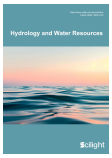
Hydrology and Water Resources
https://www.sciltp.com/journals/hwr



Article

Predicting Streamflow Regimes in Ungauged Catchments with Process-Informed Machine Learning

Hongxing Zheng [1,*], Ruirui Zhu [2], Lu Zhang [3] and Francis Chiew [1]

[1] CSIRO Environment, GPO BOX 1666, Canberra, ACT 2601, Australia
[2] Fenner School of Environment and Society, Australian National University, Canberra, ACT 2601, Australia
[3] State Key Laboratory of Water Resources Engineering and Management, Wuhan University, Wuhan 430072, China
* Correspondence: Hongxing.zheng@csiro.au

How To Cite: Zheng, H.; Zhu, R.; Zhang, L.; et al. Predicting Streamflow Regimes in Ungauged Catchments with Process-Informed Machine Learning. *Hydrology and Water Resources* **2026**, *1*(1), 5. https://doi.org/10.53941/hwr.2026.100005

Received: 13 November 2025
Revised: 1 January 2026
Accepted: 9 January 2026
Published: 23 January 2026

Abstract: Predicting daily flow duration curves (FDCs) in ungauged catchments remains a major challenge in hydrology and is critical for effective water resources management. The FDCs typical were predicted by relating FDC parameters or percentiles to catchment properties using statistical or machine learning-based models. Such models often suffer from limited interpretability and transferability across hydroclimatic conditions. In this study, we propose a process-informed, interpretable machine learning framework for predicting daily FDCs by integrating multivariate adaptive regression splines (MARS) with the Budyko theory, which provides a physically based representation of long-term water–energy constraints on catchment behaviour. Assuming the FDC follows a log-normal distribution determined by three parameters, MARS is used to explore relationships between FDC parameters and 19 catchment characteristics using data from 347 catchments across Australia. Results indicate that the proposed framework can satisfactorily predict the mean and deviation of daily streamflow, while prediction skill for ratio of non-zero flow days remains comparatively weaker. Incorporating hydrological constraints through Budyko theory improves the physical interpretability and robustness of model predictions, particularly in revealing the dominant hydroclimatic controls on streamflow regimes. Prediction performance is found generally higher in wetter catchments than in drier ones, mainly due to limitations of the models in predicting non-zero flow ratio and the lognormal assumption. To further improve FDC prediction in ungauged catchments, catchment characteristics more closely related to groundwater processes may be required, in addition to the adoption of more advanced modelling approaches.

Keywords: flow duration curve; process-informed machine learning (PIML); MARS model; prediction in ungauged basins (PUBs)

1. Introduction

Streamflow regime represents the temporal distribution and variability of the flow in the river and plays an important role in shaping river channels, structuring aquatic ecosystems and affecting water use. The streamflow regime generally can be characterized by its magnitude, frequency, duration and predictability. One convenient and effective way to describe streamflow regime is the use of flow duration curve (FDC), which represents the relationship between the magnitude and frequency of streamflow for a given catchment and provides a statistical method for illustrating streamflow characteristics. FDCs are widely used in hydrology to identify difference or changes of the streamflow, to configure instream flow requirement, to inform water allocation decisions, and to determine the capacity of a hydropower plant [1–6].

Flow duration curves for catchments with long-term daily streamflow observation can be derived by estimating the parameters of a candidate distribution function [7]. It becomes challenging to predict the FDC where observed daily streamflow is unavailable (ungauged basins) or in highly regulated catchments [8–12]. Numerous efforts have been made to address this challenge and different methods have been developed, some of which are summarized in Table 1. The methods developed include parametric regression [2], nearest neighbour [13,14], hydrological similarity [15,16], index model [17] and machine learning or deep learning (ML/DL) methods [2,18,19]. A more comprehensive review of FDC prediction in ungauged catchments can be found in [20].

Most existing methods are fundamentally based on developing statistical relationships between FDC parameters and the physical properties of catchments. However, the development of generic relationships between FDC parameters and their predictors is challenged by substantial hydrological heterogeneity across catchments. As an alternative, some researchers have tried to cluster or group catchments into hydrologically homogeneous regions based on land use, soil types and climate conditions first and then developed relationships for each homogeneous region [21]. A few studies have shown that such an approach can lead to poor model performance [22,23] unless the catchment clustering approach is effective. Moreover, the increasingly used ML/DL methods though have shown strong potential in improving hydrological prediction, they often provide limited physically meaningful interpretation of the relationships between the predictors and the predictands. Process-informed machine learning (PIML) that incorporates established hydrological knowledge with machine learning is expected to provide more robust and physically interpretable models for hydrological predictions [24,25].

Table 1. Summary of recent studies on predicting flow duration curves for ungauged catchments.

Predictor Types	Predictor Numbers	Methods	Study Region	Streamflow Gauges	References
Pedological	30	Region of influence	UK	653	[26]
Climatic, geomorphological, pedological, land use/cover	42	Regression	USA	29	[9]
Climatic, topographic, pedological, hydraulic, land use/cover	12	Statistical Index-model	Australia	227	[17]
Climatic, topographic, hydrological, land cover, water chemical	16	Regression and random forest	New Zealand	379	[2]
Hydrological, topographic, lithologic, hydraulic, pedological	19	Regression	Italy	19	[27]
Climatic, topographic, hydrogeological	12	Process-based modelling and regression	Nepal	25	[28]
Climatic, topographic, hydrological, land use/cover	8	ML-based regression (ANN)	Canada and USA	260	[18]
Climatic, topographic, hydrological, land use/cover	11	Functional multiple regression	Canada	109	[29]
Climatic, topographic, hydrological	10	Evolutionary polynomial regression	Brazil	11	[30]
Climatic, topographic, hydrological, pedological, hydrogeological, land use/cover	23	ML-based regression	USA	918	[19]
Climatic, hydrological, topographic,	24	Geostatistical model	Brazil	81	[31]
Climatic, topographic, hydrological, land use/cover	8/11	3-D kriging interpolation	Italy	41	[15]
Climatic, topographic, pedological, land use/cover,	11	Process-based modelling	USA	201	[32]
Climatic, topographic, pedological, land use/cover	19	Process-informed machine learning	Australia	347	This study

In this paper, a process-informed machine learning (PIML) approach is proposed to provide more robust and interpretable prediction of FDCs in ungauged or highly regulated catchments. The idea is to establish relationships between FDC parameters and catchment characteristics by combining the multivariate adaptive regression splines (MARS) [33] with the Budyko framework [34]. The MARS approach is an interpretable machine learning algorithm integrating clustering and multivariate regression, and has been proven to be an effective tool in handling forecasting and classification problems [35–37]. The Budyko framework, on the other hand, is a well-established hydroclimate relationship which can be used to predict mean annual runoff from mean annual precipitation and potential evapotranspiration. The combination of MARS and Budyko framework is expected to provide more robust and hydrologically interpretable relationships between the FDC parameters and catchment characteristics by capturing intrinsic complicated data structure and hydroclimatic characteristics.

