Supporting Information for "HydroAgent: Formalizing Forecaster Expertise into Skill-Orchestrated Flood Forecasting Workflows"

Qingyi Yang[1], Siqian Qiu[2], Bing Li[3], Xu Shan[4], Jia Feng[5], Shunan Zhou[6,7], Xudong Zhou[8], Tiantian Xing[1], Jiale Guo[1], Xiaoyi Dong[9], Gaoyu Liu[10], Xiaohuan Liu[1], Haiqing Pu[5], Qingwen Deng[11], Xun Zhang[12,13], Zhongrun Xiang[14], Haiyang Qian[15,16], Ying Yan[17], Yongkang Xu[18], Nuo Lei[13], Tianlong Jia[19], Baoying Shan[1], and Carlo De Michele[1]

[1]Department of Civil and Environmental Engineering (D.I.C.A.), Politecnico di Milano, 20133 Milano MI, Italy

[2]College of Hydraulic and Environmental Engineering, China Three Gorges University, 443002 Yichang, China

[3]Independent Researcher, Shandong, 252300, China

[4]Delft University of Technology, Delft, the Netherlands

[5]Qinhuangdao Hydrological Survey and Research Center of Hebei Province, Qinhuangdao, China

[6]School of Hydraulic Engineering, Dalian University of Technology, 116024 Dalian, Liaoning, China

[7]Institute of Photogrammetry and Remote Sensing, TU Dresden University of Technology, 01062 Dresden, Germany

[8]Institute of Hydraulic and Ocean Engineering, Ningbo University, Ningbo 315211, China

[9]China IPPR International Engineering Co., Ltd., SINOMACH, Beijing, China

[10]College of Hydrology and Water Resources, Hohai University, Nanjing, China

[11]School of Earth Science and Engineering, Nanjing University, Nanjing, China

[12]School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA

[13]College of Civil Engineering, Tongji University, Shanghai, China

[14]Independent Researcher, Maryland, 20878, USA

[15]Zhejiang Institute of Hydraulics and Estuary (Zhejiang Institute of Marine Planning and Design), Hangzhou 310020, China

[16]Hydro-Climate Extremes Lab (H-CEL), Ghent University, Ghent, Belgium

[17]Department of Earth, Ocean & Atmospheric Sciences, University of British Columbia, Vancouver, V6T 1Z4, Canada

[18]College of Water Sciences, Beijing Normal University, Beijing, China

[19]Karlsruhe Institute of Technology (KIT), Institute of Water and Environment, Karlsruhe, Germany

Contents of this file

1. Text S1 to S5
2. Figures S1 to S4
3. Tables S1 to S3

Introduction

The Supporting Information accompanies the HydroAgent skill-orchestrated flood forecasting framework described in the main text. Text S1 and Text S2 describe the Xinanjiang (XAJ) hydrological model and the Dynamically Dimensioned Search (DDS) calibration algorithm that together underpin the Step 0 scheme preparation. Text S3 specifies the priority rule used in the Step 2 scheme selection. Text S4 with Table S1, and Text S5 with Table S2, document the six-dimensional similarity scoring and the expert-rule reasoning that constitute the Step 1 prior-judgment skill. Figure S1 shows the hydrographs of two flood events excluded from the validation set (N.15_20240129 and N.16_20241229), with the reasons for exclusion explained in Section 2.4 of the main text. Figure S2 compares flood-type-specific and global XAJ calibrations across the calibration and validation periods. Figure S3 documents the round-to-round stability of the Step 1 prior judgment under GPT-5.4, and Figure S4 reports five-fold cross-validation results for four additional LLMs. Table S3 lists the per-event performance before and after Step 2 scheme selection for all 14 validation events (2020–2024). All analyses use hourly hydrometeorological

Corresponding author: Baoying Shan, Department of Civil and Environmental Engineering (D.I.C.A.), Politecnico di Milano, 20133 Milano MI, Italy. (baoying.shan@polimi.it)

records from the CAMELSH dataset (1995–2024) for the South Yamhill River basin at USGS gauge 14194150.

