Supplementaries for

HydroForecast: A Global Deep Learning Model for Probabilistic Flood Forecasting

Bohan Huang[1,2], Zhu Liu[1,2,3], Wentao Li[1,2,3*], Baoxiang Pan[4], Ather Abbas[5], Hylke Beck[5], and Qingyun Duan[1,2,3*]

[a]The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing China

[b]Department of Hydrology and Water Resource, Hohai University, Nanjing China

[c]China Meteorological Administration Hydro-Meteorology Key Laboratory, Hohai University, Nanjing China

Contents of this file

Introduction

This supporting information provides detailed mathematical formulations and supplementary data descriptions to support the methodology and analysis presented in the main manuscript. Text S1 details the specific algorithms used for calculating hydrological signatures, including the flood characteristic metrics and the implementation of the Lyne–Hollick digital filter for baseflow separation. Text S2 presents the theoretical framework of the diffusion model, providing a comprehensive derivation of the forward diffusion (noising) process, the reverse denoising process, and the sampling strategy used for runoff generation. Additionally, this file contains three supplementary tables. Table S1 lists the comprehensive set of 31 catchment attributes (hydrological, climatic, meteorological, and topographic) used for catchment characterization. Table S2 provides the mathematical definitions of the evaluation metrics used to assess model performance.

Finally, Table S3 details the 21 specific attributes selected as inputs for the HydroForecast model.

Text S1.

The three flood characteristics used in this study—average peak flow, flood volume, and lag time—were derived for all flood events exceeding the two-year return period threshold. For each event, peak flow represents the maximum discharge within the event window, while flood volume quantifies the cumulative discharge above the baseflow level. Lag time is defined as the temporal offset between the centroid of effective precipitation and the timing of peak discharge. All event-level metrics were computed using the formulas provided below.

$$\bar{Q}_p = \frac{1}{M}\sum_{i=1}^{M} Q_{p,i} \tag{1}$$

$$\bar{V} = \frac{1}{M}\sum_{i=1}^{M}\sum_{t=t_{start}}^{t=t_{end}} \left(q_{obs}(t) - q_{base}(t)\right)\Delta t \tag{2}$$

$$\bar{T}_{lag} = \frac{1}{M}\sum_{i=1}^{M}(T_{Q,i}^{cen} - T_{P,i}^{cen}) \tag{3}$$

Where $Q_{p,i}$ denotes the peak discharge of the i-th flood event, $q_{base}(t)$ represents the baseflow at time t, and t_{start}, t_{end} indicated the start and end times of the flood event, respectively. $T_{Q,i}^{cen}$ denotes the centroid time of the discharge hydrograph for the i-th event, and $T_{P,i}^{cen}$ denotes the centroid time of the corresponding precipitation hyetograph.

To isolate the quick-flow component from the observed hydrograph and ensure reliable computation of flood characteristics, we applied the Lyne–Hollick digital filter to derive baseflow for each catchment. This recursive filter separates discharge into slow and fast components by suppressing high-frequency variability while preserving the long-term recession behavior.

$$b_t = \alpha b_{t-1} + \frac{1+\alpha}{2}(Q_t - Q_{t-1}) \tag{4}$$

Text S2

The forward process progressively perturbs the observed discharge sequence by injecting Gaussian noise over a series of diffusion steps. At each step, a small amount of noise is added according to a predefined variance schedule, gradually reducing the influence of the original hydrological signal:

$$Q_{(t)} = \sqrt{\overline{\alpha_t}}Q_{obs} + \sqrt{1-\overline{\alpha_t}}\epsilon \tag{5}$$

Where Q_{obs} represents the observed daily discharge; $Q_{(t)}$ is the perturbed discharge at timestep t. It can be interpreted as a version of the original discharge in which the structural information has been partially obscured by noise. $\overline{\alpha_t}$ is a predefined noise schedule parameter that controls the signal-to-noise ratio at timestep t. This forward noise-injection process is fixed and does not involve learnable parameters.