2. Method

2.1. Probability Distribution of Flow Duration Curve

The daily flow duration curve can be parameterized by different probability distribution functions, such as lognormal distribution [17,38], log-logistic distribution [39] and the extended three-parameter Burr XII system [40]. There is no general agreement on the selection of the distribution functions in the literature, and the most appropriate function may depend on the region. In this study, we use the lognormal distribution to represent the FDC. The form of lognormal distribution has some variants in hydrology owing to the characteristic of the random variables and the purpose of the application. Herein, the three-parameter lognormal distribution is used and expressed as:

$$Q(p,\mu,\sigma.\tau) = \begin{cases} exp\left[\sigma\Phi^{-1}\left(1-\frac{p}{\tau}\right)+\mu\right], 0 \le p \le \tau \\ 0, \qquad\qquad\qquad p > \tau \end{cases} \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function; μ and σ are mean and standard deviation of the logarithms of the non-zero daily streamflow respectively, which describe the central tendency and variation of non-zero flows. The parameter τ is a ratio of non-zero flow days to total flow days and it equals 1 for perennial catchments and less than 1 for ephemeral ones. When observed daily streamflow data is available, the maximum likelihood estimation of the distribution parameters μ and σ, and τ is:

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} q_i \tag{2}$$

$$\hat{\sigma} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(q_i-\hat{\mu})^2} \tag{3}$$

$$\tau = {}^{m}/_{n} \tag{4}$$

where q_i (=$logQ_i$) is the logarithm of non-zero streamflow Q_i, m is the number of non-zero flow days, and n is the number of total days of the streamflow time series. The FDC parameters estimated using these equations are considered as benchmark and used to develop relationships between the FDC parameters and catchment characteristics, which can then be used to predict the FDCs for ungauged catchments. It is noted that the lognormal distribution parameters (μ and σ) of an FDC can be expressed by mean ($\bar{Q}$) and standard deviation (s) of daily streamflow respectively as [41]:

$$\hat{\mu} = log\left[\frac{\bar{Q}^2}{\sqrt{s^2+\bar{Q}^2}}\right] \tag{5}$$

$$\hat{\sigma}^2 = log\left[1+\frac{s^2}{\bar{Q}^2}\right] \tag{6}$$

Hence, FDC prediction relies on the estimation of the three parameters including streamflow mean $\bar{Q}$, streamflow deviation s and the ratio of non-zero flow days τ.

2.2. FDC Prediction Using Process-Informed Machine Learning

To predict FDCs of ungauged catchments, the multivariate adaptive regression splines (MARS) proposed by [33] is adopted in this study. MARS is an interpretable machine learning approach, which combines clustering and regression algorithm and automatically models nonlinearities and interactions between predictors [33]. The form of a MARS model can be expressed as a weighted sum of basis functions $B_i(x)$ with c_i as the weights:

$$f(x) = \sum_{i=1}^{k} c_i B_i(x) \tag{7}$$

The basis function $B_i(x)$ can be a constant (interception), a hinge function or the product of hinge functions:

$$B(x) = \prod_{j=1}^{m} h_j(x_j - u_j) \tag{8}$$

The hinge function is the critical part of the MARS and is defined as:

$$h(x-u) = \max(0, x-u) \tag{9}$$

where the constant u is called knot. The knot of a variable can be regarded as an inflection point along the range of a predictor. Given a knot value, the data x is partitioned into two parts and each is fitted by an independent linear regression function. When more knots are specified for a predictor variable, the MARS model can approximate the complex non-linear relationships between a response and its predictor. A MARS model is built using forward

and backward procedures. In forward procedures, candidate knots can be placed at any position within the range of each predictor to define a pair of basis functions. A knot value will be accepted if it reduces significantly the residual sum of squares. The forward procedure usually builds an overfit model and the backward procedure is to prune the model by removing the least effective basis functions one by one until the best model is found [33,42].

Two different MARS-based models are developed in this study to predict FDCs in ungauged catchments, distinguished by their integration with hydrological process understanding. The first model (MA) is a purely data-driven approach, without explicit process-informed components. It predicts mean ($\bar{Q}$) and standard deviation (s) of daily streamflow and non-zero flow day ratio (τ) from catchment characteristics (X) using MARS as:

$$[\bar{Q}, s, \tau] = MA(X) \tag{10}$$

The second model (MB) is a process-informed one that combines the Budyko framework with MARS to estimate the FDC parameters. The Budyko framework assumes that mean annual evapotranspiration ($\overline{ET}$) from a catchment will approach rainfall ($\bar{P}$) under very dry conditions (water limiting) and potential evapotranspiration ($\bar{E}_0$) under very wet conditions (energy limiting) [34]. The Budyko framework can be expressed in various forms, among which the Fu's equation is one of the most widely adopted [43,44]. In this study, Fu's equation is applied and formulated as:

$$\frac{\overline{ET}}{\bar{P}} = 1 + \varphi - (1 + \varphi^{\alpha})^{1/\alpha} \tag{11}$$

where, $\varphi = \bar{E}_0/\bar{P}$ is the aridity index, α is a model parameter related to catchment characteristics [43,45]. At mean annual time scale, neglecting the changes of water storage, one can assume that $\overline{ET} = \bar{P} - \bar{Q}$, the mean annual streamflow can then be estimated as:

$$\bar{Q} = \bar{P}\left[(1 + \varphi^{\alpha})^{1/\alpha} - \varphi\right] \tag{12}$$

This relationship indicates that mean annual streamflow can be determined from mean annual precipitation and potential evapotranspiration if parameter α is known. In other words, predicting mean annual (or daily) streamflow for ungauged catchments using the Budyko framework is equivalent to predicting the parameter α. Thus, the second model (MB) predicts the parameter α rather than streamflow mean ($\bar{Q}$), daily streamflow deviation (s) and non-zero flow day ratio (τ) from catchment characteristics using MARS as:

$$[\alpha, s, \tau] = MB(X) \tag{13}$$

Once the parameter α of an ungauged catchment is predicted using Equation (13), it is then used to predict mean annual streamflow $\bar{Q}$ from Equation (12) with observed mean annual precipitation and potential evapotranspiration. Similar to MA, parameters of the lognormal distribution can then be determined from Equations (5) and (6). It needs point out that, in each of the two proposed MARS models (MA and MB), all the three parameters are estimated jointly and share the same basis functions.