Text S1. Xinanjiang (XAJ) model.

The XAJ model is a physically conceptual hydrological model widely used in humid and semi-humid regions (Zhao, 1992). It consists of four modules: evapotranspiration calculation, runoff generation, water source separation, and flow concentration calculation, each governed by several adjustable parameters. The XAJ model has demonstrated strong performance in simulating flood processes and requires only standard meteorological forcing data that can be readily obtained and processed (Luo et al., 2024; Gong et al., 2025; Wang et al., 2025).

Text S2. Dynamically Dimensioned Search (DDS) algorithm.

In this study, we adopt the three runoff component version of the XAJ model, which contains 14 parameters that are calibrated using the DDS algorithm. The DDS algorithm is a global optimization method designed for the efficient calibration of complex models with high-dimensional parameter spaces (Tolson & Shoemaker, 2007). Its search strategy gradually shifts from broad global exploration to local refinement as the number of function evaluations approaches the predefined maximum. Owing to its robust performance in hydrological applications, DDS is used here to calibrate the 14 XAJ parameters for each predefined event group.

Text S3. Best matching scheme selection rule.

In Step 2, all candidate flood types defined in the Step 0 forecasting scheme are simulated using their corresponding pre-calibrated parameter sets. The simulated peak flow and flood volume of each candidate are then compared against the prior ranges inferred in Step 1. The best matching type is selected according to the following priority rule.

1. **P0**: both peak flow and flood volume fall within their respective prior ranges.

2. **P1**: only peak flow falls within its prior range.

3. **P2**: only flood volume falls within its prior range.

4. **P3**: neither peak flow nor flood volume falls within its prior range.

Within categories P0 and P3, preference is given to the candidate type whose simulated flood volume lies closest to the midpoint of the prior flood volume range.

Text S4. Six-Dimensional Similarity Scoring in Step 1.

The Step 1 analog retrieval uses a weighted six-dimensional similarity score to identify the top historical cases for expert comparison. The six dimensions are total rainfall, rainfall duration, soil moisture, event type, maximum short-duration rainfall, and initial flow. The maximum total score is 13 points, with total rainfall assigned the largest weight because it is the dominant control on runoff magnitude.

The detailed scoring rules are summarized in Supplementary Table S1. Total rainfall is scored with a continuous piecewise-linear function of the relative difference and contributes 3 points. Rainfall duration is scored with a continuous piecewise-linear rule and contributes 2 points. Maximum cumulative rainfall over k hours contributes 2 points and is scored continuously. Initial discharge contributes 2 points and is scored continuously. Antecedent soil moisture contributes 2 points through an ordinal mapping in which the score decreases as the categorical distance increases. Weather type contributes 2 points through categorical similarity, with closer matches receiving higher scores. This continuous scoring design improves discriminability relative to a purely categorical retrieval rule.

Text S5. Detailed expert-rule reasoning used in Step 1.

The detailed rule definitions above are used to refine the similarity score into a physically plausible prior interval. Analog correction accounts for differences in rainfall amount, antecedent wetness, rainfall intensity, storm duration, and basin-scale runoff controls.

Boundary handling then uses an anchor case to constrain the upper bound when a close analog exists and falls back to controlled extrapolation when the target event lies outside the support of the case library.

