https://doi.org/10.62836/iaet.v5i1.0009

Global Science Publishing

IAET

Innovations in Applied Engineering and Technology

https://ojs.sgsci.org/journals/iaet

Article

Event-Conditioned Residual Transfer for Streamflow Forecasting in Data-Limited Catchments

Yue Zhang *, Xue Song, Jintao Li and Dongsong Xue

School of Software, Henan Polytechnic University, 2001 Century Avenue, Jiaozuo 454099, China

Abstract: A short local discharge record may contain too few floods to characterize how a catchment responds to rainfall. We evaluate a forecast that adds an event-conditioned regional residual to a conceptual baseline and calibrates the remaining error using a short local record. The hourly model learns from peak magnitude, timing, and event volume as well as pointwise errors. The weekly model weights high-flow observations more heavily and shrinks the forecast toward the latest observed weekly flow. In 584 flood windows from 17 held-out CAMELS-GB v2 catchments, the six-hour model reduced peak magnitude error to 9.93% and achieved an average joint peak capture rate of 85.14% across catchments. The regional gated recurrent unit yielded 11.72% and 74.30%, respectively. In nine held-out HydroMLYR catchments in the Yellow River Basin, the one-week model lowered high-flow mean absolute error from 1.243 for persistence to 1.219 on the standardized flow scale. Recall rose from 71.57% to 73.10%, with lower high-flow error in seven of nine catchments. Removing the event-level loss degraded hourly peak forecasts, and removing shrinkage produced the largest loss of weekly skill among the tested ablations. The two evaluations support targeted gains in flood peak estimation and weekly high-flow detection with limited local calibration.

Keywords: streamflow forecasting; flood peak; high flow; catchment regionalization; residual learning; few-shot adaptation

1. Introduction

1.1. Flood-Generating Processes and Forecast Objectives

Similar rainfall totals can produce markedly different hydrographs. A short, intense storm can drive a sharp rise, whereas sustained rainfall can prolong high flows. Both responses belong to the flood types identified in Austrian catchments [1]. In Chinese headwaters, rainfall intensity and volume help explain differences among flood classes [2]. Which class can be assigned to an event also depends on when and where observations are available [3]. At forecast issue time, recent precipitation describes the forcing so far, antecedent wetness indicates how readily additional rain may generate runoff, and a rise in discharge shows that the outlet is responding. The useful distinction for a forecast is between hydrological conditions that can already be recognized from these observations.

Peak magnitude and timing respond differently to errors in the simulated hydrograph. Routing a response too quickly can place an otherwise plausible peak several hours early. Underestimating runoff can yield a small peak even when it arrives on time. Conventional calibration measures differ in how well they represent annual

Received: 3 November 2026; Accepted: 18 November 2026.

* Corresponding: Yue Zhang (y_zhang@hpu.edu.cn)

peak flows [4]. Peak forecast skill can also vary by flood type. A long short-term memory (LSTM) model, for instance, reproduced peak days more often for rainfall-driven floods than for snowmelt-driven floods, although errors in peak magnitude were similar [5]. At weekly resolution, averaging smooths short rises and peaks. The forecast question then becomes whether the next weekly mean will reach a high-flow threshold. The useful performance measures therefore depend on the temporal resolution of the observations.

1.2. Regional Learning with Limited Local Calibration

Pooling observations offers more examples of hydrological extremes than a short local record can provide. Caravan brings together forcing, discharge, and catchment attributes from multiple regions [6], and CAMELSH supplies hourly records for watersheds in the contiguous United States [7]. A larger archive also increases diversity in the processes that a regional model must represent. In a study using CAMELS-GB, transfer from two hydrologically similar catchments outperformed transfer from the full source pool for a target catchment with one year of training data [8]. Holding out geographical groups tests whether relationships learned in the source catchments transfer to distinct areas.

Regional recurrent models can use static attributes to describe differences in rainfall–runoff response [9], and comparisons across 516 CAMELS catchments have broadened the choice of neural architectures for hourly forecasting [10]. At coarser resolutions, sharing information across forecast scales can provide a more coherent runoff trajectory [11, 12]. Large regional archives have also made it possible to forecast extreme flows at ungauged sites worldwide [13]. For a target catchment with some discharge observations, the question is how much of the remaining regional error can be estimated from the local record.

Local discharge observations serve two distinct purposes. Recent flow describes the state from which each forecast starts. Earlier observations can be used to estimate persistent errors, provided that the calibration period contains enough relevant conditions. The forecasting value of recent discharge is evident at continental scale [14], and pretraining on data-rich regions can support transfer to regions with fewer observations [15]. Fine-tuning a pretrained LSTM has also improved forecasts in held-out basins using limited local records [16]. A short record may contain only a few floods, so estimating a full set of network parameters from it can be difficult. A regularized correction to a fixed regional prediction confines local fitting to a small set of coefficients that can account for recurring site errors.

1.3. Conceptual Baselines and Event-Conditioned Residuals

Rainfall can contribute to discharge after precipitation has ceased, while water already in the channel can sustain flow through a recession. A conceptual baseline gives those effects distinct roles in the forecast and makes its assumptions about catchment response explicit [17]. Hybrid models in Denmark have improved streamflow prediction by learning residuals from a national hydrological model or passing its output to an LSTM [18]. A differentiable Xin'anjiang model combined with transfer learning has also been tested for flood forecasting in data-limited basins in Anhui [19]. Experiments with hybrid models on CAMELS-GB show that a learned component can dominate predictive skill even when the model imposes conceptual constraints [20]. The value of a conceptual baseline and the value of its residual correction therefore need separate examination at the target horizon.