2.3. Model Evaluation

The MARS models for predicting FDCs involve two main steps: estimating the FDC parameters and generating the FDC based on these estimates. Model performance is therefore evaluated across gauged catchments in terms of (i) accuracy in predicting the FDC parameters and (ii) agreement between the predicted and empirical FDCs. For model evaluation regarding prediction of FDC parameters, the mean annual runoff, standard deviation of non-zero flow and the non-zero flow ratio derived from the gauged catchments are considered as the benchmark values. The performance of the model is then evaluated by comparing the predicted values against the benchmark values, where the goodness-of-fit is illustrated by coefficient of determination (R^2) and bias. Meanwhile, model evaluation regarding agreement between predicted and empirical FDCs is assessed using the Nash-Sutcliffe coefficient of efficiency [46], which is expressed as:

$$NSE = 1 - \frac{\sum_{i=1}^{m}(Q_{sim,i} - Q_{obs,i})^2}{\sum_{i=1}^{m}(Q_{obs,i} - \bar{Q}_{obs})^2} \tag{14}$$

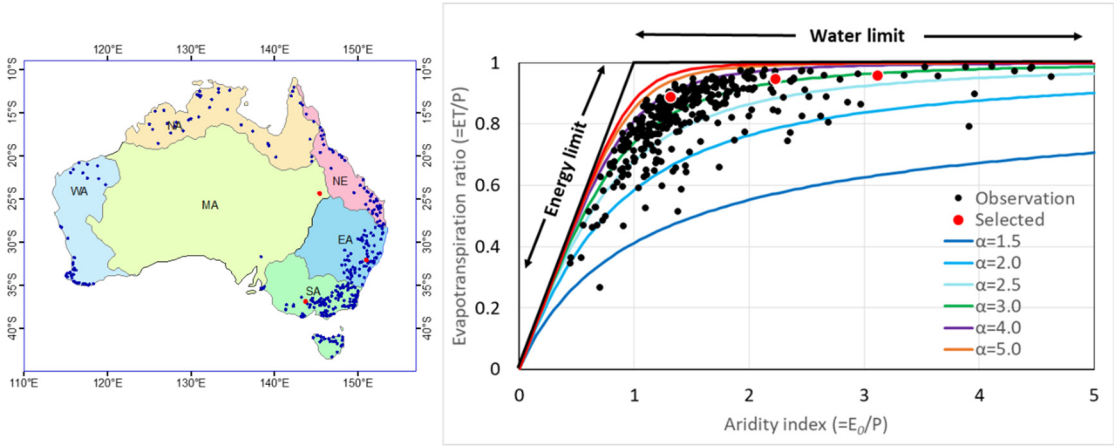
where Q_{sim} and Q_{obs} are predicted and observed streamflow percentiles, m is the number of non-zero flow days, i is the percentile of flow.

Model performance is validated using k-fold cross-validation, where all the gauged catchments are divided randomly into five groups of similar size and the cross-validation is done in five iterations. At each iteration, one of the five groups is assumed to be the "ungauged" catchments, while the remaining four groups are assumed to

be "gauged" catchments and are used to train the MARS models. The predictions of the "ungauged" groups from the five iterations are then pulled together to assess model performance.

3. Data

This study uses daily streamflow data from [47] for 347 catchments across Australia with at least ten years continuous daily streamflow observations over the period 1975–2012 (Figure 1). The streamflow dataset is the most comprehensive Australian streamflow dataset and has been used in many hydrological studies [17,48]. The streamflow data are for relatively unimpaired catchments (i.e., not significantly affected by regulation structures or streamflow extraction). The catchment size ranges from 50 to 60,184 km^2. The catchments cover different hydroclimatic conditions with the aridity index varying from 0.4 to 6.3.



(**a**) (**b**)

Figure 1. Locations of the 347 ungauged catchments used in the study (**a**), and ratio of mean annual evapotranspiration to precipitation (ET/P) as a function of the aridity index (E_0/P) as represented by the Budyko framework (**b**). The three red dots are catchments selected to demonstrate the performance of FDC prediction.

For the purpose of predicting FDCs, observed daily streamflow data from all the selected catchments are first used to estimate the model parameters based on Equations (2) to (4). Table 2 provides a summary of the parameter values for the catchments considered. The observation-based estimates of the FDC parameters are then used to develop their relationships with catchment characteristics using MARS.

Table 2. Summary of catchment characteristics and the observed FDC parameters of the 347 studied catchments.

Variables		Definition	Min	Median	Max
Climate	MDP	Mean daily rainfall (mm/day)	0.8	2.4	8.3
	SDP	Standard deviation of daily rainfall (rain days only)	3.8	6.7	19.5
	MDE	Mean daily potential evapotranspiration (PET, mm/day)	2.5	3.5	5.4
	SDE	Standard deviation of daily potential evapotranspiration	0.9	1.6	1.9
	TRD	Ratio of rain days in a year	0.2	0.6	0.9
	AI	Aridity index (=MDE/MDP)	0.4	1.4	6.3
	MDS	Mean duration of no rain days (days)	1.6	3.1	8.4
	MXS	Maximum duration of no rain days (days)	7.0	28.0	156.0
Location	Latitude	Centroid latitude of the catchment	−43.3	−33.6	−12.3
	Longitude	Centroid longitude of the catchment	114.4	147.3	153.2
	Elevation	Mean elevation of the catchment	37.6	366.9	1351.0
Topography	Area	Catchment area (km^2)	51	363	60184
	DE	Difference between maximum and minimum elevation (m)	8.8	257.8	1180.2
	SDELV	Standard deviation of elevation	2.0	62.3	310.3
	Slope	Mean slope of the catchment (degree)	0.6	11.6	41.0
Land cover	LAI	Leaf area index	0.0	1.6	4.9
	W	Fraction of total native woody vegetation	0.0	3.5	43.0
Soil	PAWC	Plant available water holding capacity (mm)	27.1	89.2	186.0
	BD	Bulk density	0.5	1.4	1.7
FDC parameters	$\bar{Q}$	Mean daily streamflow at natural scale (mm/day)	0.0	0.4	5.3
	s	Standard deviation of daily streamflow at natural scale	0.1	1.4	10.6
	μ	Mean of daily streamflow at logarithmic scale	−7.8	−2.8	1.1
	$\boldsymbol{\sigma}$	Standard deviation of daily streamflow at logarithmic scale	0.7	2.2	5.8
	$\boldsymbol{\tau}$	Non-zero flow ratio	0.1	1.0	1.0
	$\boldsymbol{\alpha}$	Parameter of the Fu's equation (Equation (8))	1.3	3.2	5.3

Nineteen catchment characteristics of climate, location, topography, land cover and soil are obtained for each catchment to develop the FDC prediction model (see Table 2). Daily rainfall and potential evapotranspiration data are obtained from the SILO Data Drill 0.05° (~5 km), which is a gridded climate dataset across Australia [49]. Climate metrics of catchments are derived from the daily climate data for the period same to the streamflow observations. The leaf area index data are obtained from the Advanced Very High Resolution Radiometer (AVHRR) generated by Boston University [50]. The topography data come from [47]. The soil data are obtained from [51], where the plant available water holding capacity (PAWHC) is an indication of effective soil depth, while the bulk density indicates the mineral make up of soil and the degree of compaction.