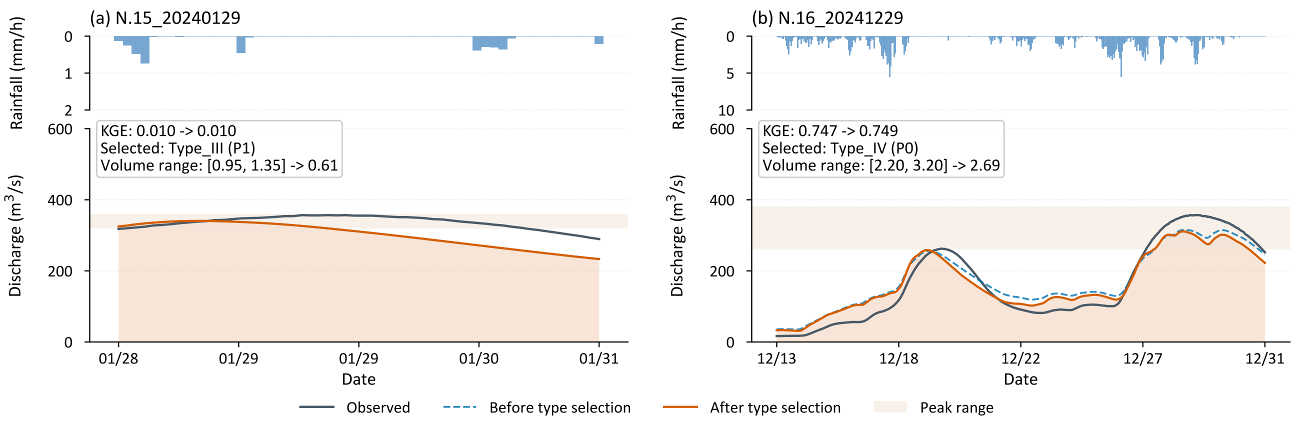

Figure S1. Hydrographs of two flood events excluded from the validation set: (a) N.15_20240129 and (b) N.16_20241229.

