152

742 50. DiStefano C, Zhu M, Mindrila D (2009) Understanding and using factor scores. *Pract*
743 *Assess Res Eval* 14(20). https://doi.org/10.7275/da8t-4g52
744 51. McLatchie Y, Vehtari A (2024) Efficient estimation and correction of selection-induced
745 bias with order statistics. *Stat Comput* 34:132. https://doi.org/10.1007/s11222-024-10442-4
746 52. Zrnic T, Fithian W (2025) A flexible defense against the winner's curse. *Ann Stat*
747 53(6):2516–2535. https://doi.org/10.1214/25-aos2553
748 53. Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models, Part I. *J*
749 *Hydrol* 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6
750 54. Granata F, Di Nunno F (2023) Neuroforecasting of daily streamflows in the UK for short-
751 and medium-term horizons: a novel insight. *J Hydrol* 624:129888.
752 https://doi.org/10.1016/j.jhydrol.2023.129888
753 55. Braddock RD, Kremmer ML, Sanzogni L (1998) Feed-forward artificial neural network
754 model for forecasting rainfall-runoff. *Environmetrics* 9:419–432.
755 https://doi.org/10.1002/(SICI)1099-095X(199807/08)9:4<419::AID-ENV312>3.0.CO;2-D
756 56. Meng E, Huang S, Huang Q, Fang W, Wu L, Wang L (2019) A robust method for non-
757 stationary streamflow prediction based on improved EMD-SVM model. *J Hydrol* 568:462–
758 478. https://doi.org/10.1016/j.jhydrol.2018.11.015
759 57. Han D, Kwong T, Li S (2007) Uncertainties in real-time flood forecasting with neural
760 networks. *Hydrol Process* 21:223–228. https://doi.org/10.1002/hyp.6184
761 58. de Vos NJ, Rientjes THM (2008) Multiobjective training of artificial neural networks for
762 rainfall-runoff modeling. *Water Resour Res* 44(8):W07405.
763 https://doi.org/10.1029/2007WR006734