4. Results

4.1. MARS Models in Predicting FDC Parameters

Table 3 lists the two MARS models (MA and MB) calibrated by using all the 347 catchments, showing the overall relationships between catchment characteristics and the parameters of FDC. For MA, the FDC parameters predicted here include the mean ($\bar{Q}$) and standard deviation (s) of daily non-zero streamflow, and non-zero flow ratio (τ). For MB, the parameters predicted are α, s and τ. Note that the three predictands in Table 3 share the same predictors and basis functions because the predictands in both the MA and MB models are predicted jointly instead of separately by three individual MARS models.

Table 3. Calibrated basis functions and their corresponding coefficients of MARS models for FDC parameter prediction.

Items	MA				MB			
	Basis Functions	$\bar{Q}$	s	τ	**Basis Functions**	α	s	τ
1	*Intercept*	0.482	1.156	0.921	*Intercept*	4.569	−0.316	0.790
2	*h(2.85-MDP)*	−0.259	1.013	−0.109	*h(2.85-MDP)*	−0.158	0.797	−0.100
3	*h(MDP-2.85)*	0.854	−0.097	−0.034	*h(3.66-MDE)*	−0.778	0.385	0.048
4	*h(3.05-MDE)*	1.308	2.305	−0.065	*h(MDE-3.66)*	−0.283	−1.110	−0.188
5	*h(MDE-3.61)*	−0.051	−1.185	−0.226	*h(MDE-4.61)*	−0.791	1.856	−0.082
6	*h(MDE-4.73)*	0.313	2.018	−0.024	*h(7.58-SDP)*	0.160	−0.556	−0.018
7	*h(7.68-SDP)*	−0.030	−0.634	−0.014	*h(SDP-7.58)*	−0.147	0.891	−0.003
8	*h(SDP-7.68)*	−0.045	0.879	−0.001	*h(1.85-SDE)*	0.011	1.747	0.479
9	*h(SDP-10.86)*	0.178	0.553	0.001	*h(PAWC-48.1)*	−0.046	0.064	0.004
10	*h(SDP-12.74)*	−0.114	−0.840	0.024	*h(71.1-PAWC)*	−0.039	0.066	0.003
11	*h(1.8-SDE)*	−0.149	2.513	0.518	*h(PAWC-71.1)*	0.054	−0.074	−0.004
12	*h(74.79-PAWC)*	0.004	0.020	−0.001	*h(1.4-BD)*	−1.259	1.745	−0.171
13	*h(124.1-Area)*	0.001	0.005	−0.001	*h(BD-1.4)*	−1.176	3.069	−0.271
14	-	-	-	-	*h(BD-1.6)*	5.302	−17.788	0.746

Table 3 shows that, for the MA model, the most important catchment characteristics determining FDC parameters are mean daily precipitation and potential evaporation (MDP and MDE), standard deviation of daily precipitation and potential evaporation (SDP and SDE), plant available water capacity (PAWC) and catchment area (Area). For the MB model, most of the salient catchment characteristics are the same as those of the MA scheme except that the catchment area is replaced by bulk density (BD).

However, the knot values (which partition the predictor variables into different parts) and the regression coefficients are different in the two MARS models. In the MA model, there is only one knot for each the predictor MDP, SDE, PAWC and Area, but three knots for MDE and SDP. In the MB model, there is one knot for MDP, SDP and SDE, but two knots for MDE, PAWC and BD. The predictor with more knots indicates that it could lead to more substantial different hydrological response among the catchments categorized by the knot values. For example, in the MA model, according to the knot values of mean daily precipitation (MDP), the catchments are partitioned into two categories: MDP ≥ 2.85 mm and MDP < 2.85 mm. For the catchments with MDP ≥ 2.85mm, the regression coefficient of $\bar{Q}$ against MDP is 0.854, which is much higher than that for the catchments with MDP < 2.85 mm. This indicates that runoff in wetter catchments (higher precipitation) is more sensitive to precipitation change than in drier catchments. It is interesting to note that for wetter catchments (MDP > 2.85mm), the regression coefficients of s and τ against MDP are relatively small, suggesting that streamflow variability (s) in wetter catchments is only weakly related to MDP. In the MB model, it even suggests that the three parameters α, s and τ have no statistically significant correlations with MDP for wetter catchments (MDP > 2.85mm) as the basis function *h(MDP-2.85)* is no longer statistically significant and is not included in the MB model.

4.2. Model Performance in Predicting FDC Parameters

The cross-validation performance of the MARS models in predicting the FDC parameters is shown in Figure 2 and Table 4. As shown in Figure 2, the MARS models predict reasonably well the mean ($\bar{Q}$) and standard deviation (s) of the daily streamflow, where the coefficient of determination (R^2) between the predicted and observed $\bar{Q}$ are about 0.9 for both the MA and MB models, and R^2 between the predicted and observed s are about 0.79 and 0.80 for the MA and MB model, respectively (Table 4). The non-zero flow ratio (τ) is more difficult to predict with the correlation coefficient between predicted and observed values is around 0.5 for both the MA and MB models.

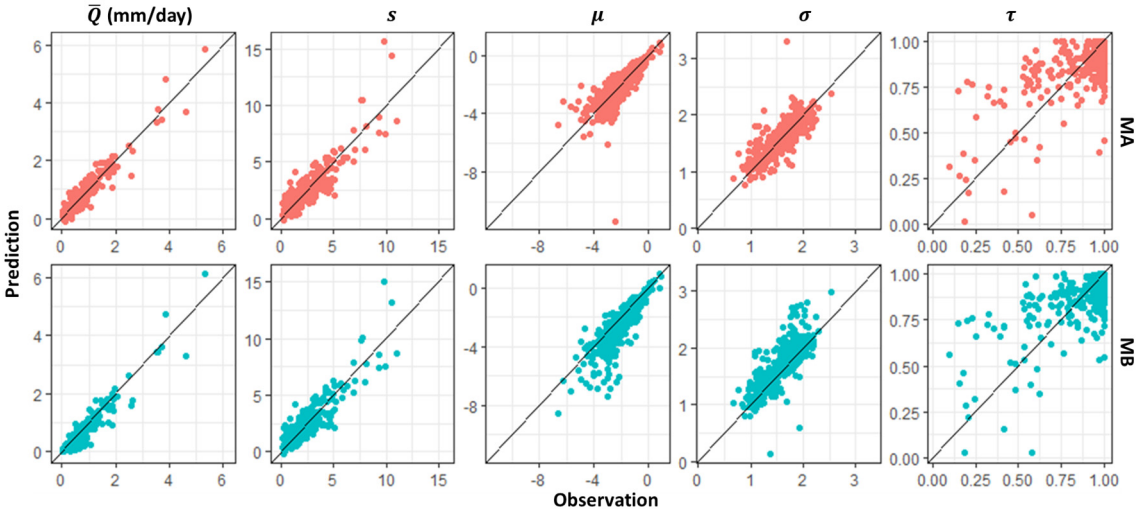


Figure 2. Parameter estimation using the MARS models. The black solid lines are the 1:1 line.