The reverse process reconstructs discharge sequences by iteratively removing the noise injected during the forward phase. At each diffusion step, the model predicts the

noise component contained in the current state using the denoising network introduced in the main text

$$\hat{\epsilon}_t = G_\theta\left(Q_t, M_{past}, M_{prediction}, s\right) \tag{6}$$

Where $\hat{\epsilon}_t$ denotes the predicted noise for step t. Q_t denotes the noise-injected discharge at time step t. G_θ represents the encoder–decoder LSTM. M_{past} and $M_{prediction}$ denotes the antecedent and prediction horizon meteorological forcing respectively. s denotes the attributes of catchments.

$\hat{\epsilon}_t$ than is used to compute the mean of distribution for the next time $t-1$ (reverse arrange):

$$Q_{t-1} = \frac{1}{\sqrt{\overline{\alpha_t}}}\left(Q_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\hat{\epsilon}_t\right) + \sigma_t z \tag{7}$$

Q_{t-1} than is used as the input for the next time step, allowing the sequence to gradually move from an initial Gaussian sample toward a realistic discharge trajectory.

The sampling procedure generates probabilistic discharge trajectories by progressively transforming random Gaussian noise into hydrologically realistic sequences:

$$Q_N \sim P_\theta\left(X_t \middle| M_{past}, M_{future}, s\right) \tag{8}$$

Where Q_N represents an ensemble predicted discharge sequences. $X_t \sim N(0,1)$.

The procedure begins by drawing an initial latent state X_t from a standard Gaussian distribution. For each sample step $t = T, \ldots, 0$, the model applies the reversing denoising update. This procedure follows the same formulation as Eq. (6) and Eq. (7), except that the noise term is obtained by combining a conditional and an unconditional prediction:

$$\bar{\epsilon}_t = G_\theta\left(Q_t, M_{past}, M_{future}, s\right) + \omega\left(G_\theta\left(Q_t, M_{past}, M_{future}, s\right) - G_\theta\left(Q_t, M_{past}, s\right)\right) \tag{9}$$

Where ω is a scalar controlling the strength of the guidance. Larger values of ω increase the model's responsiveness to the conditioning inputs. The final state Q_0 is treated as the member of the forecast ensemble.