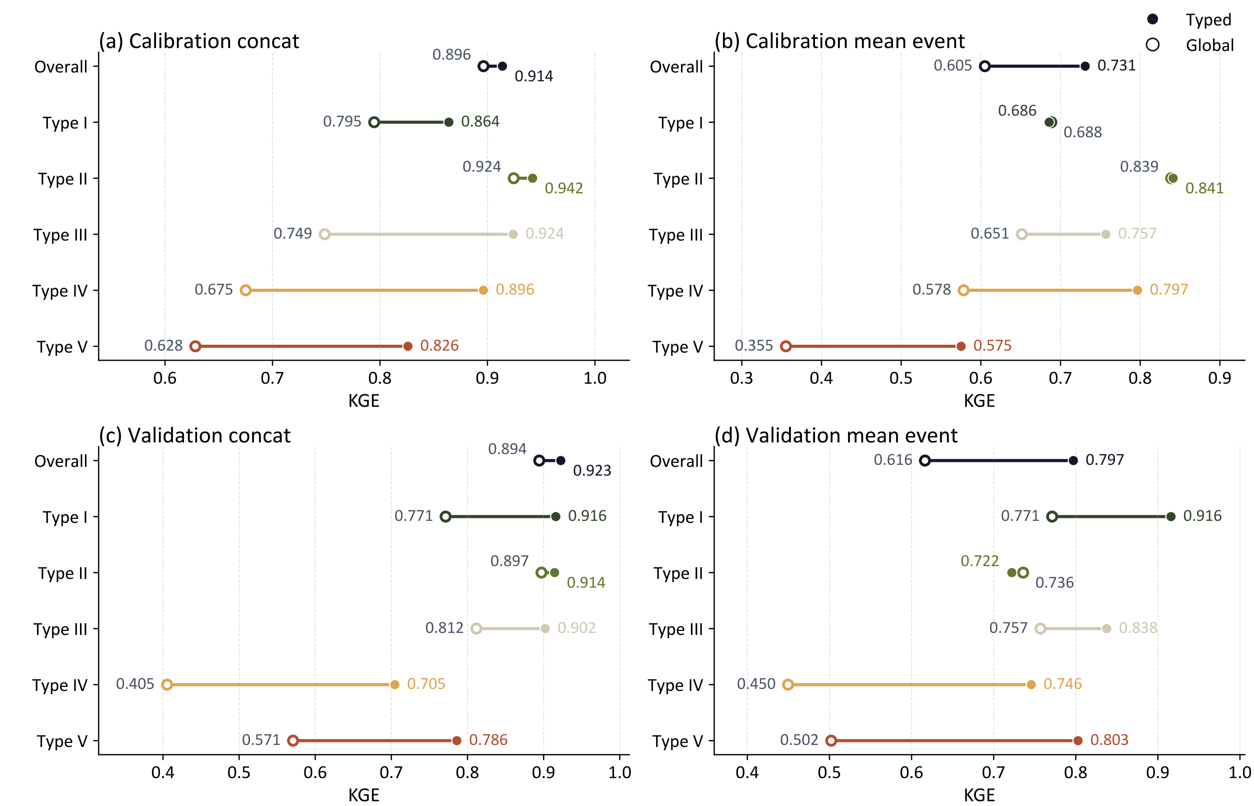

Figure S2. Comparison of KGE between five flood type specific (Typed) and single parameter set (Global) calibrations during the calibration and validation periods. (a, c) Concatenated KGE; (b, d) mean event KGE.