As mentioned earlier, the distribution parameters (μ and σ) of the FDC can be estimated using Equations (5) and (6) when the $\bar{Q}$ and s are known. The results show R^2 between predicted and observed μ of 0.62 and 0.72 respectively for the MA and MB models, and R^2 between predicted and observed σ of 0.61 and 0.59 for the MA and MB models. It is noted that the performance of the MARS models in predicting the distribution parameters (μ and σ) of FDC is not as good as that in predicting $\bar{Q}$ and s. The results suggest that good prediction in $\bar{Q}$ (or s) does not necessarily guarantee good prediction in μ (or σ). This is because μ or σ is determined jointly by both $\bar{Q}$ and s as indicated by Equations (5) and (6). Bias in estimating both $\bar{Q}$ and s propagates into the predicted errors in μ and σ.

Overall, the two MARS models perform similarly in terms of predicting $\bar{Q}$, s and τ. According to coefficients of determination (R^2), MB outperforms MA in predicting μ but underperforms in predicting σ. With respect to mean absolute error (MAE) and root mean squared error (RMSE), MA performs slightly better than MB in predicting both μ and σ. However, since the FDC is determined collectively by the three parameters (μ, σ and τ), the good prediction of an individual parameter is necessary but not a sufficient condition to reasonable FDC prediction. Hence, the performance of the MA and MB models ultimately needs to be assessed with respect to the predicted FDC shown in the following section.

Table 4. Performance of MARS models in predicting FDC parameters (MAE: mean absolute error; RMSE: root mean squared error).

Indices	MA					MB				
	$\bar{Q}$ (mm/day)	s	μ	σ	τ	$\bar{Q}$ (mm/day)	s	μ	σ	τ
MAE	0.15	0.61	0.51	0.16	0.1	0.14	0.6	0.6	0.21	0.09
RMSE	0.21	0.87	0.82	0.22	0.14	0.22	0.84	0.92	0.29	0.14
R^2	0.90	0.79	0.62	0.61	0.48	0.90	0.80	0.72	0.59	0.47

4.3. Model Performance in Predicting FDC

The overall performance of the model in reproducing FDCs for the 347 catchments is shown in Figure 3 for both calibration (i.e., using all catchments in developing MARS models) and 5-fold cross-validation. As a benchmark, the performance of the FDC reproduced by using the observed mean daily streamflow ($\bar{Q}$), standard deviation (s) and none-zero flow ratio (τ) is also shown as OBS in Figure 3. The plot shows that the NSE is above 0.85 for the FDC estimated from OBS for more than 95% of the catchments, indicating that the lognormal distribution assumption is appropriate for most of the gauged catchments investigated in this study.

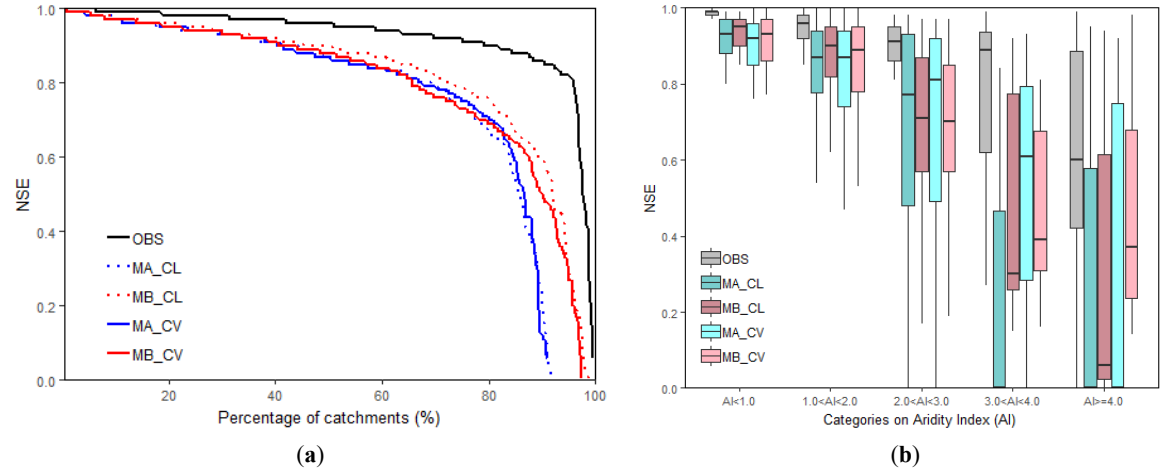


Figure 3. Performance of MARS model in reproducing empirical FDCs from both calibration (CL) and cross-validation (CV). OBS represents the FDC generated based on the parameters derived directly from observed daily streamflow. (**a**) cumulative frequency of performance index; (**b**) boxplot of performance index in relation to aridity index (AI).

The performances of the MARS models as expected are inferior to the benchmark (i.e., the OBS) due to the bias in predicting the $\bar{Q}$, s and τ. Nevertheless, the two MARS models have shown to accurately predict the FDCs of the "ungauged" catchments, with the NSE above 0.8 for 75% of catchments and above 0.65 for 85% of catchments. The two MARS models show comparable results in FDC predictions for most catchments, but MB outperforms MA for those catchments with lower NSE (<0.6). The performance difference between MA and MB is consistent in both the calibration and cross-validation results. The results also show that (Figure 3b) the performance measured by NSE of both MARS models decreases with higher aridity index, indicating that it is more challenging to predict the FDCs of the drier catchments. Overall, MB outperforms MA particularly for catchments with aridity index below 2.0. In addition, MB shows more robust prediction than MA as indicated by the relative smaller ranges of NSE for all catchments (Figure 3b).

Figure 4 shows the predicted FDCs of three selected catchments with different aridity index (see Figure 1). The predicted FDCs are compared against the empirical flow duration curve (EFDC). In general, the FDCs produced by the two MARS models are in good agreement with the EFDC for the wetter catchments that are with lower aridity index (e.g., catchment 210014). For the drier catchments, the predicted FDCs could be different to the EFDC (e.g., catchment 407211). Nevertheless, the predicted FDCs could be quite close to the FDC reproduced according to the observed FDC parameters (i.e., $\bar{Q}$, s and τ). This is because that the two MARS can predict well FDC parameters, particularly MB performs much better than MA. For drier catchments, the discrepancies between the predicted FDCs and the EFDC are primarily due to biases in the estimated non-zero flow ratio (τ) and the limitations of assuming a lognormal distribution for FDC (e.g., catchment 003303).