The errors left for a residual model to correct depend on what the baseline can represent and on how the target is observed. At hourly resolution, a rainfall response that arrives too early can distort the flood hydrograph even when a model fits much of the series. An objective defined over the trajectory can distinguish peak magnitude, timing, and event volume from ordinary pointwise error. Weekly averaging removes that within-event structure. Weekly residual learning instead has to address the scarcity of high-flow observations and the information carried by the latest observed flow.

1.4. Study Objectives and Contributions

The two datasets allow these questions to be examined at different resolutions. Hourly CAMELS-GB v2

observations support an evaluation of magnitude, timing, and joint capture in six-hour flood windows, as well as point forecasts from regular issue times [21]. In the Yellow River source region, ensemble prediction with error correction has been evaluated at three hydrological stations [22]. The HydroMLYR analysis addresses a related question across held-out catchments using the standardized weekly flow series in CCAM [23]. It examines high-flow error and detection over one-, two-, and three-week windows, with detailed model comparisons for the next-week screening task.

Both analyses pair a conceptual forecast with an event-conditioned regional residual and a regularized local correction estimated from a short support period. In the hourly model, the loss explicitly scores flood hydrograph properties. In the weekly model, high-flow observations receive more weight and the final prediction is anchored to the latest flow. Comparisons with persistence, conceptual, statistical, and neural baselines test whether the approach improves the measures that matter at each resolution. Ablations examine the contributions of the regional training objective and the adjustments made after regional prediction.

2. Materials and Methods

2.1. Forecasting Task and Framework

Let b index the catchment, t denote the forecast issue time, and h denote the forecast step. The inputs are the observation history $\boldsymbol{x}_{b,t}$, the static catchment attributes $\boldsymbol{s}_b$, and the event class $c_{b,t}$. The target $y_{b,t+h}$ is expressed as transformed unit-area discharge for the hourly forecasts and on the dataset's standardized flow scale for the weekly forecasts. We start with a conceptual baseline $B_{b,t,h}$ and add an estimated residual $f_{\theta,h}$.

$$\tilde{y}_{b,t+h} = B_{b,t,h} + f_{\theta,h}\left(\boldsymbol{x}_{b,t}, \boldsymbol{s}_b, c_{b,t}, B_{b,t}\right). \tag{1}$$

Figure 1 shows how the hourly and weekly forecasts are constructed. We estimate the regional parameters θ from source catchments. At each target catchment, we fit a low-dimensional correction using labels from a separate support period. Both fits remain fixed during testing. At each forecast issue time, the model takes the available observations from the target catchment, including historical meteorology.

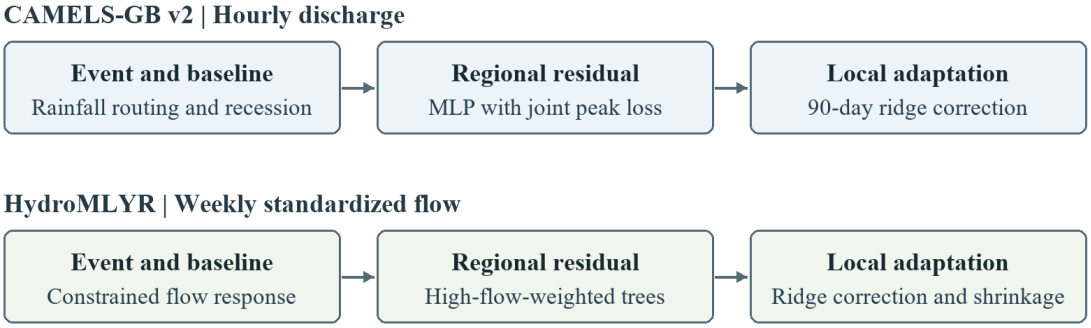


Figure 1. Hourly and weekly implementations of the regional residual framework.

2.2. Hourly Conceptual Routing and Peak-Aware Learning

For the UK implementation, observed discharge Q_t is converted to unit-area flow q_t in millimeters per hour using catchment area A_b in square kilometers. The learning target is z_t. Catchment subscripts are omitted below when the site is clear.

$$q_t = \frac{3.6 Q_t}{A_b}, \qquad z_t = \ln(1 + 10 q_t). \tag{2}$$

The inputs include 48 h of precipitation and transformed discharge, a rainfall availability indicator, an antecedent precipitation index, rolling precipitation totals, discharge changes, seasonal features, and catchment attributes. Let P_t denote precipitation in the latest hour, $P_t^{(k)}$ cumulative precipitation over the last k hours, and $\Delta_3 z_t = z_t - z_{t-3}$ the recent rise in transformed discharge. We apply the event rules in order. A window is classified as heavy rainfall or rapid rise if $P_t^{(6)} \geq 5$ mm, or if $\Delta_3 z_t \geq 0.22$ and $P_t^{(24)} \geq 2$ mm. Among windows that meet neither criterion, $P_t^{(24)} \geq 5$ mm indicates sustained rainfall. Of the remaining windows, those with $P_t^{(48)} < 0.8$ mm and

$\Delta_3 z_t < 0.035$ are classified as dry or stable. All other windows are classified as ordinary.