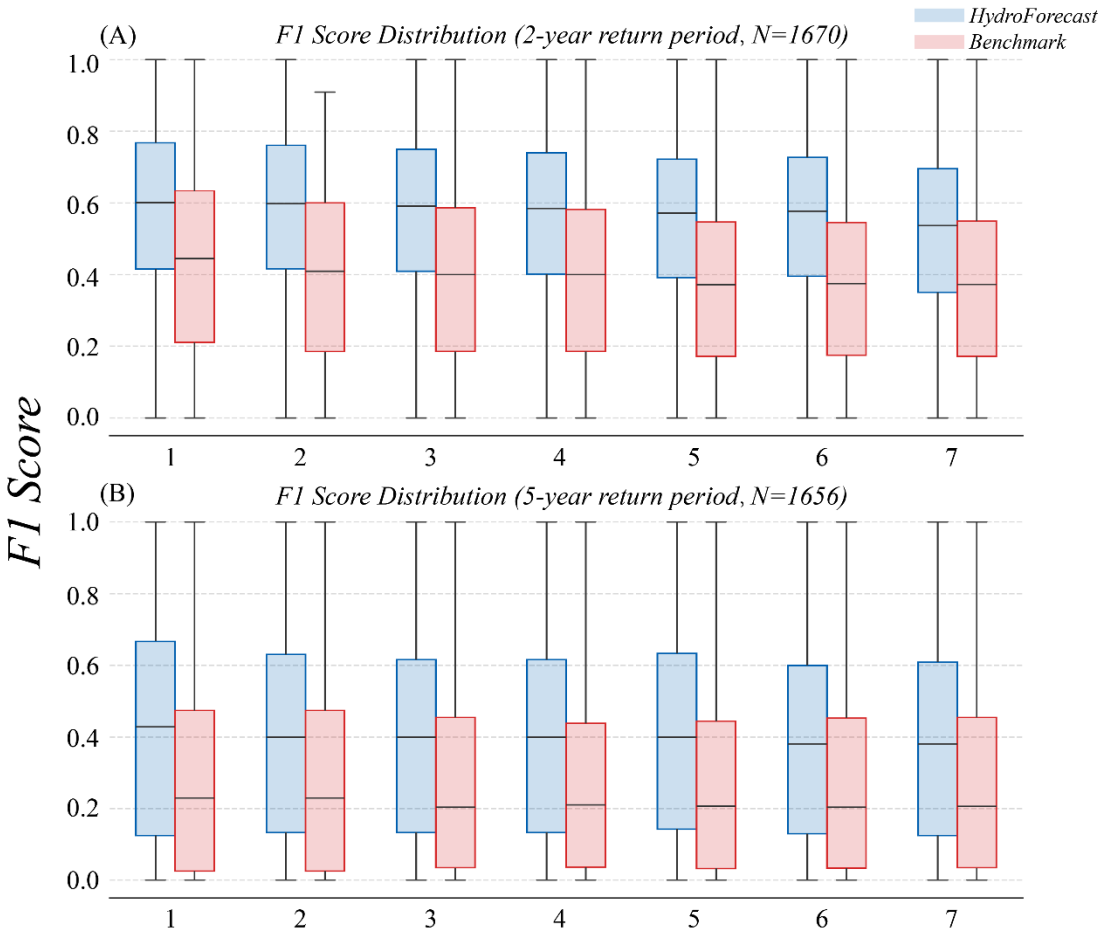

Figure 1. Assessment of flood event detection skill across forecast lead times. Boxplots illustrating the distribution of F1 scores for HydroForecast and the benchmark at lead times ranging from 1 to 7 days. *N* denotes the number of catchments.