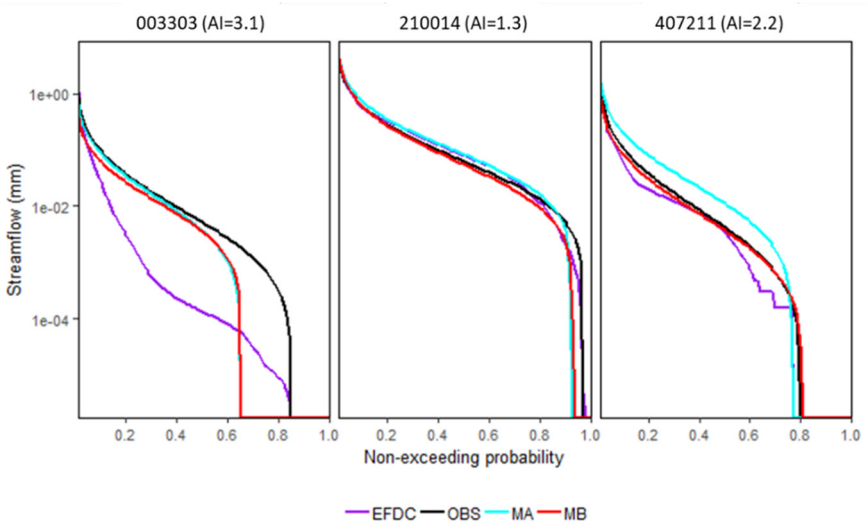


Figure 4. Performance of MARS model in predicting FDCs for both calibration (CL) and cross-validation (CV) procedures. OBS represents the FDC created based on FDC parameters derived from observed daily streamflow.

5. Discussion

5.1. Benefits of Embedding Hydrological Knowledge in ML Models

To predict FDCs in ungauged catchments using machine learning algorithms such as MARS, one approach is to directly model the mean and standard deviation of daily streamflow along with the non-zero flow ratio (the MA model). While this purely data-driven model can yield satisfactory results, it does not necessarily adhere to established hydroclimatic principles. By embedding established hydroclimatic principles (herein, the Budyko theory represented by the Fu's equation) into the machine learning framework (the MB model), the process-informed ML approach (PIML) introduces physically meaningful constraints, resulting in more robust and consistent predictions. As noticed, even though the performances of MA and MB are comparable for most studied catchments, MB tends to be more robust with the relationships between the predictors and the predictands are more physically interpretable and applicable than MA.

By integrating the Budyko theory, the MB model also provides reasonably accurate estimation of the parameter α in the Fu's equation, particularly when α is lower than 4.0 (Figure 5). This indicates that MB can be applied to predict changes in α in the future research work, capturing hydrological non-stationarity of a catchment driven by climate change or land use and land cover change. Importantly, it is noticed that the predicted mean streamflow aligns well with observations even though the predicted α seems to have a relative higher bias at its higher end ($\alpha > 4.0$), resulting in more robust FDC prediction. This robustness arises because, under the physical constraint informed by the Budyko theory, the mean annual streamflow is relatively insensitive to α when α is large, which is especially true for very humid (aridity index smaller than 0.5) and dry (aridity index higher than 2.5) catchments [45].

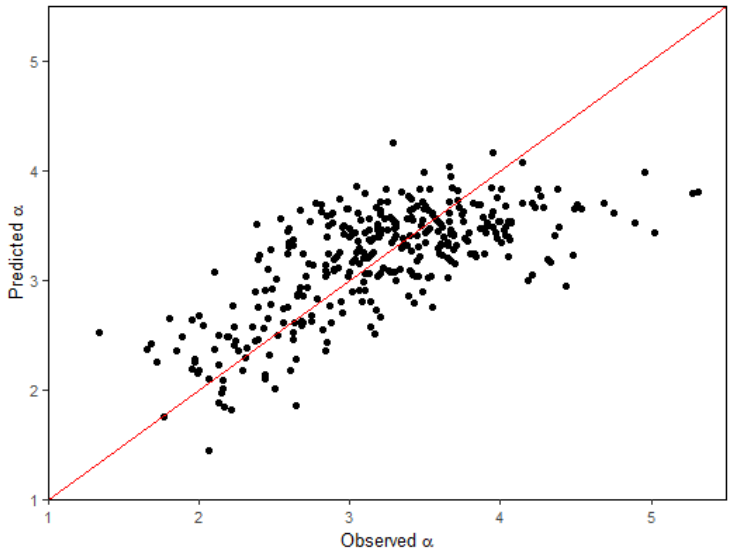


Figure 5. Performance of the process-informed machine learning model (MB) in predicting parameter α in Fu's equation.

5.2. Limitations and Uncertainties

Various machine learning (ML) and deep learning (DL) approaches have been widely adopted for hydrological prediction, reflecting their strong ability to capture nonlinear relationships between hydrological responses and catchment properties. Commonly used ML and DL methods include artificial neural networks (ANNs), support vector machines (SVMs), random forests (RF), gradient boosting methods, multivariate adaptive regression splines (MARS), long short-term memory (LSTM) networks and convolutional neural networks (CNNs). As an interpretable machine learning approach, MARS-based models presented in this study explicitly represent the relationships between predictors and predictands, even though their predictive performance is not necessarily superior to that of more complex ML or DL models. Nevertheless, MARS offers valuable insights into the dominant controls on hydrological responses, which is particularly relevant for applications in ungauged catchments. However, more research efforts are needed to further validate and understand the physical implications in the MARS's basis functions and knots. Future research could also further improve FDC prediction by integrating richer hydrological knowledge with more advanced ML and DL approaches, thereby balancing predictive skill with physical interpretability.

It is worth noting that the two MARS models are developed to predict the mean ($\bar{Q}$) and standard deviation (s) of daily streamflow, which are subsequently transformed into the lognormal parameters μ and σ using Equations (5) and (6). These equations are derived under the assumption that daily streamflow follows a lognormal distribution. Consequently, errors may arise when this assumption is violated. This highlights that prior knowledge of the FDC distribution for a catchment is important for reducing uncertainty in predicted FDCs. However, such information is generally unavailable for ungauged catchments and may require further investigation even in gauged basins.

In addition to uncertainties arising from the lognormal assumption, both the MA and MB models exhibit limitations in predicting the non-zero flow day ratio (τ). Biases in the predicted τ can substantially affect the agreement between predicted and empirical FDCs, particularly in drier catchments. This limitation reflects the fact that the 19 candidate predictors considered in this study do not sufficiently capture the contribution of groundwater discharge to sustaining baseflow. Incorporating catchment properties more closely related to groundwater processes, such as groundwater levels, aquifer characteristics, or soil permeability, could potentially improve the estimation of τ if such data become available for the studied catchments. Moreover, in this study, the three distribution parameters are estimated jointly using shared predictors and basis functions, rather than being modelled separately for each parameter. While separate estimation of individual distribution parameters may improve FDC prediction accuracy, it would likely come at the expense of reduced interpretability within a unified modelling framework.