The conceptual baseline routes effective rainfall through a discrete gamma-shaped unit hydrograph. The antecedent precipitation index I_t has a decay time of 72 h. Catchment area and slope determine the response time τ_b, which is bounded between 2 and 15 h. The normalized kernel $K_b(j)$ spans lags zero through 143 and is zero outside that range.

$$I_t = P_t + e^{-1/72} I_{t-1}, \qquad E_t = P_t \left(0.11 + \frac{0.39 I_t}{18 + I_t} \right). \tag{3}$$

$$K_b(j) = \frac{(j+0.5) e^{-(j+0.5)/\tau_b}}{\sum_{l=0}^{143} (l+0.5) e^{-(l+0.5)/\tau_b}}. \tag{4}$$

At the forecast issue time, the baseline partitions observed flow into a routed rainfall response and a recession component. The factor a_t scales the routed response so that it stays within the observed discharge. The forecast continues the recession with a time scale of 120 h and routes the remaining response to rain that has already fallen.

$$R_t = \sum_{j=0}^{143} E_{t-j} K_b(j), \qquad a_t = \min \left(1, \frac{q_t}{\max(R_t, 10^{-6})} \right). \tag{5}$$

$$q_{t+h}^{\text{base}} = \max(q_t - a_t R_t, 0) e^{-h/120} + a_t \sum_{j=0}^{143} E_{t-j} K_b(j+h). \tag{6}$$

Equation (2) puts the conceptual forecast $B_{b,t,h}$ on the same transformed scale as the target by applying the discharge transform to q_{t+h}^{base}. The multilayer perceptron (MLP) keeps hourly resolution in the latest 24 h so that the encoder can distinguish rapid changes near the forecast issue time. It pools the preceding 24 h into eight three-hour averages to retain antecedent context with fewer input values. The decoder uses the encoded history, tabular summaries, source-normalized catchment attributes, the one-hot event class, and every third step of the conceptual trajectory to predict a 24-step residual. This combination allows a similar recent rise to imply different changes to the baseline under different rainfall conditions or catchment responses.

The training objective combines a pointwise loss with three terms evaluated over the first 6, 12, and 24 forecast hours. Let L_{point} denote the pointwise Smooth L1 loss with a transition parameter of 0.2. For each prefix H in $H = \{6, 12, 24\}$, the event terms measure peak magnitude error A_H, event volume error V_H, and soft peak timing error T_H. The event weights are 1.00 for heavy rainfall or rapid rise, 0.70 for sustained rain, and 0.35 otherwise.

$$L = L_{\text{point}} + \frac{0.60}{3} \sum_{H \in H} \text{mean} \left[w(c) \left(0.65 A_H + 0.20 V_H + 0.30 T_H \right) \right]. \tag{7}$$

The peak magnitude term applies Smooth L1 to the difference between the maxima of the predicted and observed trajectories on the transformed scale. The volume term applies the transform in Equation (2) to the mean unit-area flow of each trajectory and uses Smooth L1 with a transition parameter of 0.15. Let $\boldsymbol{v}$ be a transformed trajectory and σ_j the normalized softmax weight at index j. The soft peak position $C_H(\boldsymbol{v})$ gives greater weight to the time steps with the highest values.

$$C_H(\boldsymbol{v}) = \sum_{j=1}^{H} \frac{j-1}{H-1} \, \sigma_j(12\boldsymbol{v}). \tag{8}$$

The timing term multiplies the absolute difference between the two soft peak positions by the range of the observed transformed flow, clipped to the interval from zero to one. For the auxiliary calculations, transformed predictions are bounded between zero and seven. After local calibration, we clip negative predictions to zero on the transformed scale and convert the forecasts back to physical discharge. The fitted hourly model also has a rise enhancement for rainfall cases after the ninth forecast hour. Its contribution is zero throughout the six-hour window evaluated here.

2.3. Weekly High-Flow Learning and Shrinkage

In HydroMLYR, the prediction target is standardized seven-day mean streamflow. The input features include recent weekly flow and its lags, precipitation totals from the preceding $1-28$ days, antecedent wetness, mean temperature, a temperature-based melt proxy, evaporation, seasonal position, and catchment attributes. The melt proxy sums positive daily maximum temperatures on days when the minimum is below freezing, so it flags conditions conducive to thaw.

We classify weekly windows in order using only historical observations. A window enters the heavy-rain or rising-flow class when seven-day rainfall reaches 25 mm or flow rises by at least 0.65 standardized units from the previous week. The next class is sustained wetness, defined by at least 30 mm of precipitation over 14 days or an antecedent index of at least 16. A melt proxy of at least 8 together with a 28-day mean temperature below 5 °C indicates potential thaw. All remaining windows enter the ordinary class.