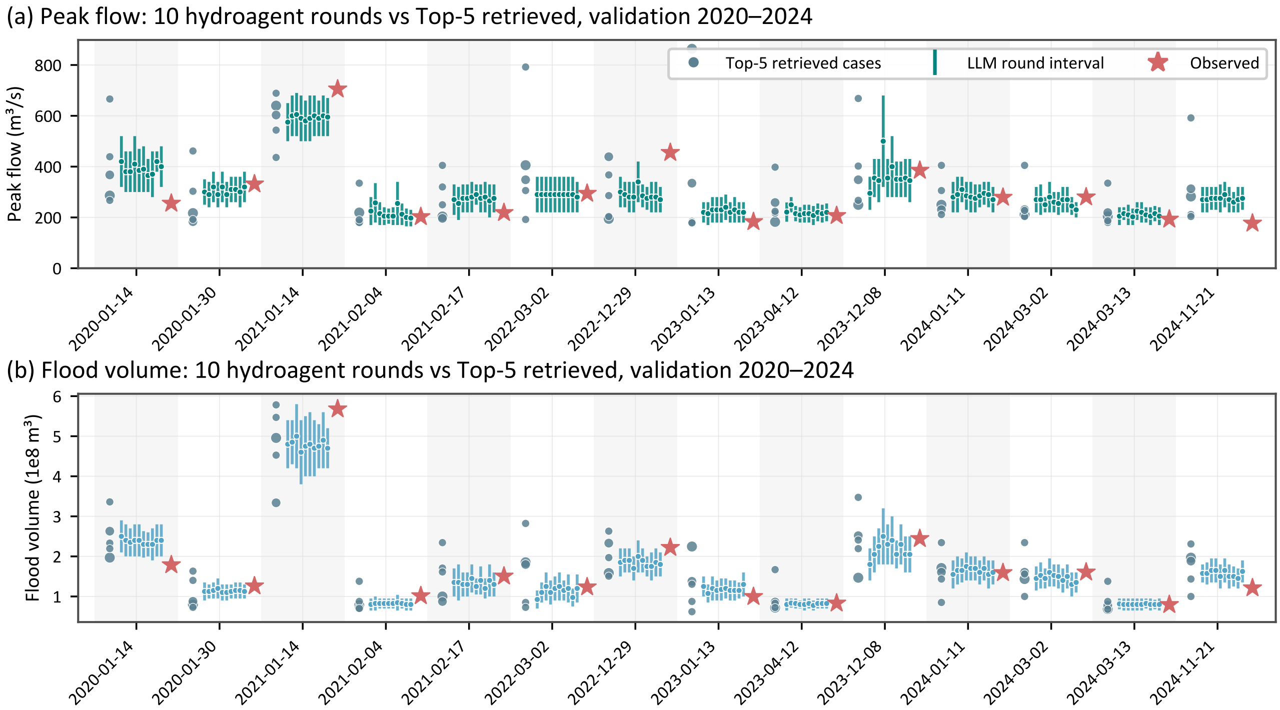

Figure S3. Round-to-round stability of the Step 1 prior judgement produced by HydroAgent (LLM backend: GPT-5.4) across 14 independent flood events from the 2020–2024 validation set on the South Yamhill River. For each event the same Step 1 skill was executed in ten independent rounds; the resulting peak-flow (**a**) and flood-volume (**b**) judgement intervals (coloured vertical bars; circular markers denote the interval midpoint) are compared against (i) the spread of the Top-5 retrieved analogue cases used as the retrieval prior (left grey dots, sized by similarity rank), and (ii) the observed value (right red star). Alternating background bands separate adjacent events. The intervals tighten markedly relative to the Top-5 envelope while still bracketing the observation in the great majority of events, indicating that the LLM's reasoning step contributes informative narrowing on top of the retrieval prior, and that the round-to-round dispersion induced by sampling stochasticity is small relative to the prior width.