6. Conclusions

The daily flow duration curve (FDC) is widely used to characterise streamflow regimes and support water resources planning and management. To address the challenge of predicting FDCs in ungauged catchments, this study proposes a process-informed machine learning (PIML) framework that integrates Budyko theory with a multivariate adaptive regression splines (MARS) model. By embedding hydroclimatic process constraints within an interpretable machine learning approach, the framework predicts FDC parameters from 19 climatic and physical catchment characteristics. The proposed method was evaluated using data from 347 catchments across Australia.

Results show that the process-informed model (MB) generally outperforms and is more robust than the purely data-driven model (MA), highlighting the value of incorporating physical constraints through Budyko theory. While both models perform well for most catchments, predictive skill is consistently higher in wetter catchments than in drier ones. This reduced performance in dry catchments is primarily attributable to limitations of the lognormal distribution assumption and to uncertainties in estimating the non-zero flow day ratio, which plays a critical role in shaping FDCs under arid conditions.

As an interpretable machine learning method, MARS offers the advantage of jointly capturing clustering and prediction, enabling the models to reflect heterogeneous relationships between FDC parameters and catchment characteristics. Both MARS-based models identify mean and variability of daily precipitation and potential evaporation, plant available water capacity, bulk density, and catchment area as the most influential factors controlling FDC shape. The weaker performance in predicting non-zero flow ratios and low-flow characteristics suggests that incorporating catchment attributes more directly related to groundwater processes could further improve FDC prediction as such data become available.

Overall, this study demonstrates that integrating interpretable machine learning with process-based constraints provides a promising and physically consistent pathway for improving streamflow regime prediction in ungauged catchments. The proposed approach is broadly applicable and can be tested across diverse catchments worldwide. Future improvements in FDC prediction may be achieved by integrating richer hydrological knowledge with more advanced machine learning and deep learning approaches, thereby balancing predictive skill with physical interpretability.

Author Contributions

H.Z.: conceptualization, methodology, writing—original draft preparation; R.Z.: data curation, methodology, visualization, writing—reviewing and editing; L.Z.: investigation, writing—reviewing and editing; F.C.: investigation, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

The research was funded through a partnership between the Murray-Darling Basin Authority (MDBA) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) under the Australian National Collaboration Framework. The authors would like to acknowledge the contributions and support of the project's

Advisory Group, comprised of experienced modellers from Basin jurisdictions, and the project's Steering Committee.

Data Availability Statement

Data used in this study are publicly available. The SILO climate data can be downloaded from www.nrw.gov.au/silo. The streamflow data can be downloaded from https://www.bom.gov.au/waterdata/ or requested from [47]. The processed data and the relative codes are available upon request to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Abril, M.; Muñoz, I.; Casas-Ruiz, J.P.; et al. Effects of water flow regulation on ecosystem functioning in a Mediterranean river network assessed by wood decomposition. *Sci. Total Environ.* **2015**, *517*, 57–65.
2. Booker, D.J.; Snelder, T.H. Comparing methods for estimating flow duration curves at ungauged sites. *J. Hydrol.* **2012**, *434–435*, 78–94.
3. Brown, A.E.; McMahon, T.A.; Podger, G.M.; et al. *A Methodology to Predict the Impact of Changes in Forest Cover on Flow Duration Curves*; CSIRO Land and Water Science Report 8/06; CSIRO: Canberra, NSW, Australia, 2006.
4. Castellarin, A.; Galeati, G.; Brandimarte, L. Regional flow duration curves: Reliability for ungauged basins. *Adv. Water Resour.* **2004**, *27*, 953–965.
5. Pyron, M.; Neumann, K. Hydrologic alterations in the Wabash River watershed, USA. *River Res. Appl.* **2008**, *24*, 1175–1184.
6. Smakhtin, V.U. Low flow hydrology: A review. *J. Hydrol.* **2001**, *240*, 147–186. 10.1016/S0022-1694(00)00340-1.
7. Vogel, R.M.; Fennessey, N.M. Flow duration curves I: New interpretation and confidence intervals. *J. Water Resour. Plan. Manag.* **1994**, *120*, 485–504.
8. Hrachowitz, M.; Savenije, H.H.G.; Blöschl, G.; et al. A decade of predictions in ungauged basins (PUB)—A review. *Hydrological sciences journal* 2013, 58, 1198–1255.
9. Mohamoud, Y.M. Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrol. Sci. J.* **2008**, *53*, 706–724.
10. Razavi, T.; Coulibaly, P. Streamflow prediction in ungauged basins: Review of regionalization methods. *J. Hydrol. Eng.* **2013**, *18*, 958–975.
11. Shu, C.; Ouarda, T. Improved methods for daily streamflow estimates at ungauged sites. *Water Resour. Res.* **2012**, *48*, WO2523.
12. Sivaplan, M.; Takeuchi, K.; Franks, S.W.; et al. IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrol. Sci. J.* **2003**, *48*, 857–880.
13. Mosley, M.P. Delimitation of New Zealand hydrologic regions. *J. Hydrol.* **1981**, *49*, 173–192.
14. Vandewiele, G.L.; Elias, A. Monthly water balance of ungauged catchments obtained by geographic regionalization. *J. Hydrol.* **1995**, *170*, 277–291.
15. Castellarin, A. Regional prediction of flow-duration curves using a three dimensional kriging. *J. Hydrol.* **2014**, *513*, 179–191.
16. Robson, A.; Reed, D. *Flood Estimation Handbook—3: Statistical Procedures for Flood Frequency Estimation*; Institute of Hydrology: Wallingford, UK, 1999.
17. Li, M.; Shao, Q.X.; Zhang, L.; et al. A new regionalization approach and its application to predict flow duration curve in ungauged basins. *J. Hydrol.* **2010**, *389*, 137–145.
18. Atieh, M.; Taylor, G.; Sattar, A.M.; et al. Prediction of flow duration curves for ungauged basins. *J. Hydrol.* **2017**, *545*, 383–394.
19. Fouad, G.; Loáiciga, H.A. Independent variable selection for regression modeling of the flow duration curve for ungauged basins in the United States. *J. Hydrol.* **2020**, *587*, 124975. https://doi.org/10.1016/j.jhydrol.2020.124975.
20. Leong, C.; Yokoo, Y. A step toward global-scale applicability and transferability of flow duration curve studies: A flow duration curve review (2000–2020). *J. Hydrol.* **2021**, *603*, 16. https://doi.org/10.1016/j.jhydrol.2021.126984.
21. Burn, D.H.; Boorman, D.B. Estimation of hydrological parameters at ungauged catchments. *J. Hydrol.* **1993**, *143*, 429–454.
22. Gupta, H.V.; Bastidas, L.A.; Sorooshian, S.; et al. Parameter estimation of a land surface scheme using multicriteria methods. *J. Geophys. Res.* **1999**, *104*, 19491–19503. https://doi.org/10.1029/1999JD900154.