Let y_t denote the latest standardized weekly flow. The predictors $P_t^{(7)}$, I_t, and $M_t^{(7)}$ are the seven-day precipitation total, the daily exponentially weighted precipitation index with an eight-day half-life scaled by seven, and the seven-day melt proxy. Let ϕ_t denote the seasonal phase. We fit a constrained linear baseline separately for each lead.

$$B_{t,h} = \beta_{0,h} + \beta_{q,h} y_t + \beta_{p,h} P_t^{(7)} + \beta_{i,h} I_t + \beta_{m,h} M_t^{(7)} + \beta_{s,h} \sin\phi_t + \beta_{c,h} \cos\phi_t. \tag{9}$$

The current-flow coefficient $\beta_{q,h}$ is bounded between zero and one, which allows the baseline to carry forward or damp the latest flow. The nonnegative coefficients for rainfall, antecedent wetness, and melt allow these inputs to raise the baseline when the corresponding forcing is larger. A histogram gradient-boosting model [24] learns the remaining error separately at each of the three forecast steps. Equation (10) gives larger observed flows more weight, increasing their influence on the fitted residual.

$$w_{t,h} = 1 + 0.8 \min\left(\max(y_{t+h} - 0.6, 0), 3\right). \tag{10}$$

An auxiliary boosted tree predicts the maximum of the next three weekly flows. Let d_t denote that prediction and m_t the maximum of the initial forecast trajectory. We cap the absolute difference at 1.5 standardized units and use an exponential factor to concentrate the adjustment near the highest predicted value. The first step of the adjusted trajectory gives the one-week forecast.

$$u_{t,h} = \tilde{y}_{t+h} + 0.35 \operatorname{clip}(d_t - m_t, -1.5, 1.5)\, e^{2.5(\tilde{y}_{t+h} - m_t)}. \tag{11}$$

After local adaptation, we shrink the weekly forecasts toward the latest observed flow y_t. Let $\hat{r}_{b,t,h}$ be the local ridge correction clipped between minus one and one. The model selected on the validation catchments retains one tenth of the difference between the corrected forecast and persistence.

$$\hat{y}_{b,t+h} = 0.90\, y_t + 0.10\left(u_{t,h} + 0.50\, \hat{r}_{b,t,h}\right). \tag{12}$$

2.4. Local Adaptation and Event Analogues

We fit a ridge regression [25] in each target catchment to correct the remaining forecast error. For a given lead h, let $\boldsymbol{a}_i$ contain an intercept and local hydrometeorological features for support sample i. The response r_i is the observed value minus the forecast before local correction. We construct the features in the same way for support and test windows.

$$\hat{\boldsymbol{\gamma}}_{b,h} = \arg\min_{\boldsymbol{\gamma}} \left[\sum_{i \in S_h} (r_i - \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{\gamma})^2 + \lambda \left\| \boldsymbol{\gamma} \right\|_2^2\right]. \tag{13}$$

The UK correction uses 90 days of historical labels and $\lambda = 100$. The regressors summarize recent rainfall, current flow, the recent rise in flow, and the event class. We cap the absolute correction at 0.6 transformed units and add it to the forecast with a multiplier of 0.50. The Yellow River correction uses 26 eligible support windows and $\lambda = 18$. Its regressors describe current flow, the recent rise in flow, antecedent rainfall, and whether the window falls into the heavy-rain or rising-flow class. We fit the corrections separately at each target catchment.

The initial models included a correction based on regional forecast errors from similar historical events. We first selected source catchments by comparing standardized attributes. The UK search also compared hydrological response statistics from the local support period. Within the selected catchments, we searched for nearby event signatures, giving priority to events of the same type. We then added a distance-weighted mean residual to the forecast. The configurations selected on the validation catchments set the weight of this correction to zero in both datasets. We examine the analogue correction using the initial configurations and their ablations.

2.5. Dataset Splits and Evaluation

The datasets use separate source, validation, and geographically held-out test catchments. Table 1 summarizes their samples and calibration records. Dates are given as half-open intervals.

Table 1. Data partitions and evaluation windows.

Protocol Component	CAMELS-GB v2	HydroMLYR
Source, validation, and test catchments	35/12/17	17/6/9
Source training interval	Mar 2006–Oct 2013	Jan 2000–Jan 2012
Target adaptation	90 days from 1 Oct 2014	Last 26 eligible windows, Jan 2013–Jan 2015
Validation interval	Jan 2015–Jan 2017	Jan 2015–Jan 2017
Test interval	Jan 2017–Jan 2020	Jan 2017–Jan 2021
Regular test windows	70,074	1440
Peak or high-flow windows	584	197
Forecast window reported	Six hours	One week

CAMELS-GB v2 provides catchment-scale meteorology, observed flow, quality flags, and landscape attributes [21]. We use precipitation estimated from gauges and radar, together with quality-screened discharge. Each forecast requires 48 h of complete discharge history, with at least 43 of the last 48 and five of the last six hourly precipitation values present. The selected catchments cover 25–1500 square kilometers. We split them into source, validation, and test groups by hydrometric area, reserving six areas for testing.

HydroMLYR provides daily meteorology, standardized weekly streamflow, and catchment descriptors for 102 catchments in the Yellow River Basin [23]. We retain 32 sites after checking temporal coverage and excluding candidate pairs whose overlap exceeds 5% of the area of the smaller basin. We cluster the gauge locations into nine spatial blocks and assign three blocks to each split. An eligible window requires observed flow in the issue week and the previous week, together with valid targets for the next three weeks.