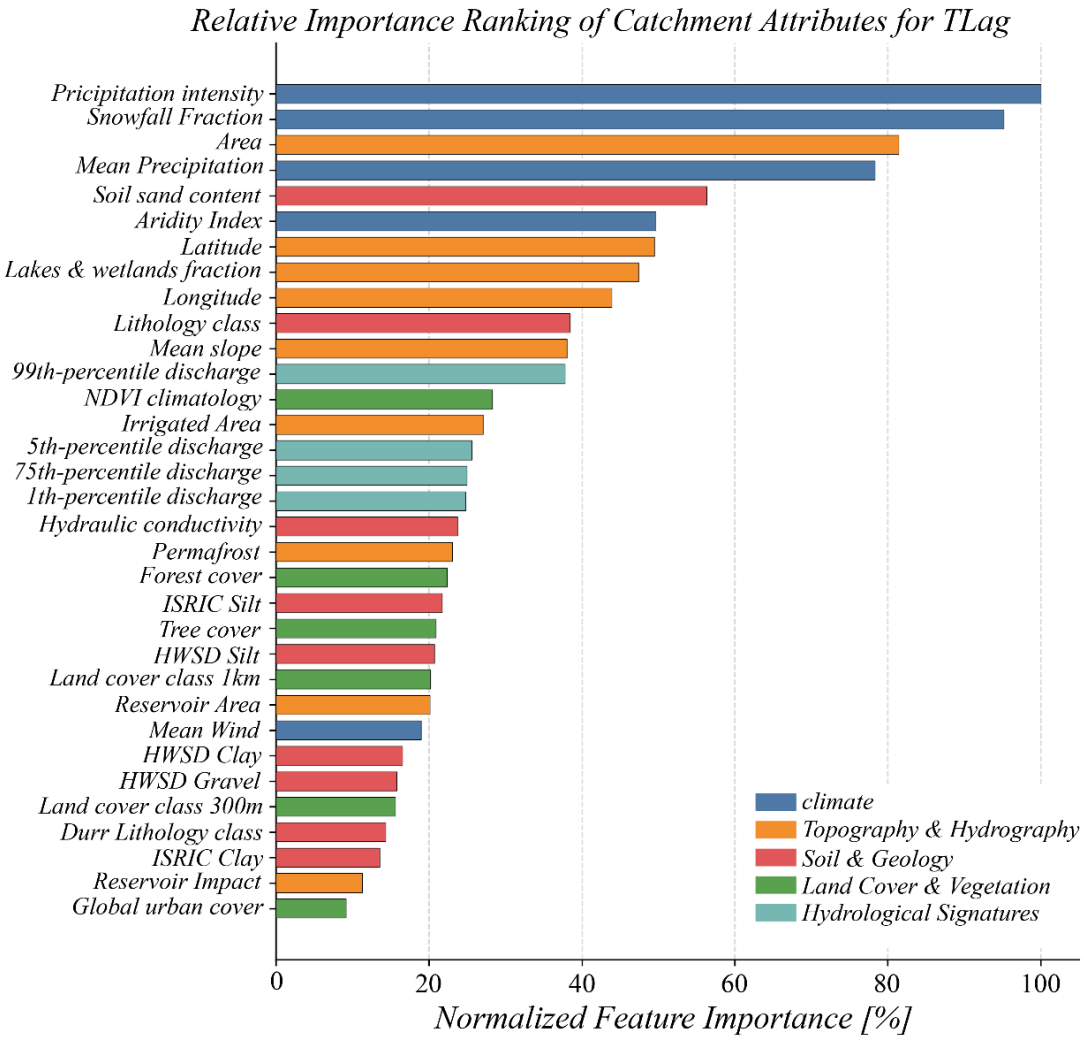

Figure S2. Normalized global feature importance ranking for time lag.