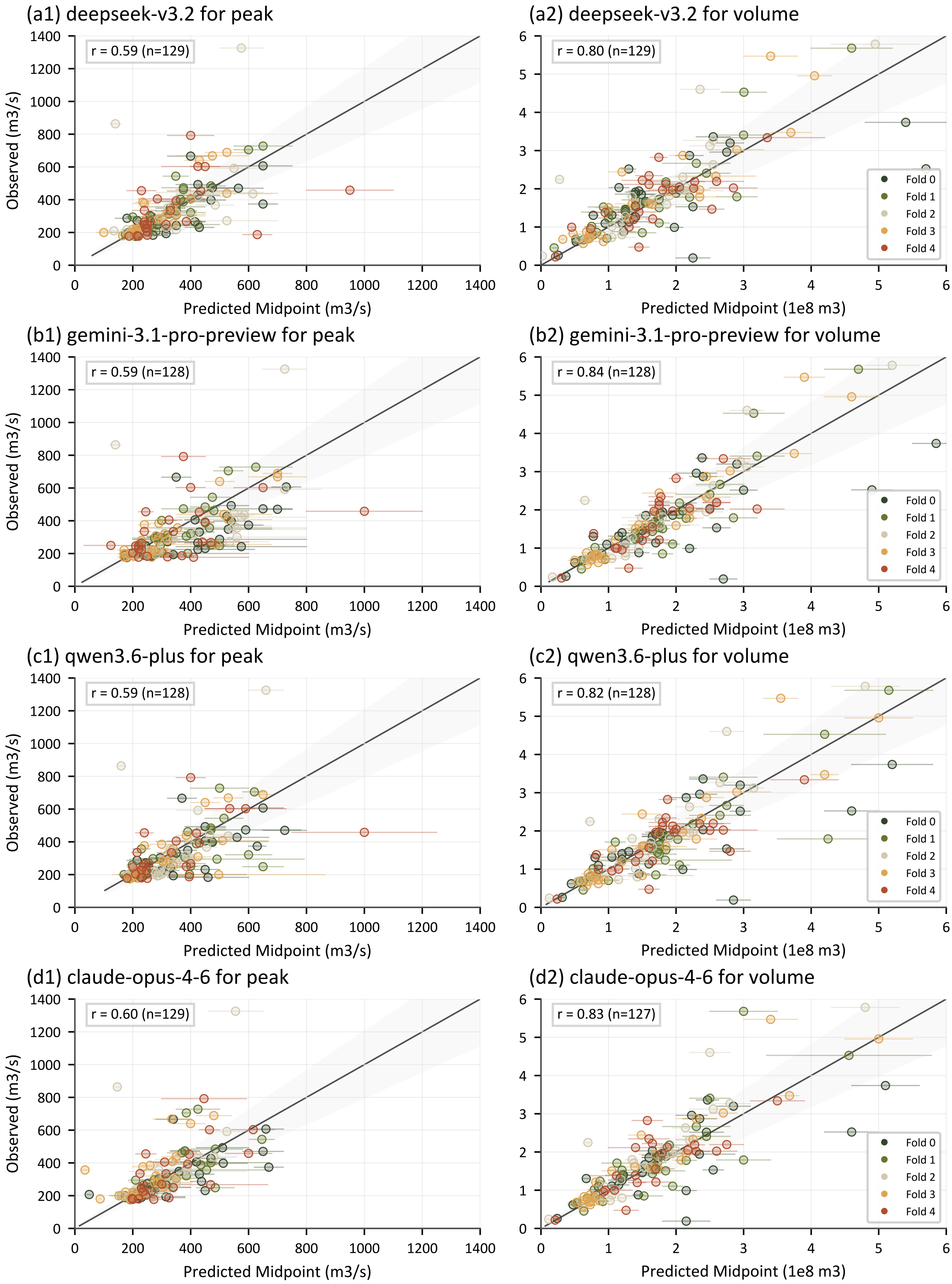

Figure S4. Five-fold cross-validation results for four additional LLMs (DeepSeek-V3.2, Gemini-3.1-Pro-Preview, Qwen3.6-Plus, and Claude-Opus-4.6), presented in the same format as Figure 3 b1 and b2.

Table S1. Six-dimensional weighted similarity metric used in Step 1. The total similarity score is the sum across all six dimensions (maximum: 13 points). P, H, Q_0, and M_k denote total rainfall, rainfall duration, initial discharge, and k-hour maximum accumulated rainfall of the current query event (subscript "cur") and the candidate historical case (subscript "case"), respectively.

Dimension	**Variable**	**Max**	**Scoring rule**	**Notes**
Total rainfall	P (mm)	3	$s_1 = 3\max(0,\, 1 - r_P/0.5)$, $r_P = \dfrac{\lvert P_{\text{cur}} - P_{\text{case}}\rvert}{P_{\text{case}}}$	Largest weight; zero score at $\geq 50\%$ relative difference.
Rainfall duration	H (h)	2	$s_2 = 2\max(0,\, 1 - r_H/0.8)$, $r_H = \dfrac{\lvert H_{\text{cur}} - H_{\text{case}}\rvert}{\max(H_{\text{cur}}, H_{\text{case}})}$	Wider tolerance than rainfall; zero score at $\geq 80\%$ relative difference.
Soil moisture	SM (class)	2	$s_3 = \begin{cases} 2.0 & \Delta c = 0 \\ 1.3 & \Delta c = 1 \\ 0.7 & \Delta c = 2 \\ 0 & \Delta c \geq 3 \end{cases}$	Δc is the absolute difference in soil-moisture class (dry / normal / wet / saturated).
Weather type	T (category)	2	$s_4 = \begin{cases} 2 & \text{exact match} \\ 1 & \text{adjacent type} \\ 0 & \text{different type} \end{cases}$	Categories: frontal, typhoon, plum rain, convective, and mixed.
k-h maximum rainfall	M_k (mm)	2	$s_5 = 2\max(0,\, 1 - r_M/0.5)$, $r_M = \dfrac{\lvert M_{k,\text{cur}} - M_{k,\text{case}}\rvert}{\max(M_{k,\text{cur}}, M_{k,\text{case}})}$	Window $k \in \{12, 24, 36, 48, 64\}$ h is adaptively selected based on H_{cur}.
Initial flow	Q_0 (m^3/s)	2	$s_6 = 2\max(0,\, 1 - r_Q)$, $r_Q = \dfrac{\lvert Q_{0,\text{cur}} - Q_{0,\text{case}}\rvert}{\max(Q_{0,\text{cur}}, Q_{0,\text{case}})}$	Zero score when one event begins near-dry and the other near-bankfull.
Total		**13**	$S = \sum_{i=1}^{6} s_i$	Top-5 highest-scoring historical events are retrieved as analogs.