For the UK event evaluation, we identify observed local maxima at or above the 95th percentile of discharge in the adaptation period for each catchment. Peaks must be at least 72 h apart. We place the forecast origin three hours before each observed peak to construct a retrospective evaluation sample. For overall point metrics, we issue regular forecasts every six hours and score the sixth forecast hour. In HydroMLYR, a window counts as high flow when the observed next-week mean reaches 1.0 on the standardized scale. Each one-week target is a single weekly mean, so the evaluation measures high-flow error and detection.

Let P_i and $\hat{P}_i$ denote the maximum observed and forecast unit-area flows in UK event window i. Let T_i and $\hat{T}_i$ denote the first forecast steps at which these maxima occur. A forecast captures a peak if its magnitude error is at most 25% and its timing error is at most two hours. For each catchment, we compute the median absolute percentage error in peak magnitude, the mean absolute timing error in hours, and the percentage of peaks captured. We then average each statistic across catchments.

$$\mathrm{MdAPE}_{\mathrm{peak}} = 100\ \mathrm{median}_i\left(\frac{\left|\hat{P}_i - P_i\right|}{\max(P_i, 10^{-6})}\right), \qquad \mathrm{MAE}_{\mathrm{time}} = \mathrm{mean}_i|\hat{T}_i - T_i|. \tag{14}$$

In both datasets, the Nash–Sutcliffe efficiency (NSE) [26] compares the squared forecast error with the error from predicting the mean observed flow. Let y_i, $\hat{y}_i$, and $\bar{y}$ denote the observed flow at the evaluated lead, the corresponding forecast, and the catchment mean over the evaluation period. We calculate NSE and overall error scores separately for each catchment and average them across the test catchments.

$$\mathrm{NSE} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{15}$$

For HydroMLYR, we calculate high-flow MAE within each catchment using the observed high-flow windows. For the two- and three-week windows, the observed and predicted quantities used for high-flow error and detection are the maxima of the weekly means through the stated horizon. Recall is the fraction of observed high-flow windows in which the forecast reaches the threshold. Precision is the fraction of windows with forecasts at or above the threshold in which the observation also reaches it. We pool the windows to calculate recall and precision. The baselines include persistence, conceptual models, regional regression models, and local

ridge regression. For the hourly forecasts, we also compare with a gated recurrent unit (GRU) [27].

We select configurations using the validation catchments and freeze them before testing. For the hourly models, the search balances peak magnitude and timing errors while monitoring peak capture and overall accuracy. For the weekly models, configuration selection averages a criterion based on high-flow magnitude and timing errors across all three windows, subject to limits on how much overall forecast skill can fall below persistence. The CAMELS-GB v2 results describe sub-daily flood peaks at six hours. The HydroMLYR results compare the tuned model with persistence over all three weekly windows, with the one-week window used for the detailed model and ablation comparisons.

2.6. Implementation Details

The hourly MLP uses two encoder layers of widths 96 and 64 and two decoder layers of widths 128 and 96 before the output layer. Training runs for 12 epochs with batches of 1536 and AdamW, using an initial learning rate of 0.0015, weight decay of 0.003, and cosine annealing to 0.0001. The weekly gradient-boosting models use 210 iterations, a learning rate of 0.055, at most 17 leaves, and at least 35 samples per leaf. Their L2 regularization parameter is 5 and early stopping is disabled.

3. Results

3.1. Six-Hour Flood Peaks in Great Britain

The clearest six-hour gains were in peak magnitude and joint capture (Table 2). The tuned model achieved the lowest peak MdAPE (9.93%) and the highest capture rate (85.14%) among the models in Table 2. It improved on both persistence (19.67% MdAPE and 70.00% capture) and the regional GRU (11.72% and 74.30%), with a 10.84 percentage point gain in capture over the GRU. The paired bootstrap across six held-out hydrometric areas gave a 95% interval of −3.01 to −0.76 percentage points for the difference in peak MdAPE between the tuned model and the GRU. This interval describes variation under resampling of the six observed areas.

Table 2. Six-hour CAMELS-GB v2 test performance. Bold marks the best value for each metric among the models listed.

Model	Peak MdAPE (%)	Timing MAE (h)	Capture (%)	NSE	RMSE (m^3/s)
Persistence	19.67	1.99	70.00	0.719	4.366
Conceptual routing	22.87	1.94	66.27	0.660	4.632
Local ridge	11.77	**1.50**	72.97	0.845	3.178
Regional MLP	11.86	1.64	74.51	0.871	2.882
Regional GRU	11.72	1.70	74.30	0.871	2.916
Regional MLP with local adaptation	11.48	1.55	76.62	**0.881**	**2.776**
Initial prototype	12.13	1.67	77.06	0.878	2.867
Tuned hourly model	**9.93**	1.74	**85.14**	0.875	2.875

Recent rainfall can continue to reach the outlet after the forecast starts. The routed baseline describes this delayed contribution, and the joint loss gives errors in peak magnitude and timing explicit weight in the residual fit. The resulting forecasts also retained high efficiency at regular issue times (NSE of 0.875). The regional MLP with local adaptation led on overall NSE and RMSE, and local ridge regression had the smallest standalone timing error. Peak timing alone consequently gives a different ranking from the joint capture metric.