23. Nijssen; B; Schnur, R.; Lettenmaier, D.P. Global retrospective estimation of soil moisture using the variable infiltration capacity land surface model, 1980–1993. *J. Clim.* **2001**, *14*, 1790–1808. https://doi.org/10.1175/1520-0442(2001)014,1790:GREOSM.2.0.CO;2.
24. Bhasme, P.; Vagadiya, J.; Bhatia, U. 2022. Enhancing predictive skills in physically-consistent way: Physics informed machine learning for hydrological processes. *J. Hydrol.* **2022**, *615*, 128618.
25. Mohammadi, B.; Gao, H.; Pilesjö, P.; et al. Integrating machine learning with process-based glacio-hydrological model for improving the performance of runoff simulation in cold regions. *J. Hydrol.* **2025**, *656*, 132963. https://doi.org/10.1016/j.jhydrol.2025.132963.
26. Holmes, M.G.R.; Young, A.R.; Gustard, A.; et al. A region of influence approach to predicting flow duration curves within ungauged catchments. *Hydrol. Earth Syst. Sci.* **2002**, *6*, 721–731. https://doi.org/10.5194/hess-6-721-2002.
27. Mendicino, G.; Senatore, A. Evaluation of parametric and statistical approaches for the regionalization of flow duration curves in intermittent regimes. *J. Hydrol.* **2013**, *480*, 19–32. https://doi.org/10.1016/j.jhydrol.2012.12.017.
28. Müller, M. F. and Thompson, S. E.: Comparing statistical and process-based flow duration curve models in ungauged basins and changing rain regimes, Hydrol. Earth Syst. Sci., 20, 669–683, https://doi.org/10.5194/hess-20-669-2016, 2016.
29. Requena, A.I.; Chebana, F.; Ouarda, T.B. A functional framework for flow-duration-curve and daily streamflow estimation at ungauged sites. *Adv. Water Resour.* **2018**, *113*, 328–340. https://doi.org/10.1016/j.advwatres.2018.01.019.
30. Costa, V.; Fernandes, W. Regional modeling of long-term and annual flow duration curves: Reliability for information transfer with evolutionary polynomial regression. *J. Hydrol. Eng.* **2021**, *26*, 12. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002051.
31. Wolff, W.; Duarte, S.N. Toward geostatistical unbiased predictions of flow duration curves at ungauged basins. *Adv. Water Resour.* **2021**, *152*, 13. https://doi.org/10.1016/j.advwatres.2021.103915.
32. Lan, T.; Zhang, J.; Li, H.; et al. Flow duration curve prediction: A framework integrating regionalization and copula model. *J. Hydrol.* **2025**, *647*, 132364. https://doi.org/10.1016/j.jhydrol.2024.132364.
33. Friedman J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–141.
34. Budyko, M.I. *The Heat Balance of the Earth's Surface*; U.S. Department of Commerce: Washington, DC, USA, 1958.
35. Friedman, J.H.; Roosen, C.B. An introduction to multivariate adaptive regression splines. *Stat. Methods Med. Res.* **1995**, *4*, 197–217. https://doi.org/10.1177/096228029500400303.
36. Shao, Q.; Traylen, A.; Zhang, L. Nonparametric method for estimating the effects of climatic and catchment characteristics on mean annual evapotranspiration. *Water Resour. Res.* **2012**, *48*, W03517. https://doi.org/10.1029/2010WR009610.
37. Sharda, V.N.; Prasher, S.O.; Patel, R.M.; et al. Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data. *Hydrol. Sci. J.* **2008**, *53*, 1165–1175. https://doi.org/10.1623/hysj.53.6.1165.
38. Beard, L.R. Statistical analysis in hydrology. *Trans. Am. Soc. Civ. Eng.* **1943**, *108*, 1110–1160.
39. Post, D.A. A new method for estimating flow duration curves: An application to the Burdekin River Catchment, North Queensland, Australia. In Proceedings of the iEMSs 2004 International Congress ''Complexity and Integrated Resources Management", Osnabrueck, Germany, 14–17 June 2004.
40. Shao, Q.; Wong, H.; Xia, J.; et al. Models for extremes using the extended three-parameter Burr XII system with application to flood frequency. *Hydrol. Sci. J.* **2004**, *49*, 702. https://doi.org/10.1623/hysj.49.4.685.54425.
41. Johnson, N.L.; Kotz, S.; Balakrishnan, N. Lognormal Distributions, Continuous univariate distributions. In *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 1994; Volume 1; ISBN 978-0-471-58495-7.
42. Milborrow, S. Package "Earth". 2018. Available online: https://cran.r-project.org/web/packages/earth/earth.pdf (accessed on 1 July 2025).
43. Fu, B.P. On the calculation of evaporation from land surface in mountainous areas. *Sci. Meteorol. Sin.* **1996**, *16*, 328–335. (In Chinese)
44. Zhang, L.; Dawes, W.R.; Walker, G.R. The response of mean annual evapotranspiration to vegetation changes at catchment scale. *Water Resour. Res.* **2001**, *37*, 701–708.
45. Zhang, L.; Hickel, K.; Dawes, W.R.; et al. A rational function approach for estimating mean annual evapotranspiration. *Water Resour. Res.* **2004**, *40*, W02502. https://doi.org/10.1029/2003WR002710.
46. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I: A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290.
47. Zhang, Y.Q.; Viney, N.; Frost, A.; et al. *Collation of Australian Modeller's Streamflow Dataset for 780 Unregulated Australian Catchments;* CSIRO: Canberra, NSW, Australia, 2013; 115p.
48. Ukkola, A.M.; Prentice, I.C.; Keenan, T.F.; et al. Reduced streamflow in water-stressed climates consistent with CO_2 effects on vegetation, *Nat. Clim. Chang.* **2015**, *6*, 75–78.

49. Jeffrey, S.J.; Carter, J.O.; Moodie, K.M.; et al. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Model. Softw.* **2001**, *16*, 309–330.
50. Zhu, Z.; Bi, J.; Pan, Y.; et al. Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011. *Remote Sens.* **2013**, *5*, 927–948. https://doi.org/10.3390/rs5020927.
51. MaKenzie, N.J.; Jacquier, D.W.; Ashton, L.J.; et al. *Estimation of Soil Properties Using the Atlas of Australian Soils*; CSIRO Land and Water, Report 11/00; CSIRO: Canberra, NSW, Australia, 2000.