Table S2. Detailed expert-rule reasoning used in Step 1. The table groups the rules into analog correction and boundary handling.

Rule	**Purpose**	**Summary**
Difference-based adjustment	Correct the analog estimate using event differences	Adjust expected runoff according to differences in total rainfall and antecedent precipitation response.
Diminishing marginal effects	Prevent unrealistic linear extrapolation	Represent the weaker increase in runoff efficiency as soil moisture approaches saturation.
Rainfall-intensity and hydrograph-shape association	Link storm structure to peak response	Persistent low-intensity rainfall tends to produce attenuated hydrographs, whereas short-duration intense rainfall tends to produce sharper peaks.
Rainfall-duration and hydrograph-shape association	Relate duration to temporal concentration	Shorter-duration events generally produce more concentrated rainfall and higher peak flow than longer events with similar totals.
Flood-volume dominance	Constrain volume by basin-scale controls	Total flood volume is governed primarily by total rainfall, runoff coefficient, and catchment area, while intensity and duration mainly shape the hydrograph.
Anchor-case constraint	Prevent upper-bound underestimation	Reference a selected case from a higher flood class when a close analog is available.
Out-of-distribution detection and extrapolation	Handle events outside the archive support	Detect events that deviate substantially from the retrieved analogs and fall back to controlled extrapolation of peak flow and flood volume.

Table S3. Performance comparison before and after Step 2 scheme selection for all 14 validation events (2020–2024). Columns show the Kling–Gupta Efficiency (KGE), peak flow (m^3/s), and flood volume (10^8 m^3), along with the corresponding Step 1 judgment ranges for peak and volume. Blue shading indicates the better KGE or values within the judgement range.

ID	KGE		Peak (m^3/s)			Volume (10^8 m^3)		
	Before	**After**	**Judgement range**	**Observed**	**Simulated**	**Judgement range**	**Observed**	**Simulated**
n.1_20200114	**0.517**	0.278	350–520	**255.5**	**417.2**	2.10–2.80	**1.789**	**2.519**
n.2_20200130	**0.855**	**0.855**	240–380	**331.3**	**350.8**	1.15–1.60	**1.261**	**1.326**
n.3_20210114	**0.916**	**0.916**	500–660	**705.1**	**586.0**	4.20–5.80	**5.679**	**5.705**
n.4_20210204	**0.764**	**0.764**	170–240	**202.7**	**171.7**	0.65–0.90	**1.016**	**0.776**
n.5_20210217	0.651	**0.674**	220–340	**218.6**	**220.4**	1.00–1.80	**1.506**	**1.503**
n.6_20220302	0.834	**0.921**	220–360	**294.5**	**290.9**	0.90–1.60	**1.236**	**1.246**
n.7_20221229	**0.589**	**0.589**	220–320	**455.2**	**304.9**	1.60–2.10	**2.221**	**1.900**
n.8_20230113	**0.654**	0.218	180–260	**183.0**	**202.2**	1.00–1.45	**0.995**	**0.949**
n.9_20230412	0.833	**0.871**	185–255	**206.9**	**198.6**	0.68–0.90	**0.832**	**0.773**
n.10_20231208	0.830	**0.906**	260–430	**385.1**	**369.2**	1.50–2.40	**2.442**	**2.631**
n.11_20240111	**0.781**	**0.781**	260–360	**278.9**	**265.0**	1.30–1.80	**1.592**	**1.727**
n.12_20240302	**0.873**	**0.873**	205–270	**280.2**	**243.2**	1.20–1.75	**1.606**	**1.553**
n.13_20240313	0.720	**0.874**	190–250	**192.8**	**194.9**	0.70–0.95	**0.795**	**0.824**
n.14_20241121	**0.715**	0.509	240–320	**177.1**	**239.7**	1.35–1.95	**1.213**	**1.583**