3.2. One-Week High Flow in the Yellow River Basin

At one week, the gains in high-flow prediction were larger against the regression baselines than against persistence (Table 3). High-flow MAE was 1.219, an improvement of 8.9% over the local ridge model and

18.8% over the regional gradient-boosting model. Persistence was a much closer benchmark. Its high-flow MAE was 1.243, and it detected 141 of the 197 observed high-flow windows. The tuned model detected 144 windows and achieved the highest recall in Table 3 (73.10%), with precision of 74.61% compared with 73.44% for persistence. Across all windows, the local ridge model led on NSE and persistence retained a slightly lower MAE.

Table 3. One-week HydroMLYR test performance. Errors are in standardized weekly streamflow units. Bold marks the best value for each metric among the models listed.

Model	High-Flow MAE	Overall MAE	NSE	Recall (%)	Precision (%)
Persistence	1.243	**0.316**	0.271	71.57	73.44
Conceptual baseline	1.699	0.535	0.166	10.15	80.00
Source ridge	1.560	0.488	0.264	25.38	**87.72**
Regional boosted trees	1.501	0.492	−0.081	32.49	65.98
Local ridge	1.338	0.333	**0.387**	63.45	82.78
Initial prototype	1.618	0.506	0.022	30.96	67.78
Tuned weekly model	**1.219**	0.318	0.299	**73.10**	74.61

High-flow MAE was lower than persistence in seven of the nine test catchments. The advantage also appeared in two of the three calendar groups formed by taking every third issue week. Weekly averaging smooths brief rises, and the latest observed flow often remains informative through the next week. This helps explain both the strength of persistence and the limited margin available to a forecast that adds information from rainfall, antecedent wetness, and other predictors.

At two and three weeks, high-flow MAE was 1.548 and 1.716 for the tuned model, compared with 1.596 and 1.790 for persistence. Recall was 59.77% and 52.44%, compared with 57.81% and 50.49% for persistence, respectively. NSE at the final weekly lead was 0.026 and 0.023 for the tuned model, compared with −0.032 and −0.067 for persistence.

3.3. Component Ablations

At six hours, the joint loss improved both peak magnitude and capture (Table 4). Removing it increased peak MdAPE from 9.93% to 12.10% and reduced capture from 85.14% to 78.32%. Local correction mainly improved magnitude and overall efficiency. Removing it increased MdAPE to 10.47% and lowered NSE to 0.866, although capture rose to 85.81%. A 30-day calibration record gave an intermediate magnitude error of 10.39%. These responses show that ridge correction and the joint loss affect different parts of the forecast.

Table 4. Selected ablations at the reported windows. Event error is UK peak MdAPE or Yellow River high-flow MAE, as indicated. Detection is UK capture or Yellow River recall.

Dataset and configuration	Event error	Detection (%)	NSE
UK, tuned	9.93%	85.14	0.875
UK, with pointwise loss only	12.10%	78.32	0.876
UK, without local correction	10.47%	85.81	0.866
UK, 30-day adaptation	10.39%	85.55	0.871
Yellow River, tuned	1.219	73.10	0.299
Yellow River, without local correction	1.210	72.59	0.299
Yellow River, without high-flow weighting	1.219	71.07	0.311
Yellow River, without current-flow shrinkage	1.487	49.24	−0.404

At the weekly scale, shrinkage toward the latest flow had the largest effect on performance. Removing it raised high-flow MAE to 1.487, cut recall to 49.24%, and lowered NSE to − 0.404. Removing high-flow weighting reduced recall from 73.10% to 71.07%, with little change in high-flow MAE. Removing the local correction lowered high-flow MAE from 1.219 to 1.210 and slightly reduced recall. The analogue correction offered little benefit at either reported horizon. When it was removed from the initial models, UK peak MdAPE fell from 12.13% to 11.98% and Yellow River high-flow MAE fell from 1.618 to 1.607. Both validation-selected models assigned zero weight to this term.

4. Discussion

4.1. What the Two Implementations Establish

The hourly results suggest that the regional model can recover part of the flood-peak structure by correcting a routed response with a loss defined over the trajectory. A rainfall pulse and the antecedent conditions affect the rise together. Giving peak magnitude, timing, and event volume explicit roles in training changes how the model fits the largest deviations from the baseline. The remaining timing errors point to catchment-specific travel times and delays that are still difficult to represent from a short calibration period.

The event class shapes both the hourly forecast and its training objective. Event indicators enter the regional residual network and the local ridge correction. The class weight in Equation (7) gives errors in peak magnitude, timing, and event volume more influence in training during heavy rainfall or rapid flow rises. In the regional residual models trained with pointwise loss, adding the event indicators alone changed peak MdAPE from 12.45% to 12.66%. The gains observed with the weighted trajectory loss therefore reflect the combined training objective, whose class weighting and trajectory terms were evaluated together.

At the weekly scale, catchment storage and prolonged wetness can carry information from one week into the next. The strong persistence baseline and the deterioration after removing shrinkage are consistent with this hydrological memory. High-flow weighting changes which windows the regional model emphasizes, while shrinkage limits departures from the observed discharge state. The modest gain over persistence came from small adjustments to the latest flow. The ablations suggest that a larger departure from that observation can easily erase the gain.

The two experiments evaluate transfer to new catchments within each dataset. Both use a conceptual baseline, an event-conditioned residual, and a short local record for calibration. The UK analysis measures errors in hourly peaks and peak timing from observed discharge. HydroMLYR measures errors in standardized weekly flow. The different targets and metrics limit comparisons of error magnitudes across the two datasets.