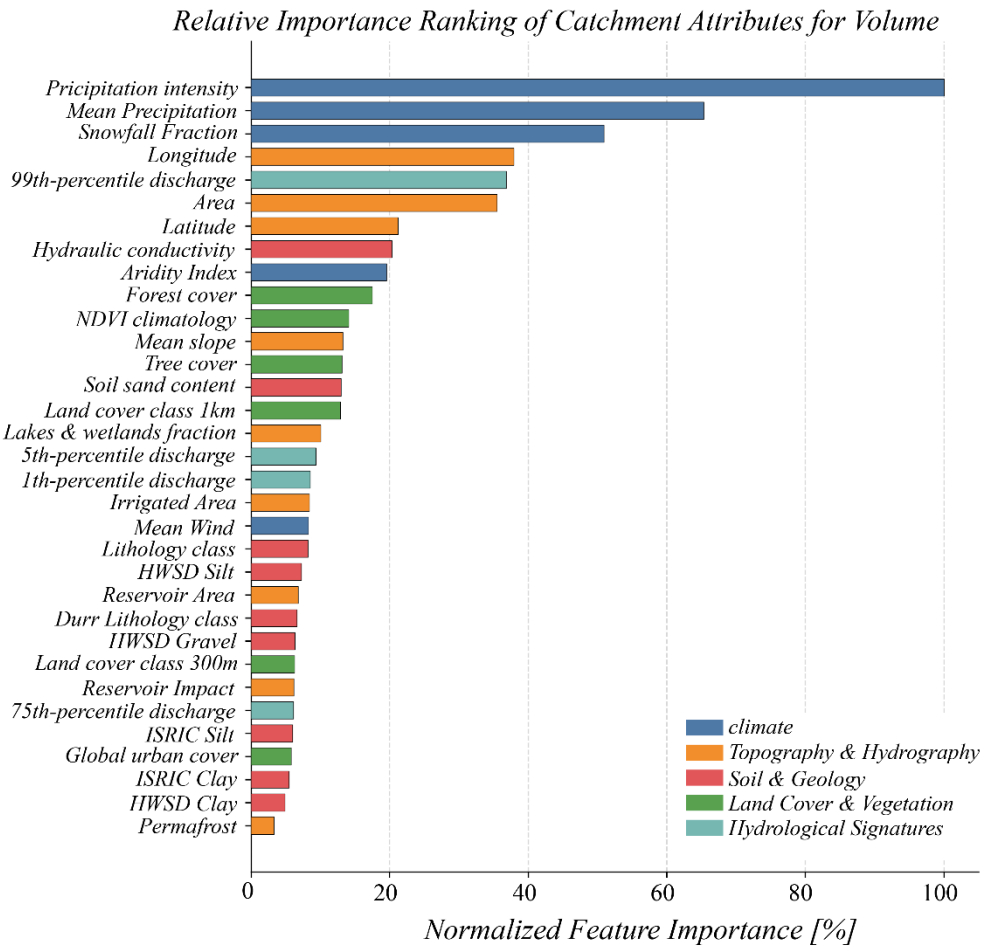

Figure S3. Normalized global feature importance ranking for volume.

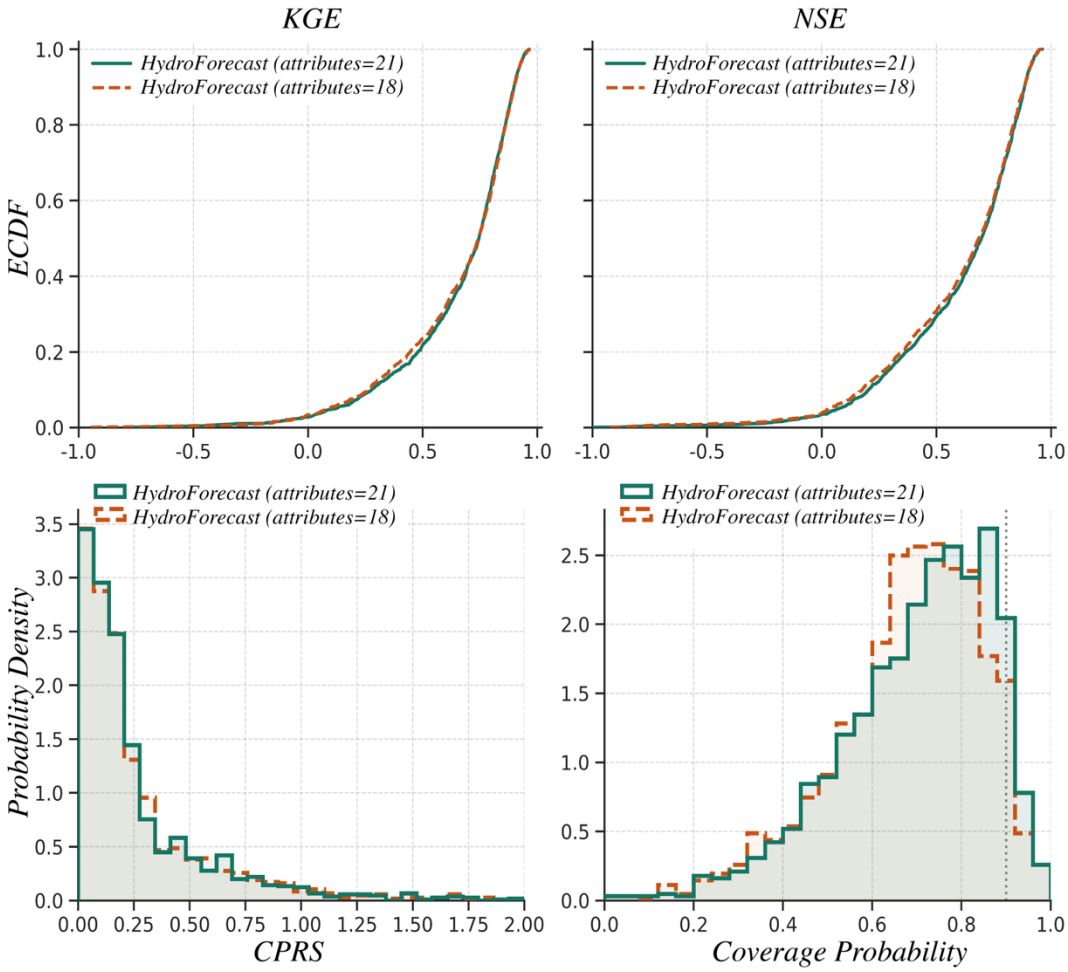

Figure S4. Analysis of HydroForecast performance to the number of input attributes. Comparison of deterministic (NSE, KGE) and probabilistic (CRPS, 90% Reliability) metrics between the model trained with a subset of 18 attributes versus the final set of 21 attributes. The boxplots compare deterministic (NSE, KGE) and probabilistic (CRPS, 90% Reliability) metrics. The reduced 18-attribute set excludes discharge quantiles ($\boldsymbol{Q_1}$, $\boldsymbol{Q_5}$, $\boldsymbol{Q_{99}}$) to simulate strictly ungauged scenarios. Deterministic metrics show statistically detectable but practically negligible performance penalties; the mean difference in NSE is only -0.034 (effect size = -0.029), while KGE shows no statistically significant difference ($\boldsymbol{p} = \boldsymbol{2.39 \times 10^{-4}}, \boldsymbol{d} = \boldsymbol{-0.013}$). In contrast, the exclusion of quantile signatures has a more pronounced impact on uncertainty quantification. The degradation in CRPS is minimal (mean change -0.005, $\boldsymbol{p} = \boldsymbol{3.97 \times 10^{-6}}$, d=-0.096), but the 90% Reliability shows a highly significant decline ($\boldsymbol{p} = \boldsymbol{5.53 \times 10^{-32}}$, $\boldsymbol{d} = \boldsymbol{0.281}$). This indicates that extreme flow

quantiles are primarily beneficial for constraining probabilistic bounds rather than improving point estimation accuracy.