4.2. Scope and Limitations

The evaluation is retrospective. The detailed HydroMLYR comparison emphasizes the one-week window after inspection of the test results and should be interpreted as exploratory. In the UK event sample, we set the forecast origin relative to the observed peak, so the capture rate applies to those event windows. Operational studies can use official forecasts as a point of comparison, as McEachran et al. did using U.S. National Weather Service flood forecasts [28]. An operational evaluation of our approach would also need a rule for issuing warnings based on information available at the time, with both missed events and false alarms counted. The estimates from resampling and sensitivity analysis describe the six UK hydrometric areas and three Yellow River spatial blocks used for testing.

Both models use observed meteorology up to the forecast issue time. Forecast rainfall, event-triggered switching between time scales, and continuous updates to the local correction remain untested. The standardized weekly observations in HydroMLYR support forecasts of weekly high flow. Testing hourly transfer in Chinese catchments, including peak magnitude and timing in physical discharge, would require observations at the event scale. A larger archive and features that describe the rising limb more closely could help assess when analogue transfer is useful.

4.3. Data and Reproducibility

The CAMELS-GB v2 data used in this study are openly available from the Environmental Information Data Centre [29]. Their attribution is "Contains data supplied by UK Centre for Ecology & Hydrology, British Geological Survey, Environment Agency, Natural Resources Wales and Scottish Environmental Protection Agency." HydroMLYR is distributed with the CCAM archive [30]. The experimental records include catchment assignments, frozen model configurations, per-catchment evaluations, and predictions. These records support independent checking of the aggregate results and the subset definitions.

5. Conclusions

Regional residual learning improved the high-flow measures targeted in both held-out evaluations. At six hours in CAMELS-GB v2, the model achieved 9.93% peak magnitude error and 85.14% joint peak capture, with an NSE of 0.875 on regular forecast windows. The joint trajectory loss accounted for much of the gain in magnitude and capture, and short-record local calibration further reduced magnitude error. At one week in HydroMLYR, high-flow MAE fell from 1.243 for persistence to 1.219 and recall rose to 73.10%. High-flow MAE was lower in seven of nine catchments. High-flow MAE and recall also favored the tuned model over persistence over the two- and three-week windows. Weekly ablations show that shrinkage toward the latest flow kept errors in check, while high-flow weighting helped recover more events. These results describe retrospective performance in the evaluated catchments. Operational warnings require prospective evaluation under a real-time issue rule.

Funding

This research was funded by the Department of Science and Technology of Henan Province through its Key Scientific and Technological Research Program (grant no. 252102210100) for the project "Research on Key Technologies for Multiscale Forecasting of Hydrological Data."

Author Contributions

Y. Z. developed the study concept and methodology, secured funding, and prepared the original draft. X. S. curated the hydrological datasets and contributed to validation. J. L. implemented the forecasting models and conducted the experiments. D. X. analyzed the results and prepared the figures. All authors contributed to reviewing and editing the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The source datasets are distributed by EIDC and Zenodo under the identifiers listed in References [29,30].

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Merz R, Blöschl G. A Process Typology of Regional Floods. *Water Resources Research* 2003; **39**: 1340. https://doi.org/10.1029/2002wr001952.

2 Zhang Y, Zhang Y, Zhai X, *et al.* Understanding Meteorological and Physio-Geographical Controls of Variability of Flood Event Classes in Headstream Catchments of China. *Hydrology and Earth System*

Sciences 2025; **29**: 3257–3275. https://doi.org/10.5194/hess-29-3257-2025.

3 Tarasova L, Merz R, Kiss A, *et al*. Causative Classification of River Flood Events. *Wiley Interdisciplinary Reviews-Water* 2019; **6**: e1353. https://doi.org/10.1002/wat2.1353.

4 Mizukami N, Rakovec O, Newman AJ, *et al*. On The Choice of Calibration Metrics for "High-Flow" Estimation Using Hydrologic Models. *Hydrology and Earth System Sciences* 2019; **23**: 2601–2614. https://doi.org/10.5194/hess-23-2601-2019.

5 Bakke SJ, Barna DM, Engeland K, *et al*. The Ability of LSTM to Model Snowmelt Versus Rainfall Generated Floods. *Hydrology and Earth System Sciences* 2026; **30**: 6207–6233. https://doi.org/10.5194/hess-30-6207-2026.

6 Kratzert F, Nearing G, Addor N, *et al*. Caravan—A Global Community Dataset for Large-Sample Hydrology. *Scientific Data* 2023; **10**: 61. https://doi.org/10.1038/s41597-023-01975-w.

7 Tran VN, Xu D, Van Nguyen T, *et al*. CAMELSH: A Large-Sample Hourly Hydrometeorological Dataset and Attributes at Watershed-Scale for CONUS. *Scientific Data* 2025; **12**: 1307. https://doi. org/10.1038/s41597-025-05612-6.

8 Gao Y, Liang D, Borgomeo E, *et al*. Multi-Source Adaptive-Fusion Transfer Learning for Streamflow Forecasting in Data-Scarce Catchments. *Water Resources Research* 2026; **62**: e2025WR042495. https://doi.org/10.1029/2025WR042495.