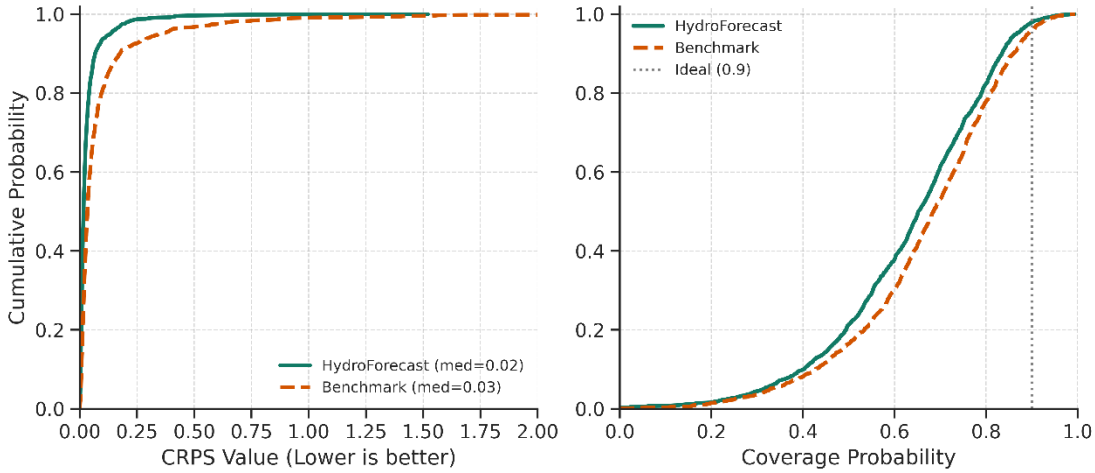

Figure S5. (a) Cumulative distributions of Threshold-Weighted CRPS (twCRPS), focusing on accuracy in high-flow regimes. (b) Reliability of 90% prediction intervals for these extreme events. The figures compare the performance of HydroForecast against the benchmark. HydroForecast demonstrates a closer fit to the extreme value distributions (indicated by lower twCRPS and better-calibrated reliability), highlighting its superior capability in capturing extreme flood conditions compared to the benchmark.

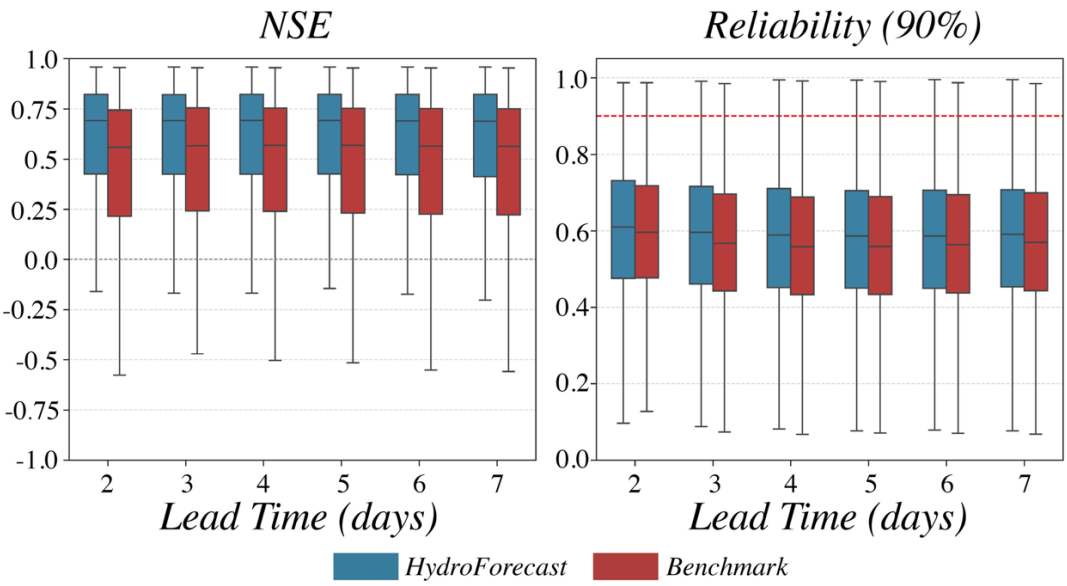

Figure S6. Performance comparison between HydroForecast and the benchmark model across different lead times.

Table S1. Comprehensive set of 31 original catchment attributes hydrological, climatic, meteorological, and topographic.

Variable	Source

Precipitation intensity	Computed from MSWX
Mean Precipitation	WorldClim v2.1
Area	GRDC
Longitude	GRDC
99th percentile of discharge	Computed from long-term discharge (GRDC)
Snowfall Fraction	WorldClim v2.1
Latitude	GRDC
Aridity Index	Global Aridity Index & PET Database v3
Soil sand content	SoilGrids (ISRIC).
Hydraulic conductivity	GLHYMPS v2
Mean slope	GMTED2010
NDVI climatology	Copernicus PROBA-V NDVI climatology
Lakes & wetlands fraction	Global Lakes and Wetlands Database (GLWD-3)
Reservoir Area	GeoDAR v1.1 Reservoir Database
Tree cover	MODIS Vegetation Continuous Fields (MOD44B)
5th percentile of discharge	Computed from long-term discharge (GRDC)
1th percentile of discharge	Computed from long-term discharge (GRDC)
Forest cover	FAO FRA 2000
Irrigated Area	GMIA v5 (FAO Global Map of Irrigated Areas)
Lithology class	Global Lithological Map (GLiM v1)
Land cover class 1km	UMD Global Land Cover (University of Maryland)
HWSD Silt	Harmonized World Soil Database
Mean Wind	WorldClim v2.1
Reservoir Impact	GRanD v1.3 (Global Reservoir and Dam Database)
HWSD Gravel	Harmonized World Soil Database
Global urban cover	GlobCover 2009 (ESA)
ISRIC Silt	Lithological class (Dürr et al., 2005)
Durr Lithology class	Dürr et al. (2005) Global Lithological Map
Land cover class 300m	ESA GlobCover 2009
ISRIC Clay	SoilGrids (ISRIC)
Permafrost	NSIDC Permafrost Data (National Snow and Ice Data Center)
75th percentile discharge	Computed from long-term discharge (GRDC)
HWSD Clay	Harmonized World Soil Database

Table S2. Evaluation metrics and their mathematical formulations.

Metric	Formula
Nash–Sutcliffe Efficiency (NSE)	$NSE = \frac{\sum_{t=1}^{n}\left(Q_t^{obs} - Q_t^{sim}\right)^2}{\sum_{t=1}^{n}\left(Q_t^{obs} - \bar{Q}^{obs}\right)^2}$
Kling–Gupta Efficiency (KGE)	$KGE = 1 - \sqrt{(\gamma-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$
Percentage Bias (PBIAS)	$PBIAS = \frac{\left(Q_C^{sim} - Q_C^{ob}\right)}{Q_C^{ob}}$
Continuous Ranked Probability Score (CRPS)	$CRPS_t = \frac{1}{M}\sum_{m=1}^{M}\left\|x_t^{(m)} - Q_t^{obs}\right\| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\left\|x_t^{(m)} - x_t^{(n)}\right\|$
$Reliabilty_{90}$	$Reliabilty_{90} = \frac{1}{N}\sum_{t=1}^{N}(L_t \le Q_t^{obs} \le U_t)$
F1 Score	$2 * \frac{Precision * Recall}{Precision + Recall}$
Precision	$Precision = \frac{TP}{TP + FP}$
Recall	$Recall = \frac{TP}{TP + FN}$

Notes:

Q_t^{obs} denotes the observed discharge at time step t;

Q_t^{simi} denotes the forecasted discharge;

N is the total number of time steps;

$x_t^{(m)}$ is the m-th ensemble forecast at time step t;

M is the number of ensemble members;

TP is the number of correctly predicted flood event.

FP is the number of false-alarm flood events.

FN is the number of observed flood events that the model fails to predict.

Table S3. the 21 attributes for HydroForecsat.

Variable	Source
Precipitation intensity	Computed from MSWX
Mean Precipitation	WorldClim v2.1
Area	GRDC
Longitude	GRDC
99th percentile of discharge	Computed from long-term discharge (GRDC)

Snowfall Fraction	WorldClim v2.1
Latitude	GRDC
Aridity Index	Global Aridity Index & PET Database v3
Soil sand content	SoilGrids (ISRIC).
Hydraulic conductivity	GLHYMPS v2
Mean slope	GMTED2010
NDVI climatology	Copernicus PROBA-V NDVI climatology
Lakes & wetlands fraction	Global Lakes and Wetlands Database (GLWD-3)
Reservoir Area	GeoDAR v1.1 Reservoir Database
Tree cover	MODIS Vegetation Continuous Fields (MOD44B)
5^{th} percentile of discharge	Computed from long-term discharge (GRDC)
1^{th} percentile of discharge	Computed from long-term discharge (GRDC)
Forest cover	FAO FRA 2000
Irrigated Area	GMIA v5 (FAO Global Map of Irrigated Areas)
Lithology class	Global Lithological Map (GLiM v1)
Land cover class 1km	UMD Global Land Cover (University of Maryland)