9 Kratzert F, Klotz D, Shalev G, *et al*. Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine Learning Applied to Large-Sample Datasets. *Hydrology and Earth System Sciences* 2019; **23**: 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

10 Zhang B, Ouyang C, Xu Q, *et al*. Temporal Inductive Biases in Hourly Flood Forecasting: A Comparative Analysis of Recurrent, Attention-Based, and State-Space Neural Networks. *Journal of Hydrology* 2026; **677**: 135727. https://doi.org/10.1016/j.jhydrol.2026.135727.

11 Gauch M, Kratzert F, Klotz D, *et al*. Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network. *Hydrology and Earth System Sciences* 2021; **25**: 2045–2062. https://doi.org/10.5194/hess-25-2045-2021.

12 Jahangir MS, Quilty J. Hierarchical Deep Learning for Consistent Multi-Timescale Hydrological Forecasting. *Water Resources Research 2025*; **61**: e2024WR038105. https://doi.org/10.1029/2024WR038105.

13 Nearing G, Cohen D, Dube V, *et al*. Global Prediction of Extreme Floods in Ungauged Watersheds. *Nature* 2024; **627**: 559–563. https://doi.org/10.1038/s41586-024-07145-1.

14 Feng D, Fang K, Shen C. Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks with Data Integration at Continental Scales. *Water Resources Research* 2020; **56**: e2019WR026793. https://doi.org/10.1029/2019wr026793.

15 Ma K, Feng D, Lawson K, *et al.* Transferring Hydrologic Data across Continents–Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. *Water Resources Research* 2021; **57**: e2020WR028600. https://doi.org/10.1029/2020wr028600.

16 Ougahi JH, Rowan JS. Investigating Deep Learning Knowledge Transfer in Streamflow Prediction from Global to Local Catchment. *Water Resources Research* 2026; **62**: e2025WR041194. https://doi.org/10.1029/2025WR041194.

17 Kirchner JW. Getting the Right Answers for the Right Reasons: Linking Measurements, Analyses, and Models to Advance the Science of Hydrology. *Water Resources Research* 2006; **42**: W03S04. https://doi.org/10.1029/2005wr004362.

18 Liu J, Koch J, Stisen S, *et al*. A National-Scale Hybrid Model for Enhanced Streamflow Estimation–Consolidating a Physically Based Hydrological Model with Long Short-Term Memory (LSTM) Networks. *Hydrology and Earth System Sciences* 2024; **28**: 2871–2893. https://doi.org/10.5194/hess-28-2871-2024.

19 Chai Y, Ouyang W, Gu X, *et al*. A Differentiable Xin'anjiang Model Incorporating Transfer Learning for Cross-Basin Flood Prediction under Data Scarcity. *Journal of Hydrology* 2026; **679**: 136321. https://doi.org/10.1016/j.jhydrol.2026.136321.

20 Álvarez Chaves M, Acuña Espinoza E, Ehret U, *et al*. When Physics Gets in the Way: An Entropy-Based

Evaluation of Conceptual Constraints in Hybrid Hydrological Models. *Hydrology and Earth System Sciences* 2026; **30**: 629–658. https://doi.org/10.5194/hess-30-629-2026.

21 Coxon G, Zheng Y, Barbedo R, *et al*. CAMELS-GB V2: Hydrometeorological Time Series and Landscape Attributes for 671 Catchments in Great Britain. *Earth System Science Data* 2026; **18**: 4345–4371. https://doi.org/10.5194/essd-18-4345-2026.

22 Wang J, Li X, Wu R, *et al*. A Runoff Prediction Approach Based on Machine Learning, Ensemble Forecasting and Error Correction: A Case Study of Source Area of Yellow River. *Journal of Hydrology* 2025; **658**: 133190. https://doi.org/10.1016/j.jhydrol.2025.133190.

23 Hao Z, Jin J, Xia R, *et al*. CCAM: China Catchment Attributes and Meteorology Dataset. *Earth System Science Data* 2021; **13**: 5591–5616. https://doi.org/10.5194/essd-13-5591-2021.

24 Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 2001; **29**: 1189–1232. https://doi.org/10.1214/aos/1013203451.

25 Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970; **12**: 55–67. https://doi.org/10.1080/00401706.1970.10488634.

26 Nash JE, Sutcliffe JV. River Flow Forecasting through Conceptual Models Part I: A Discussion of Principles. *Journal of Hydrology* 1970; **10**: 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

27 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning Phrase Representations Using RNN Encoder – Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.

28 McEachran Z, Ghosh R, Renganathan A, *et al*. Knowledge-Guided Machine Learning for Operational Flood Forecasting. *Water Resources Research* 2025; **61**: e2024WR039064. https://doi.org/10.1029/2024WR039064.

29 Coxon G, Zheng Y, Barbedo R, *et al*. Catchment Boundaries, Daily and Sub-Daily Hydrometeorological Time Series, Groundwater Level Time Series and Attributes for 671 Catchments in Great Britain (CAMELS-GB V2). *Environmental Information Data Centre* 2025. https://doi.org/10.5285/9a46d428-958f-4ac1-86eb-94eee70c0955.

30 Hao Z, Jin J, Xia R, *et al*. CCAM: China Catchment Attributes and Meteorology Dataset, Version 2.0. *Zenodo* 2021. https://doi.org/10.5281/zenodo.5729444.

© The Author(s) 2026. Published by Global Science Publishing (*GSP*).

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.